UNIVERSIDAD POLITÉCNICA DE MADRID
Escuela Técnica Superior de Ingeniería de Sistemas Informáticos

New Deep Learning techniques for image analysis: enhancements through information fusion, ensembling and explainability

DOCTORAL THESIS

Submitted for the degree of Doctor by:

Helena Liz López
Master's Degree in Biocomputing and Computational Biology

Madrid, 2023

UNIVERSIDAD POLITÉCNICA DE MADRID
Escuela Técnica Superior de Ingeniería de Sistemas Informáticos

Doctoral Degree in Doctorado en Ciencias y Tecnologías de la Computación para Smart Cities

New Deep Learning techniques for image analysis: enhancements through information fusion, ensembling and explainability

DOCTORAL THESIS

Submitted for the degree of Doctor by:

Helena Liz López
Master's Degree in Biocomputing and Computational Biology

Under the supervision of:
Dr. David Camacho Fernández(Supervisor)
Dr. Javier Huertas Tato(Cosupervisor)

Madrid, 2023

Title: New Deep Learning techniques for image analysis: enhancements through information fusion, ensembling and explainability

Author: Helena Liz López

Doctoral Programme: Doctorado en Ciencias y Tecnologías de la Computación para Smart Cities

Thesis Supervision:

Dr. David Camacho Fernández, Catedrático de Universidad, Universidad Politécnica de Madrid(Supervisor)

Dr. Javier Huertas Tato, Ayudante Doctor, Universidad Politécnica de Madrid(Co-supervisor)

External Reviewers:

Thesis Defense Committee:

Thesis Defense Date:

This thesis has been partially supported by Grant PLEC2021-007681 funded by MCIN/AEI/ 10.13039/501100011033 and, as appropriate, by "ERDF A way of making Europe", by the "European Union" or by the "European Union NextGenerationEU/PRTR", and by Comunidad Autónoma de Madrid under S2018/TCS-4566 (CYNAMON) grant.

Agradecimientos

En primer lugar, me gustaría dar las gracias a mis directores por haberme acompañado y apoyado todos estos años. En primer lugar, a David, por haber puesto su voto de confianza en mí para comenzar esta aventura, por haberme apoyado y dado una oportunidad a una bióloga que deseaba ser informática. Gracias por todos tus consejos, confianza, por haberme ayudado a llegar hasta aquí y prepararme para el camino que queda por delante. A Javi, por todos sus consejos, por todo lo que me has enseñado, por no rendirte conmigo y por haberme apoyado y animado a lo largo de todo este camino. No tengo palabras suficientes para agradeceros todo lo que habéis hecho por mi, y espero continuar este camino con vosotros. Aunque parezca el final, solo es el inicio.

También quiero agradecer a todos los integrantes de AIDA que me han acompañado estos años, no solo por el apoyo y la amistad. Para mí, no sois solo compañeros de trabajo, sino que os habéis convertido en mi segunda familia. Gracias a JaviT (por estar siempre ahí y apoyarme en todo momento), a Ángel (por todo lo que me has enseñado y por ser capaz de sacarme una sonrisa siempre), a Sergio (por esos descansos y cafés que tanto me han ayudado), a Guille (por animarme incluso con la mascarilla puesta), a Macu (porque aunque llevas poco tiempo te has vuelto imprescindible), a Cristian (por todos tus consejos y también los de Pokémon), a Víctor (por mostrarme el camino de las IAs generativas), a Adri (aunque lleves poco, eres un gran compañero), a Alex (por tu ayuda y enseñarme que soy capaz de mucho más de lo que creía). También me gustaría dar un abrazo a Raquel, Antonio, Emma, Carlos, Luis, Aurea y al resto del grupo de AIDA. Todos vosotros habéis contribuido para ayudarme a llegar hasta aquí.

Quiero agradecer a mis amigos, tanto a Mapu, Marina, Marta y Gato, que siempre han estado ahí desde mucho antes de emprender este viaje; no sé cómo darles las gracias. A mis amigos de Dungeons & Dragons, Relinx, Guinot, Puch, Oscar, por todos esos fines de semana que lograron que desconectara un poco de la tesis. En especial a Víctor, gracias por todo lo que me has ayudado estos meses... También quiero recordar esos días de Arkham Horror con Patri y Juan. Por último, y no menos importante, Ares y Dani, me animaron a perseguir un sueño que al final se ha hecho realidad. Gracias por aquel día en que me dijisteis que me atreviera a cambiar de rumbo.

A Alfonso, porque has estado ahí desde el primer momento, por tus palabras de ánimo, por todos tus consejos y por todas tus bromas. Gracias por sacarme siempre una sonrisa y por no soltar mi mano en todo este tiempo. Te quiero.

Finalmente, quiero agradecer a mi familia. A mis padres, por enseñarme tanto en estos años, por no dudar de mí en ningún momento y por los valores que me han inculcado. A mi tía Pili y mi tío Toño, que han sido como mis segundos padres y me han cuidado desde que era pequeña. A mi tía Marimar, mi tío Mariano y mis primos Javi, Iván y Héctor. A mi tía Cris, mis primas Alba y Malena, y mis abuelos Montse

y Juan. Todos vosotros habéis formado parte de este camino, habéis estado ahí en todo momento y habéis sido siempre un apoyo para mí. Fanuca y Jose, no podía olvidarme de vosotros. Por último, quiero dar las gracias a una persona que, aunque ya no esté, siempre me acompañará, mi tío Marco. Sé que estarías muy orgulloso y me sonreirías como siempre lo hacías.

Resumen, conclusiones y trabajo futuro

Resumen

El procesamiento de imagen se enfrenta a un grave problema, las deficiencias de los datasets. En muchas situaciones los dataset no tienen la calidad necesaria para ser procesados mediante técnicas de Deep Learning (DL) y obtener los resultas deseados. Por ello, los investigadores han tenido que explorar diferentes técnicas que permitan superar estas limitaciones sin necesidad de ampliar los datasets, ya que en muchas situaciones es imposible como puede ser el campo de la medicina. Por ello, los investigadores tienen que explorar diferentes vías que permitan crear sistemas para resolver problemas de procesamiento de imagen de manera precisa y efectiva.

En esta tesis se abordaran diferentes problemas de procesamiento de imagen mediante técnicas de fusión de información y ensembles. Además se desarrollaran diferentes técnicas de IA explicativa que permitan tanto a los investigadores como a los usuarios finales de los sistemas de DL a comprender cómo estos sistemas han llegado a la solución final y comprender si están usando la información adecuada para tomar dicha decisión.

En primer lugar se han utilizado técnicas de ensembles para resolver problemas de clasificación dentro del campo de la medicina, más específicamente utilizando radiografías torácicas. Dentro de este dominio se han diseñados tanto sistemas de clasificación binarios como problemas de clasificación multietiqueta. Por otro lado, debido al dominio de aplicación se han creado diferentes técnicas de visualización basadas en mapas de calor que permitan al personal sanitario entender cómo se ha llegado al resultado final. Las técnicas de visualización son especialmente importantes en este dominio debido al peligro que supone un resultado erróneo del sistema de clasificación, puede provocar daños en los pacientes.

En segundo lugar se han utilizado técnicas de fusión de información para resolver tareas de regresión dentro del campo de aplicación de los incendios forestales. Se ha creado un sistema de regresión basado en Redes Neuronales Convolucionales que sea capaz de predecir los recursos necesarios en caso de incendios en una comunidad autónoma de España. Para facilitar la aplicación de estas técnicas se generaron unas visualizaciones donde se predicen los recursos que serían necesarios en caso de incendio a nivel de todo el área de estudio diariamente.

Finalmente, con el objetivo de ampliar el área de aplicación de las técnicas de procesamiento de imagen utilizando técnicas de ensembles y fusión de información a otras modalidades de datos multimedia: vídeo, entendido como una secuencia de imágenes; audio y datos multimodales, que combina los dos anteriores. Para ello se realizó una revisión del estado del arte de las técnicas de detección de manipulación para los datos multimedia mencionados. Para ello se analizaron tanto las técnicas de manipulación de datos multimedia como las técnicas de detección lo que permite crear una visión completa del campo y analizar las posibles áreas menos exploradas y más susceptibles de ser analizadas con las técnicas desarrolladas a lo largo de esta tésis doctoral.

Conclusiones y Trabajos Futuro

En este último capítulo, se expondrán las diferentes conclusiones extraídas de la investigación presentada en esta tesis doctoral. Se dará respuesta a las diferentes Preguntas de Investigación, planteadas en el Capítulo 1, con el objetivo de dar una idea general y los detalles necesarios para futuros trabajos dentro del procesado de imágenes que permitan superar los problemas presentes en los conjuntos de datos de imágenes. Por último, se proponen posibles líneas de trabajo futuras dentro de este dominio que sería interesante explorar.

Conclusiones

A lo largo de los diferentes capítulos de esta tesis se han ido resolviendo diferentes problemas derivados de los conjuntos de datos de imágenes. Estos conjuntos de datos presentan en la mayoría de los casos una serie de problemas o limitaciones que no pueden resolverse mejorando la calidad de los mismos mediante la adición de nuevas muestras. Para resolver estas dificultades se han utilizado dos enfoques principales: las técnicas de fusión de información y los *ensembles*. Debido a los dominios de aplicación abordados a lo largo de esta tesis, otro de los pilares de la misma ha sido el desarrollo de técnicas de visualización basadas en mapas de calor para facilitar su aplicación en los diferentes dominios.

En el capítulo 2, se analizaron en primer lugar los problemas relacionados con los conjuntos de datos de pequeño tamaño y baja calidad en problemas de clasificación binaria utilizando un conjunto de datos de radiografías pediátricas de tórax compuesto por 950 muestras. Para ello, se creó un sistema basado en *ensembles* compuesto por cinco modelos creados desde cero, aplicando técnicas básicas de combinación, en particular, utilizando la media de las predicciones. También se demostró que, cuando los datos disponibles no contienen suficiente información, los modelos simples obtienen mejores resultados que los modelos más complejos, como CheXNet. Los modelos más complejos requieren un conjunto de datos de calidad mínima, lo que no es posible en muchas situaciones.

Tras resolver problemas de clasificación binaria, se siguió explorando problemas más complejos relacionados con la calidad del conjunto de datos. En el capítulo 3, se resolvió un problema de clasificación multietiqueta, en el que el número de clases era muy elevado (más de 30 clases diferentes) dentro del ámbito médico. Los conjuntos de datos con un número tan elevado de clases tienden a presentar problemas de desequilibrio, es decir, algunas clases presentan un número de muestras muy inferior a otras clases. En el campo de la medicina, este problema es muy común porque no todas las enfermedades tienen la misma incidencia en la población. Para resolver este problema relacionado con el desequilibrio, se exploraron diferentes técnicas de *ensemble* centradas en problemas de clasificación multietiqueta. Se crearon diferentes sistemas basados en conjuntos utilizando modelos preentrenados del estado del arte, lo que demostró que las técnicas de ensemble permiten un mejor rendimiento y una mayor capacidad de generalización.

En el capítulo 4, se propuso una metodología basada en la fusión temprana para resolver problemas de regresión en el ámbito de la gestión de incendios forestales. Para resolver este problema se disponía de tres fuentes de información, dos de ellas se utilizaron para crear las muestras, las variables Greenness Index (GI) y atmosféricas, y otra fuente de información se utilizó para crear las etiquetas, la información de los incendios. Combinando estas fuentes de información se consiguió diseñar un sistema de regresión capaz de predecir los recursos necesarios en caso de

incendio en una localización concreta de Castilla y León.

Como se ha mostrado a lo largo de esta tesis doctoral, esta investigación está muy centrada en la aplicación de técnicas de DL o relacionadas con la DL en otros dominios alejados de la DL, como la medicina o la desinformación. La aplicación de estas técnicas en dominios ajenos al DL es extremadamente difícil debido a la opacidad de estos sistemas y modelos. Por ello, un pilar de esta tesis, más concretamente del capítulo 5, ha sido el diseño de técnicas de visualización de mapas de calor que nos permitan mostrar las características que los modelos o sistemas han utilizado para generar el resultado final. Demostrando que dentro de las visualizaciones obtenidas mediante el algoritmo grad-CAM, se pueden obtener una gran variedad de visualizaciones en función de los problemas a resolver y de las preferencias de los usuarios finales. Demostrando que estas técnicas no sólo facilitan su aplicación en dominios externos sino que también nos permiten detectar errores y sesgos en nuestros sistemas debidos en gran medida a limitaciones debidos a los conjuntos de datos utilizados.

Por último, el capítulo 6 ha sido el punto de partida para la extensión de esta tesis a otras modalidades de datos, concretamente vídeo y audio. Debido a su naturaleza, son especialmente susceptibles de ser analizados mediante técnicas de ensemble y fusión de información. Para comprender adecuadamente el campo y descubrir las posibles lagunas que aún no han sido explotadas, se consideró que una revisión de las publicaciones de los últimos años nos permitiría hacernos una idea general del estado del dominio y descubrir en qué áreas podríamos centrarnos y aplicar las técnicas objeto de esta investigación. Esta investigación nos ha permitido comprobar que, al igual que en otros dominios, en el campo de la manipulación de información multimedia y multimodal, los conjuntos de datos aún presentan muchas debilidades y limitaciones que pueden superarse con las técnicas analizadas a lo largo de esta tesis.

Respuestas a las preguntas de investigación

Este apartado se centrará en dar respuesta a las diferentes preguntas de investigación planteadas en el Capítulo 1, gracias a los análisis experimentales realizados a lo largo de los Capítulos que componen esta tesis doctoral.

RQ1: *¿Pueden las técnicas de ensemble mejorar el rendimiento de los sistemas de clasificación binaria con conjuntos de datos limitados por el tamaño y la calidad de la muestra en el campo de la medicina?*

En el capítulo 2, se analizó el rendimiento de un enfoque ensemble que utiliza la media para combinar las probabilidades obtenidas de los distintos modelos creados desde cero, en comparación con los modelos individuales y con un modelo del estado de la técnica ampliamente utilizado en tareas similares, CheXNet. Además, también se aplicó la metodología presentada en otro conjunto de datos similar del estado del arte.

Los experimentos realizados mostraron una mejora en AUC del 11% en comparación con el mejor modelo individual propuesto y una mejora del 6% en TPR para el conjunto de datos propuesto en esta investigación. Si comparamos los resultados obtenidos por el ensemble (AUC = 0,92 y TPR = 0,73) con el modelo utilizado como punto de comparación del estado del arte, la mejora en AUC es del 16% y una mejora en TPR del 29%. En otras palabras, aunque el modelo del estado del arte presentó un AUC adecuado, aunque ligeramente inferior al esperado, no fue capaz de detectar correctamente la clase de interés, consolidación. Esto demostró que la técnica ensemble propuesta es capaz de obtener resul-

tados mejores y más robustos que los modelos individuales y el modelo del estado del arte establecido en esta investigación. En segundo lugar, se puede observar que los modelos creados a partir de cero pueden presentar resultados notables cuando el conjunto de datos no es lo suficientemente grande y no presenta suficientes características para ser analizado con modelos más complejos, como CheXNet. En segundo lugar, se utilizó otro conjunto de datos de última generación para comprobar la solidez del enfoque propuesto en esta investigación. El conjunto de datos presentaba dos problemas de clasificación diferentes: normal frente a neumonía y neumonía bacteriana frente a neumonía vírica. En el primer problema de clasificación, el sistema basado en ensemble mostró una mejora del 2% y del 6,8% en AUC y TPR, respectivamente, mejorando los resultados obtenidos en el artículo original. En el segundo problema de clasificación, aunque no se mejoró el TPR obtenido en el artículo original, se obtuvo una mejora en el AUC del 2,4%.

Los resultados obtenidos a lo largo de esta investigación muestran, en primer lugar, que el sistema basado en ensemble propuesto es capaz de superar problemas de clasificación binaria con conjuntos de datos que presentan dos debilidades, el tamaño reducido y la calidad limitada de las muestras. En segundo lugar, muestra resultados robustos que se mantienen para otros conjuntos de datos similares en el estado del arte. Esto confirma nuestra hipótesis inicial de que un sistema basado en conjuntos con modelos creados desde cero, menos complejos que los modelos habituales del estado del arte, es capaz de resolver problemas de clasificación binaria, aunque los conjuntos de datos no presenten las características esperadas.

RQ2: *¿Pueden los enesembles mejorar el rendimiento de los problemas de clasificación multietiqueta con conjuntos de datos con un desbalanceo extremo en el campo de la medicina?*

En el capítulo 3, se propone en esta investigación una metodología de aprendizaje profundo para tareas de clasificación con conjuntos de datos desequilibrados multietiqueta. Con esta metodología se ha construido un ensemble compuesto de cinco arquitecturas de última generación: DenseNet-201, EfficientNet B0, Inception, InceptionResNet y Xception. Se empleó la función de pérdida "Weighted crossentropy with logit" para mitigar el desequilibrio de los datos. En cuanto a las técnicas de combinación, hemos optado por utilizar dos enfoques específicos para problemas multietiqueta: Combine Then Predict (CTP) y Predict Then Combine (PTC), dentro de esta se han utilizado dos técncias PTC-mode y PTC-lw.

Los resultados de los experimentos son prometedores y han superado las expectativas. En primer lugar, a diferencia de los artículos existentes sobre el estado de la técnica, hemos establecido punto de referencia metodológicamente robusto para futuras investigaciones, independientemente de si se utilizan etiquetas específicas o generales. Este enfoque nos permite evaluar el rendimiento de estos modelos con un número variable de etiquetas. Nuestro sistema alcanza valores de AUC elevados para el número de clases utilizadas. En concreto, en el caso de las etiquetas específicas, el sistema muestra un rendimiento excepcional, con un AUC de 0,84. En el caso de las etiquetas generales, obtenemos un AUC de 0,819. Para las etiquetas generales, obtenemos un AUC de 0,819. Este valor inferior puede atribuirse a la amplitud de las clases y a la diversidad de signos radiológicos dentro de la clasificación general. En consecuencia, la variabilidad es mayor, lo que dificulta la clasificación.

Se han analizado los resultados obtenidos con las diferentes técnicas de combinación, el

enfoque CTP obtiene mejores resultados que los enfoques PTC, ya que es más informativo y en las clases en las que las probabilidades son intermedias el PTC no permitirá minimizar los errores individuales. Además, esta técnica ha sido capaz de mejorar los resultados individuales de los distintos modelos. Los resultados de esta investigación muestran cómo el uso de ensembles, con la técnica combinatoria adecuada, puede superar los problemas de los conjuntos de datos multietiqueta, como el desequilibrio extremo.

RQ3: *¿Pueden las técnicas de fusión de información ayudar a superar problemas de regresión en el ámbito de los incendios forestales, cuando el tamaño del conjunto de datos es extremadamente limitado?*

La arquitectura del Wildfire Assessment Model (WAM) se propone en el capítulo 4 para predecir los recursos de incendios forestales utilizando variables atmosféricas y el Índice de verdor. Debido al número restringido de muestras etiquetadas, se utilizó un autoencoder basado en autoaprendizaje con muestras no etiquetadas, lo que le permitió discernir tendencias y patrones dentro de las variables de un área geográfica específica. Este conocimiento adquirido se aplicó después a una tarea de regresión, que predice los recursos necesarios, el tiempo de control y extinción, y la superficie potencial quemada en caso de incendio. Además se examinó su capacidad de generalización a diferentes áreas con condiciones meteorológicas diferentes, utilizando un conjunto de datos de Andalucía. En último lugar, como la aplicabilidad del modelo a una localización concreta está limitada, se generaron mapas de predicción para cada etiqueta, ilustrando los recursos necesarios en diferentes lugares de la comunidad autónoma de Castilla y León para una fecha concreta.

Los resultados son prometedores y destacan el carácter innovador de la metodología en el ámbito de los incendios forestales. A diferencia de la mayoría de los artículos publicados que emplean exclusivamente algoritmos Machine Learning (ML), mientras que este estudio explora un enfoque de DL. En concreto, la arquitectura del autoencoder residual supera en rendimiento a la secuencial. Esto se atribuye principalmente a las conexiones, que ofrecen rutas alternativas para los gradientes durante el proceso, permitiendo a las capas más profundas extraer información de las capas iniciales. Significativamente, se obtuvieron mejores resultados cuando se aplicó la arquitectura residual tras reentrenar el codificador en la tarea de regresión. Esto se debe probablemente a que la introducción de nuevas muestras facilitó que el modelo de regresión captara patrones que el autocodificador inicial había pasado por alto en un principio.

Es importante señalar que nuestro modelo se entrenó inicialmente en una región específica, Castilla y León, que es una comunidad autónoma de España. En consecuencia, realizamos un análisis de su adaptabilidad a otras zonas con condiciones meteorológicas diferentes, como Andalucía, otra comunidad autónoma dentro del mismo país. Los resultados indicaron que el WAM predijo eficazmente dos etiquetas, tiempo de control y extinción. Sin embargo, no predijo correctamente el resto de etiquetas. En concreto, el WAM no pudo adaptarse a otras regiones sin un reentrenamiento. Como ya se ha mencionado, el modelo mostró un mejor rendimiento cuando se reentrenó el codificador en el contexto de la tarea de regresión, incluso cuando las muestras procedían de la misma región autónoma. La dificultad del modelo para generalizar a zonas con condiciones meteorológicas distintas puede atribuirse a su preentrenamiento original en una región específica.

Esta investigación ha demostrado que la aplicación de técnicas de fusión previa puede aumentar la información de las muestras, incluso si el conjunto de datos contiene un número

limitado de muestras. La combinación de la fusión de información con un entrenamiento previo con muestras no etiquetadas ha permitido resolver la tarea de regresión en Castilla y León con un conjunto de datos compuesto por 445 muestras. Sin embargo, no se ha podido generalizar a otras zonas como Andalucía, posiblemente porque el entrenamiento previo se ha realizado en una zona limitada con unas condiciones específicas.

RQ4: *¿Cómo pueden adaptarse las técnicas de XAI a diferentes problemas de clasificación binaria y multietiqueta para facilitar el uso de modelos DL en el ámbito médico?*

En el capítulo 5 se muestran diferentes técnicas de visualización basadas en mapas de calor generados mediante el algoritmo grad-CAM. Las diferentes técnicas se han generado para diferentes problemas de clasificación, binarios y multietiqueta, y para diferentes preferencias del usuario final, desde visualizaciones muy sencillas que sólo muestran las áreas utilizadas por el modelo para llegar al resultado final hasta técnicas más complejas donde se representa toda la información obtenida del modelo, es decir, la visualización de cada clase diferente, la probabilidad obtenida por el modelo o modelos y en caso de aplicación de técnicas ensemble el acuerdo entre los diferentes modelos (tanto numérica como visualmente).

Las técnicas de visualización presentadas en este capítulo se derivan de los problemas de clasificación de los capítulos 2 y 3, donde se resolvieron dos tareas de clasificación diferentes: binaria y multietiqueta. También se realizaron varias visualizaciones para cada problema de visualización, desde la opción más sencilla hasta la visualización más compleja posible, conteniendo la mayor cantidad de información posible. Esta amplia variedad de representaciones se realizó con el objetivo de representar todas las visualizaciones posibles dentro del espectro permitido, ya que cada usuario final puede tener diferentes preferencias. A lo largo de esta investigación ha sido posible validar estas visualizaciones por un grupo de médicos, tanto con médicos seniors como junios. Domínguez-Rodríguez et al. [1] ha analizado la concordancia entre los médicos y las visualizaciones. En el caso de los médicos seniors, coincidieron con las visualizaciones obtenidas por 70%, y en el caso de los residentes, en todos los casos analizados, su precisión en la lectura de las radiografías mejoró cuando se utilizaron las visualizaciones. Esto demuestra que la aplicación de estos sistemas en medicina puede mejorar el diagnóstico.

Sin embargo, es importante señalar que la calidad de las visualizaciones depende totalmente del rendimiento del sistema de clasificación, por lo que es muy importante que los sistemas tengan un alto rendimiento para obtener visualizaciones de clasificación.

RQ5: *¿Cómo pueden aplicarse las técnicas de fusión de información y ensembles analizadas en esta tesis a otras modalidades de información multimedia, dentro del dominio de los trastornos de la información?*

A lo largo de este capítulo se ha realizado una revisión del estado del arte de las técnicas de detección de manipulación en vídeo, audio y datos multimodales. El objetivo es aplicar las técnicas y conocimientos de esta tesis a otras modalidades de información multimedia, como el vídeo y el audio.

En cuanto a los conjuntos de datos en este campo, se han detectado una serie de limitaciones que pueden ser solventadas con las técnicas y conocimientos de esta tesis, tales como el desequilibrio entre clases, muchos conjuntos de datos presentan un mayor número de muestras falsas, especialmente en los conjuntos de datos de vídeo y multimodales. En segundo lugar, no todos los conjuntos de datos tienen el tamaño adecuado teniendo en cuenta

la complejidad del problema a resolver y la gran variedad de técnicas de manipulación que se pueden encontrar.

En cuanto a las técnicas de detección, hay una fuerte presencia de ensembles, no sólo aplicados a información multimedia multimodal, sino también en vídeo tratado como secuencias de imágenes. Por otro lado, la fusión de información, aunque presente en numerosos artículos, ha sido menos explorada en este dominio. Teniendo en cuenta la naturaleza de la información, la multimodalidad, es susceptible de ser analizada mediante técnicas de ensemble y fusión de información, lo que permitiría utilizar toda la información disponible, mejorar el rendimiento de los sistemas y la capacidad de generalización.

Trabajos futuros

A lo largo de esta disertación, hemos intentado resolver varios problemas derivados de las debilidades de los conjuntos de datos y facilitar su aplicabilidad en diferentes dominios; sin embargo, también hemos observado algunos puntos en los que esta investigación puede ampliarse:

- Dentro del análisis de radiografía torácica nos gustaría explorar la aplicación de ensembles y técnicas de fusión para combinar diferentes vistas de una misma radiografía. Es muy habitual en este dominio realizar diferentes ángulos para tener mayor información de los pulmones. Este enfoque nos permitiría aumentar la información disponible para los modelos sin aumentar el número de muestras y con información habitualmente disponible. Esto permitiría el rendimiento de cualquier tarea de clasificación, binaria o multietiqueta.

- Como se ha demostrado en el capítulo 3, forzar al modelo a centrarse en las áreas relevantes para el problema mejora el rendimiento, mejora la capacidad de generación y reduce los errores, por ello se quieren explorar nuevas técnicas como pueden ser los mecanismos de atención u otras técnicas de segmentación y recortes basadas en segmentación. Mediante el uso de estas técnicas se espera mejorar el rendimiento de las tareas de clasificación.

- En el Capítulo 4 hemos observado problemas de generalización a otros áreas con diferentes condiciones atmosféricas, para mejorar el rendimiento del modelo se quiere crear un autoencoder basado en aprendizaje autosupervisado con datos procedentes de un área más amplia que la comunidad autónoma de Castilla y León. Además se pueden incluir nuevas variables atmosféricas que permitan la aplicación del modelo a otros problemas de ecología, como pueden ser las sequías.

- Para mejorar el rendimiento del autoencoder del Capítulo 4 se quieren explorar nuevas opciones de arquitecturas para el encoder como por ejemplo transformers [2] o ConvNeXT [3]. La creación de nuevos modelos capaces de reconstruir de manera más precisa las representaciones será capaz de comprender en mayor detalle los patrones y las tendencias y por ello crear modelos de regresión más precisos.

- Otra línea de investigación dentro del campo de la fusión de información que deseamos explorar o desarrollar son técnicas que permitan conocer la relevancia de cada fuente de información, con el fin de seleccionar aquellas fuentes de información realmente relevantes para la tarea y evitar incluir información que puedan confundir o no aportar información relevante. Un ejemplo sería el análisis de las diferentes variables y su influencia en el problema que permita seleccionar o ponderar las diferentes fuentes de información.

- Dentro del campo de la explicabilidad, queremos explorar otros algoritmos y técnicas de

eXplainable AI (XAI) para mejorar la calidad de las visualizaciones. También queremos explorar la posibilidad de combinar diferentes algoritmos de XAI para visualizar mejor las características utilizadas en los diferentes modelos. Esto facilitaría la aplicabilidad de los sistemas en campos externos y también una herramienta para detectar posibles errores durante la generación de los sistemas.

- Se quiere desarrollar una metodología que permita aplicar las técnicas de XAI desarrolladas a lo largo de esta técnica en el proceso de desarrollo de modelos de DL que permita una detección precoz de errores y o selección inadecuada de las áreas de la imagen a lo largo del entrenamiento. Esta metodología permitiría conocer las debilidades de los modelos y dataset de manera rápida y más efectiva que el análisis de evaluaciones únicamente numéricas.
- En último lugar se quiere ampliar la aplicación de las metodologías de esta tesis a otros datos multimodales como puede ser el vídeo, por su componente visual y acústico, continuando la investigación realizada en el Capítulo 6, especialmente centrándonos en sistemas de detección de manipulación de grano fino, más específico e informativo que la mayoría de trabajos publicados en el área, que se centran en la clasificación a nivel de muestra.

Abstract

Image processing faces a serious problem, the deficiencies of datasets. In many situations, datasets do not have the quality necessary to be processed using DL and obtain the desired results. Therefore, researchers have had to explore different techniques to overcome these limitations without the need to extend the datasets, as in many situations this is impossible, such as in the field of medicine. Therefore, researchers have to explore different ways to create systems to solve image processing problems accurately and effectively.

In this thesis, different image processing problems will be addressed by means of information fusion and ensemble techniques. In addition, different explanatory AI techniques will be developed to allow researchers and end users of the systems to understand how these systems have arrived at the final solution and to understand whether they are using the right information to make that decision.

Ensemble techniques have first been used to solve classification problems within the medical field, more specifically by using chest radiographs. Within this domain, both binary classification systems and multi-label classification problems have been designed. On the other hand, due to the application domain, different visualisation techniques based on heat maps have been created to allow healthcare personnel to understand how the final result has been reached. Visualisation techniques are especially important in this domain due to the danger of an erroneous result of the classification system, which can cause harm to patients.

Secondly, information fusion techniques have been used to solve regression tasks within the wildfire application domain. A regression system based on Convolutional Neural Networks has been created that is capable of predicting the resources needed in the event of wildfires in an autonomous community in Spain. To facilitate the application of these techniques, visualisations were generated to predict the resources that would be necessary in the event of a fire in the entire study area on a daily basis.

Finally, with the aim of extending the area of application of image processing techniques using ensembles and information fusion techniques to other multimedia data modalities: video, understood as a sequence of images; audio and multimodal data, which combines the two previous ones. For this purpose, a review of the state of the art of manipulation detection techniques was carried out for the aforementioned multimedia data. For this purpose, both multimedia data manipulation techniques and detection techniques were analysed, which allows us to create a complete vision of the field and to analyse the possible areas that are less explored and more susceptible to be analysed with the techniques developed throughout this doctoral thesis.

Contents

List of Figures

List of Tables

List of acronyms

Part I

PhD Dissertation

Chapter 1

Introduction

Technology is our second nature.

— Lynn Margulis

This chapter, which heads this dissertation, aims to present the context of this research in order to provide the knowledge necessary to understand the underlying problem and the reasons that drive research are to offer knowledge and solutions. The first two sections provide an overview of the context and issues that have motivated this doctoral thesis. Section 1.1 describes the motivation of this work, whereas Section 1.2 defines the problem presented. Then, Section 1.3 presents the five Research Questions that this dissertation will try to answer. Section 1.4 shows the structure of this thesis, and finally Section 1.5 provides a summary of the main contributions and publications associated with this dissertation.

1.1 Context and Motivation

DL, since its origin in the 1950s, has drastically improved the state of the art in several tasks, such as speech recognition and computer vision. Within the field of image processing, since the early 2000s, CNN has demonstrated outstanding results on numerous problems, such as classification [4], regression [5], detection [6], and more recently image generation [7]. This has motivated many researchers to create new image datasets that allow the creation of systems based on deep learning that improve quality and performance. In other words, it is essential to have large and high-quality datasets to create more precise systems.

Although the trend in recent years has been towards an improvement in the quality of datasets, they still present a number of problems or weaknesses. DL systems, unlike ML algorithms, need a large number of samples, as can be seen in some of the most widely used datasets in the state of the art, such as ImageNet [8], CIFAR100 [9], MNIST [10], Fashion MNIST [11]. The datasets have to have a number of key qualities:

- ***Representativeness***: The dataset should be representative of the variability of real-world data that the model will eventually face. It should include a wide range of examples that cover different classes, scenarios, and conditions.

- ***Adequate size***: The dataset must contain a sufficient number of examples for the deep learning model to learn relevant patterns. In general, the more data there is, the better the performance of the model. However, it must be balanced with quality and computational capacity.

- ***Accurate and consistent labels***: Each example in the dataset must be correctly labelled or annotated. This is essential for supervised learning, where the model needs to know the correct answers to learn to make accurate predictions and generalise correctly.

- ***Class balance***: When dealing with a classification problem involving multiple classes, ensuring a balanced dataset with an equal distribution of examples per class is crucial. Imbalanced datasets can introduce bias and affect the model's performance. This is especially important in multilabel problems.

- ***Intra-class variability***: The dataset should capture the natural variability in real-world data. This means that the examples should represent different conditions, variations in lighting, noise, and other factors that the model may face in real-world situations.

- ***Visual quality and resolution***: samples should have adequate quality and resolution, allowing the model to learn relevant details.

All these attributes cannot always be satisfied in all datasets, so strategies and systems DL will have to be designed to overcome these difficulties. This research focusses on solving two image processing problems: the quality of the datasets, understood as the *number of samples and their quality* and the *class imbalance*. On the other hand, I will also focus on the development of new visualisations based on *XAI techniques*, which will facilitate the application of these systems in domains outside of Computer Science (CS). For this purpose, this thesis focuses on different tasks within image processing: classification, binary and multiclass, and regression. The objective is to overcome the main problems within the field that are limiting or hindering the development of new quality Deep Learning systems to solve different tasks within image processing in different application domains, such as medicine, ecology, or information disorders.

1.2 Problem statement

This work tries to tackle problems related to different image datasets from different perspectives through the application of DL techniques; more specifically, I will analyse different ***information fusion*** and ***ensembles*** techniques to overcome these difficulties. This research seeks to explore the potential of these techniques when used to address this issue, providing the necessary tools to combine them, and developing systems based on ensembles and information fusion for different image processing tasks. This research has focused more specifically on supervised learning tasks, which are those in which models are trained on labelled samples to subsequently predict classes of new unlabelled samples [12]. Specifically, focus on classification tasks, both binary and multilabel, and regression.

Two main approaches, ensemble and information fusion techniques, have been used to achieve the first objective of this work for the different tasks mentioned above. Ensemble techniques exploit the knowledge of several models to improve the overall performance of the system [13]. This approach is not only seen in the field of CS; for example, it is common for medical tests to be checked by several doctors to reduce the possibility of error, just as ensemble techniques

exploit the knowledge of different models. On the other hand, information fusion techniques combine different sources of information to provide the necessary features for the model to solve the task [14]. This approach increases the complexity of the information provided and reduces the need to include new samples to the original dataset.

The second objective of this dissertation is to facilitate the application of these DL-based systems outside the field of CS, through the use of visualisations that facilitate the interpretation of end users. Although it may seem easy, this task is not trivial. DL models and systems are considered "Black Box algorithms", the input and output of the system is known, but the underlying inference process cannot be understood. ***XAI techniques*** increase the confidence of the end user in DL models, reduce the risk of errors and help regulatory compliance, such as in the medical field, where decisions must be explainable and justifiable. These techniques also help in the field of CS, as they allow the finding of errors or biases that were not initially considered. The application of ensemble techniques to combine different models can make it difficult to create useful visualisations for end users, so it will be especially important to study in detail how to apply XAI techniques to systems based on information fusion or ensembles.

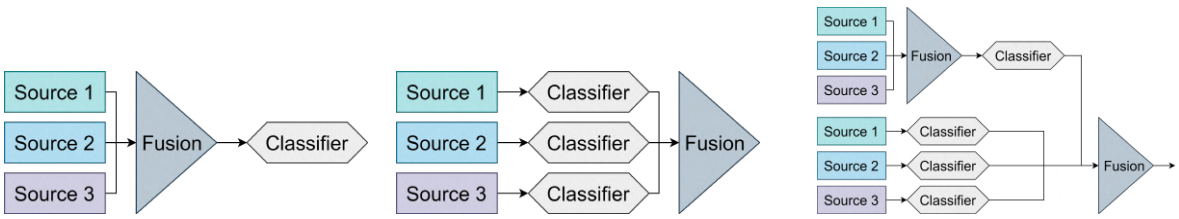
The DL systems that have been designed for this thesis are composed of different stages that form complex systems that can be divided as follows: Data processing, Design of the classification or regression system, Training and validation, and Creation of visualisations and testing. To study the use of ensembles and information fusion techniques to solve both main tasks, the problems of image datasets and the lack of explainability of DL models, this research has been structured around the following objectives, all of them aimed at extending current research:

- Create ensemble-based systems capable of overcoming intrinsic dataset problems for different classification problems.
- Explore the ability of information fusion techniques to improve the performance of regression problems for small datasets.
- Analyse different XAI-based visualisation techniques applied to different classification problems: binary and multilabel.
- Study the applicability of the techniques developed in other more complex problems, such as video, audio, or multimodal data.

1.2.1 Information Fusion

Information fusion techniques in the field of AI refer to methods and approaches used to combine and use information from multiple sources to improve the performance, reliability, and relevance of the results obtained by the systems. These techniques are essential in situations where available information is scattered across multiple sources and needs to be integrated in order to make accurate decisions. The flexibility to represent data at different levels of abstraction is considered the most critical aspect of the combinatorial approach. When an intermediate formalism is used, the learnt information can be combined into two or more modalities for a particular hypothesis [15]. To improve the generalisation performance of complex cognitive systems, it is necessary to capture and merge a suitable set of informative features from multiple modalities using standard techniques, Figure 1.1.

Depending on whether the fusion is generated before or after the classifier, we can divide the techniques into early, late, and hybrid fusion [16]. Early fusion, Figure 1.1a, corresponds to



(a) Early fusion **(b)** Late fusion **(c)** Hybrid fusion

Figure 1.1: Conventional methods for data fusion

fusion at the feature level. This technique stands out for providing a wealth of information due to the heterogeneity and disparity of information sources. The main inconvenience of this technique is that it can generate prediction errors by generating a single large representation [17]. Late fusion, Figure 1.1b, corresponds to the fusion of already processed information; each source of information will be processed independently, producing a prediction. Unlike the previous technique, late fusion combine the final results reducing the general performance of the system [18], this method is also known as ensembles, which will be explained in more detail in the next section. Finally, hybrid fusion, Figure 1.1c, is the most difficult to combine due to the intermediate representations of the different information sources, Joze et al. [19] designed a multimodal transfer module, MMTM, to hierarchically aggregate knowledge from multiple sources in CNN.

These techniques can be seen in two chapters of this dissertation. First, early fusion has been applied in Chapter 4, where two sources of information, the GI and different geolocated atmospheric variables, are combined. The result is a combination of nine matrices, each of which matches a variable or GI, creating pseudoimages that can be processed by CNN. We can also observe, in Chapter 6, a variety of information fusion techniques to combine visual information from videos with information from audio.

1.2.2 Esemble techniques

Ensemble or late fusion techniques are a set of methods that combine multiple ML models to improve predictive performance and generalisation. Instead of relying on a single model, ensembles use the collective knowledge of multiple models to make more accurate and robust decisions. Ensembles have been shown to be effective in a wide range of ML and DL problems. They have also been applied to a variety of tasks with excellent results [20]. The main ensemble techniques can be divided into two groups: simple and complex techniques. Simple techniques, although very basic, are still very common within the DL field and perform well in different tasks. These techniques use max-voting, average, and weighted average to combine the predictions of the different models that make up the ensemble [21]. In Chapter 2 the Averaging Ensemble technique was used to combine several CNNs created from scratch. Different classical ensemble techniques were also observed in Chapter 6, for processing multimodal and video samples.

Within advanced ensemble techniques, we can find classic ones, such as bagging, which involves training multiple models on different subsets of data and then combining their predictions; boosting, where models are run independently; boosting is based on the idea of iteratively improving model accuracy [22]. Specific ensemble techniques have also been designed for multilabel classification tasks, such as Nguyen et al. [23], which proposes two approaches depending on whether to combine the probabilities produced by the different models, called CTP, or to first predict the classes and then fuse to generate a global prediction, called PTC. The application of ensem-

ble techniques to multilabel classification tasks is an underexplored field due to its complexity compared to other classification problems. These tasks can be analysed at the level of individual labels or complete predictions, so Nguyen et al. [23] proposes two ensemble techniques within the PTC approach: PTC-label-wise voting (lw) where the max voting technique is applied at the label level; and PTC-mode, which applies the max voting technique at the complete prediction level; in case the maximum is not unique, the best-performing model is chosen. It was therefore decided to test three different ensemble techniques in Chapter 3, the CTP technique, and the two PTC techniques, both at the label and prediction level, to test the performance of each of the methods on a problem with a large number of labels (35 and 54 labels) and a degree of imbalance.

1.2.3 Explainable AI

XAI techniques refer to a variety of methods and approaches used to make artificial intelligence systems, especially deep learning and complex models, more transparent and understandable to humans. These techniques are developed with the aim of providing clear and coherent explanations of how artificial intelligence models make decisions or predictions [24]. Explainability in the context of DL is particularly relevant due to the inherent complexity of models. DL models, especially deep neural networks, consist of many layers of interconnected nodes, which allows them to learn complex patterns and high-dimensional data representations. While these networks are incredibly effective for many tasks, such as image recognition and natural language processing, their complexity also makes them difficult to understand and explain. These opaque algorithms are called black-box algorithms.

DL techniques have obtained very good results in different areas, processing different types of data: images [25], text [26], audio [27] or video [28]. However, without understanding the process of the models, their application is very limited. This set of techniques improves the understanding of models on two levels: first, it allows researchers to check whether models use the right features, whether there are biases or errors due to a poor training set, i.e. poor dataset or incorrect subset splits. Second, it allows us to apply DL models in fields outside CS, improving confidence and making them easier to use for end users [29].

Within the field of image processing, XAI techniques show which areas of the image are relevant to solve the task. The main XAI techniques within the processing field are: Local Interpretable Model-agnostic Explanations (LIME), Shapley Additive Explanations (SHAP), Gradient-weighted Class Activation Mapping (grad-CAM) and Saliency Maps [30]. The first two techniques are generic and are not specific to image processing. The SHAP technique, based on game theory, provides a coherent and global explanation for the predictions of the ML model. This technique assigns a contribution value to each feature, in the case of images, to each pixel, based on the concept of shapley values, indicating how much each contributes to the final result; however, it has a disadvantage, the high computational and time cost to obtain visualisations [31]. LIME is an XAI technique used to explain predictions at the local level for specific data instances. This technique focuses on local explanations by fitting local interpretable models to specific data points. The main limitation is the choice of the neighbouring points and the model, which can affect the visualisation quality [32].

The saliency map and grad-CAM techniques are unique XAI imaging techniques; however, they have a number of differences. The first technique, Saliency maps, is based on the calculation of the partial derivatives of the neural network output with respect to the input pixels of the sample,

highlighting areas relevant to the model at a general level [33]. The grad-CAM technique is an extension of the previous technique that improves the ability to localise specific areas of the sample that are relevant to the different classes of the model, i.e. it does not extract a general visualisation of the areas relevant to the system but will generate independent visualisations for each class. This method uses the output gradients of the last convolutional layer with respect to the target class to weight the activations of that layer. This allows the generation of an activation map of the class of interest, highlighting the precise regions of the image that contribute most to the classification of that particular class [34].

In this dissertation, it was decided to focus on the grad-CAM technique and how to adapt it to different classification problems, both binary and multiclass, and how to adapt it depending on the preferences of the end users; therefore, in Chapter 5, different heatmap techniques based on the grad-CAM technique are presented depending on the task and preferences of the users.

1.3 Research questions

The problems presented in the previous section present a number of issues and challenges that this thesis aims to address, with the objective of developing new ensemble-based systems and information fusion mechanisms for image processing problems. To solve the different problems related to image datasets presented in the previous section, we have focused on specific case studies from three different application domains: medicine, ecology, and information disorders. This section consists of a series of questions and objectives that will be solved throughout this dissertation. To achieve this goal, the different objectives have been articulated in five research questions that will be addressed and answered.

- **RQ 1.** Can ensemble techniques improve the performance of binary classification systems with datasets limited by sample size and quality in medicine field?
- **RQ 2.** Can enesembles improve the performance of multilabel classification problems with data sets with an extreme imbalance within the medical field?
- **RQ 3.** Can data fusion techniques help overcome regression problems within the forest fire domain, when the size of the dataset is extremely limited?
- **RQ 4.** How can XAI techniques be adapted to different binary and multilabel classification problems to facilitate the use of DL models within the medical domain?
- **RQ 5.** How can the techniques of information fusion and ensembles analysed in this dissertation be applied to other multimedia information modalities, within the domain of information disorders?

1.4 Structure of the thesis

This thesis is presented as a compendium of publications and is structured into two parts. **Part I**, describes the main lines of this work, providing an overview of the necessary context, summarising the main results obtained, and presenting a series of results emerging from the research. **Part II** contains four articles published in international journals, which provide a description of the central focus of this dissertation. These articles also serve as the core for Part I. The first part is composed of the following chapters.

- **Chapter 1: Introduction.** This section provides a contextualisation and presentation of the issues on which this dissertation is based. Then a series of proposed research questions, the structure that this dissertation will be structured to follow. Finally, a list of publications that integrate this research.

- **Chapter 2: Ensembles and deep learning techniques for poor-quality datasets.** This chapter will cover the problem of small size and low quality datasets in binary classification tasks. This chapter is related to RQ1: "*Can ensemble techniques solve the problem of dataset size and quality in binary classification tasks?*". To solve this problem, an ensemble technique will be used with models created from scratch due to the limited information available in the dataset. For this purpose, a paediatric chest X-rays dataset composed of 950 samples was used [35].

- **Chapter 3: Multilabel imbalance datasets and esembles techniques.** Datasets with a large number of classes, such as the one used in this chapter consisting of 35 and 54 different classes, often lead to an extreme imbalance between them, which can make it difficult to perform classification tasks. This chapter proposes the use of ensembles together with pre-trained state-of-the-art models to overcome these difficulties. This research is directly related to RQ2: "*Can ensemble techniques improve multilabel imbalance classification problems?*" For this purpose, we have studied different ensemble techniques specialised in multilabel classification tasks. For this we used a dataset of adult chest X-rays [36].

- **Chapter 4: Information fusion for regression tasks.** In this chapter I will explore how information fusion techniques, more specifically early fusion techniques, allow one to improve the quality of datasets to obtain DL models with better performance, especially when the dataset has a low number of samples. For this purpose, a dataset related to forest fire resource management has been created where three sources of information have been combined: the GI, atmospheric variables, and wildfire information. This allows us to obtain adequate results for a dataset with fewer than 500 samples. This chapter is related to RQ3: "*How do information fusion and ensemble techniques help regression tasks?*".

- **Chapter 5: Explainable Artificial Intelligence for image classification** DL models are considered black-box algorithms due to their opacity. For this reason, their application in domains outside CS is not a trivial issue, XAI techniques are needed to explain how the models have reached the final result. For this reason, in this dissertation we have explored different heatmap generation techniques that facilitate their application and increase their confidence; for this purpose, we have created different visualisation techniques for chest X-ray classification problems. This chapter is directly related to RQ4: *"How can we take advantage of techniques to facilitate the usability of classification and regression tasks?*". In addition to creating different heatmap techniques within the medical domain [35, 36, 37], the effectiveness of XAI techniques has been validated in a hospital, demonstrating that the use of DL models together with XAI techniques can have benefits for the medical field [1].

- **Chapter 6: Application of information fusion, ensembles and XAI to other domains.** Finally, we wanted to expand the field of study to other modalities in addition to images, specifically video (understood as a sequence of images), audio, and multimodal. To this end, we began by reviewing the state of the art in the field of manipulation detection in these modalities, allowing us to create a complete vision of the field and establishing a starting point for further research. This will allow the application of ensembles, information

fusion, and XAI techniques to different classification, detection, or regression tasks. This chapter is directly related to RQ5: "*How can these techniques of information fusion be applied to other modalities such as video?*".

- **Chapter 7: Conclusions and future work.** This is the last chapter of the dissertation, which aims to present the conclusions based on the results and findings obtained in the course of this dissertation. Different possible avenues for future work are also presented to continue this research.

1.5 Publications of the compendium and Contributions

In this section, the compilation of publications that form the basis of this thesis is presented, with detailed descriptions of the quality indices and contributions made by the PhD candidate for each of them.

J1: Liz, H., Sánchez-Montañés, M., Tagarro, A., Domínguez-Rodríguez, S., Dagan, R., & Camacho, D. (2021). Ensembles of Convolutional Neural Network models for pediatric pneumonia diagnosis. Future Generation Computer Systems, 122, 220-233.
DOI: 10.1016/j.future.2021.04.007
Impact factor: 7.187 (JCR, 2021) [Q1, 10/110 CS, Artificial Intelligence]

- **Overall contribution**: This article presents a new approach to diagnosing paediatric pneumonia using ensembles of Convolutional Neural Network (CNN) models and explainable AI XAI techniques. Ensembles of models are used to improve classification performance, as presented in Chapter 2 while XAI techniques are used to overcome the lack of interpretability in CNN "black-box" algorithms. Specifically, the authors propose a new XAI technique based on combining individual heatmaps obtained from each model in the ensemble, which highlights the areas of the image that are most relevant to generate the classification, as presented in Chapter 5.

- **Contribution of the PhD candidate**:

  - First author of the article.

  - Contribution to the conception of the presented idea.

  - Design and execution of the experiments.

  - Co-author of the interpretation and discussion of the results provided.

  - Elaboration of the manuscript and visualisations.

J2: Liz, H., Huertas-Tato, J., Sánchez-Montañés, M., Del Ser, J., & Camacho, D. (2023). Deep learning for understanding multilabel imbalanced Chest X-ray datasets. Future Generation Computer Systems, 144, 291-306.
DOI: 10.1016/j.future.2023.03.005
Impact factor: 7.5 (JCR, 2022) [Q1, 10/110 CS, Artificial Intelligence]

- **Overall contribution**: This article proposes a methodology for improving the performance imblalance classification tasks. The authors address the challenges associated with imbalanced multilabel datasets by applying ensemble techniques of models and

using a specific loss function for imbalanced data, as presented in Chpater 3. The ensemble technique involves training multiple pretrained models and combining their predictions to obtain a final result. The authors also introduce a heatmap-based visualisation technique to highlight the most important areas for detecting each disease represented in the dataset. This technique generates a report that includes the visualisation of the heatmap, the probability produced by the system, and the agreement between the ensemble models, which is described in Chapter 5.

- **Contribution of the PhD candidate**:
  - First author of the article.
  - Contribution to the conception of the presented idea.
  - Implementation of the experiments.
  - Co-author of the interpretation and discussion of the results provided.
  - Elaboration of the manuscript and visualisations.

J3: Generation and detection of manipulated multimodal audiovisual content: advances, trends and open challenges". Helena Liz-López, Mamadou Keita, Abdelmalik Taleb-Ahmed, Abdenour Hadid, Javier Huertas, David Camacho. Information Fusion. ISSN: 1566-2535, eISSN: 1872-6305. Vol. xx, pp. 1–44, 2023. Submitted, 26th July 2023.
DOI: url10.1016/j.inffus.2023.102103
Impact factor: 18.6 (JCR, 2022) [Q1, 4/145 CS, Artificial Intelligence]

- **Overall contribution**: The main contribution of this article is a comprehensive review of the current state of the art in multimodal multimedia data forensics, specifically focuses on audio and video data (visual and multimodal approach), as described in Chapter 6. It includes manipulation techniques, available datasets, and detection tools. The article also presents a summary of the main challenges and future trends in the field.

- **Contribution of the PhD candidate**:
  - First author of the article.
  - The conception of the idea presented.
  - Elaboration of the methodology.
  - Co-author of the manuscript, figures, and tables.
  - Co-author of the interpretation and discussion of the results provided.

1.6 Other publications and Contributions

1.6.1 Journal papers

J4: Domínguez-Rodríguez, S., Liz-López, H., Panizo, A., Ballesteros, Á., Dagan, R., Greenberg, D., ... & Camacho, D. (2023). Testing the performance, adequacy, and applicability

of an Artificial intelligent model for pediatric pneumonia diagnosis. Computer Methods and Programs in Biomedicine, 107765.
DOI: 10.1016/j.cmpb.2023.107765
Impact factor: 6.1 (JCR, 2022) [Q1, 15/110 CS, Theory & Methods; Q1, 22/96 Engineering, Biomedical; Q1, 7/31, Medical Informatics]

- Overall contribution: A study is conducted on the use of artificial intelligence in the diagnosis of paediatric pneumonia. The main contributions of the study are the demonstration that convolutional neural networks (CNNs) can be effective in detecting pneumonia in chest X-rays of children, and the identification of the most important features for pneumonia detection in these images. Furthermore, the study provides a comparison of the results of pneumonia diagnosis with artificial intelligence with traditional methods and discusses the clinical implications of these findings for the diagnosis and treatment of pneumonia in children.

- Contribution of the PhD candidate:

  – Second author of the paper.

  – Implementation of the presented system.

  – Co-author of the manuscript, figures and tables.

J5: "Spain on Fire: A novel wildfire risk assessment model based on image satellite processingand atmospheric information". Helena Liz-López, Javier Huertas-Tato, Jorge Pérez-Aracil, Carlos Casanova-Mateo, Julia Sanz-Justo and David Camacho. Knowledge-Based Systems. ISSN: 0950-7051, eISSN: 1872-7409. Volume xx, 2023, 109265, pp. 1-19. **Submitted, 26th July 2023**. Impact factor: 8.8 (JCR, 2022) [Q1, 19/145 Computer Sicence, Artificial Intelligence]
Impact factor: 8.8 (JCR, 2022) [Q1, 19/145 Computer Sicence, Artificial Intelligence]

- **Overall contribution**: The main contributions of this article are the presentation of a forest fire risk assessment model that uses satellite image processing and atmospheric data to predict the impact of forest fires, as presented in Chapter 4. The proposed model can accurately estimate the costs of wildfire management and take proactive measures to prevent wildfires.

- **Contribution of the PhD candidate**:

  – First author of the article.

  – The conception of the idea presented.

  – Design and execution of the experiments.

  – Co-author of the interpretation and discussion of the results provided.

  – Elaboration of the manuscript and visualisations.

1.6.2 Conference papers

C1: "A Computer-Assisted Tool can categorize radiographs with pneumonia in childhood". C. Moraleda, S. Domínguez-Rodríguez, A. Panizo-LLedot, H. Liz, R. Dagan, D. Greenberg,

P. Rojo, M. Serna-Pascual, L. Gutiérrez, D. Camacho, A. Tagarro. 38th Annual Meeting of the European Society for Paediatric Infecftious Diseases (ESPID 2020). Rotterdam, the Netherlands, 26-29 October 2020. Accepted (poster)

- Overall contribution: Exploration of different DL models for paediatric chest x-ray classification, including state of the art models, such as VGG16, and a model from scratch. In this initial approach we mazimise the sensitivity for consolidation class. In addition, a basic heatmap-based visualisation technique was designed to enable clinicians to understand the DL models.

- Contribution of the PhD candidate:

  - Contributions made in the design and execution of the experiments.

  - Co-author of the manuscript, figures and tables.

  - Co-author of the interpretation and discussion of the results provided.

C2: Huertas-García, Á., Liz, H., Villar-Rodríguez, G., Martín, A., Huertas-Tato, J., & Camacho, D. (2022, July). AIDA-UPM at semeval-2022 task 5: Exploring multimodal late information fusion for multimedia automatic misogyny identification. In Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022) (pp. 771-779).

- Overall contribution: Exploration of different late information fusion methods to improve the performance of the combination based on a Transformer-based model and convolutional neural networks (CNN) for text and image, respectively, in the task of automatic identification of misogyny in multimedia.

- Contribution of the PhD candidate:

  - Second author of the paper.

  - Contributions made in the conception of the idea presented.

  - Contributions made in the design and execution of the experiments.

  - Co-author of the manuscript, figures and tables.

  - Co-author of the interpretation and discussion of the results provided.

C3: Liz-López, H., Huertas-Tato, J., & Camacho, D. (2023, September). Transparency in Medicine: How eXplainable AI is Revolutionizing Patient Care. In 2023 International Conference on Network, Multimedia and Information Technology (NMITCON) (pp. 1-6). IEEE.

- Overall contribution: This paper analyzes different visualization options, based on XAI techniques, that favor the interpretability of different classification problems: binary, multi-class and multi-label. This paper focuses on both single-model and ensemble-based systems.

- Contribution of the PhD candidate:

  - First author of the article.

- The conception of the idea presented.
- Analysis of the main XAI techniques for classification tasks.
- Co-author of the manuscript, figures, and tables.
- Co-author of the interpretation and discussion of the results provided.

C4: Transparency in Ensembles: XAI-Based visualisation for Understanding Multilabel Classification from Within. Helena Liz-López, Javier Huertas-Tato, David Camacho. In 2023, The 12th International Conference on Smart Media and Applications (SMA). **Submitted, 15th September 2023**.

- Overall contribution: The main contribution of this paper is a new visualization technique for ensemble-based multilabel classification systems. This visualization technique shows the most relevant areas for the different models, the areas that present higher variability, are less certain, and the agreement of the models in the predictions.
- Contribution of the PhD candidate:
  - First author of the article.
  - The conception of the idea presented.
  - Elaboration of the methodology.
  - Co-author of the manuscript, figures, and tables.
  - Co-author of the interpretation and discussion of the results provided.

Chapter 2

Ensembles and Deep Learning techniques for poor-quality datasets

Don't adventures ever have an end?
I suppose not.
Someone else always has to carry on the story.

— J.R.R. Tolkien

DL classification algorithms have been used in different domains, including those related to medicine, with excellent performance. However, as explained in Chapter 1, the available data do not always have the number of samples and quality necessary to apply DL techniques, which can compromise the results of these algorithms. This chapter presents an **ensemble-based classification system capable of solving binary classification tasks for limited datasets**: low sample number and limited quality. A dataset of paediatric chest radiographs with fewer than a thousand samples was used for this purpose.

2.1 Problem definition and objective

This chapter proposes an ensemble-based system composed of five models designed from scratch to solve a binary classification problem of paediatric chest X-rays, whose dataset has two main limitations: a limited number of samples, fewer than a thousand samples; and a low quality of the samples. On a medical level, it presents two challenges: the age range of the patients is very wide, between one month and 16 years, which will lead to a large variability in the samples; and the classification will be between the presence of consolidation or other infiltrates, which can be difficult to distinguish. This chapter focusses on solving the ***RQ1: "Can ensemble techniques improve the performance of binary classification systems with datasets limited by sample size and quality in medicine field?"***

The hypothesis proposed is that an ensemble-based system, built with five low-complex architectures created from scratch, will perform better on the dataset than more complex state-of-the-art

models, such as CheXNet. For this purpose, two experiments have been proposed: the first compares the performance of different ensemble-based systems, with different hyperparameters, and CheXNet; second, the performance of the proposed system was tested with a similar dataset and compares the results with those of the original research.

To overcome the limitations imposed by the dataset, I will focus on two potential training elements: the architecture of the system and the pre-processing of the samples. Regarding the *architecture*, an ensemble composed of five CNN was chosen, with a low number of convolutional layers. This system, being less complex than most state-of-the-art models, will favour generalisability. The second reason for creating the model from scratch was that most state-of-the-art models focus on adult patient x-rays, which differ significantly from paediatric chest x-rays. Concerning the preprocessing of the samples, the number of channels was reduced in order to reduce redundancy, since, being black and white images, all three channels contain the same information. Second, *data augmentation* was applied to improve the generalisability of the system. Due to the low number of available samples, it was applied quite significantly.

In short, this research can be divided into the following phases.

- Improving the quality of the dataset through preprocessing techniques.
- Design of an ensemble-based system for binary classification tasks for chest x-rays.
- Verification of performance improvement compared to the state of the art.
- Testing its performance on other datasets, coming from the state of the art.

2.2 Datasets

In this research, two different datasets were used to carry out the experiment. The first one, the XrPP dataset, is a private dataset provided by Ben-Gurion University of Israel; and the second one, published by Kermany et al. [38]. Both datasets correspond to paediatric chest x-rays and have been meticulously annotated by expert professionals and are classified into different classes.

The first dataset, XrPP, is composed of 950 annotated chest x-rays, ranging in age from one month to 16 years. Within this dataset, there are 403 and 547 samples of consolidation and non-consolidation, respectively. An expert panel composed of two senior paediatricians and a radiologist from the Hospital 12 de Octubre Research Institute (Madrid, Spain). Although the classes are balanced, the number of samples is quite limited, and the average image size is approximately 200,000 pixels, which is quite low for x-rays, indicating that the resolution of the samples is also limited.

- **Consolidation**: this class includes samples that show signs of consolidation, corresponding to *alveolar pneumonia*.
- **Non-consolidation**:, denoting image samples with signs of other infiltrates, corresponding to *non-alveolar pneumonia*.

The second dataset published by Kermany et al. [38] contains a total of 5,857 chest X-rays, with an age range of one to five years. This age range is much narrower, which reduces the variability within the dataset. Unlike our dataset, this has two different classification tasks: pneumonia and normal (1,583 samples), i.e., healthy patients; and bacterial pneumonia (2,780 samples) versus

viral (1,493 samples). Another difference from the original dataset of this work is the average size of the images, between 1,000,000 and 2,000,000 pixels.

2.3 Methodology

This section will explain the process used to create the classification system proposed in this research. The phases that composed it are: ***preprocessing***, where the images will be normalised and apply data augmentation; ***architecture design***, where I will create different CNNs; ***application of ensemble techniques***, which will combine the predictions of the different models; and external validation, where the performance of the system will be checked on another dataset.

2.3.1 Preprocessing

The X-ray images provided by the medical centres were in JPG format, which encodes colour information using three channels known as the "RGB" components. However, in our dataset, these RGB components are redundant for greyscale images, so only the first component was used. Additionally, the original images vary in size, so I standardise them to a resolution of 150x150 pixels.

Furthermore, to ensure uniformity in the pixel values of each image, I normalise them by dividing by the average pixel value of the respective image. As mentioned earlier, Convolutional Neural Networks (CNNs) typically require a large number of training images to avoid overfitting. Unfortunately, our dataset is relatively small. To address this limitation, a popular technique called Data Augmentation was used, which allows us to expand our dataset effectively [39].

Data Augmentation involves generating batches of images with real-time data modifications. During each training epoch, a diverse set of variations is created from the original training images using various types of transformation [40]. In this study, shear transformations (0.2), zooming transformations (0.05), rotation (0.2), horizontal shifting (0.1), vertical shifting (0.1), and horizontal flipping transformations were applied, all with a batch size of 32.

In the last step, to thoroughly assess the performance of our diverse model architectures, I established multiple sets of training, validation, and test partitions for the dataset. First, the dataset is divided into two subsets: train (70%) and test (30%) using stratified partitioning. The train subset was then divided into train (80%) and validation (20%) using the same partitioning technique. The training subset will be employed to train the CNN by adjusting its weights, while the validation subset will serve the purpose of continuously monitoring the CNN's metrics during the model's learning process, thereby preventing overfitting. Ultimately, the test subset will be used to assess the model's generalisation capabilities, providing an evaluation of its performance when handling new radiographs.

2.3.2 Convolutional Neural Networks architectures

In the CNN model selection, different architectures were explored, focusing on a balanced range of convolutional layers, within the 3-4 layer range. This choice is made because a smaller number of convolutional layers may not capture all the relevant image features, while too many can increase the risk of overfitting.

Each convolutional layer in our architectures is equipped with 32 kernels and employs a Rectified Linear Unit (ReLU) activation function. The output of the final convolutional layer is then

flattened and subjected to Dropout with a rate of 70%. Subsequently, this processed information is passed through a dense layer with a variable number of neurones depending on the specific architecture, always with a ReLU activation.

Finally, the classifier of our CNN has a dense layer with two neurones and uses Softmax activation for the classifier. This results in a total of six distinct architectures, each with unique layer configurations. To enhance the model's generalisation and reduce overfitting, Kernel L2 regularisation was applied with a strength of 0.01 to the FC dense layer. During training, each CNN is optimised using the Adam optimiser with a learning rate set to 1e-4. These practises collectively contribute to the building of robust and effective CNN models for our classification task. All parameters used in the training are summarised in Table 2.1.

CheXNet is the state-of-the-art model selected as a baseline. This model was designed for an adult chest X-ray classification problem with 14 different classes, corresponding to 14 thoracic diseases, including pneumonia. Due to the similarity and relevance to the state of the art, CheXNet was selected as the baseline.

The baseline is designed for a different classification task, so CheXNet will have to be modified to compare it with the proposed system. First, I adjusted the number of neurones in CheXNet's classifier from 14 (the number of classes in the original paper) to two, aligning it with the two classes in our work. Once this final layer was modified, the remaining layers were froze in CheXNet, except for the last two layers. The last two layers were then retrained using our target dataset.

This approach, known as transfer learning, involves taking a pretrained model from a related domain and fine-tuning it with the current dataset. It is a widely used strategy in Deep Learning to leverage knowledge from existing models and adapt them to specific tasks, consolidation vs. non-consolidation classification problem, ensuring a valid and robust comparison between the two models.

2.3.3 Ensemble technique application

To enhance the performance and robustness of our system, an ensemble-based strategy was adopted. Each ensemble is composed of five distinct CNNs, each built using a different dataset partition with varying training/validation subsets while maintaining the same test subset. This deliberate diversity among models is essential to enhance ensemble performance [41].

The partitioning process involved the following steps: Initially, I randomly divided the dataset into a construction subset (70%) and a test subset (30%). Subsequently, I generated five separate sets of training/validation partitions within the construction subset, using a 80% / 20% split. For each of these five partitions, an individual CNN was trained from scratch. Then I calculated the predictions for the test subset using the ensemble formed by these five CNNs and evaluated their performance metrics. The ensemble's predictions for the probability of consolidation (or non-consolidation) were computed by averaging the probability predictions from the five CNNs within the ensemble.

In summary, to ensure robustness and performance for the ensemble, I repeated this entire process with five different construction/test divisions. Consequently, for each division, I built a total of five CNN models, resulting in the construction of a total of 25 models.

Table 2.1: Summary of parameters used in the architectures.

Topology	
# Convolutional layers	[3,4]
# Neurones in Dense layers	[64,128,256]
# Neurones in convolutional layers	32
Block configuration	
Activation function (convolutional layers)	ReLu
Activation function (fully connected layers)	ReLu
Activation function (classifier)	Softmax
Dropout	70%
Regularisation	
Regulariser (type)	L2
Regulariser (strength)	0.01
Optimisation	
Optimiser	Adam
Learning rate	1e-4
Maximum epochs	500
Batch size	32

2.4 Experimentation

Six different CNN architectures were evaluated, as detailed in Table 2.2. These architectures were compared with CheXNet, which was re-trained in the dataset by transfer learning. Table 2.3 shows the performance for the ddiferent models. AUC and TPR were computed for all architectures.

Table 2.2: Configuration of the different architectures.

	Arch 1	Arch 2	Arch 3	Arch 4	Arch 5	Arch 6
# convolutional layers	4	4	4	3	3	3
# neurones in Dense layers	64	128	256	64	128	256

Regarding the baseline, although the AUC is similar to that obtained in the proposed architectures, the TPR is significantly lower than expected. These results may be due to the differences between the x-rays of adults and children; and to the fact that in this task I classify between consolidation and other infiltrates, which may be a more complex task than the original one. Secondly, we can see how architecture 1 presents the best performance, with an improvement of 2-5% in the AUC and 5-7% in the TPR, with respect to the rest of the architectures.

To perform a comprehensive analysis of Arch 1, the dataset was divided into five distinct sets of training and validation/test subsets, resulting in five unique training/validation splits. Subsequently, I trained a model from scratch using Arch 1 for each of these training/validation/test

Table 2.3: AUC and TPR values of our six architectures and CheXNet.

Arch	AUC	TPR
Architecture 1	**0.80 ± 0.03**	**0.62 ± 0.04**
Architecture 2	0.77± 0.02	0.56± 0.06
Architecture 3	0.78± 0.02	0.55± 0.08
Architecture 4	0.76± 0.02	0.57± 0.01
Architecture 5	0.75± 0.01	0.55± 0.07
Architecture 6	0.77± 0.02	0.55± 0.09
CheXNet	*0.76± 0.02*	*0.43± 0.08*

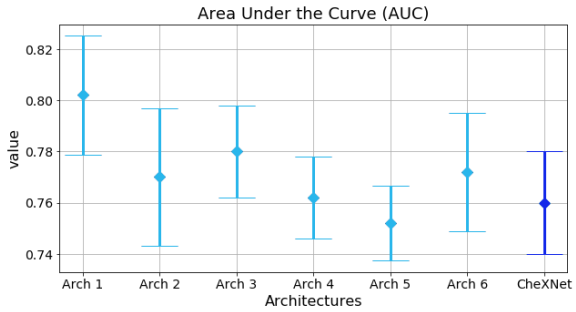


Figure 2.1: AUC values for CheXNet-based model (dark blue) and the CNN models trained from scratch (light blue).

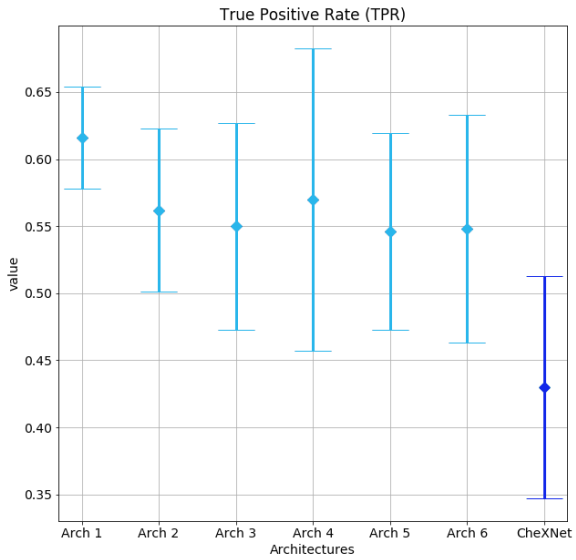


Figure 2.2: TPR values for CheXNet-based model (dark blue) and the CNN models trained from scratch (light blue).

partitions, resulting in a total of 25 models. This approach allowed to assess the robustness of the architecture across the different partitions, as illustrated in Table 2.4 and Figures 2.3 and 2.4. They clearly demonstrate that the performance of the architecture remains highly consistent across train and validation/test partitions, with minimal variance in the metrics. As a result, I can confidently conclude that this architecture shows robustness when applied to our dataset.

Table 2.4: AUC and TPR values for architecture 1 across the five different training and validation/test partitions. For each of them, five different training/validation splits were generated. A total of 25 different models are considered.

	AUC	TPR
Partition 1	0.78 ± 0.01	0.60 ± 0.05
Partition 2	**0.81 ± 0.01**	**0.67 ± 0.06**
Partition 3	0.80 ± 0.02	0.62 ± 0.07
Partition 4	0.79 ± 0.01	0.64 ± 0.10
Partition 5	**0.80 ± 0.02**	**0.71 ± 0.09**
Average	0.80 ± 0.01	0.65 ± 0.04

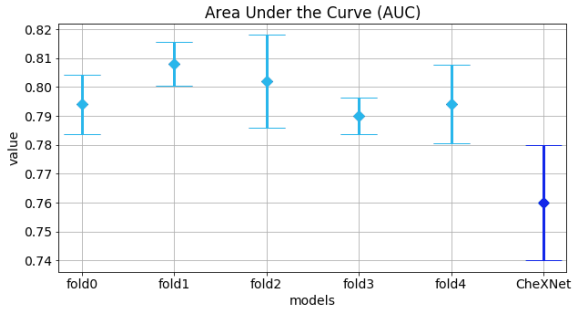


Figure 2.3: AUC for CheXNet (dark blue) and our different models from architectgure 1 (light blue).

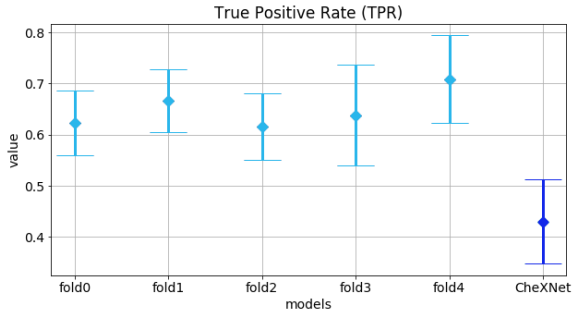


Figure 2.4: TPR for CheXNet (dark blue) and our different models from architectgure 1 (light blue).

Table 2.5: AUC and TPR values for Arch 1 ensembles.

	AUC	Improvement	TPR	Improvement
Partition 1	**0.89**	**11%**	**0.71**	**11%**
Partition 2	**0.92**	**11%**	**0.73**	**6%**
Partition 3	0.88	8%	0.65	3%
Partition 4	0.88	9%	0.73	9%
Partition 5	0.87	7%	0.79	8%
Average	0.89 ± 0.02	9%	0.72 ± 0.04	7%

Although the results of the individual models are promising, they are still lower than desired, especially considering that a model error could cause damage to patients. Finally, five ensembles were built, one per train and validation/test partition, with the aim of improving the individual models. Table 2.5 shows that in the case of Ensemble 2, the improvement is 11% for the AUC and 6% for the TPR, reaching an AUC of 0.92 and a TPR of 0.73. This shows how the proposed approach is able to overcome some intrinsic problems of the datasets and obtain satisfactory results considering the complexity of the problem.

To test the performance of the proposed approach, using a public dataset provided by Kermany et al. [38]. As the original article proposed two classification problems: normal versus pneumonia and bacterial versus viral pneumonia. To check the ensemble, it is compared with the results of the original article and the individual models. In Table 2.6, the results pertaining to the classification problem between normal and pneumonia cases. It is evident that individual architecture 1 models achieve AUC values that are in line with those reported by the authors. However, individual models show a slightly lower TPR (0.91 compared to 0.932) in comparison. Remarkably, the ensemble composed of the individual models delivers exceptional results, with an AUC of 0.976 and a TPR of 1.0. This underscores the robustness and performance of the ensemble compared to the individual models and even the original results, which based on DenseNet and transfer learning.

Table 2.7 shows the results for the classification problem that distinguishes between cases of bacterial and viral pneumonia. Once again, we can observe that individual architecture 1 models achieve AUC values that are comparable to the original one. However, the ensemble achieves an AUC of 0.964. These results are particularly noteworthy, as they demonstrate that the simple CNN ensembles outperform very deep CNN transfer learning techniques, such as those presented by Kermany et al. [38], where transfer learning was used with a DenseNet architecture with 121 convolutional layers. This highlights the efficacy of our approach in achieving superior AUC performance in this classification task.

Table 2.6: Kermany et al. dataset, normal versus pneumonia classification: comparison of AUC and TPR values originally reported by Kermany et al. to results obtained by our models.

	AUC	TPR
Kermany et al. model	0.968	0.932
Individual Arch 1 models	0.966 ± 0.007	0.91 ± 0.02
Arch 1 ensemble	0.976	1

Table 2.7: Kermany et al. dataset, classification bacterial versus virical: comparison of AUC and TPR values originally reported by Kermany et al. to results obtained by our models.

	AUC	TPR
Kermany et al. model	0.940	0.886
Individual Arch 1 models	0.94 ± 0.02	0.78 ± 0.05
Arch 1 ensemble	0.964	0.791

Chapter 3

Multilabel imbalance datasets and ensembles techniques

No water, no life. No blue, no green.

— Sylvia Earle

Another problem that can be found in datasets, as explained in Chapter 1, is imbalance, a situation in which data are unevenly represented and this can cause the classification system to be biased towards majority classes. Therefore, this chapter proposes an ensemble-based classification system for datasets with extreme imbalance. For this purpose, a dataset consisting of more than 30 different classes has been used.

3.1 Problem definition and objective

This chapter proposes an ensemble-based system composed of five state of the art pretrained models. For this purpose, a dataset of adult, multilabel, and extremely unbalanced chest X-rays has been used. The imbalance of the classes is due to an unequal representation of the different classes in the real world. In medicine, this is a common occurrence, as not all diseases have the same incidence in the population. The number of classes in the dataset is extremely high, as it tries to detect a wide variety of different radiological signs, which are not mutually exclusive. This will increase the complexity of the problem and also exacerbate the imbalance of the classes. This chapter is directly related to and aims to resolve the ***RQ2: "Can enesembles improve the performance of multilabel classification problems with data sets with an extreme imbalance within the medical field?"***

The hypothesis of this research is that an ensemble-based system, composed of five architectures widely used in the state of the art, together with a suitable preprocessing will be able to solve the multilabel classification task. The dataset used to perform this research contains a term tree that organises the different radiological signs into more generic categories, so it was proposed to create two classification systems: the first one for the radiological signs and the second one for the generic classes, see Figure 3.1. In addition, two further experiments were designed to test the effectiveness of the proposed preprocessing.

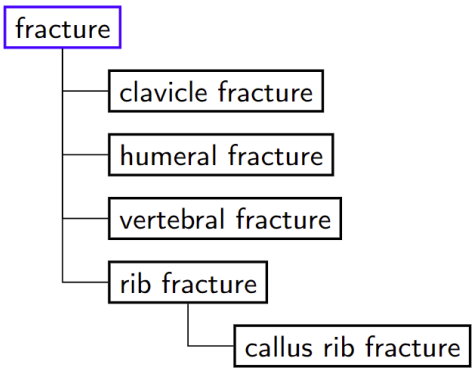


Figure 3.1: An illustration of a segment of the dataset's term tree, where general labels are enclosed in blue boxes, while specific labels are indicated in black.

To solve the problems of the dataset, a high number of classes, and extreme imbalance, we will focus on the same elements that were the focus of Chapter 2, the preprocessing and the design of the architecture. Regarding *preprocessing*, we will apply a segmentation-based cropping technique to eliminate non-relevant areas and force the system to focus on the areas where the radiological signs are. Second, data augmentation techniques were applied to try to alleviate the imbalance. In terms of architecture, an ensemble system was created consisting of five models: EfficientNet-B0 [42], DenseNet-201 [43], InceptionV3 [44], InceptionResNetV2 [45], and Xception [46], and three different probability combination techniques were tested: Predict then Combine label-wise (PTC-lw), PTC-mode and Combine then Predict (CTP) [23].

In short, this research can be divided into the following phases.

- Remove irrelevant areas.
- Alleviating imbalance through data augmentation techniques.
- Retraining of state-of-the-art models.
- Comparison of the performance of different combination techniques.

3.2 Dataset

In this article, the PadChest dataset [47] was utilised. It is an imbalanced and multilabel dataset published in January 2019 by the University of Valencia in collaboration with BIMCV. The dataset comprises samples collected at the Hospital de San Juan (Spain) between 2009 and 2017. It includes 160,868 clinical images obtained from 67,625 patients, categorised into 174 distinct labels, representing various manifestations of thoracic diseases. This dataset encompasses chest X-rays captured from different angles, including posteroanterior (PA), anteroposterior (AP), and lateral views. However, for our experimentation, only PA X-rays were used, corresponding to 91,728 clinical images from the original dataset.

The authors provided a term tree, in which all labels are organised into broader categories, as demonstrated in Figure 3.1. For example, in this illustration, the general label is "fracture," while the specific labels include "clavicle fracture", "humeral fracture", "vertebral fracture", and "rib and callus rib fractures"

Consequently, we designed two experiments: the first one employed specific labels for classification, and the second one used more general labels, each grouping one or more specific labels. We established a minimum sample requirement for each class to be included in the classification system. The broader classification system consists of a greater number of classes that exhibit more heterogeneity, whereas the more specific classification system features fewer classes but offers higher precision.

Table 3.1 provides the specifics of the two classification systems, including the number of samples, classes, and the sizes of the training, validation, and test sets. In the train/test/validation split, we applied stratification based on both classes and patient IDs to mitigate biases and issues across subsets.

Table 3.1: Summary table of the two types of experiments conducted, one using general labels and the other using specific labels. The table includes the total number of labels and the total number of samples in each of the splits.

	# classes	# samples	train size	val. size	test size
General labels	54	90687	63475	9069	18143
Specific labels	35	85367	59753	8532	17082

3.3 Methodology

This section explains the methodology used to develop the classification system presented in this research. It comprises the following phases: ***Preprocessing***, where segmentation-based crop and data augmentation techniques are applied; ***train models***, from the state of the art; ***applying combination techniques*** to create ensemble-based systems.

3.3.1 Label selection

As mentioned above, multi-label datasets often present imbalances, as some classes have a limited number of instances. Consequently, it is imperative to establish a criterion for the selection of labels in our classification system, especially when considering datasets with an exceptionally large number of classes, as is the case in our study. Initially, I determined that a label should only be included in the classification task if it had a minimum of 200 X-ray samples associated with it. Given a dataset size of 90,000 samples, this represents 0.22% of the total. This step was vital because our model is not designed to function effectively with under-represented labels, as it is not a few-shot system.

In this research, I conducted two distinct experiments. In the first experiment, the class labels originally proposed by the authors of the dataset, which corresponded to specific medical labels. This approach led to a reduction in both the number of samples and labels, primarily due to the removal of under-represented labels. Despite having fewer instances, this classification system offered a more fine-grained and precise categorisation. In the second experiment, a general approach was otped. I created classes by grouping specific labels based on their shared features. This resulted in a classification system with a greater number of samples and classes, although at the expense of precision. However, this general categorisation allowed us to encompass a more extensive range of distinct classes within our analysis.

3.3.2 Preprocessing

To enhance the model's training efficiency, a series of preprocessing steps on the raw images was implemented. These steps are described below. ***Channel Reduction***, initially, it reduced the number of colour channels from three (RGB) to one. Despite the original files being RGB images, X-ray images are inherently greyscale, and all three channels contain the same information. ***Size Normalisation***: it standardised the image dimensions to 512x512 pixels, ensuring uniformity in size across all X-ray images. ***Pixel Value Normalisation***, to facilitate subsequent processing, the pixel values were normalised within the range of 0 to 1, as depicted in Figure 3.2 (first image).

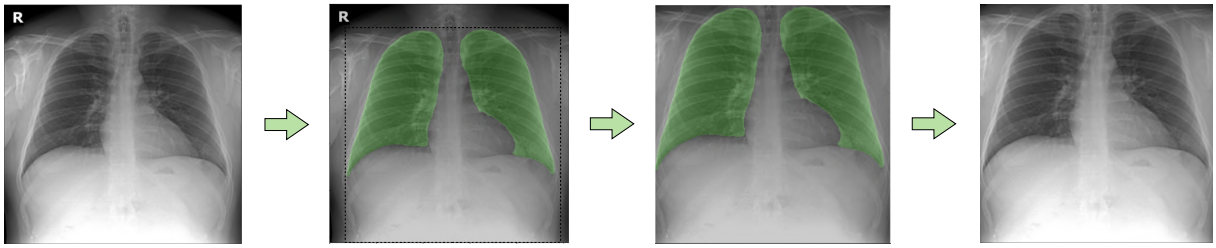


Figure 3.2: It represents a segmentation-based cropped sample. The initial image depicts the original X-ray. The second image shows the lung segmentation mask, outlining the boundaries of the lungs. The third image reveals the cropped X-ray, taking into account the lung mask. Finally, the last image represents the input to our system, showcasing the result of the preprocessing stage.

Chest X-rays typically capture a larger area than the region of interest (ROI) for our specific problem. Irrelevant areas, such as arms or the neck, are not related to the problem I am trying to address. To address this, I performed cropping based on segmentation masks, which forced the system to focus exclusively on pertinent areas. This cropping process unfolded in three steps:

- **Lung Mask Generation**: Initially, I generated lung masks using a segmentation model based on the U-Net architecture [48], as shown in Figure 3.2 (second image). Additionally, I incorporated the region under the lungs into these masks, because it could contain radiological signs of interest.

- **Mask Post-processing**: Given that segmentation models can produce imperfect results, generating more than two masks or leaving gaps within the masks, I opted for a mask post-processing system [49]. This system used an algorithm to fill potential gaps within the masks. It analyses neighbouring pixels to determine whether they belonged to the mask or not, subsequently filling any gaps as necessary. If more than two masks were generated (one per lung), those with an area below a predetermined threshold were eliminated. In cases where lung masks were connected, they were separated.

- **Image Cropping**: Finally, the image was cropped using the mask coordinates and the lower boundary of the sample, resulting in the image shown in Figure 3.2 (third image). To standardise the image size for compatibility with state-of-the-art models, the images were normalised to 224x224 pixels, as shown in Figure 3.2 (last image).

3.3.3 Image classification with CNN

I chose five state-of-the-art architectures, all pre-trained with ImageNet, based on their relevance:

EfficientNet-B0 [42]. This architecture uses distinct scaling coefficients to adjust width, depth, and resolution. Within the EfficientNet family, this architecture represents the smallest variant. It is based on the concept that when dealing with larger images, the network requires an increased number of layers to effectively extract pertinent information.

DenseNet-201 [43]. Instead of introducing additional layers to the architecture, this approach increases the number of connections between units by establishing a link between each unit and the preceding one. In contrast to ResNet50, which connects only one unit to the subsequent output, this architecture offers several benefits: It mitigates the vanishing gradient issue, promotes feature propagation and reuse, and diminishes the overall parameter number.

Inception V3 [44]. This architecture diverges from its predecessors by decomposing convolutions into smaller, potentially asymmetric convolutions, trying to diminishing computational costs. Additionally, this design incorporates an auxiliary classifier placed between layers that serves as a regularisation component.

InceptionResNet V2 [45]. This architecture merges elements from both ResNet and InceptionV3, comprising multiple Inception units interconnected by shortcut connections to reinforce its capacity.

Xception [46]. It incorporates depth-wise separable convolutions, involving a two-step process: first, depth-wise convolution, which operates exclusively on a single channel, and then point-wise convolution, applying a 1x1 convolution across all channels. Additionally, like ResNet50, this architecture incorporates shortcut connections.

In the experimentation, Transfer Learning was applied to the five aforementioned architectures and reconfigured them using the PadChest dataset. I substituted the classifier in all cases with two dense layers. To optimise the training process, it freezes the initial 10% of the convolutional layers, as they mainly detect fundamental patterns that do not require further adaptation. The remaining convolutional layers underwent retraining to acquire patterns specifically relevant to our problem. Table 3.2 summarises the key training parameters. Additionally, I implemented a checkpoint mechanism to save the best performing model based on the validation loss. Furthermore, an early stopping algorithm was used to stop training when the validation loss failed to improve over the last 25 epochs by a margin exceeding the threshold of 0.001.

3.3.4 Ensemble technique

Ensemble learning is recognised as an effective approach to improving the performance and resilience of deep learning algorithms. In this case, the results of all trained models were combined to create a system composed of five different architectures evaluated on the same test set. Two distinct ensemble aggregation methods were employed, as outlined by Nguyen et al. [23]:

- **Combine then Predict (CTP)**: In this approach, the label probabilities predicted by each individual model are first calculated. Subsequently, the average probability for each label is calculated, which is then used to determine the prediction of the ensemble label.
- **Predict then Combine (PTC)**: The PTC method combines the binary predictions generated by the individual models to arrive at the ensemble prediction. Two variations of PTC are considered:

Table 3.2: An overview of the hyperparameters employed during training, including optimization techniques, data augmentation methods, and training methodologies.

Optimization	
Optimizer	Adam
Learning rate	1e-4
Loss	weighted cross entropy with logits
Feed-forward classifier	
# Neurones	512
Activation	ReLu
Dropout	0.2
Data Augmentation	
Shear range	0.1
Zoom range	0.1
Rotation range	45
Width shift range	0.1
Height shift range	0.1
Horizontal flip	True
Fill mode	nearest
Brightness range	0.7-1.1
Channel shift range	0.05
Training methodology	
Maximum epochs	350
Early stopping patience	25
Early stopping threshold	0.001
Batch size	32
Image size	224x224

- ***Label-wise Voting (PTC-lw)***: In this variant, the number of positive and negative predictions made by each model for each label is tallied. The majority decision is adopted to determine the prediction for each label. Consequently, PTC-lw treats the prediction of each label independently, regardless of the others.

- ***Mode-based Predictions (PTC-mode)***: In this version of the PTC, the set of labels predicted by each individual model is calculated and the most frequently occurring set of labels is identified as the ensemble prediction.

3.4 Experimentation

In this section, the results obtained with the proposed methodology are described, and the performance on a multilabel and imbalanced problem, the PadChest dataset, is evaluated. Two strategies for handling the classes were considered: one involved the direct utilisation of labels provided by the dataset creators, while the other involved grouping them into more generic classes that encompassed similar radiological signs.

The evaluation process began with an assessment of whether preprocessing had a positive impact on the ensemble's performance. Subsequently, the performance of both the individual models and the ensemble was assessed. To measure the performance of the different models, three metrics suitable for multilabel problems were used: Area Under the Curve (AUC), Hamming Loss, and F-measure, as recommended by Charte et al. [50].

3.4.1 Impact of preprocessing techniques

Initially, models were trained using images without segmentation-based cropping or data augmentation. The results, as shown in Tables 3.3 and 3.4, reveal that only two individual models, EfficientNet and DenseNet, managed to learn effectively. On the contrary, the remaining models failed to learn, showing a consistently flat training curve with an AUC of 0.5. As anticipated, the ensemble approach did not yield the desired results, underscoring the necessity of the preprocessing step.

The results of training with segmentation-based cropping but without applying data augmentation techniques are presented in Tables 3.5 and 3.6. In particular, Inception did not demonstrate effective learning, possibly due to its limited generalisation capabilities in the absence of data augmentation techniques. Among the models, InceptionResNet exhibited the most favourable results across most classes, while EfficientNet achieved the highest overall performance, boasting an AUC of 0.792 compared to InceptionResNet, which achieves 0.779. A comparison between Table 3.7 and these results highlights the improvement in system performance attributable to the application of data augmentation techniques. Furthermore, when examining the results for different ensembles, it is evident that the CTP technique consistently outperforms the two PTC methods on all labels. Additionally, CTP surpasses the individual models in most cases, except for three instances where it equals their performance and two instances where it performs worse. In conclusion, data augmentation significantly enhances the system's performance.

3.4.2 Performance analysis

The first step is to compare the individual architectures used. These baseline results serve as a reference to evaluate the performance of the ensemble system. I address two types of classification problem: the first utilises the original labels proposed by the authors of the dataset, called "specific labels," while the second employs general labels formed by grouping specific labels. In the first problem, a fine-grain classification is performed, albeit with a limited number of labels, specifically 35. Many of the original 144 labels do not meet the minimum sample threshold (200) and are excluded. In the second problem, the focus shifts to classifying general radiological patterns, leading to a larger number of labels, totalling 54, as the grouping of labels results in more classes meeting the minimum sample threshold of 200.

Tables 3.7 and 3.8 present the results obtained through the proposed methodology for the first case study, which involves classification using specific labels. Among the models, DenseNet exhibits the highest global AUC value at 0.818, followed closely by EfficientNet at 0.804. The remaining models (Inception, InceptionResNet, and Xception) do not achieve an AUC of 0.8.

These results are further detailed by class. Interestingly, labels with fewer samples do not exhibit worse results on average compared to classes with more samples, indicating the successful mitigation of data imbalance issues. Additionally, the table reveals that certain models excel with majority classes, such as Inception, while others perform best with minority classes, such as EfficientNet and Xception. However, DenseNet 201 and InceptionResNet deliver commendable performance across both scenarios.

Secondly, I conducted an analysis of the results obtained using ensemble techniques, with the individual models serving as baselines. Interestingly, only the CTP technique demonstrates an improvement over the individual models, a trend consistent with the observations in Table 3.5.

In focussing on the performance of this ensemble technique, it is noteworthy that there are two classes, namely "Pleural effusion" and "pacemaker," where the results of the individual models are not enhanced. These two classes consist of 658 and 336 samples, respectively, indicating that they are not majority classes. Consequently, one hypothesis could be that the ensemble's performance may be weaker in minority classes. However, the number of labels where the ensemble underperforms in comparison to the individual models is quite small relative to the total number of labels.

Furthermore, the ensemble manages to achieve an AUC exceeding 0.85 for more than 40% of the labels, exceeding initial expectations. In particular, the ensemble achieves AUC values above 0.9 for classes such as "hemidiaphragm elevation," "hiatal hernia," and "sternotomy," all of which contain fewer than 300 samples. This suggests that class imbalance does not significantly impede the performance of our system. Given that the model is trained for 35 distinct classes and faces an imbalance between majority and minority classes at a ratio of 1:172, I can conclude that the system's performance is sufficiently high, considering its inherent features.

In the second case study used to validate the proposed methodology, the different radiological signs were grouped into more general classes of higher level, as exemplified in the fracture case, as shown in Figure 3.1. Following this grouping, the number of labels satisfying the minimum requirement of 200 samples increased to 54 (compared to only 35 labels in the specific labels experiment), resulting in a more realistic representation of health centres.The system is now trained with a larger number of labels, aligning it more closely with the conditions found in real health care settings.

Regarding individual models, the best model is observed to be EfficientNet B0, followed by DenseNet, achieving AUC values of 0.767 and 0.761, respectively. On the contrary, the remaining models register values lower than 0.75. When measuring the performance per class for each model, it becomes apparent that Xception, EfficientNet, and DenseNet excel in majority classes, while Inception and ResNet perform better in minority classes.

Turning to the results obtained through the ensemble technique, similar to the previous case, CTP emerges as the best performer, achieving an AUC of 0.819, marking a notable improvement of 0.052 over EfficientNet. There are four classes where the ensemble performs equally well as the best individual model, but there is no class where the individual models surpass the ensemble. In this case, it is evident that the ensemble further enhances the individual models, as the improvement over the best individual model is notably high. The incorporation of diverse architectures serves to mitigate overfitting and enhance generalisation capacity, particularly in a challenging classification problem characterised by a high number of classes, multilabel and class imbalance. These results underscore the efficacy of this methodology in addressing highly imbalanced and multilabel datasets, despite slightly lower overall performance compared to the previous case, with a difference in the overall AUC of 0.022.

Table 3.3: Specific labels experiment: results obtained by training the models without segmentation-based cropping or data augmentation. For each label, the individual models with the best performance and the ensembles that outperform all individual models are marked in bold. The best ensemble result is marked in italics unless it ties the random classifier.

		DenseNet		**EfficientNet**		**Inception**		**InceptionResNet**		**Xception**		**PTC-mode**		**PTC-lw**		**CTP**	
	# Samples	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1
Normal	34327	**0.589**	0.470	0.500	0.374	0.500	0.374	0.500	0.374	0.500	0.374	0.500	0.374	0.500	0.374	0.589	0.374
Copd signs	13419	0.500	0.457	0.500	0.457	0.500	0.457	0.500	0.457	0.500	0.457	0.500	0.457	0.500	0.457	0.500	0.457
Cardiomegaly	8412	**0.620**	0.551	0.611	0.563	0.500	0.475	0.500	0.475	0.500	0.475	0.500	0.475	0.500	0.475	**0.633**	0.475
Aorticelongation	1399	0.538	0.509	**0.553**	0.526	0.500	0.479	0.500	0.479	0.500	0.479	0.500	0.479	0.500	0.479	**0.558**	0.479
Unchanged	1311	**0.535**	0.483	0.526	0.504	0.500	0.480	0.500	0.480	0.500	0.480	0.500	0.480	0.500	0.480	**0.543**	0.480
Scoliosis	1073	0.500	0.484	**0.550**	0.522	0.500	0.484	0.500	0.484	0.500	0.484	0.500	0.484	0.500	0.484	**0.550**	0.484
Chronic changes	873	**0.581**	0.481	0.578	0.451	0.500	0.487	0.500	0.487	0.500	0.487	0.500	0.487	0.500	0.487	**0.585**	0.487
Costophrenic angle blunting	703	**0.556**	0.525	0.541	0.532	0.500	0.490	0.500	0.490	0.500	0.490	0.500	0.490	0.500	0.490	0.545	0.490
Air trapping	663	0.500	0.490	0.498	0.510	0.500	0.490	0.500	0.490	0.500	0.490	0.500	0.490	0.500	0.490	0.498	0.490
Pleural effusion	658	0.655	0.573	**0.656**	0.567	0.500	0.490	0.500	0.490	0.500	0.490	0.500	0.490	0.500	0.490	**0.676**	0.490
Pneumonia	651	0.626	0.556	**0.629**	0.566	0.500	0.490	0.500	0.490	0.500	0.490	0.500	0.490	0.500	0.490	**0.645**	0.490
Interstitial pattern	594	**0.597**	0.544	0.582	0.547	0.500	0.491	0.500	0.491	0.500	0.491	0.500	0.491	0.500	0.491	0.594	0.491
Infiltrates	591	**0.615**	0.540	0.594	0.542	0.500	0.491	0.500	0.491	0.500	0.491	0.500	0.491	0.500	0.491	0.612	0.491
Laminar atelectasis	578	0.500	0.491	**0.508**	0.491	0.500	0.491	0.500	0.491	0.500	0.491	0.500	0.491	0.500	0.491	0.508	0.491
Vertebral degenerative	575	0.500	0.491	**0.573**	0.485	0.500	0.491	0.500	0.491	0.500	0.491	0.500	0.491	0.500	0.491	**0.573**	0.491
Kyphosis	526	**0.602**	0.558	0.538	0.520	0.500	0.492	0.500	0.492	0.500	0.492	0.500	0.492	0.500	0.492	**0.606**	0.492
Apical pleural thickening	469	0.500	0.493	0.499	0.488	0.500	0.493	0.500	0.493	0.500	0.493	0.500	0.493	0.500	0.493	0.499	0.493
Vascular hilar enlargement	463	0.584	0.510	**0.587**	0.475	0.500	0.493	0.500	0.493	0.500	0.493	0.500	0.493	0.500	0.493	**0.602**	0.493
Fibrotic band	449	0.500	0.493	0.489	0.484	0.500	0.493	0.500	0.493	0.500	0.493	0.500	0.493	0.500	0.493	0.489	0.493
Nodule	449	0.500	0.493	0.500	0.493	0.500	0.493	0.500	0.493	0.500	0.493	0.500	0.493	0.500	0.493	0.500	0.493
Calcified granuloma	388	0.500	0.494	0.499	0.494	0.500	0.494	0.500	0.494	0.500	0.494	0.500	0.494	0.500	0.494	0.499	0.494
Callus rib fracture	360	0.500	0.495	0.500	0.495	0.500	0.495	0.500	0.495	0.500	0.495	0.500	0.495	0.500	0.495	0.500	0.495
Pacemaker	336	0.627	0.543	**0.646**	0.523	0.500	0.495	0.500	0.495	0.500	0.495	0.500	0.495	0.500	0.495	***0.663***	0.495
Aortic atheromatosis	318	0.500	0.495	**0.616**	0.457	0.500	0.495	0.500	0.495	0.500	0.495	0.500	0.495	0.500	0.495	*0.616*	0.495
Volume loss	294	0.500	0.496	**0.512**	0.496	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	*0.512*	0.496
Sternotomy	292	0.530	0.517	**0.539**	0.506	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	***0.545***	0.496
Bronchiectasis	290	0.500	0.496	0.480	0.496	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	0.480	0.496
Hiatal hernia	287	0.500	0.496	**0.533**	0.506	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	***0.533***	0.496
Pseudonodule	275	0.500	0.496	0.498	0.500	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	0.498	0.496
Hemidiaphragm elevation	254	0.515	0.496	**0.531**	0.496	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	***0.544***	0.496
Alveolar pattern	248	**0.664**	0.531	0.663	0.503	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	***0.695***	0.496
Increased density	239	0.528	0.513	**0.536**	0.502	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	***0.547***	0.496
Vertebral anterior compression	214	0.546	0.510	**0.548**	0.487	0.500	0.497	0.500	0.497	0.500	0.497	0.500	0.497	0.500	0.497	***0.559***	0.497
Suture material	210	0.500	0.497	**0.542**	0.509	0.500	0.497	0.500	0.497	0.500	0.497	0.500	0.497	0.500	0.497	*0.542*	0.497
Supra aortic elongation	200	0.500	0.497	**0.503**	0.497	0.500	0.497	0.500	0.497	0.500	0.497	0.500	0.497	0.500	0.497	***0.504***	0.497
Global		0.543	0.508	**0.547**	0.502	0.500	0.488	0.500	0.488	0.500	0.488	0.500	0.488	0.500	0.488	***0.558***	0.488

Table 3.4: Specific labels experiment: global results obtained by the individual models and the ensemble without using segmentation-based cropping or data augmentation techniques.

	DenseNet	EfficientNet	Inception	InceptionResNet	Xception	PTC-mode	PTC-lw	CTP
Hamming Loss	0.067	0.107	0.046	0.046	0.046	0.046	0.046	0.046
AUC	0.543	0.547	0.500	0.500	0.500	0.500	0.500	0.558
F1	0.508	0.502	0.488	0.488	0.488	0.488	0.488	0.488

Table 3.5: Specific labels experiment: results obtained by training the models with segmentation-based cropping, but without data augmentation. For each label, the individual models with the best performance and the ensembles that outperform all individual models are marked in bold. The best ensemble result is marked in italics.

		DenseNet		**EfficientNet**		**Inception**		**InceptionResnet**		**Xception**		**PTC-mode**		**PTC-lw**		**CTP**	
	# Samples	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1
Normal	34327	0.5	0.374	0.802	0.725	0.453	0.374	**0.819**	0.723	0.5	0.374	0.528	0.444	0.500	0.374	*0.806*	0.374
Copd signs	13419	0.777	0.682	0.785	0.672	0.500	0.457	**0.799**	0.690	0.777	0.682	0.538	0.534	0.648	0.676	***0.825***	0.674
Cardiomegaly	8412	0.900	0.768	0.898	0.774	0.641	0.474	**0.918**	0.767	0.917	0.762	0.596	0.628	0.814	0.792	***0.938***	0.795
Aortic elongation	1399	0.863	0.700	0.874	0.686	0.500	0.479	**0.875**	0.719	0.837	0.705	0.594	0.623	0.767	0.724	***0.898***	0.724
Unchanged	1311	0.612	0.556	**0.625**	0.549	0.500	0.480	0.597	0.549	0.602	0.544	0.506	0.495	0.531	0.539	***0.642***	0.537
Scoliosis	1073	0.823	0.678	0.808	0.690	0.500	0.484	**0.830**	0.702	0.500	0.484	0.591	0.628	0.674	0.711	***0.863***	0.708
Chronic changes	873	0.707	0.537	**0.731**	0.553	0.500	0.487	0.696	0.547	0.695	0.538	0.515	0.518	0.625	0.568	***0.738***	0.568
Costophrenic angle blunting	703	0.810	0.698	0.837	0.691	0.500	0.489	**0.842**	0.655	0.810	0.704	0.558	0.587	0.729	0.713	***0.884***	0.712
Air trapping	663	0.500	0.490	0.671	0.568	0.500	0.490	0.500	0.490	**0.688**	0.553	0.508	0.506	0.500	0.490	***0.705***	0.490
Pleural effusion	658	0.925	0.839	0.942	0.818	0.479	0.046	**0.943**	0.770	0.927	0.838	0.818	0.542	0.901	0.823	*0.942*	0.825
Pneumonia	651	0.759	0.675	0.803	0.671	0.500	0.490	**0.808**	0.655	0.806	0.657	0.572	0.603	0.704	0.691	***0.851***	0.692
Interstitial pattern	594	0.799	0.638	0.795	0.650	0.500	0.491	**0.813**	0.637	0.812	0.615	0.562	0.576	0.714	0.678	***0.858***	0.680
Infiltrates	591	0.733	0.620	0.776	0.635	0.500	0.491	**0.802**	0.597	0.771	0.627	0.563	0.583	0.668	0.639	***0.831***	0.639
Laminar atelectasis	578	0.500	0.491	**0.806**	0.639	0.500	0.491	0.754	0.630	0.745	0.646	0.560	0.587	0.572	0.607	***0.837***	0.607
Vertebral degenerative	575	**0.730**	0.544	0.721	0.540	0.500	0.491	0.725	0.564	0.718	0.533	0.571	0.560	0.620	0.568	***0.771***	0.568
Kyphosis	526	0.796	0.611	**0.813**	0.644	0.500	0.492	0.794	0.628	**0.813**	0.615	0.569	0.585	0.683	0.664	***0.860***	0.664
Apical pleural thickening	469	**0.798**	0.591	0.787	0.573	0.500	0.493	0.775	0.569	0.758	0.575	0.574	0.567	0.701	0.619	***0.838***	0.619
Vascular hilar enlargement	463	0.679	0.562	**0.741**	0.506	0.500	0.493	0.717	0.559	0.715	0.547	0.531	0.533	0.596	0.578	***0.771***	0.582
Fibrotic band	449	0.756	0.568	**0.772**	0.599	0.500	0.493	0.767	0.608	0.758	0.593	0.573	0.585	0.688	0.636	***0.813***	0.638
Nodule	449	0.616	0.557	0.677	0.567	0.500	0.493	**0.688**	0.547	0.626	0.558	0.535	0.545	0.566	0.574	***0.719***	0.572
Calcified granuloma	388	0.741	0.651	0.752	0.641	0.500	0.494	**0.757**	0.622	0.689	0.611	0.578	0.601	0.645	0.654	***0.819***	0.656
Callus rib fracture	360	0.682	0.600	**0.773**	0.594	0.500	0.495	0.500	0.495	0.497	0.495	0.529	0.543	0.500	0.495	***0.799***	0.495
Pacemaker	336	**0.996**	0.948	**0.996**	0.945	0.500	0.495	**0.996**	0.946	**0.996**	0.949	0.741	0.799	0.992	0.951	*0.996*	0.951
Aortic atheromatosis	318	**0.812**	0.538	0.810	0.542	0.500	0.495	0.791	0.567	0.742	0.577	0.544	0.545	0.672	0.605	***0.852***	0.607
Volume loss	294	0.855	0.687	0.862	0.717	0.500	0.496	**0.882**	0.677	0.830	0.691	0.560	0.581	0.762	0.729	***0.910***	0.731
Sternotomy	292	0.991	0.945	0.991	0.939	0.500	0.496	**0.993**	0.872	**0.993**	0.918	0.756	0.814	0.983	0.948	***0.996***	0.948
Bronchiectasis	290	0.673	0.593	**0.726**	0.597	0.500	0.496	0.719	0.587	0.725	0.576	0.541	0.563	0.594	0.613	***0.784***	0.614
Hiatal hernia	287	0.912	0.852	0.920	0.826	0.500	0.496	0.939	0.843	**0.945**	0.726	0.747	0.784	0.877	0.870	***0.962***	0.872
Pseudonodule	275	0.632	0.524	**0.705**	0.547	0.500	0.496	0.536	0.514	0.639	0.540	0.530	0.536	0.540	0.544	***0.718***	0.545
Hemidiaphragm elevation	254	0.902	0.706	0.879	0.697	0.500	0.496	**0.911**	0.696	0.891	0.687	0.749	0.709	0.816	0.751	***0.951***	0.751
Alveolar pattern	248	0.791	0.626	0.834	0.603	0.500	0.496	**0.853**	0.580	0.810	0.604	0.568	0.579	0.715	0.621	***0.887***	0.622
Increased density	239	0.580	0.551	0.586	0.521	0.500	0.496	**0.619**	0.521	0.569	0.526	0.501	0.500	0.533	0.537	***0.634***	0.539
Vertebral anterior compression	214	0.640	0.536	**0.645**	0.530	0.500	0.497	0.644	0.537	0.623	0.517	0.516	0.522	0.524	0.524	***0.702***	0.525
Suture material	210	0.798	0.663	0.791	0.649	0.500	0.497	**0.824**	0.628	0.786	0.665	0.622	0.622	0.742	0.679	***0.833***	0.680
Supra aortic elongation	200	0.697	0.569	0.778	0.564	0.500	0.497	**0.832**	0.561	0.738	0.554	0.579	0.574	0.613	0.577	***0.861***	0.578
Global		0.751	0.633	**0.792**	0.648	0.502	0.475	0.779	0.636	0.750	0.622	0.584	0.586	0.677	0.650	***0.831***	0.651

Table 3.6: Specific labels experiment: global results obtained by the individual models and the ensemble with preprocessing (segmentation-based cropping) but without data augmentation.

	Densenet201	EfficientNet	Inception	InceptionResnet	Xception	PTC-mode	PTC-lw	CTP
Hamming Loss	0.077	0.079	0.072	0.070	0.077	0.056	0.057	0.057
AUC	0.751	**0.792**	0.502	0.779	0.750	0.584	0.677	**0.831**
F1-score	0.633	0.648	0.475	0.636	0.622	0.586	0.650	0.651

Table 3.7: Specific labels experiment: results obtained with by training the models with segmentation-based cropping and data augmentation. For each label, the individual models with the best performance and the ensembles that outperform all individual models are marked in bold. The best ensemble result is marked in italics.

		Densenet201		**EfficientNet**		**Inception**		**InceptionResnet**		**Xception**		**PTC-mode**		**PTC-lw**		**CTP**	
	# Samples	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1
Normal	34327	0.820	0.722	0.811	0.716	**0.832**	0.727	0.820	0.709	0.827	0.732	0.725	0.731	0.725	0.731	***0.837***	0.730
Copd signs	13419	**0.823**	0.681	0.785	0.644	0.816	0.678	0.815	0.675	0.800	0.666	0.588	0.610	0.647	0.675	***0.833***	0.672
Cardiomegaly	8412	**0.927**	0.773	0.907	0.749	0.926	0.767	**0.927**	0.777	0.922	0.779	0.746	0.765	0.825	0.789	***0.937***	0.791
Aortic elongation	1399	0.885	0.690	0.846	0.655	0.882	0.676	**0.888**	0.698	**0.888**	0.702	0.690	0.682	0.777	0.705	***0.894***	0.707
Unchanged	1311	0.636	0.553	0.614	0.543	0.638	0.549	**0.642**	0.544	0.636	0.547	0.531	0.537	0.545	0.551	*0.641*	0.551
Scoliosis	1073	0.759	0.636	0.712	0.598	0.745	0.605	0.732	0.602	**0.774**	0.661	0.630	0.637	0.671	0.659	***0.793***	0.664
Chronic changes	873	0.759	0.518	0.720	0.549	**0.768**	0.546	0.762	0.519	0.752	0.533	0.621	0.556	0.684	0.545	***0.772***	0.547
Costophrenic angle blunting	703	**0.862**	0.674	0.845	0.674	0.832	0.662	0.831	0.665	0.855	0.663	0.685	0.676	0.739	0.693	***0.877***	0.693
Air trapping	663	**0.692**	0.557	0.687	0.560	0.515	0.490	0.469	0.490	0.506	0.490	0.525	0.532	0.500	0.490	***0.704***	0.490
Pleural effusion	658	**0.959**	0.827	0.951	0.811	0.956	0.823	0.955	0.830	0.945	0.816	0.862	0.822	0.886	0.839	***0.967***	0.840
Pneumonia	651	0.815	0.671	**0.821**	0.660	0.810	0.663	**0.821**	0.672	**0.821**	0.668	0.681	0.668	0.703	0.687	***0.850***	0.687
Interstitial pattern	594	0.834	0.625	0.828	0.636	**0.846**	0.651	0.843	0.616	0.830	0.613	0.727	0.651	0.743	0.650	***0.858***	0.651
Infiltrates	591	0.812	0.629	0.803	0.617	0.803	0.626	**0.815**	0.633	0.808	0.639	0.649	0.635	0.662	0.644	***0.840***	0.646
Laminar atelectasis	578	**0.843**	0.670	0.812	0.637	0.827	0.643	0.827	0.654	0.833	0.654	0.666	0.658	0.690	0.677	***0.858***	0.678
Vertebral degenerative changes	575	0.779	0.545	0.730	0.544	0.774	0.546	**0.785**	0.518	0.779	0.547	0.627	0.560	0.670	0.557	***0.797***	0.556
Kyphosis	526	**0.867**	0.640	0.834	0.589	0.845	0.587	0.849	0.609	0.839	0.625	0.691	0.617	0.736	0.639	***0.870***	0.640
Apical pleural thickening	469	**0.808**	0.553	0.801	0.568	0.789	0.573	0.509	0.493	0.500	0.493	0.592	0.574	0.661	0.619	***0.830***	0.621
Vascular hilar enlargement	463	**0.746**	0.549	0.742	0.568	0.769	0.522	0.755	0.544	0.745	0.515	0.618	0.556	0.651	0.562	***0.783***	0.563
Fibrotic band	449	**0.831**	0.583	0.809	0.600	0.813	0.614	0.575	0.493	0.806	0.611	0.641	0.604	0.716	0.658	***0.848***	0.659
Nodule	449	**0.706**	0.578	0.675	0.554	0.561	0.493	0.574	0.493	0.551	0.493	0.518	0.526	0.500	0.493	*0.704*	0.493
Calcified granuloma	388	**0.808**	0.653	0.802	0.649	0.542	0.494	0.554	0.494	0.496	0.494	0.572	0.593	0.500	0.494	***0.833***	0.494
Callus rib fracture	360	0.717	0.606	**0.765**	0.557	0.614	0.495	0.609	0.495	0.571	0.495	0.550	0.549	0.500	0.495	***0.787***	0.495
Pacemaker	336	0.993	0.927	**0.997**	0.942	0.996	0.919	0.996	0.931	0.984	0.926	0.984	0.930	0.993	0.946	***0.997***	0.946
Aortic atheromatosis	318	0.856	0.521	0.847	0.516	0.862	0.550	0.852	0.541	**0.871**	0.559	0.739	0.556	0.786	0.558	***0.885***	0.557
Volume loss	294	**0.917**	0.693	0.902	0.657	0.904	0.636	0.896	0.672	0.902	0.640	0.789	0.670	0.809	0.693	***0.928***	0.697
Sternotomy	292	0.992	0.898	0.987	0.920	0.990	0.926	**0.995**	0.936	0.992	0.865	0.961	0.912	0.984	0.941	***0.997***	0.941
Bronchiectasis	290	0.801	0.549	0.775	0.561	0.796	0.578	**0.805**	0.562	0.794	0.573	0.682	0.580	0.690	0.587	***0.820***	0.588
Hiatal hernia	287	0.939	0.856	0.941	0.697	0.947	0.824	**0.964**	0.801	0.947	0.851	0.871	0.786	0.876	0.867	***0.967***	0.867
Pseudonodule	275	0.612	0.496	**0.670**	0.550	0.598	0.496	0.589	0.496	0.545	0.496	0.536	0.543	0.500	0.496	***0.672***	0.496
Hemidiaphragm elevation	254	0.893	0.667	0.882	0.679	**0.915**	0.670	0.894	0.702	0.898	0.684	0.759	0.692	0.784	0.725	***0.934***	0.724
Alveolar pattern	248	0.876	0.627	0.877	0.589	0.885	0.607	**0.895**	0.606	0.871	0.594	0.748	0.612	0.774	0.623	***0.911***	0.622
Increased density	239	0.643	0.541	0.641	0.509	0.651	0.534	0.633	0.526	**0.668**	0.535	0.545	0.531	0.547	0.544	***0.673***	0.547
Vertebral anterior compression	214	**0.749**	0.532	0.696	0.524	0.736	0.517	0.743	0.527	0.734	0.535	0.573	0.530	0.597	0.540	***0.752***	0.539
Suture material	210	0.819	0.652	0.820	0.663	0.818	0.639	0.811	0.662	**0.822**	0.612	0.759	0.663	0.768	0.685	***0.847***	0.684
Supra aortic elongation	200	0.865	0.576	0.821	0.541	0.880	0.546	**0.882**	0.563	0.857	0.562	0.628	0.558	0.681	0.576	***0.894***	0.575
Global		**0.818**	0.642	0.804	0.629	0.797	0.625	0.780	0.621	0.782	0.625	0.677	0.637	0.701	0.647	***0.840***	0.647

Table 3.8: Specific labels experiment: global results obtained from the individual models and the ensembles.

	Densenet201	EfficientNet	Inception	InceptionResnet	Xception	PTC-mode	PTC-lw	CTP
Hamming Loss	0.082	0.078	0.081	0.077	0.074	0.063	0.065	0.065
AUC	0.818	0.804	0.797	0.780	0.782	0.677	0.701	0.840
F1	0.642	0.629	0.625	0.621	0.625	0.637	0.647	0.647

Table 3.9: General labels experiment: results obtained by training the models with segmentation-based cropping and data augmentation. For each label, the individual models with the best performance and the ensembles that outperform all individual models are marked in bold. The best ensemble result is marked in italics.

		Densenet201		**EfficientNet**		**Inception**		**InceptionResnet**		**Xception**		**PTC-mode**		**PTC-lw**		**CTP**	
	# Samples	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1
Normal	34327	0.735	0.685	0.707	0.652	**0.750**	0.691	0.723	0.658	0.732	0.674	0.702	0.693	0.690	0.691	***0.770***	0.691
Copd signs	13419	0.771	0.629	0.761	0.649	0.771	0.618	**0.793**	0.665	0.779	0.645	0.596	0.621	0.615	0.645	***0.816***	0.640
Cardiomegaly	8120	0.899	0.736	**0.904**	0.746	0.890	0.723	0.892	0.751	0.898	0.741	0.766	0.747	0.816	0.760	***0.923***	0.762
Thoracic cage deformation	7778	0.706	0.603	**0.728**	0.627	0.500	0.478	0.675	0.595	0.708	0.609	0.577	0.586	0.601	0.612	***0.745***	0.614
Aortic elongation	7436	0.858	0.691	0.853	0.690	0.842	0.661	**0.866**	0.683	**0.866**	0.687	0.701	0.690	0.758	0.697	***0.886***	0.700
Infiltrates	6706	0.794	0.686	**0.802**	0.664	0.791	0.663	0.797	0.668	0.794	0.663	0.703	0.676	0.721	0.690	***0.827***	0.692
Unchanged	6487	0.630	0.538	**0.636**	0.552	0.618	0.543	0.633	0.545	0.631	0.552	0.536	0.543	0.536	0.546	***0.652***	0.547
Chronic changes	4312	**0.759**	0.548	0.754	0.542	0.752	0.525	0.734	0.520	0.740	0.522	0.656	0.567	0.685	0.543	***0.768***	0.545
Surgery	3928	0.813	0.730	0.815	0.766	0.750	0.722	0.766	0.713	**0.829**	0.724	0.726	0.739	0.739	0.762	***0.845***	0.765
Atelectasis	3565	**0.798**	0.628	0.756	0.636	0.698	0.570	0.729	0.596	0.759	0.628	0.587	0.598	0.647	0.632	***0.804***	0.630
Costophrenic angle blunting	3306	**0.845**	0.655	0.807	0.638	0.758	0.604	0.784	0.638	0.828	0.652	0.660	0.634	0.700	0.656	***0.864***	0.658
Calcified densities	3253	0.719	0.638	**0.751**	0.639	0.500	0.491	0.500	0.491	0.500	0.491	0.520	0.527	0.500	0.491	***0.764***	0.491
Vertebral degenerative changes	3203	**0.744**	0.502	0.726	0.528	0.676	0.497	0.733	0.487	0.730	0.512	0.643	0.532	0.664	0.514	***0.751***	0.514
Hilar enlargement	3162	**0.755**	0.549	0.732	0.544	0.699	0.551	0.738	0.533	0.731	0.538	0.618	0.562	0.649	0.560	***0.765***	0.565
Pleural thickening	3010	0.753	0.586	**0.773**	0.585	0.737	0.572	0.743	0.525	0.763	0.562	0.651	0.587	0.671	0.586	***0.790***	0.587
Mediastinal enlargement	2813	0.795	0.643	0.798	0.668	0.774	0.688	0.778	0.675	**0.822**	0.657	0.705	0.689	0.710	0.697	***0.841***	0.700
Air trapping	2765	0.654	0.528	0.669	0.536	0.500	0.492	0.665	0.495	**0.672**	0.536	0.520	0.523	0.602	0.544	***0.692***	0.546
Fracture	2529	**0.749**	0.663	0.725	0.599	0.5	0.493	0.640	0.507	0.732	0.611	0.574	0.590	0.579	0.615	***0.792***	0.617
Pleural effusion	2436	**0.942**	0.738	0.927	0.782	0.930	0.720	0.935	0.735	0.937	0.771	0.878	0.762	0.900	0.775	***0.956***	0.775
Granuloma	2306	0.500	0.493	**0.777**	0.652	0.500	0.493	0.500	0.493	0.500	0.493	0.513	0.519	0.500	0.493	***0.777***	0.493
Nodule	1936	0.653	0.583	**0.679**	0.571	0.617	0.542	0.643	0.547	0.622	0.575	0.560	0.569	0.558	0.575	***0.707***	0.573
Fibrotic band	1781	0.738	0.530	**0.747**	0.531	0.712	0.522	0.500	0.495	0.727	0.519	0.606	0.546	0.654	0.556	***0.770***	0.556
Electrical device	1772	0.992	0.959	0.992	0.913	0.992	0.889	**0.994**	0.871	0.992	0.929	0.990	0.935	0.992	0.942	***0.997***	0.942
Pneumonia	1652	0.804	0.594	0.790	0.567	0.804	0.549	**0.813**	0.577	0.799	0.599	0.712	0.602	0.728	0.606	***0.854***	0.607
Aortic atheromatosis	1581	0.834	0.502	0.830	0.540	0.813	0.477	0.840	0.524	**0.843**	0.519	0.713	0.554	0.769	0.522	***0.866***	0.523
Pseudonodule	1451	0.693	0.561	**0.727**	0.553	0.500	0.496	0.500	0.496	0.708	0.562	0.557	0.555	0.576	0.589	***0.759***	0.593
Bronchiectasis	1430	0.795	0.544	0.776	0.571	0.789	0.548	**0.814**	0.539	0.779	0.561	0.639	0.575	0.696	0.573	***0.833***	0.574
Hiatal hernia	1362	0.916	0.813	0.892	0.796	0.906	0.773	0.918	0.804	**0.927**	0.788	0.857	0.810	0.889	0.852	***0.959***	0.852
Hemidiaphragm elevation	1231	0.814	0.651	**0.841**	0.680	**0.841**	0.596	0.823	0.649	0.811	0.645	0.733	0.670	0.746	0.683	***0.890***	0.683
Increased density	1133	0.633	0.497	**0.640**	0.524	0.596	0.492	0.609	0.509	0.606	0.511	0.539	0.516	0.533	0.514	***0.661***	0.515
Diaphragmatic eventration	757	0.500	0.498	**0.775**	0.586	0.500	0.498	0.500	0.498	0.500	0.498	0.525	0.534	0.500	0.498	***0.775***	0.498
Volume loss	684	0.802	0.542	0.776	0.580	0.809	0.531	**0.814**	0.513	0.776	0.560	0.728	0.561	0.761	0.564	***0.865***	0.564
Adenopathy	659	0.500	0.498	**0.697**	0.538	0.500	0.498	0.548	0.520	0.583	0.543	0.500	0.498	0.521	0.528	***0.715***	0.528
Bronchovascular markings	602	0.712	0.570	0.738	0.537	**0.777**	0.576	0.765	0.545	0.704	0.585	0.685	0.592	0.703	0.585	***0.802***	0.584
Mass	574	0.707	0.621	0.715	0.608	0.744	0.570	0.732	0.574	**0.746**	0.616	0.700	0.615	0.707	0.641	***0.806***	0.641
Artificial heart valve	562	0.969	0.658	0.953	0.730	0.975	0.696	**0.977**	0.727	0.972	0.713	0.941	0.736	0.968	0.730	***0.981***	0.731
Catheter	545	0.871	0.740	0.874	0.721	0.866	0.673	**0.878**	0.639	0.861	0.717	0.799	0.724	0.848	0.773	***0.905***	0.773
Suboptimal study	544	0.743	0.524	0.693	0.510	**0.754**	0.522	0.697	0.531	0.727	0.506	0.666	0.526	0.681	0.539	***0.784***	0.540
Pulmonary fibrosis	523	0.850	0.584	0.834	0.587	0.837	0.551	**0.864**	0.577	0.862	0.565	0.795	0.577	0.810	0.591	***0.892***	0.591
Heart insufficiency	520	0.875	0.541	0.877	0.555	0.896	0.546	**0.884**	0.538	0.870	0.547	0.819	0.553	0.856	0.551	***0.920***	0.551
Hypoexpansion	476	0.838	0.541	0.745	0.545	**0.846**	0.534	0.768	0.571	0.500	0.499	0.651	0.556	0.677	0.571	***0.900***	0.573
Gynecomastia	437	0.852	0.527	0.810	0.552	0.852	0.501	**0.858**	0.507	0.806	0.550	0.772	0.540	0.825	0.554	***0.917***	0.555
Emphysema	410	0.780	0.508	0.715	0.520	0.801	0.512	**0.809**	0.506	0.724	0.521	0.684	0.529	0.732	0.524	***0.862***	0.525
Sclerotic bone lesion	352	**0.506**	0.511	0.500	0.499	0.500	0.499	0.500	0.499	0.500	0.499	0.500	0.499	0.500	0.499	***0.506***	0.499
Fissure thickening	336	0.816	0.533	0.802	0.573	0.819	0.518	**0.842**	0.526	0.798	0.539	0.746	0.547	0.806	0.557	***0.891***	0.558
Hilar congestion	318	0.785	0.503	0.798	0.519	0.790	0.514	**0.827**	0.519	0.808	0.520	0.734	0.526	0.756	0.523	***0.896***	0.522
Osteopenia	318	0.659	0.508	0.688	0.507	0.659	0.483	0.695	0.466	**0.701**	0.497	0.611	0.500	0.647	0.500	***0.752***	0.500
Tuberculosis	299	0.852	0.534	0.861	0.567	**0.869**	0.561	0.824	0.559	0.805	0.597	0.760	0.577	0.848	0.592	***0.909***	0.592
Bullas	290	**0.746**	0.520	0.685	0.532	0.739	0.524	0.715	0.512	0.651	0.549	0.667	0.543	0.714	0.547	***0.777***	0.547
Hyperinflated lung	272	0.715	0.506	0.630	0.502	**0.719**	0.504	0.645	0.485	0.659	0.501	0.649	0.513	0.658	0.512	***0.728***	0.512
Cavitation	243	0.780	0.556	0.834	0.575	**0.856**	0.546	0.789	0.539	0.823	0.590	0.679	0.555	0.823	0.585	***0.934***	0.585
Mediastinic lipomatosis	212	0.648	0.499	**0.654**	0.551	0.5	0.499	0.500	0.499	0.500	0.499	0.520	0.514	0.500	0.499	***0.681***	0.499
Pneumothorax	210	0.705	0.572	0.717	0.530	0.717	0.540	**0.721**	0.518	0.620	0.592	0.596	0.546	0.630	0.572	***0.847***	0.573
Vascular redistribution	204	**0.774**	0.499	0.752	0.526	0.705	0.508	0.694	0.516	0.667	0.507	0.635	0.514	0.676	0.519	***0.837***	0.519
Global		0.761	0.589	**0.767**	0.600	0.732	0.566	0.739	0.572	0.739	0.589	0.669	0.594	0.696	0.601	***0.819***	0.602

Table 3.10: General labels experiment: global results obtained by the individual models and the ensembles.

	Densenet201	EfficientNet	Inception	InceptionResnet	Xception	PTC-mode	PTC-lw	CTP
Hamming Loss	0.070	0.065	0.070	0.075	0.065	0.052	0.057	0.056
AUC	0.761	0.767	0.732	0.739	0.739	0.669	0.696	0.819
F1-score	0.589	0.600	0.566	0.572	0.589	0.594	0.601	0.602

Chapter 4

Information fusion for regression tasks

What is truth?
The negation of lies?
Or the statement of a fact?
And if the fact is a lie,
what then is the truth?.

— Andrzej Sapkowski

In the previous two chapters we have solved two different problems in classification tasks, both binary and multi-label, in this chapter we want to solve the problem of extremely limited datasets, due to the number of samples, in a regression task. In this chapter, a regression model for resource management in forest fires is proposed. The proposed system is composed of two different phases: an autoencoder based on self-supervised learning that will learn trends and patterns from the data; and a regression model that will exploit the knowledge of the encoder to perform the prediction of the resources needed in case of a forest fire. For this purpose, a dataset consisting of 445 samples was used.

4.1 Problem definition and objectives

This chapter proposes a regression system for forest fire management. To solve it, a dataset was created from a set of fire information from Castilla y León (Spain), composed of 445 samples, an extremely limited dataset, even more limited than the dataset used in Chapter 2 of this dissertation. However, a set of atomospheric and GI information was available with a larger number of samples but not labelled. Due to the limited number of radiographs, it was decided to generate different labelled and unlabelled datasets to solve the task. Therefore, it was decided to train an autoencoder based on self-supervised learning to learn patterns and trends and apply it to the regression model. This chapter focuses on solving ***RQ3: "Can data fusion techniques help overcome regression problems within the forest fire domain, when the size of the dataset is extremely limited?"***

The hypothesis of this research is that the combination of different sources of information together with pre-training based on self-supervised learning, which does not need a labelled dataset, we will be able to develop a regression model capable of predicting the resources needed in case of a forest fire, the labels we are trying to predict are: the burned area, control and extinction time and the human, heavy and aerial resources needed.

The hypothesis of this research is that by combining different sources of information together with pre-training based on self-supervised learning, which does not need a labelled dataset, we will be able to develop a regression model capable of predicting the resources needed in case of a forest fire, the labels we are trying to predict are: the burned area, control and extinction time and the human, heavy and air resources needed. In order to achieve this objective, two different encoder architectures are presented, the first basic one formed by three convolutional blocks and the second one including skip connections. To facilitate its application, the predictions are not only calculated for a specific point but for the whole study area, in this case Castilla y León (Spain).

4.2 Dataset

The objective of this paper is the development of WAM with the ability to predict the resources required, the time required to control and extinguish wildfires, and the extent to which it will be affected during the event. This predictive model aims to assist in resource management and mitigate economic and environmental losses. To achieve this goal, three different sources of information were used: ***Wildfire Information***, these data were used to create labels, including details about fire locations, resources necessary for extinguishment, duration of control and extinguishment, and the total area burnt; ***Atmospheric Variables***, these variables, such as the 10-meter U wind component and total column ozone, are pertinent to fire spread; ***Greenness Index***, this index, indicating vegetation health, is closely associated with forest fire management. The last two sources were combined using early fusion techniques to generate samples representing different fires (variable X), while information on fire management constituted the labels for these samples. This integration of data sources forms the basis for the regression model, enabling predictions crucial for effective fire management strategies.

4.2.1 Wildfires in Spain

In this study, a dataset consisting of 597 wildfire records was used, sourced from two Spanish autonomous regions: Castilla y León (446 records) and Andalucía (151 records), as shown in Figure 4.1. Each record contains details of the wildfires, including coordinates and dates. The coordinates of Castilla y León are latitudes 40 to 43.3 and longitudes -7.2 to -1.8, while the coordinates of Andalucía are latitudes 35.5 to 38.5 and longitudes -7.6 to 0. Despite the proximity of these regions, there are significant differences, particularly in their weather conditions. Castilla y León experiences colder temperatures with annual differences of more than 5 degrees compared to Andalucía. Andalucía presents a wider range of environments, from arid areas in the east to more humid regions in the west. In contrast, Castilla y León shows a more uniform weather, with lower average temperatures, except in mountainous areas. In addition, disparities in radiation and insolation levels are observed, Andalucía recording higher values compared to Castilla y León. Consequently, the different conditions in these regions require adapted wildfire management strategies. In Castilla y León, wildfires are concentrated in specific subregions such as Zamora and León, with minimal occurrences in areas like Valladolid, Soria, or Palencia. On the other

hand, in Andalucía, wildfires are more evenly distributed, with notable concentrations in Huelva, Málaga, Jaén, and Almería, particularly in terms of the number of events and the burnt area. It is important to note that forest fires are more widespread in Andalucía compared to Castilla y León.

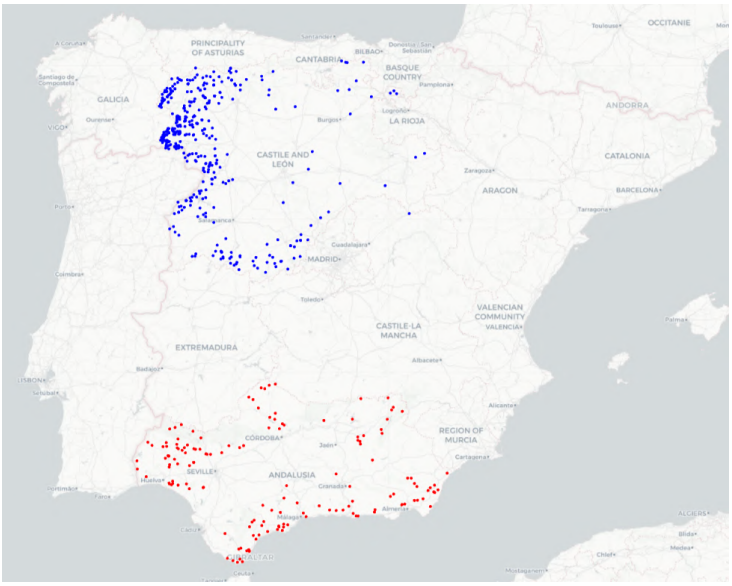


Figure 4.1: Map of Spain showing the locations of all wildfires recorded between 2001 and 2012. Wildfires in Castilla y León are highlighted in blue and those in Andalucía in red.

As stated previously, each wildfire data record contains crucial details regarding fire management, including parameters such as extinction time and burnt area. These specific pieces of information serve as the labels used in our study, and each is detailed below.

1. **Burnt area** (metres): total area affected by wildfire.
2. **Control time** (min): time required to enter the control phase, that is, when the fire conditions have changed enough to prevent its propagation.
3. **Extinction time** (min): time until the fire is extinguished, that is, when there are no active hotspots and the technicians verify that there is no possibility of reignite.
4. **Human resources** (units): number of people involved in extinguishing the fire.
5. **Aerial resources** (units): number of aerial vehicles involved in the extinguishing of the fire.
6. **Heavy resources** (units): number of heavy vehicles involved in the extinguishing of the fire.

4.2.2 Atmospheric variables

As mentioned above, the second data source utilised comprises atmospheric variables associated with the spread of forest fires. These data sources are georeferenced, allowing them to be organised into two-dimensional matrices, akin to images. The variables employ a decimal coordinate system, where the distance between measured latitudes is 111 km. During our study period,

these data were measured daily. Depending on the variable, measurements were taken an hourly or at specific times throughout the day (consistently at the same times). Variables are described as follows [51]:

- *10 metre U wind component*: is the horizontal speed of air moving toward the east at a height of 10 metres above the Earth's surface [52].
- *10 metre V wind component*: it is the vertical speed of the air moving north at a height of 10 metres above the surface of the Earth [52].
- *2 metre dewpoint temperature* (K, kelvin): is the temperature that would have to be cooled for saturation to occur at 2 metres above the Earth's surface. It is a measure that combines humidity, temperature, and pressure [53].
- *Surface net solar radiation*(J/m^2): is the low-wave solar radiation incident on the Earth's surface, direct and indirect, less the reflected radiation. It is the radiation that crosses a horizontal plane to the Sun's direction [54].
- *Surface net thermal radiation* (J/m^2): is the radiation emitted by the atmosphere, clouds, and the surface of the Earth. It is the difference between downward and upward thermal radiation on the Earth's surface [55].
- *Surface solar radiation downwards* (J/m^2): describes the amount of shortwave solar radiation that reaches the earth's surface in a horizontal plane [56].
- *Surface thermal radiation downwards* (J/m^2): is a type of thermal radiation that describes the thermal radiation reaching the Earth's surface emitted by the atmosphere and clouds [55].
- *Total column ozone* (kg/m^2): is the total ozone along a column from the surface to the top of the atmosphere. It provides information on the densities in the atmosphere [57].

4.2.3 Greenness index

The final data source used in this study is the GI or Green Leaf Index, which represents the relationship between the reflectance in the green channel compared to the other two visible light channels, red and blue [58]. Similar to atmospheric variables, these data points are georeferenced and can be processed as images. GI measurements were made three times per month, specifically on the 1st, 11st, and 21st days, in contrast to daily measurements for atmospheric variables. The coordinate system employed for GI data is UTM (Universal Transverse Mercator), with the distance between two contiguous latitudes being 25 km, as opposed to 111 km for the atmospheric variables.

$$\text{Greenness index} = \frac{2Green - Blue - Red}{2Green + Blue + Red} \tag{4.1}$$

4.2.4 Data preparation

Having defined the various data sources for our study, it was necessary to apply early fusion techniques to leverage all available information. All these variables are georeferenced with X and Y coordinates for each value, forming two-dimensional matrices where columns represent

longitudes and rows represent latitudes. However, differences existed between these variables, including the coordinate systems and spatial/temporal resolutions, necessitating preprocessing for effective integration.

For instance, the greenness index uses the UTM (Universal Transverse Mercator) system, while atmospheric variables utilise the decimal coordinate system. To ensure consistency, the GI variable was converted to the decimal coordinate system. Another dissimilarity was in resolution; the atmospheric variables had a distance of 111 km between latitude values, whereas the GI variable had a distance of 25 km. To harmonise the resolution for integration, we adjusted the atmospheric variables to match the number of columns and rows with the GI variable, enabling the application of early fusion techniques.

Furthermore, there was a difference in temporal resolution; atmospheric variables were recorded daily, while GI was recorded three times per month (from the 1st to the 10th, 11th to the 21st, and the 21st to the end of the month). To address this, we divided the variables into two distinct groups, as follows.

- **Daily**: the variables considered are the 10-meter U wind component and the 10-meter V wind component. They are measured daily, either at 12:00 or 18:00, depending on the specific variable. If both measurements are available, the data from 12:00 are chosen. However, if the 12:00 measurement is unavailable, the data from 18:00 are used. The values recorded on the day of the fire were used for these variables due to their direct relevance to the incidence of wildfires, as incorporated in the Fire Weather Index [59].

- **Trend**: in this category, we included the greenness index, the dewpoint temperature at 2 metres, evaporation, net surface solar radiation, net surface thermal radiation, downward surface thermal radiation, and total ozone columns. The greenness index remained unchanged. For atmospheric variables, the discrete differences in the time series were calculated between the dates of measurement of the greenness index. The rationale for employing the trends of these variables is as follows.

To generate the matrices described earlier, I centre each wildfire and extracted the neighbouring data. Each set of fire coordinates produced a 128x128 matrix for each variable. These matrices overlapped, resulting in samples with dimensions of 128x128x9. The labels were represented as a list of six values corresponding to the wildfire data records. Finally, the data were normalised. Z-score normalisation was applied to all subsets (both labelled and unlabelled samples), while min-max normalisation was used for the labels.

4.2.5 Data

As detailed in Section 4.2.1, the available number of wildfire data records is limited. This limitation restricts the availability of labelled samples, particularly for Deep Learning applications. To address this, we opted to create two unlabelled subsets along with two labelled subsets corresponding to the two autonomous regions for which we have wildfire data records.

- **Unlabelled subsets**: that corresponds to the subsets for which we do not have labels.

  - *Castilla y León*: a set of 445 labeled samples from the autonomous region was utilized. This set was used to test the autoencoder, fine-tuning, and testing the WAM model, which comprised 70% and 30% of the samples, respectively.

– *Andalucía*: a set of 151 labelled Andalucía samples was used specifically to test the WAM model. This set of samples enabled us to assess the system's generalisability when the study area was changed, validating the model's performance in a different geographical context.

4.3 Methodology

The proposed methodology comprises four distinct modules. The initial module involves data pre-processing, wherein I prepare the images and labels for various models. In the second module, we construct the AutoEncoder, training with a random dataset and subsequent testing using the Castilla y León dataset. Then, we take advantage of the trained encoder to create a regression model aimed at estimating wildfire cost variables. With this regression model in hand, we proceed to generate individual maps for each variable, with each pixel value representing a corresponding prediction.

4.3.1 Autoencoder pretraining

We developed an autoencoder for meteorological comprehension, adopting a self-supervised masked image modelling (MIM) approach. Our strategy involves dividing the 128x128 matrix into an 8x8 grid of patches, with each patch randomly masked with a probability of 0.5 (Figure 4.2, second column). Using the unmasked patches (Figure 4.2, third column), the system attempts to predict the average values for the masked regions in each patch (Figure 4.2, last column).

In contrast to most papers, such as those referenced in [60, 61, 62, 63], where autoencoders aim to directly reconstruct the original image, our autoencoder focusses on learning to discern patterns and trends within different channels (representing atmospheric variables). We do not find complete image reconstruction to be sufficiently valuable to justify the added complexity in our modelling. Consequently, discrete categorical labels have been assigned to the mean values of the patches.

In the process of selecting the optimal number of bins for our experiments, we conducted preliminary trials with various bin counts: 4, 8, 16, 32, and 64 categories. Furthermore, for each experiment, we explored different learning rates (5e-5, 1e-4, 2e-4, 5e-4, 1e-3) in order to determine the combination that exhibited the best performance for this task. After preliminary experiments we settled on the following hyper-parameters for WAM pretraining as summarized in Table 4.1.

Encoder architecture As explained previously, the features of the input are mapped into a latent representation by the encoder. In this paper, two different encoder architectures are considered, with the purpose of developing an encoder that can better understand the patterns and trends of the variables.

- *Sequential architecture*: is composed of three different convolutional blocks, as shown in Figure 4.3. In each block, a convolutional layer is used with 128, 256, and 512, respectively, with a kernel size of (3,3), and the "same" padding method is applied to adjust the input size to our requirements. The second layer is the BatchNormalization layer, followed by the activation function, which is ReLU. Finally, a MaxPooling2D layer with a pool size of

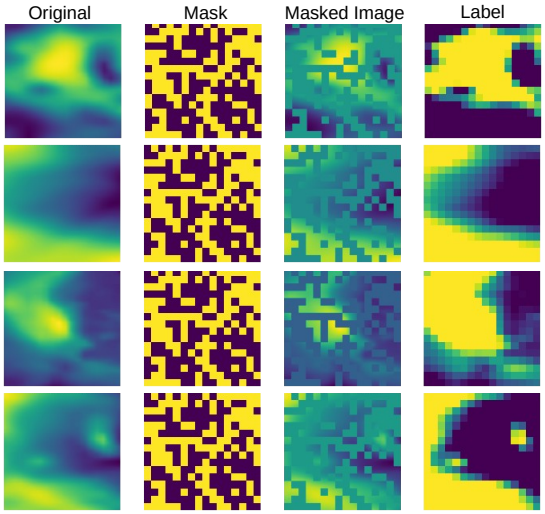


Figure 4.2: Example of the preprocessing showing four channels of a sample, where the first column represents the atmospheric variables; the second column represents the mask for that channel; the third represents masked image and finally the fourth sample represents the label for the autoencoder.

(2,2) is included. Except for the convolutional layer, which has varying values depending on the block, the remaining layers have consistent values across all convolutional blocks.

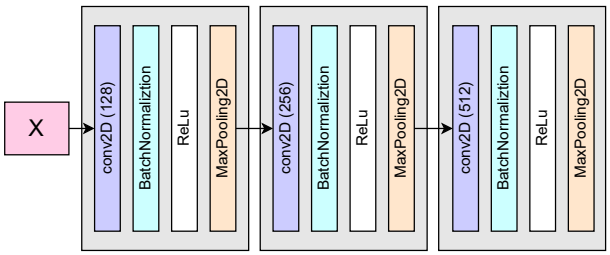


Figure 4.3: Visual representation of architecture 1, the different convolutional blocks can be seen in grey colour.

- *Residual architecture*: unlike the previous architecture, this one incorporates skip connections or shortcuts that enable bypassing certain layers, as shown in Figure 4.4. These connections facilitate the creation of deeper networks while mitigating the vanishing gradient problem. Our architecture consists of three convolutional blocks, similar to the previous one. Each block comprises four convolutional layers with an equal number of neurones, 128, 256, and 512, respectively, along with a batch normalisation layer and a Gelu activation layer. At the end of each convolutional block, a Max pooling layer is included. The convolutional layer uses a (3,3) kernel size with padding. Within each convolutional block, the activation layer receives input from both the preceding layer and the last activated layer, although there are no connections established between the different convolutional blocks.

Table 4.1: Summary of parameters used in pretrain: optimization and training methodology.

Optimization	
Optimizer	Adam
Learning rate	1e-4
Loss function	Sparse Categorical Cross Entropy
Metric	Sparse Categorical Accuracy
Training methodology	
Maximum epochs	2000
Checkpoint monitor	Sparse Categorical Accuracy
Batch size	64
Image size	128 x 128
# bins	64
Patch size	16x16

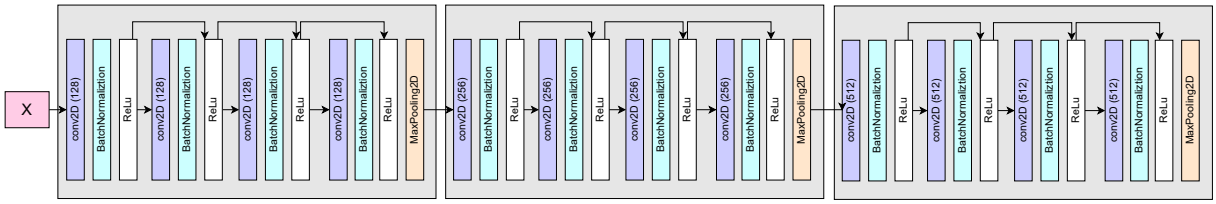


Figure 4.4: The design of Architecture 2 is represented visually, with the individual convolutional blocks highlighted in grey. The skip connections are indicated by arrows, which illustrate their paths within the network.

Decoder architecture The original image is reconstructed by the decoder, which involves filling in the masked patches using the latent representation. To achieve this, the decoder is designed to return the categorical values of these patches. Its architecture is streamlined, consisting of only two layers. The first layer is a dense layer with a neurone count equal to the product of the number of channels in the image (9 in our case) and the number of categories the patches can have (64). This results in a total of 576 neurones in the dense layer. Following this, there is a reshape layer responsible for altering the output dimensions. The desired output shape is (number of patches, number of patches, number of channels, categories). This reshaping ensures that the system generates an image with the same dimensions as the input, thereby completing the reconstruction process.

4.3.2 The WAM

In this approach, wildfires are recognised as inevitable and unpredictable events, their spread influenced by natural factors and mitigated by human intervention. To assist experts in estimating the necessary actions, a model was developed to predict the potential damage caused by a wildfire starting on a specific date. The labels used for this prediction include the burnt area, the control and extinction time, and the human, vehicle, and aerial resources needed for extinguishing efforts. Due to the lack of samples available for deep learning techniques, a transfer learning strategy was used. Pre-trained weights from a previous task were leveraged to fine-tune a regression model.

Initially, the latent representation is flattened and subjected to a dropout operation with a rate

of 70%. The output is then passed through a dense layer comprising 512 neurones activated using the GELU activation function. The final classification layer consists of a dense layer with six neurones and linear activation. During training, the model was trained a maximum of 6,000 epochs using the Adam optimiser with a learning rate of $1e-5$. Mean absolute error (MAE) served as the evaluation metric, while mean squared error (MSE) was employed as the loss function to monitor the model's progress. MAE was chosen as a checkpoint to save the model weights.

For training, the Castilla y León dataset, consisting of 446 samples, was used. To evaluate the generalisation of the model to different environmental conditions, it was tested on the Andalucía dataset, assessing its performance in areas with different characteristics.

4.3.3 Baselines

A private dataset that has not been published yet was used, making it impossible to compare it with datasets used in other studies. Additionally, a novel approach was taken in treating atmospheric variables as two-dimensional matrices. These matrices were centred on the geographical location of the fire, incorporating the surrounding area. This methodology differs from existing approaches in the field. Previous research articles in this domain predominantly employed classical machine learning techniques to tackle tasks such as predicting susceptibility and estimating the area vulnerable to wildfires.

To ensure a fair comparison with state of the art techniques, we selected methodologies identified from reviews in the literature in this field [64, 65, 66], including Decision trees, GBoost, Random forests, Support vector regression and XGBoost.

In the context of the deep learning model, statistics such as mean, standard deviation, and centre point were extracted from the same arrays, creating an array of 27 values for each sample. These matrices were derived from the original samples, which were normalised according to the z-score as mentioned above and were analysed using the explained techniques. In contrast, for the Deep Learning models, Min-Max normalisation was applied to the labels. To evaluate the true performance of the regression models, the predictions were first denormalised, and the Mean Absolute Error (MAE) was measured using the denormalised predictions against the actual labels.

4.3.4 Visualization

In the final step of the methodology of this manuscript, maps of the autonomous region were visualised for each unit of time and label, including burnt area, control time, extinction time, human resources, vehicle resources, and aerial resources. To generate these maps, a sample was created for each pixel of the map, using the input dimensions outlined. In total, 2,970,000 samples were prepared to generate the map. However, there remains a frame on the outside of the map that could not be generated due to input size requirements. Six predictions were extracted from each of these samples. These predictions were then organised into a two-dimensional matrix to reconstruct the maps, one for each specific label.

4.4 Experimentation

In this section, we present the results obtained using the proposed methodology and evaluate its performance on the dataset described in Section 4.2, covering the autonomous regions of Castilla y León and Andalucía from 2001 to 2012. The results are categorised into four main sections. Firstly, we assess the autoencoder's performance, conducting a parameter selection test and comparing the results for the two proposed architectures using the Sparse Categorical Accuracy metric. Second, we evaluate the performance of the proposed regression model and compare it with widely used techniques for similar state-of-the-art problems, employing the Mean Square Error (MAE) metric. The third section presents the generated prediction map. Lastly, the section discusses the known limitations of the proposed approach and the results obtained.

4.4.1 Performance of Autoencoder models

As previously mentioned, an initial hyper-parameter search was conducted before training the final models. The objective of this search was not to maximize accuracy but to identify the most challenging objective the model could still solve with reasonable accuracy. As detailed in Section 4.2, various combinations of pixel value categories and learning rate values were explored. The results are summarized in Table 4.2. Notably, models with a lower number of bins achieved better accuracy due to the limited number of epochs used. However, it was observed that all models were able to learn patterns and fill the hidden patches. Based on the results obtained for the combination of learning rate = 1e-4 and 64 bins, achieving an accuracy of 0.6, these parameters were selected for training the models.

For subsequent experiments, a higher number of epochs and a larger batch size were used. This choice was informed by the model's ability to learn fine-grained patterns and trends in different atmospheric variables, resulting in a more robust intermediate representation of the input data. Additionally, the model trained with a learning rate of 1e-4 demonstrated superior performance.

Table 4.2: Accuracy of each parameters combination in the AE training.

	Number of bins				
Learning rate	*4*	*8*	*16*	*32*	*64*
1e-4	0,975	0,939	0,852	0,736	0,600
5e-5	0,975	0,946	0,893	0,786	0,528
2e-4	0,966	0,949	0,876	0,731	0,567
5e-4	0,969	0,939	0,874	0,767	0,542
1e-3	0,968	0,949	0,878	0,759	0,558

Upon selecting the training parameters, the proposed architectures, both sequential and residual, were trained. The results of the labelled training dataset, trained and validated with the random unlabelled dataset, are presented in Table 4.3. Notably, the residual architecture outperformed the sequential one, achieving a score of 0.861 compared to the sequential architecture's 0.773. These results are further validated by Figure 4.5, where the residual architecture successfully reconstructs an image from the latent representation. Despite the differences in their results, regression models were created using both encoders to assess their respective performance.

The results, illustrated in Figure 4.5 and summarised in Table 4.3, indicate that the autoencoder models effectively comprehend patterns and trends within various atmospheric variables and the greenness index. In particular, the residual architecture outperforms the sequential one.

Table 4.3: Results of the two AutoEncoder architectures evaluated with the labelled training set.

	Accuracy
Sequential architecture	0.773
Residual architecture	**0.861**

This discrepancy can be attributed to the differences in the number of convolutional layers in each block and the incorporation of skip connections or shortcuts. The residual architecture, with its larger number of trainable parameters and the utilisation of skip connections, effectively mitigates vanishing gradient issues, aligning with anticipated behaviour.

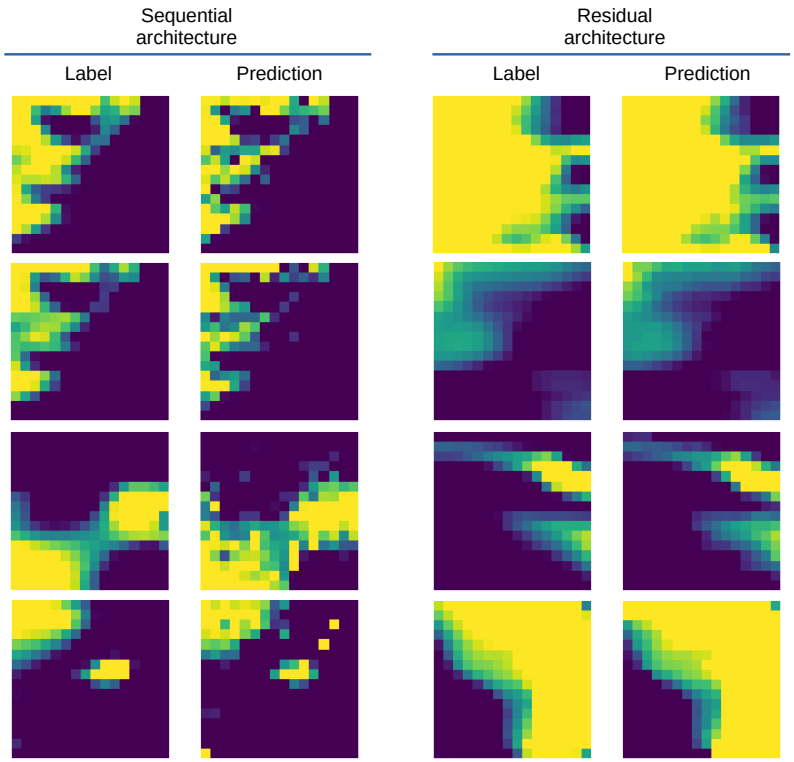


Figure 4.5: The results of the two AutoEncoder architectures are demonstrated in the examples provided for different samples from the labeled training set. The first and third columns represent the AutoEncoder labels, while the second and fourth columns display the predictions generated by the two proposed architectures: sequential and residual.

4.4.2 WAM

The performance analysis of the proposed regression models, as shown in Table 4.5, reveals valuable information. These models were tested using the two encoders generated in the previous step, both with and without fine-tuning. To validate their efficacy, comparisons were made with five different baseline techniques, chosen as the most common state-of-the-art methods available at the time of writing. The analysis indicates that Support Vector Regression consistently performs the poorest, displaying the highest MAE values in most classes except for aerial resources and control time. The decision tree model also underperforms, particularly in control time and

aerial resources. Among the baselines, GBoost and random forest exhibit the best results, outperforming models with frozen encoders for the human and heavy-resource classes. In particular, all baseline predictions surpass the average of the training values.

Comparing our techniques with the baselines, it is evident that the sequential architecture occasionally fails to outperform the baselines. However, in three classes (burnt surface, control time, and extinguishing time), it performs similarly to the fine-tuned encoder. On the other hand, fine-tuning the entire encoder for the task consistently achieves the best performance. When comparing the regression models generated with the encoders from the sequential and residual architectures with fine-tuned encoders, the residual architecture consistently outperforms the others. It achieves the lowest MAE for three classes: burnt area, control time, and aerial resources. Consequently, it emerges as the model with the best overall results among the nine models presented in Table 4.5.

These results underscore how deep learning techniques can surpass the classical machine learning models prevalent in the field of fire management. Furthermore, the results demonstrate the potential of few-shot learning techniques in achieving remarkable results in fire management even with a limited number of samples. In particular, the fine-tuned residual encoder emerges as the most promising approach, showcasing the potential for further enhancements despite the already promising results.

Table 4.4: Results of the four regression models generated, frozen encoder and fine-tune encoder from both architectures, compared with the average and five models from the state of the art. Best result in bold.

	Average baseline	Decission Tree	GBoost	Random Forest	Support Vector Regression	XGBoost	Frozen encoder		Fine-tune encoder		Improvement (%)
							Sequential architecture	Residual architecture	Sequential architecture	Residual architecture	
Burnt Area (m)	406,189	434,473	333,814	287,051	747,340	302,683	280,800	255,800	278,200	**233,100**	18,8%
Control Time (min)	1846,537	1875,236	1274,760	1416,345	1660,959	1295,846	1218,000	1212,000	1192,000	**1114,000**	21%
Extintion Time (min)	3255,813	2566,528	2630,425	2538,686	2797,823	2654,167	2234,000	2338,000	**2230,000**	2280,000	10,2%
Human resources (units)	87,582	66,289	54,785	53,373	82,152	57,140	73,700	**51,100**	73,100	52,600	1,4%
Heavy resources (units)	6,047	5,056	4,139	4,011	6,450	4,528	5,918	**3,672**	5,930	3,863	3,7%
Aerial resources (units)	5,197	4,309	3,283	3,169	3,626	3,318	3,684	2,988	3,744	**2,883**	9%

4.4.3 Visualization

As detailed in Section 4.3, in addition to testing our model in the designated test set, we generated a prediction map for a specific date using three test samples to analyse its functionality and performance. These visualisations were intended to assess risk assessment across the autonomous region during a fire outbreak. As depicted in Figure 4.7, individual maps were created for each label, indicating the resources required in the event of a fire.

Upon examination of these maps, it is evident that the areas near the black spots, which denote specific wildfire incidents, require the most resources. Comparing these results with forest and wildfire maps (Figures 4.1 and 4.6 respectively) reveals an alignment. For example, three wildfires in the western part of the autonomous region coincide with the areas that show the highest values for the six labels. Moreover, these regions often overlap with dense vegetation areas. Although these visualisations lack a direct ground truth for comparison, they provide indications

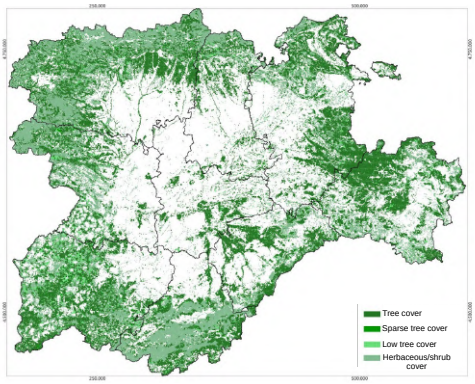


Figure 4.6: The map of Castilla y León illustrates the forest cover in varying shades: dark green denotes dense tree cover, green represents sparse tree cover, light green indicates low tree cover, and grey-green signifies herbaceous-shrub cover. This adaptation is based on the original forest map provided by the Ministry for Ecological Transition and Demographic Challenge in Spain.

of potential fire damage to specific areas. The identification of high risk regions is based on conditions conducive to fire spread, serving as indicators of increased fire probability in those areas. The goal is not to pinpoint exact fire locations, but to provide information on the potential impact and severity of a fire in different areas.

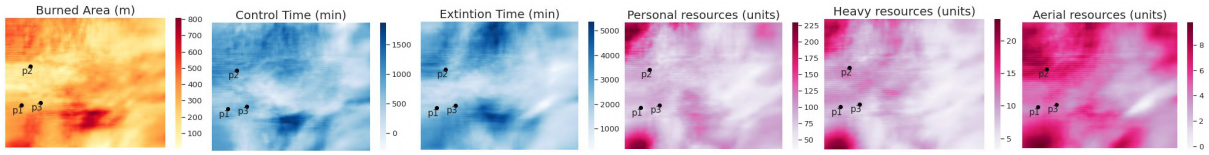


Figure 4.7: Prediction maps, visualization of burnt area, control and extinction time and needed resources.

4.4.4 Known limitations

As we explained earlier in Section 4.2, we used the Andalucía dataset to check the generalisation of the model without fine-tuning with samples from the new area with different environmental conditions. For this purpose, we used the regression model that obtained the best performance in the Castilla y León dataset, the fine-tuned residual architecture encoder. Like in the previous section, we compare our model with five different baselines and the average.

First of all, we can observe that the baseline results are similar in both experiments, except for the last label, aerial resources, where all baselines perform worse than the average baseline. Second, it is notable that the support vector regression and the XGBoost models obtain the same results for both datasets. The rest of the models for the other five labels achieve better results than the average. About our model, only two labels have a lower MAE than average one: control and extinction time. The other variables have a higher MAE than average ones, burnt area, human, heavy, and aerial resources. It is important to note that there are some other factors, such as: distance between the water sources and the fire location, type of plane or the orography, that may have an influence on the resource prediction.

The model we used to test the new Andalucía samples was fine-tuned with a set of 446 samples from Castilla y León, a limited region with different meteorological features and vegetation than

Andalucía, so the model is not able to correctly predict the resources that would be needed in case of fires, except for the control and extinction time classes. The model has been trained by means of the atmospheric variables in a region surrounding the Castilla y León region. Thus, different regions are affected by changes in meteorological conditions. The model may need to be adjusted if it is to be applied to other areas, i.e., pretraining with a small sample set of the area of interest. Another factor that could also affect its generalisation is that the encoder was trained with a set of random samples from Castilla y León.

Table 4.5: Results of residual architecture with fine-tune encoder and baselines for Andalucía dataset.

	Average baseline	Decision Tree	GBoost	Random Forest	Support Vector Regression	XGBoost	WAM
Burnt area (m)	393.209	391.556	320.148	**293.271**	747.340	302.683	751.5
Control Time (min)	1747.570	1713.208	1276.427	1371.579	1660.959	1295.846	**950.0**
Extintion Time (min)	3067.964	2798.904	2559.446	2523.677	2797.823	2654.167	**2268.0**
Human resources (units)	83.971	63.671	56.237	**53.931**	82.152	57.140	206.6
Heavy resources (units)	5.996	5.247	4.189	**4.061**	6.450	4.528	14.414
Aerial resources (units)	4.993	6.107	5.817	**5.665**	6.103	5.918	11.01

Chapter 5

Explainable AI for image classification

El mideo tienen raíces difíciles de arrancar,
si ves que se hacen cadenas,
rómpelas y échate a andar.

— J.A. Labordeta

The application of CNN in image processing has revolutionised the field of medicine, enabling significant advances in the diagnosis and treatment of various diseases. However, as these techniques gain ground in clinical settings, there is a growing need to understand and explain the decisions made by models. This is where the concept of XAI comes into play, which refers to the ability of these networks to provide clear and understandable explanations of how they achieve their results. In the medical context, the relevance of XAI techniques cannot be underestimated, as the transparency and interpretability of the models are crucial to gain the trust of healthcare professionals and ensure a safe and effective use of artificial intelligence in medical decision making. This chapter will explore the importance and applications of XAI in medicine, highlighting its potential to improve diagnostic accuracy and streamline the work of doctors. Therefore, this chapter will show six different types of visualisation that fit different classification tasks and different proposed methodologies.

5.1 Problem definition and objective

As explained in Chapter 1, DL models have proven successful in many problems, including medicine. X-rays are clinical images that can be analysed with CNN, as we have seen in the previous two chapters, Chapters 2 and 3. However, the results provided by these classification systems are opaque and insufficient in the medical field, making it difficult for doctors to trust them. It is important to highlight that DL systems in medicine are not diagnostic systems, but rather diagnostic support systems, i.e., they are tools that help healthcare professionals to reduce diagnostic errors and speed up the process. DL classification systems provide an estimate of the probability that the sample belongs to a specific category, which is not informative for doctors, making it difficult to apply these systems in the medical domain. To overcome this obstacle,

different visualisations have been created that adapt to different classification tasks, both binary and multilabel, and to possible systems designed for this purpose, simple models and ensemble systems. Therefore, this chapter will focus on solving ***RQ4: "How can XAI techniques be adapted to different binary and multilabel classification problems to facilitate the use of DL models within the medical domain?"***

The aim of this research is to perform an in-depth analysis of different visualisation techniques of the activation maps obtained by means of XAI techniques, with this aim we will address two classification tasks: binary and multilabel, although as we will see later on these techniques can also be applied to multiclass tasks; and for two different approaches: systems composed of a single CNN and for ensemble-based systems, which will be composed of several independent models. In the latter case, ensemble-based systems, a heatmap will be obtained for each model, which will have to be combined, as well as the results of the different models. For this purpose, we will use the Gradient-weighted Class Activation Mapping (Grad-CAM) technique to extract the relevant information for the model and then create the visualisation [34]. This chapter will not focus on the different methods of extracting information, but rather on how to represent this information, seeking a balance between providing as much information as possible and making it useful for the end user.

Therefore, the main objective of this research can be divided into the following specific objectives.

- Design of the different classification systems.
- Obtaining the activation maps of the layers of interest.
- Calculation of the importance of the features.
- Create the final visualisations, depending on the features of the proposed problem.
- Analysis of the advantages and disadvantages of the different options available.

5.2 Methodology

This section describes the different heatmaps representation techniques with the aim of enhancing their explainability and applicability in fields beyond computer science. This enables the intrinsic opacity of this type of technique to be overcome.

In this approach, techniques are divided into two groups to address explainability in artificial intelligence applied to medicine. The first group corresponds to classification tasks where the classes are mutually exclusive, which would encompass multiclass and binary tasks, although we will use a binary classification problem as an example. These techniques are appropriate when the goal is to assign a specific label to a patient or a medical image, such as the detection of a specific disease. The models and ensemble systems presented in Chapter 2 will be used for this purpose. The second group would coincide with tasks where the classes are not mutually exclusive, i.e. multilabel tasks. These techniques are especially useful when it is necessary to identify and assign multiple radiological conditions or signs to a patient or a medical image, such as the diagnosis of co-existing diseases or the classification of multiple abnormalities in medical images, that is, it is closer to the reality of healthcare facilities. The models and ensemble systems introduced in Chapter 3 will be employed for this objective.

5.2.1 Common steps

The CNN, from which the heatmaps are to be obtained, needs to be trained first, with one neurone corresponding to each class in the classifier, which is particularly relevant in binary classification problems, where there might be only one neurone in the classifier. This allows the extraction of the areas utilised by the system to provide the probability of each class. After the model has been trained, the activation function of the classifier or the output layer needs to be changed from softmax to linear.

Second, we have to choose from which convolutional layer we want to extract, in this case the last convolutional layer will be used because it contains high-level information about the features detected in the input image. Then, the gradients of the target class score with respect to the activations in the selected layer are calculated. This is done using the backpropagation algorithm. Gradients indicate how the target class score changes as each activation in the layer is modified. The gradients are used to weight the activations in the selected layer. This means that activations that contribute more to the network decision, according to the gradients, are highlighted more in the heatmap calculation. A weighted aggregation of the activations of the selected layer is performed using the gradients as weights. The result is an activation map that highlights the regions in the image that were most influential in the CNN decision for the target class. These matrices are superimposed with the original radiograph to generate the visualisations, since the matrices alone lack informativeness but indicate areas of the original sample. To ensure that the original image is visible, a 50% transparency is applied to the heatmap.

Once the heatmap is obtained, it can be represented in different ways depending on the proposed methodology and the requirements of the medical staff who will utilise the tool. The proposed system can consist of a single model or an ensemble of several models, each with its visualisation. Second, the visualisation needs to be adapted to the user's needs, in this case the medical staff. Striking a balance between the information provided and the readability to speed up the interpretation of medical imaging tests and reduce subjectivity is crucial. Otherwise, the task of medical staff will be unnecessarily complicated.

5.2.2 Visualisation techniques for binary classification tasks

In this type of classification problem, the model or system must return a single label per sample. However, two visualisations can be generated, one for each label. This section presents two methodologies based on the number of models that make up the classification system.

Classification system based on one model: This represents the simplest scenario, involving a single classification model tasked with selecting one of the two available classes, where the probabilities of both classes must sum up to 100%. In such systems, we acquire the matrices for the two neurones in the output layer. Depending on the preferences of healthcare staff and the nature of the classification problem, such as distinguishing between pathology and healthy patients, we can generate heatmaps for either the class of interest or both classes. As an illustrative example, we consider a binary classification problem involving chest X-rays, distinguishing between consolidation, bacterial pneumonia, and nonconsolidation, which corresponds to other infiltrates [35].

Classification system based on ensembles: These classification problems produce a matrix for each output neurone of each model within the ensemble system. These systems produce a final

result by fusing information from multiple models, necessitating the combination of visualisations based on heatmaps from these diverse models. To accomplish this, the same technique employed for probabilities is often applied, with averaging being the most commonly used method for either probabilities or matrices. Averages are computed for each pixel to derive the final matrix. In the case study, the same classification problem as in the previous scenario is addressed, but in this case, an ensemble of five distinct models is employed. For this instance, the decision was made to utilise the mean to generate heatmaps for each neurone and depict regions with less consensus, while an additional heatmap was created to represent the standard deviation among the various matrices. In this problem, heatmaps were generated for both output neurones [35].

5.2.3 Visualisation techniques for multilabel classification tasks

Multilabel classification problems are inherently more complex than binary ones, offering a wider array of visualisation possibilities. In contrast to binary classification, in multilabel scenarios, it is imperative to depict heatmaps for all classes or, at the very least, those classes detected within the sample. When creating these visualisations, two key considerations come into play, along with accommodating healthcare staff preferences: first, whether the approach involves individual models or ensemble systems and second, whether each class should be showcased through separate visualisations or consolidated into a single comprehensive visualisation.

Classification problem based on one model and combining the heatmaps In these scenarios, in contrast to binary classification problems, it's necessary to extract matrices for at least all the classes detected in the medical image. In this specific context, as we aim to amalgamate all the heatmaps into a unified visualization, only the matrices corresponding to the detected classes will be extracted. Subsequently, all the heatmaps are superimposed onto the original image with a 50% transparency. To facilitate distinction, each class is assigned a unique and distinguishable color. In close proximity to the image, a legend is incorporated, detailing the class detected in each region of the image. Depending on the preferences of the physicians, the legend may also include the probability associated with each class. For the case study illustrating this amalgamation of options, we are dealing with a collection of X-rays encompassing a total of 54 distinct classes, spanning chest and upper abdomen diseases [36].

Classification problem based on ensemble system and one heatmap per class The main difference of this approach with respect to the previous one is that it can also represent the heat maps of the classes that have not been detected in the sample, which can be interesting as long as the total number of classes is not too high. Unlike the proposed binary system, in this case two different techniques have been proposed, the first one without displaying the uncertainty heatmaps and the second one displaying them. In addition, to know the number of models that have detected that class at the top of the heatmap, we will show the average probability and the number of models that agree with that result numerically [36].

5.3 Experimentation

The results obtained with the proposed methodologies and a qualitative analysis of them are described in this section. Visualisations based on heatmaps are crucial for the application of Deep Learning techniques in various domains, with particular significance in fields like medicine, where incorrect decisions can have detrimental effects on patients. In this section, various visualisation

options will be introduced, applied to two classification problems: one involving paediatric radiography distinguishing between consolidation and nonconsolidation, and another dealing with multilabel classification in adult chest radiography involving 54 distinct classes.

5.3.1 Binary classification tasks

In the binary classification problem involving paediatric chest x-rays, three distinct visualisations have been generated. The first visualisation, as shown in Figure 5.1, is the easiest. This classification system comprises only one model. The visualisation consists of three components: on the left, the original x-ray; on the right, the heatmap for the consolidation class, with the associated probability displayed on top; and at the bottom, a colour scale representing the pixel relevance. This visualisation offers the advantage of being highly comprehensible, containing less information compared to other representations, and being exceptionally easy to interpret.

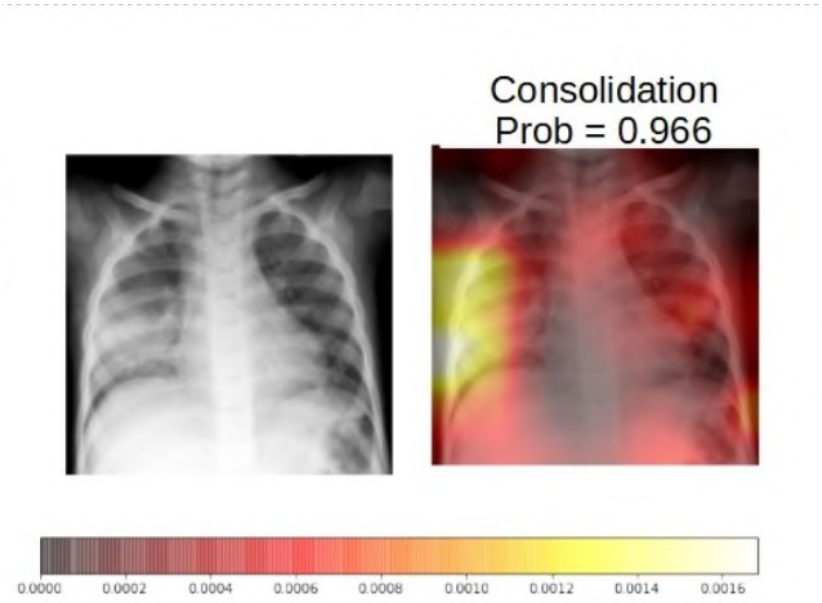


Figure 5.1: Visualization 1: On the left, the original x-ray image; on the right, heatmaps illustrating neuron consolidation with corresponding probability indicated in the title; at the bottom, a color scale representing pixel relevance.

The second visualisation is quite similar to the previous one. It also represents the result of the normal neurone, corresponding to the x-rays that show the nonconsolidation heat map, as seen in the centre of Figure 5.2. This heatmap can offer valuable information to physicians, although it may pose a greater challenge for interpretation compared to Figure 5.1. The ease of interpretation depends on the preferences of the medical staff.

Figure 5.3 illustrates a visualisation of a classification system comprising five distinct models. It shows the heatmaps for both classes along with the standard deviation of these heatmaps. In the top row, moving from left to right, you can observe the original x-ray, the heatmap for the normal class, and the heatmap for the consolidation class. The bottom row presents the heatmaps for the standard deviation of the normal and consolidation classes, respectively. In particular, the highlighted region on the consolidation class heatmap corresponds to the area indicating signs of consolidation. Additionally, on the standard deviation heatmap for the same class, it becomes

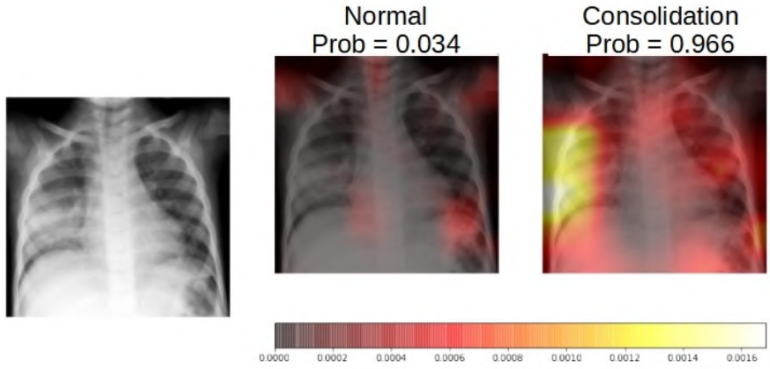


Figure 5.2: Visualization 2: On the left, the original x-ray image; in the center, a heatmap illustrating the "normal" class; on the right, heatmaps depicting the "consolidation" class; at the bottom, a color scale representing pixel relevance. The probabilities for each class are indicated in the titles of the respective heatmaps.

apparent that areas of greater uncertainty are located at the periphery of this region or outside the lung.

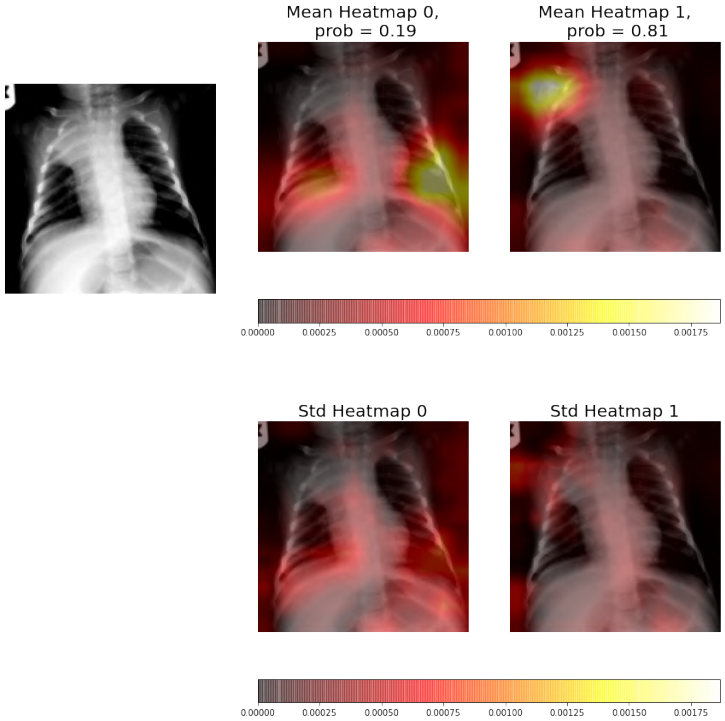


Figure 5.3: Visualization 3: Heatmaps from the ensemble system illustrating uncertainty. In the first row, from left to right: the original x-ray image, a heatmap for the "normal" class, and a heatmap for the "consolidation" class. In the second row on the left, you can observe the uncertainty associated with both heatmaps.

5.3.2 Multilabel classification tasks

Figures 5.4, 5.5 and 5.6 show the same chest X-ray, showing a total of 3 different pathological signs: cardiomegaly, pacemaker, and sternotomy. However, the visualisations are very different from each other. Figure 5.4 is the most different, where we have chosen to combine the heatmaps of the three classes into a single image and we have chosen to include only the visual information. This visualisation is the easiest to read, as it does not include any additional information and would generate less bias for the doctors reading it. Figure 5.5, on the other hand, shows a visualisation for each detected class, and above each visualisation the probability obtained from the system and the agreement between the different models that compose it. This visualisation is less clear than the previous one but provides more information than the previous one, as it reports the degree of agreement between the different models, numerically. The last visualisation, Figure 5.6, is the most complete of the three, but also the most difficult to read. The main difference from Figure 5.5 is that in addition to presenting the agreement between the models numerically, it also shows the uncertainty of the heatmaps, i.e., the areas with the highest variability (second line of visualisations), which allows medical staff to know the safest areas of the heatmaps.

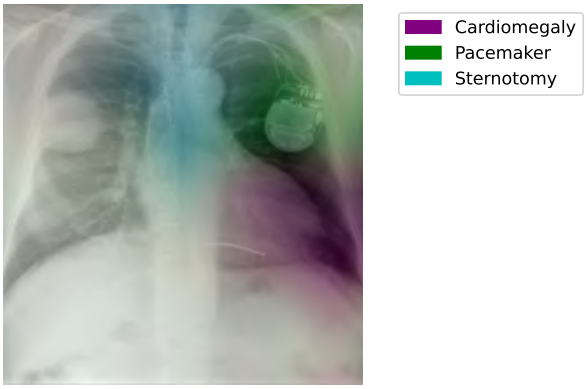


Figure 5.4: Visualisation 4: A combination of heat maps representing three different classes (cardiomegaly, pacemaker and sternotomy) each represented in a unique colour. The original radiograph is omitted in this visualisation.

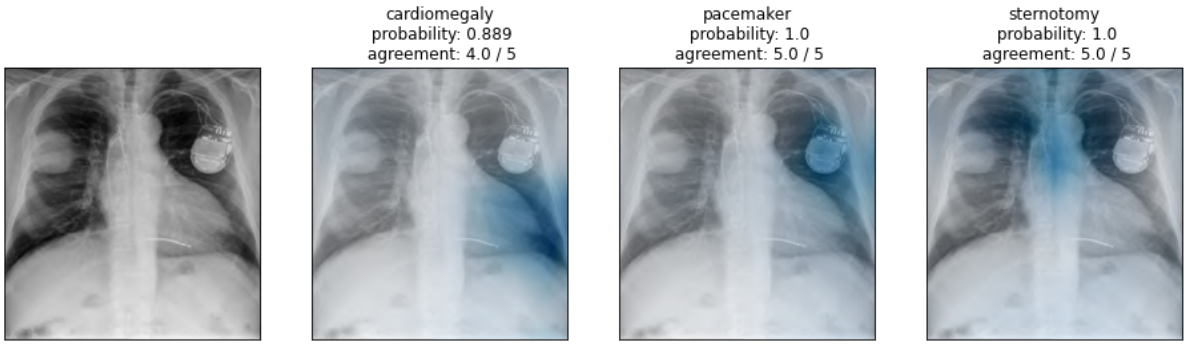


Figure 5.5: Visualization 5: Heatmaps illustrating the detection of four radiological signs: cardiomegaly, pacemaker, and sternotomy. The title of each heatmap displays the associated label, the estimated probability calculated by the ensemble, and the consensus among the ensemble models. The regions of interest for classification are highlighted in blue.

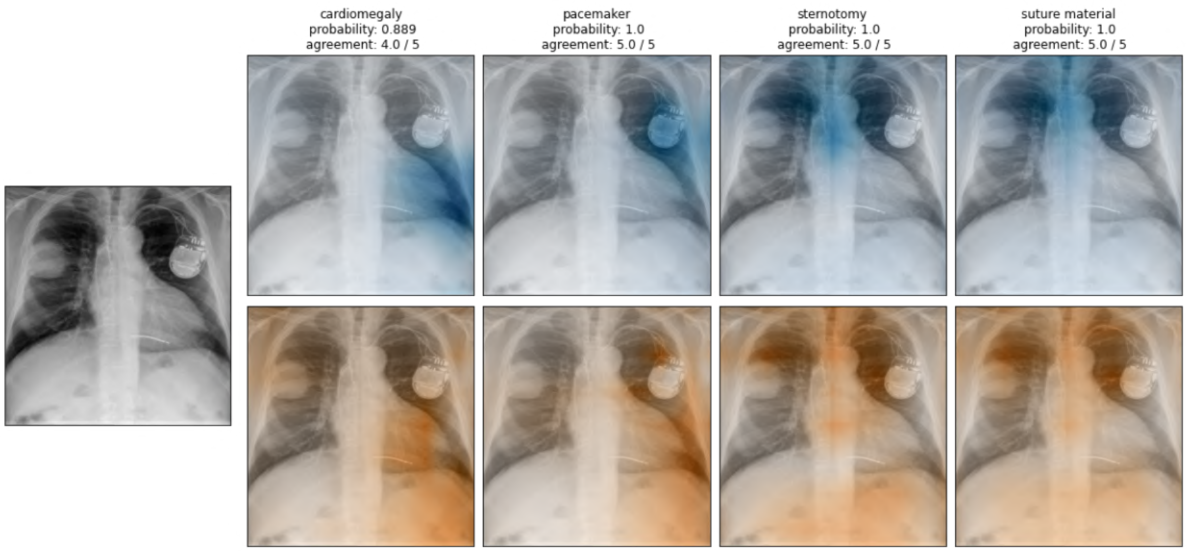


Figure 5.6: Visualization 6: From left to right: Heatmaps for cardiomegaly, pacemaker, sternotomy, and suture material. The first row shows the ensemble mean heatmap, while the second row represents the standard deviation, indicating the uncertainty between models.

5.4 Analysis

As evident throughout this research, XAI techniques offer numerous possibilities to represent the information derived from classification systems. The choice of the most suitable combination depends on the specific problem and the preferences of the end users. In this study, we have thoroughly analysed various options, highlighting the strengths and weaknesses of each in Figure 5.7.

It should be noted that all options have their advantages and disadvantages, and there is no single technique that outperforms the others. Striking the right balance between the amount of information provided and ensuring sample readability is crucial. The most informative combination might become overly complex, making it challenging to interpret due to an excess of information. Conversely, some techniques, while easy to read, may lack sufficient detail required by the end user.

The methodology presented in Figure 5.2 was evaluated by two senior physicians who analysed 40 randomly selected radiographs. Their analysis showed a 70% agreement rate with the heatmaps. Additionally, experiments were conducted to assess whether the application of these techniques improved doctor performance and in all cases, improvements were observed **?**]. This demonstrates the utility of heatmaps in the medical field. It is important to note that the 30% of cases where doctors did not agree with the heatmaps highlight that these techniques only represent how classification systems work, and the quality of the visualisation is entirely dependent on the quality of its underlying algorithms.

Figure 5.7: Strengths and weaknesses of the different components that can be included in a visualisation based on heatmaps.

Chapter 6

Application of information fusion, ensembles and to other domains

Seeing is in some respect an art,
which must be learned.

— Caroline Herschel

Throughout the previous chapters a number of solutions to different problems related to image datasets and the lack of explainability of DL-based models and systems have been presented, so it was decided to continue this dissertation by extending the field of knowledge and introducing other multimedia modalities. The aim is to apply the knowledge of this dissertation to other more complex domains, such as video, which includes temporal and spatial information, and video includes another modality that can be analysed jointly or independently, audio. As a starting point for this research line, it was decided to carry out a ***state of the art review of DL techniques*** applied to video and audio within the field of disinformation, more specifically within the field of ***multimedia data manipulation detection***.

6.1 Problem definition and objective

In this chapter, a state of the art review of the last six years in the field of multimedia, video and audio, data manipulation detection techniques has been carried out, with the aim of obtaining an overview of the main problems of datasets and the techniques most commonly used by authors to solve them, which will allow me to identify less explored areas and opportunities. By examining the temporal evolution and frequency of publications on these topics, we have gained valuable information on current research trends and possible knowledge gaps. This chapter is dedicated to addressing the ***RQ5: "How can the techniques of information fusion and ensembles analysed in this dissertation be applied to other multimedia information modalities, within the domain of information disorders?"***

Within this research, I will focus on the two main streams of investigation: the generation of manipulated multimedia data and the forensic techniques needed to detect it. Within the forensic techniques we will also analyse the tools implemented and available to social network users, who

are not experts in this area. For this reason, this research has been organised around four research questions.

- **RQ1**: What are the current topics and works on multimedia data forensics?
- **RQ2**: Which publicly available datasets/data sources are currently used in multimedia data forensics?
- **RQ3**: What techniques are used in the detection of multimedia data manipulation?
- **RQ4**: What detection tools are used in multimedia data manipulation?

The main contribution of this article, resulting from the exploration of these research questions and the methodology used to address them, can be summarised as follows.

- It offers an up-to-date overview of video and audio manipulation techniques as they relate to social networks.
- It provides an updated picture of available datasets and current video and audio manipulation techniques.
- It provides an up-to-date overview of audio and video forensic techniques, together with the availability of trained models in this field.
- The research introduces available tools designed for end users, intended to assist authors interested in conducting experiments or making advancements in this field.
- It outlines the current trends, challenges, and potential research directions within the field of forensics.

6.2 Methodology

In this section, the systematic process employed for conducting a survey on articles concerning forensics, or manipulation detection, techniques in video, audio, or both domains will be described, based on PRISMA protocol [67]. The process included the collection of articles from four major scientific databases: Scopus, ScienceDirect, IEEE Xplore, and arXiv.

The steps for conducting the manual screening process and reviewing the articles were as follows:

1. *Search in databases*: The focus of our research is on the detection of multimedia content manipulation. However, due to the exceptionally large volume of articles available and the predominant focus of existing surveys on image manipulation detection, we have opted to focus on three specific areas: *video manipulation detection*, *audio manipulation detection*, and *multimodal manipulation detection* (involving both video and audio content). Additionally, we have included the thesaurus term "social media" because these platforms are primary sources of such manipulated content. Thus, the final thesaurus used for our research comprises the following keywords: ("Video manipulation detection" OR "audio manipulation detection" OR "multimodal manipulation detection") AND ("social media"), between 2018 and 2023 (up to May).
2. *First screening*: Surveys and reviews of the literature and all articles not belonging to the

Table 6.1: Articles extracted from different databases about forensics in video, audio and multimodal.

Data source	# articles
ScienceDirect	688
Scopus	51
arXiv	30
IEEE Xplore	26
Total	**795**

field of computer science were eliminated. Only research articles related to DL techniques to detect manipulation residues in multimedia data were kept.

3. *Second screening*: The title, abstract, and methodology were examined to assess whether the review inclusion criteria were met for the articles. The summary criteria are as follows:

   - The articles must be written in English.
   - The methodology must be clear, including all the necessary details for its implementation, and the results must be clearly explained.
   - The analysis must be quantitative.
   - The articles have to apply forensics techniques for extracted multimedia information.

4. ***Third screening***: The content of the articles was reviewed more thoroughly in a second stage, where the content of each document was carefully read and those that were found not to satisfy the criteria mentioned above were excluded. After this step, 729 items were discarded and *66 articles* related to forensics remain in multimedia data, video, audio and multimodal.

5. ***Analysis of the selected articles and extracting information***: The final step involved the analysis and comparison of the information extracted from the articles on the detection of multimedia data manipulation.

6.3 Initial Analysis

In this section, the articles to be analysed throughout the survey and the evolution of the field of forensics will be described during the period defined in the methodology, from 2018 to 2023. The aim of this section is to create an overview of the state of the art and thus facilitate the understanding of the next sections.

First, a comparison will be made between the number of datasets and the forensic works published throughout the study period, allowing the analysis of trends in the different modalities investigated in the survey (video, audio and multimodal); Figure 6.1 will be referenced. It can be observed that in the initial years, from 2018 to 2021, video-related papers, particularly those focussing on the visual aspects of video content, excluding audio, dominate. Subsequently, there is a deeper exploration of audio and multimodal datasets, encompassing both visual and

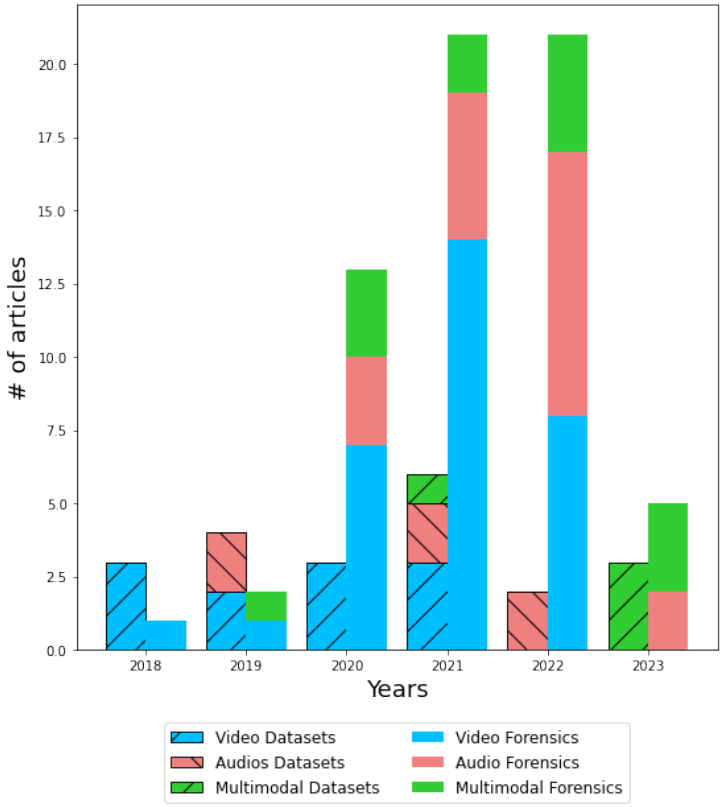


Figure 6.1: The evolution over the years of articles on manipulation and forensics techniques, divided into the three categories of the study: video, audio, and multimodal, will be examined.

audio information. In particular, the study of forensics in multimodal data gains significance in 2021, accompanied by the creation of new datasets adapted to this problem. In summary, there has been growing attention in this field in recent years, and evolving trends indicate that multimodality will assume a more prominent role in this area of study in the coming years.

Turning our attention to Figure 6.2, which examines the evolution of datasets over time, only datasets that provide information on the number of samples have been included in this analysis. Two variables are employed for this assessment: the number of manipulation techniques utilised, represented on the y-axis, and the number of samples within each dataset, depicted by the size of the circles. It is apparent that datasets tend to incorporate a growing number of manipulation techniques, and these datasets are expanding in size, indicating a clear trend towards increased complexity.

Furthermore, in terms of distinctions between multimodal, audio and video datasets, the same trend observed in Figure 6.1 is evident. Additionally, it should be noted that multimodal datasets exhibit more constrained characteristics compared to their video and audio counterparts, both in terms of the number of manipulation techniques used and the total number of samples included. Interestingly, while audio datasets exhibit similar or even larger sizes compared to video datasets, this has not been mirrored in the volume of publications pertaining to audio forensic techniques. However, it remains plausible that the number of works focused on audio forensics will grow in the coming years.

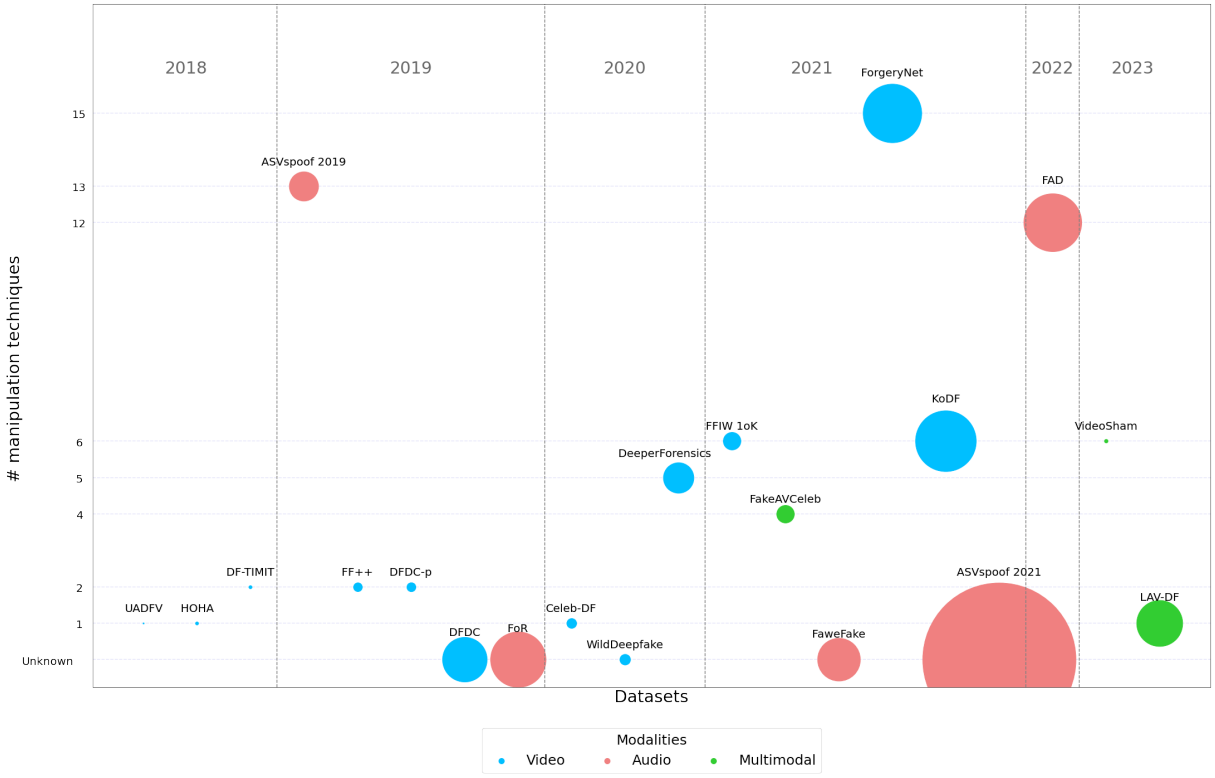


Figure 6.2: The evolution of the datasets is examined chronologically, comparing two variables: the number of manipulation techniques used, depicted on the y-axis, and the total number of samples in the dataset, represented by the size of the circles.

6.4 Datasets for forensics

The creation of high-quality and realistic datasets is of immense importance in the field of multimedia data forensics. These datasets serve as vital resources, offering a collection of authentic and manipulated content samples essential for the development and evaluation of effective detection algorithms and techniques.

Realistic datasets empower to replicate diverse realistic scenarios in which audiovisual media manipulations can occur. This is particularly vital, given the ever-evolving and increasingly sophisticated nature of manipulation techniques. Access to authentic data enables us to train and assess the capabilities of detection systems to handle a wide spectrum of manipulations, thus enhancing their ability to identify counterfeit content. Furthermore, dataset quality is essential to prevent bias and ensure outstanding results. A high-quality dataset should encompass a variety of samples from different contexts, geographic regions, cultures, and demographic groups. This approach mitigates inherent biases and ensures that detection algorithms perform consistently across all situations and user groups.

In summary, the construction of high-quality and realistic datasets serves as a fundamental tool to advance video and audio forensics. These datasets form the essential basis for the development of more resilient and efficient algorithms capable of confronting the growing sophistication of manipulation techniques.

6.4.1 Video

In this section, the attention turns to video manipulation datasets, specifically in which visual content is emphasised. Table 6.2 provides a brief summary of the video datasets used in the literature.

UADFV [68] stands as the pioneer dataset for deepfake detection, but its low quality and easy detectability limit its applicability. ***HOHA-based dataset*** [69] selects realistic videos from known movies, providing a mix of authenticity and manipulation. ***Deepfake-TIMIT*** [70] presents videos of varying quality, allowing the analysis of resolution's influence on deepfake detection. ***FaceForensics++,*** [71] an extension of FaceForensics, is widely used and offers a range of manipulation techniques to assess detector robustness.

DFDC [72] stands out for its massive size and diverse manipulation techniques, providing a significant challenge for researchers. ***Celeb-DF*** [73] aims to overcome limitations of other datasets, focusing on the presence of visible artifacts. ***DeeperForensics*** [74] enhances sample quality through image enhancement techniques, creating substantial variability in appearances and scenarios. ***WildDeepfake*** [28] excels in complexity, representing a realistic variety of situations and manipulation techniques. ***ForgeryNet*** [75] is especially useful for manipulation localization tasks and assessing resistance against perturbations. ***KoDF*** [76] introduces a wide diversity of generation techniques and post-processing applications, resulting in significant variability in quality and realism. Finally, ***FFIW 10K*** [77] provides a large amount of data with detailed annotations, being essential for research into the detection of facial manipulations in multi-person videos.

Together, these datasets offer a comprehensive overview of digital manipulations, from the simplest and easily detectable to the most complex and realistic, enabling researchers to develop and evaluate more effective and robust detection methods against emerging threats in the field of video manipulation.

Table 6.2: Comparison of available video datasets.

Name	Year	# manipulation methods	# samples			Source	Average duration	Resolution	Visual quality	Reference
			Total	Fake	Real					
UADFV	2018	1	98	49	49	Youtube	11.4 sec	294x500	Low	[68]
HOHA-based dataset	2018	1	600	300	300	Internet	varied	360x240	Low	[69]
DeepFake TIMIT	2018	2	640	640	0	Actors	4 sec	64x64 (LQ) 128x128 (HQ)	Low	[70]
FaceForensics++	2019	2	5,000	4,000	1,000	Youtube	18 sec	480p, 7200p, 1080p	Low	[71]
Celeb-DF	2019	1	6,229	5,639	590	Youtube	13 sec	varied	High	[78]
DFDC-Preview	2020	2	5,250	4,119	1,131	Actors	30 sec	180p, 2160p	High	[72]
DFDC	2020	Unknow	128,154	104,500	23,654	Actors	30 sec	180p, 2160p	High	[79]
DeeperForensics	2020	5	60,000	50,000	10,000	Actors	varied	1920x1080	High	[74]
WildDeepFake	2020	Unknow	7,314	3,805	3,509	Internet	varied	varied	High	[28]
ForgeryNet	2021	15	221,247	121,617	99,630	Internet	varied	varied	High and low	[75]
KoDF	2021	6	237,942	175,776	62,166	Actors	90 sec	1920×1080	High	[76]
FFIW 10K	2021	6	20,000	10,000	10,000	Youtube	varied	varied	High	[77]

6.4.2 Audio

In this section, attention will be directed towards audio manipulation datasets, which constitute an underexplored domain within the field of multimedia data manipulation. Therefore, it is considered important to generate high-quality datasets that enable the creation of novel systems and models for the detection of manipulation signals in audio samples and the advancement of

this domain. A concise overview of the audio datasets used in the literature is presented in Table 6.3.

The audio datasets provide a diverse resource for research in audio manipulation and automatic speaker verification (ASV). ***ASV Spoof 2019*** [80] focuses on both logical and physical access attacks, featuring recordings from 107 speakers split into training, development, and evaluation sets. ***Fake-or-Real (FOR)*** [81] datasets offer over 198,000 English sequences, including real and synthetic samples created with state-of-the-art text-to-speech methods. ***ASV Spoof 2021*** [82] extends this research with recordings from multiple speakers, while ***WaveFake*** [83] generates 117,985 simulated audio clips using various techniques, though limiting diversity to one speaker per sample. ***In-The-Wild Audio Deepfake (IWA)*** [84] presents audios of public figures from diverse sources, and ***Chinese Fake Audio Detection Dataset (FAD)*** [85] stands out for including background noise in samples, making it closer to real-world scenarios. These datasets provide significant opportunities for developing and testing speaker verification systems against audio manipulations.

In summary, these datasets offer a comprehensive exploration of various audio manipulation techniques and challenges in the field of speaker verification. Researchers can leverage these resources to advance the development of robust systems capable of detecting and countering different ways of audio spoofing.

Table 6.3: Comparison of available audio manipulation datasets. * The WaveFake dataset contains only manipulated samples. ** The number of samples is not given, but the total time of each class (real or fake) is presented.

Dataset	Language	Condition	# manipulation techniques	# samples			# Speakers	Reference
				Total	Fake	Real		
ASVspoof 2019	English	Clean	13	55,200	15,600	39,000	107	[80]
ASVspoof 2021	English	Clean, noisy	Unknow	1,513,852	130,032	1,383,820	149	[82]
FoR	English	Clean	Unknow	+ 198,000	+ 87,000	+110,000	140 real 33 fake	[81]
WaveFake*	English, Japanese	Clean	Unknow	117,985	117,985	-	+100	[83]
FAD	Chinese	Clean, noisy	12	173,800	-	-	1,024 real, 279 fake	[85]
IWA**	English	Clean, noisy	Unknow	38 h	20,8 h	17,2 h	58	[84]

6.4.3 Multimodal

Finally, the published multimodal datasets, which include both audio and video, will be described. In contrast to the preceding sections, these datasets investigate acoustic and visual features together rather than analysing them separately. These datasets offer a richness of information and facilitate a broader generalisation to a larger variety of modified samples compared to their predecessors. Table 6.4 presents a summary of the multimodal datasets used in the literature.

The ***FakeAVCeleb dataset*** [86] is a multimodal dataset featuring manipulated audio and video content, resulting in various combinations of real and fake elements. Real audio and video are combined with fake counterparts, achieved through algorithms like Faceswap and FSGAN for video manipulation and SV2TTS for audio modification. The dataset focuses on precise lip synchronization, making it a valuable resource for deepfake research. In contrast, ***VideoSham*** [87] comprises 826 videos, including both real and manipulated variants. Unlike typical deepfake datasets, VideoSham goes beyond facial alterations, encompassing changes in background, text, audio, aesthetic edits, and temporal manipulations. Six distinct attack techniques are employed, categorized into spatial, temporal, and geometric manipulations, creating a diverse and challenging dataset for deepfake detection studies. Lastly, the ***Localized Audio Visual***

DeepFake (LAV-DF) dataset [88] is adapted to detect and locate temporary deep falsifications. It incorporates speech-to-text techniques for audio modification and subsequent voice and face reenactment. Featuring over 130,000 samples, the dataset includes real and fake segments, enabling a detailed comparative analysis. It is important to note that features such as sentiment scores, sentiment changes, and segment lengths facilitate in-depth analysis and detection of manipulation signals within the samples.

These datasets, FakeAVCeleb collectively offer a comprehensive view of deepfake manipulations, encompassing precise audio-video synchronization, diverse content alterations, and localized manipulations. Researchers can leverage these datasets to advance the development of robust deepfake detection methods, taking into account the intricate challenges posed by multimodal and temporally nuanced manipulations. The richness of datasets in varied manipulations makes them invaluable resources for the ongoing efforts to fight the proliferation of misleading and malicious multimedia content in digital spaces.

Table 6.4: Comparison of available multimodal manipulation datasets.

Dataset	Year	# samples			Subject	# manipulation techniques	Average duration	Source	Reference
		Total	Fake	Real					
FakeAVCeleb	2021	20,000	19,500	500	500	4	7,8 sec	VoxCeleb2 dataset	[86]
VideoSham	2023	826	413	413		6	8 sec	Vimeo	[87]
LAV-DF	2023	136,304	99,873	36,431	153	1	0,64 sec	VocCeleb2 dataset	[88]

6.5 Multimedia data forensics

In this section, emphasis will be placed on a key facet of disinformation, namely the veracity of multimedia content, including video and audio. In the analysis of the authenticity of multimedia content, particular attention will be directed towards the signs of manipulation. The proliferation of manipulated information has become a significant challenge on social networks. Consequently, an increasing number of researchers are developing novel systems, architectures, and frameworks to identify such content and ultimately combat disinformation. Although both modalities are expected to be analysed together, most researchers still analyse them independently.

6.5.1 Techniques for video forensics

The first modality that will be analysed is video. This type of data can be analysed from different approaches, Figure 6.3: **visual techniques**, a set of techniques that are based on the information of the frames independently without using the temporal information of the samples; **visio-temporal techniques**, unlike the previous technique, it takes advantage of the temporal information of the videos looking for, for example, inconsistencies between frames; and finally **metadata techniques**, unlike the previous techniques, it will not use the information of the video content but the metadata of the files, such as compression, which will show traces that the file has been altered.

6.5.1.1 Techniques for detecting manipulated metadata

Traces of manipulation frequently manifest themselves in samples through indicators such as texture inconsistencies or compressed data within the video. Within this section, the spotlight will be on techniques devised for identifying manipulation traces in metadata. High-definition videos often undergo compression to facilitate their storage and transmission. The latest video codec

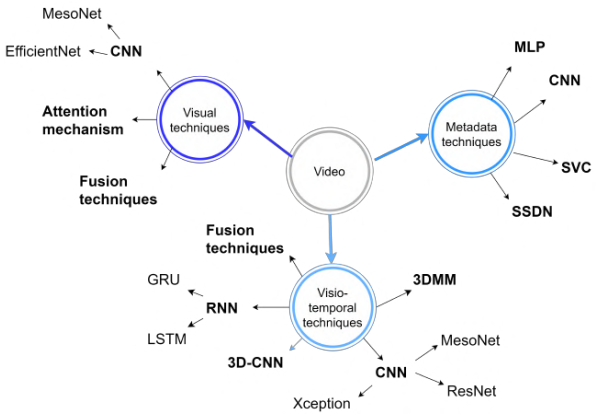


Figure 6.3: Representation of main video forensics techniques.

employed for high-definition video compression is the "High-Efficiency Video Coding (HEVC)," which effectively diminishes both spatial and temporal redundancies. In cases where high-quality videos undergo manipulation, a secondary HEVC compression step becomes inevitable, making these traces left by manipulation techniques the focal point of numerous projects. A brief summary of video manipulation techniques based on metadata information appears in Table 6.5.

This approach, which focuses on metadata information rather than traces of manipulation in the content, is less well represented in the field of study but we should not forget about it. Metadata analysis can be one of the ways to control the spread of manipulated multimedia content. This section focusses on a completely different approach; instead of focussing on the features extracted from the video content, it focusses on the properties of the video. Another striking detail is that there are no papers that combine both sources of information. We can find several examples that combine frame and video level information as well as metadata and video or frame information.

Table 6.5: Summary table of the different articles related to video forensics techniques based on *metadata information*.

	Year	Methodology	Dataset available	Code available	Task	Main contribution
Hong et al. [89]	2019	CNN, SVM, statistical and visual features, Fusion Learning	✓	X	Binary classification	Classification system based on coding patterns using MLP
Huamán et al. [90]	2020	Atom extraction algorithm, video structure	✓	X	Binary classification	Novel system, for different video formats, based on the structure of video containers and behaviour in social networks.
Zhang et al. [91]	2021	Self-supervised, Ensemble, EfficientNet-B2,	✓	X	Binary classification	Ensemble-based system, SSDN, focus on compression ratio and authenticity
Uddin et al. [92]	2022	CNN, SVM, statistical and visual features, Fusion Learning, Picture partitioning	X	X	Binary classification	Compression level classification system, between single and double compression, focussing on statistical and visual features from frame partitioning infromation

6.5.1.2 Techniques for detecting manipulated video based on frame information

In this section, an all-encompassing review of video forensic techniques will be presented, with a specific focus on *frame features*. These researchers employ distinct frames of the video as independent entities. Although this approach to video forensics has a weakness in that it does not use the temporal information inherent in video files, it does enable the analysis of samples using image detection systems that have shown success in recent years. Some authors even try to classify videos between authentic or manipulated retraining of the widely adopted state-of-the-art CNN.

One of the indicators that manipulation techniques may leave behind in a video is the *inconsistency of texture*. Some authors explore additional sources of information beyond the image content itself, the most extensively studied source being texture. As we can see in Table 6.7, other authors have focused on information fusion techniques and ensembles, combining information from different sources, to detect traces of manipulation in the samples. Other authors have focused on another approach to detect manipulation in videos, the *attention mechanisms*, based on the idea of using only the relevant features of the input.

In the domain of visual techniques for detecting manipulation in videos, where information is extracted from individual frames without considering the video's temporal aspects, it is evident that many authors not only extract data directly from frames but also employ textural features to identify inconsistencies. Regarding the most prevalent techniques and architectures, Convolutional Neural Networks (CNNs) and attention mechanisms consistently achieve impressive performance, while ensemble methods remain a favorable choice for enhancing system performance. A concise summary of video manipulation techniques reliant on visual information is provided in Tables 6.6 and 6.7.

Table 6.6: Comparative table of video forensics works based on *visual features* analysis and the datasets used. [1]: Data will be made available on request. [2]: the dataset is available in the article or in a reference given in the article.

	UADFV	DF-TIMIT	HOHA-based	FF++	DFDCp	DFDC	Celeb-DF	DeeperForensics	WildDeepFake	ForgeryNet	FakeAVCeleb	KoDF	Dessa [1]	VoxCeleb	AffectNet	Other dataset
Dang et al. [93]																X[2]
Qian et al. [94]				X												
Pokroy and Egorov [95]						X										
Chen et al. [96]				X												
Guo et al. [97]																X[2]
Kim and Cho [98]				X												X
Mitra et al. [99]				X		X										
Xu et al. [100]		X		X	X											
Zhao et al. [101]	X			X			X									
Yu et al. [102]		X		X			X									
Mazaheri and Roy-Chowdhury [103]				X											X	
Kingra et al. [104]				X		X	X	X								X[1]

6.5.2 Techniques for detecting manipulated video-temporal continuity features

Videos possess a distinctive attribute that separates them from images: *temporal continuity*. A video is composed of numerous frames, each frame exhibiting a sense of continuity and correlation with its neighbouring frames. Manipulation can alter these two fundamental properties of a video, resulting in temporal incoherence within the samples. Although current video manipulation techniques have advanced to such an extent that they may be imperceptible to the naked eye, they can still generate temporal inconsistencies, such as object displacement and rapid eye blinks, among others. This temporal information can be exceptionally valuable for identifying manipulated signals.

As demonstrated in the previous section, CNN stand out as one of the most prevalent architectures and consistently outstanding in various computer vision tasks, including manipulation detection, such as 3D-CNN, which is able to process spatial and temporal information by means of 3D convolutional filters. However, other authors continue to rely on 2D-CNNs, which are more widely used in image processing, to extract the relevant information from the frames and subsequently process it with other architectures such as Recurrent Neural Networks (RNN). While

Table 6.7: Summary table of the different articles related to video forensics techniques based on *visual information*.

Article	Year	# datasets	Methodology	Code available	Task	Main contribution
Dang et al. [93]	2020	1	CNN, Attention mechanism	✓	Binary classification	Verification system by comparison with real subjects Novel metric, Inverse Intersection Noncontainment
Qian et al. [94]	2020	1	CNN, Colaborative Learning, F^3-Net	X	Binary classification	System based on frequency domain based on two own modules
Pokroy and Egorov [95]	2021	1	CNN, EfficientNet	X	Binary classification	Establish a baseline with a single pretrained CNN
Chen et al. [96]	2021	1	CNN, Ensemble,	✓	Binary classification, Localisation	Manipulation localisation system based ensemble of frame and noise from frame
Guo et al. [97]	2021	1	CNN, AMTENnet	✓	Binary classification	Novel preprocessing model, Adaptative manipulation trace extraction network
Kim and Cho [98]	2021	1	CNN, Ensemble, XAI	X	Binary classification	Ensemble combining frames with trace features
Mitra et al. [99]	2021	2	CNN, Xception	X	Binary classification	Lower computational requirements
Xu et al. [100]	2021	3	CNN, Ensemble, MesoNet, Xception, SCNN	X	Binary classification	Novel system, Set Convolutional Neural Network, based on ensembles
Zhao et al. [101]	2021	3	Multiattention Network, Texture enhancement	✓	Binary classification	Novel Multiattention Network based in local information and texture
Yu et al. [102]	2022	3	CNN, Facial Patch Mapping, Patch-DFD, XAI	X	Binary classification	Ensemble combining patches and frames
Mazaheri and Roy-Chowdhury [103]	2022	2	CNN, Xception,XAI	X	Binary classification, Localisation	Manipulation localisation system with XAI
Kingra et al. [104]	2022	5	CNN, Local Binary Pattern, LBPNet	X	Binary classification	Novel model, LBPNet, based on facial texture irregularities

RNN are the architecture most commonly used to integrate information from individual frames in video analysis, some authors explore alternative techniques such as *channel-wise spatio-temporal aggregation*. Certain models deviate from focussing solely on frame information and instead leverage other features extracted from the frames, including *biometrical features*. These features rely on distinctive measurements and unique characteristics for each individual.

Finally, within this section, it is worth highlighting a group of techniques that independently process information from both the frame and video levels. This approach enables the exploitation of all available information resources.

Within this set of techniques that combine information at both the frame and video levels, a broader spectrum of approaches can be discerned. In contrast, the remaining articles in this section predominantly rely on a more restricted set of architectures, primarily centred on CNN and RNN. Furthermore, it is evident that the information used in the techniques of this section is notably more comprehensive than that in the previous section, where temporal information remains unexploited. A concise summary of video manipulation techniques grounded in visio-temporal information can be found in Tables 6.8 and 6.9.

6.5.3 Techniques for audio forensics

The second modality, which will be analysed in this section, is audio and, as will be demonstrated below, it is the least investigated multimedia information modality in the state of the art. As in the previous subsection, the problem of detecting manipulated audio signals has been approached by the authors from various perspectives, as shown in Figure 6.4: *approach based on the selection and extraction of features*, the authors aim to utilise the features that are most pertinent for this task; *audio processing using CNN*, encompasses all the works that have either employed CNNs or leaned on them during the development of their models; and *attention layer approach*, includes the works of all authors who have used transformers or attention layers.

Table 6.8: Comparative table of video forensics works based on *visio-temporal information* and the datasets used. [1]: code available.

	UADFV	DF-TIMIT	HOHA-based	FF^{++}	DFDCp	DFDC	Celeb-DF	DeeperForensics	WildDeepFake	ForgeryNet	FakeAVCeleb	KoDF	Dessa [2]	VoxCeleb	Other dataset
Tran et al. [105]															X
Montserrat et al. [106]						X									
Fernando et al. [107]				X											
Chintha et al. [108]				X		X									
Zi et al. [28]		X		X	X	X	X		X						
Das et al. [109]				X											
Cozzolino et al. [110]														X	
Nguyen et al. [111]		X		X											
Hu et al. [112]				X			X								
Lu et al. [113]				X	X		X								
Agarwal et al. [114]				X		X	X								
Kolagati et al. [115]						X							X		
Chamot et al. [116]				X			X								X
Pu et al. [117]				X		X	X								
Wang et al. [118]				X		X	X	X							

Table 6.9: Summary of different articles related to video forensics technique based on *visio-temporal information*.

Article	Year	# datasets	Methodology	Code available	Task	Main contribution
Tran et al. [105]	2018	4	3DCNN, skip connection	X	Binary Classification	Comparison between different 3DCNNs, some with skip connections
Montserrat et al. [106]	2020	1	CNN, RNN, EfficientNet-B5, Facial feature extraction	X	Binary Classification	New system composed of ensemble CNN-based and RNN focuss on facial features
Fernando et al. [107]	2020	1	Attention mechanism, RNN, CNN, ResNet, GRU	X	Binary Classification	Novel Hierarchical Attention Memory Model (HAMN), using knowledge stored in neural memories
Chintha et al. [108]	2020	2	CNN, RNN, Xception, LSTM	X	Binary Classification	Novel system, XceptionTemporal, and loss function
Zi et al. [28]	2020	6	3DCNN, Attention mechanism, low-level facial features	X	Binary Classification	System based on 3D CNN that used attention masks
Das et al. [109]	2021	1	3DCNN, ResNet, Attention layers, Face	✓	Binary Classification	Novel system based on 3DCNN with attention layer in convolutional blocks, pretrained with real samples
Cozzolino et al. [110]	2021	1	Biometric features, Generalisability, 3D morphable model, Temporal ID network	✓	Binary Classification	Novel system, ID-Reveal, trained only with real samples and focused on the inconsistencies between the visual identity and biometric features
Nguyen et al. [111]	2021	2	3DCNN, 3D convolution kernels, temporal face extractor	X	Binary Classification	System based on 3D CNN with 2D convolutional kernels Novel method for constructing 3D images from the faces of consecutive frames
Hu et al. [112]	2021	2	CNN, MesoNet, XAI, Face Swap, Frame and temporal features, ResNet 18, Ensemble	X	Binary Classification	Ensemble-based system, combining frames and temporal information preocessed by CNN
Lu et al. [113]	2021	3	CNN, EfficientNet-B0, Face features, Channel-wise Spatiotemporal Aggregation module	✓	Binary Classification	A novel fusion module, CWSA, which combines the information from the frames and classifies using a CNN
Agarwal et al. [114]	2021	3	CNN, Xception, XAI, cross-stitched network	X	Binary Classification	Novel system focus on frame and temporal information using cross stitch connection
Kolagati et al. [115]	2022	2	CNN, MLP, Fusion Learning, Hybrid Neural Network, Facial landmarks	X	Binary Classification	Novel hybrid neural network, that combinea CNN, frames, and MLP, temporal analysis using facila landmarks
Chamot et al. [116]	2022	3	CNN, RNN, LSTM, MesoNet, EfficientNet, XAI	✓	Binary Classification	Combination of CNN and RNN, to process the frame information, taking advantage of temporal information. Apply XAI techniques for frames.
Pu et al. [117]	2022	3	CNN, RNN, Colaborative learning, ResNet, Ensemble, Joint loss function	✓	Binary Classification	Ensemble composed of frame and video classification system with joint loss function to maximise the performance
Wang et al. [118]	2022	4	CNN, Xception, MLP, Local correlation, Facial features, Local incoherences	X	Binary Classification	Novel system, MCLCR, based on frame and patches from frames information with contrastive loss with cross-entropy loss

6.5.3.1 Audio forensics based on feature selection and extraction

In this section, I will present a thorough examination of audio forensics systems that prioritise feature selection and extraction. The authors explore various techniques to ensure that the input

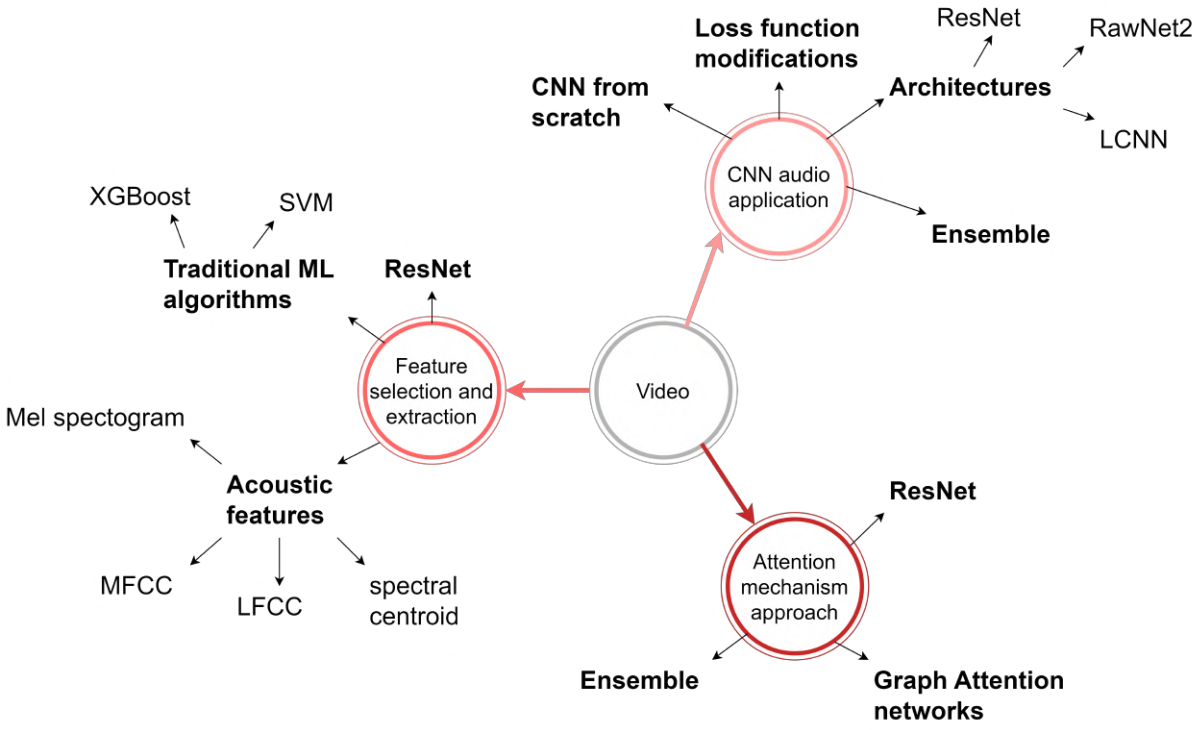


Figure 6.4: Representation of main forensics techniques in audio.

of the model or system comprises the most pertinent features to detect signs of manipulation. Such as Mel Frequency Cepstral Coefficients (MFCC), a set of parameters that describe the general shape of the spectral envelope, widely used in speech recognition; or Linear Frequency Cepstral Coefficients (LFCC), a variant of MFCC, that represents the linealy spaced frequency bands in the audio spectrum. Other authors opt for biometric features, which are specific and identifiable to each person and, unlike the previous ones, are independent of the manipulation technique used. As for the techniques used in this domain, some authors prefer traditional Machine Learning techniques, while others focus on well-known CNNs such as ResNet and techniques such as Active Learning, which allows them to select the most relevant features from the training set.

Although feature extraction has been extensively explored in research, it continues to receive significant attention due to its consistently promising outcomes. Nevertheless, there remains a substantial reliance on manual feature extraction in numerous instances, a practise that carries a noteworthy limitation in terms of its capacity to adapt to novel situations. In Tables 6.10 and 6.11, we provide a concise overview of the audio manipulation techniques rooted in feature extraction and selection.

6.5.4 Audio forensics based on CNN architectures

In this section, the attention is directed towards articles that have employed Convolutional Neural Networks (CNNs) or architectures inspired by them in the development of their audio forensics models. CNNs have consistently shown exceptional performance in various tasks, including audio classification for manipulation detection. Despite their long presence in the field, these architectures remain a reliable choice for a wide range of applications.

Table 6.10: Comparative table of audio forensics works, focus on *feature selection and extraction*, and the datasets used.

	ASVspoof	FoR	WaveFake	FAD	ITW	VoxCeleb	FakeAVCeleb	Other dataset
Iqbal et al. [119]		X						
Dongre et al. [120]	X	X						
Rahman et al. [121]	X	X						X
Pianese et al. [122]	X				X		X	
Wang and Yamagishi [123]			X					

Table 6.11: Summary of different articles related to audio forensics technique based on *feature selection and extraction*.

Article	Year	# datasets	Methodology	Code available	Task	Main contribution
Iqbal et al. [119]	2022	1	ML algorithms,PCA, mel spectrpogram	X	Binary classification	Novel ML model focusing on feature selection through principal component analysis
Dongre et al. [120]	2022	2	Adaptative channel-wise feature recalibration technique, Attention feature fusion block	✓	Binary classification	The introduction of the attentional feature fusion block of image domian, to combine LFCC and MFCC
Rahman et al. [121]	2022	3	CNN, x-ResNet, Probabilistic linear, discriminant analysis, Squeeze excitation block	X	Binary classification	Combination on x-ResNet and probabilistic linear discriminant analysis
Pianese et al. [122]	2022	3	Person-of-Interest, Centroid-based and multi-similarity testing	X	Binary classification	New system based on speaaker approach, exploring capability of person-of-interest (POI)
Wang and Yamagishi [123]	2023	1	Active Learning, wav2vec model	✓	Binary classification	Novel spoofing countermeasure using active learning to remove useless samples from a pool

To enable the analysis of audio signals using such architectures, it is imperative to convert these signals into a format that the model can interpret. In this case, we employ a multi-step process. First, we represent the signal, capturing the variation in amplitude over time. Subsequently, I applied the Fourier transform, a technique that dissects the signal into its constituent frequencies, along with their respective amplitudes. This transformation effectively shifts the signal from the time domain to the frequency domain, generating a spectral representation. The spectrogram is the result of calculating the spectrum of a signal within evenly spaced time windows. This transformation is crucial for efficient compatibility with CNN. However, as I will explore later, many researchers further converted the spectrogram into a Mel spectrogram, effectively mapping it to the Mel scale. The result of calculating the spectrum of a signal within fixed time windows results in what is known as a spectrogram. This spectrogram provides a visual depiction of how the signal's frequencies evolve over time and serves as a valuable input for CNNs. However, as we will explore further in this discussion, numerous researchers took an additional step by transforming the spectrogram into a Mel spectrogram. In this process, the spectrogram is converted into the Mel scale, which offers specific advantages for some applications.

Although certain authors persist in the creation of CNN from scratch, these networks may not has the same feature extraction capabilities found in widely adopted pre-trained CNNs used in the state of the art. However, they still yield highly satisfactory results for some tasks. Other authors choose to use pre-trained state-of-the-art architectures or based on them to create their own architectures. Another approach widely used in this field is ensembles, due to their performance in many tasks within the field of DL. This technique allows unifying the knowledge of different models, improving performance, and reducing the error of individual models. The improvement of the performance of these architectures or systems in the audio signal manipulation detection task involves modifying existing loss functions or developing new ones, such as the large-margin cosine loss function (LMCL) [124] or mean-square-error loss function with P2SGrad [125].

As for the tasks the authors focus on, most of them try to classify entire samples between

manipulated and authentic, however, some authors try to detect which segments are manipulated within full audio samples. This approach is more complex and much more informative, as it allows you to know what exact information has been manipulated and also adds explainability to the system. However, its representation within this domain is lower than expected. The information on all the articles analysed in this section and their comparison with the public datasets of the field can be seen in Tables 6.12 and 6.13.

Table 6.12: Comparison of audio forensics works *based on CNN* and the datasets used.

	ASVspoof	FoR	WaveFake	FAD	IWA	VoxCeleb	FakeAVCeleb	Other dataset
Gomez-Alanis et al. [126]	X							
Wang et al. [127]		X						
Chen et al. [124]	X							
Wang and Yamagishi [125]	X							
Tak et al. [128]	X							
Hua et al. [129]	X							
Khochare et al. [130]		X						
Zhang et al. [131]	X							X
Zhang et al. [132]								X
Kawa et al. [133]	X		X					X

Table 6.13: Summary of the different articles related to audio forensics techniques *based on CNN*.

Article	Year	# datasets	Methodology	Code available	Task	Contribution
Gomez-Alanis et al. [126]	1	2020	Fusion learning, presentation attack detection, automatic speaker verification, DNN	X	Binary classification	Novel model that combine presentation attack detection and automatic speaker verification
Wang et al. [127]	1	2020	DeepSonar, layer-wise neurone activation patterns	X	Binary classification	New approach, DeepSonar, based on the monitoring of neuronal behaviour in a DNN-based speaker recognition system with a binary classifier
Chen et al. [124]	1	2020	CNN, ResNet, Data augmentation, LMCL	X	Binary classification	New loss function, XXX, which maximises the interclass variance and minimises the intraclass variance. Include FreqAugment layer to improve generalisability
Wang and Yamagishi [125]	1	2021	mean-square-error loss function with P2SGrad, gradient similarity, LCNN	✓	Binary classification	New loss functio, mean-square-error with P2SGrad, no hyperparameter setting needed, based on mean square error metric and probability gradient to similarity
Tak et al. [128]	1	2021	RawNet2, Combined adaptative and multiplicative feature scaling approach	✓	Binary classification	Modification of the parameters of the convolutional blocks and the inclusion of combined additive and multiplicative feature scaling approach
Hua et al. [129]	1	2021	Time-domain Synthetic Speech Detection Net, Res-TSSDNet, Inc-TSSDNet	✓	Binary classification	Novel architecture, Time-domain Synthetic Speech Detection Net, with skip conenection or parallel convolutions with mixup relugarisation
Khochare et al. [130]	1	2021	Traditional machine learning algorithms, TCN, STN, ensemble	X	Binary classification	New ensemble model, combining features spatial and temporal information and comparison with traditional machine learning approach
Zhang et al. [131]	2	2022	Self-supervised, labelling granularity	X	Temporal location, segment detection	Novel dataset for segments detection Novel model for labelling at segment label
Zhang et al. [132]	1	2022	CNN, ResNet, Res-TSSDNet	✓	Binary classification	Two baselines: ResNet for processing CQCC and Res-TSSDNet for processing the raw speech waveform New mandarin dataset
Kawa et al. [133]	3	2022	SpecRNet, RawNet-2, FMS attention layer, GRU	✓	Binary classification	New architecture, SpecRNet, composed of residual blocks, FMS attention layer and GRU layers

6.5.5 Audio forensics based on attention layers

In 2017, the introduction of Transformer models marked a significant breakthrough in the realm of natural language processing (NLP). This innovative architecture, built on the foundation of attention mechanisms, has shown remarkable effectiveness in capturing complex relationships within sequences and temporal data. It has demonstrated a robust capacity to learn universal patterns and representations. These advantages make Transformers an interesting option for audio forensics. Nevertheless, despite their successes, Transformers have not reached the same level of popularity as Convolutional Neural Networks (CNNs), which still maintain two primary advantages: superior computational efficiency and a broader applicability to spatial data. Con-

sequently, in this section, we will concentrate on research papers that have used Transformer models for audio manipulation detection. Within some of these studies, we can see how the authors have exploited the power of the Transformer models.

As in the case of CNNs, we can find authors who use already implemented state-of-the-art architectures or models, while other authors prefer to include layers of attention in other architectures. We also observe the presence of ensemble-based systems to combine different models. However, we can observe a set of techniques that have not been seen before, the application of self-supervised learning techniques, which although they are not exactly forensic systems, these models have a great capacity to learn patterns and trends from samples and apply them to other specific tasks such as tamper detection. Although self-supervised learning is not widely represented in this domain, it has great potential.

A concise summary of audio manipulation techniques centered around attention mechanisms is provided in Tables 6.14 and 6.15. As evident, the current volume of articles employing Transformers for our specific task is relatively lower when compared to those utilizing CNNs. Nevertheless, it's important to note that this current scenario doesn't necessarily indicate that Transformer models won't gain more prominence in this field in the coming years.

Table 6.14: Comparative table of audio forensics works based on *attention mechanism* and the datasets used.

	ASVspoof	FoR	WaveFake	FAD	IWA	VoxCeleb	FakeAVCeleb	Other dataset
Zhang et al. [134]	X	X						
Jung et al. [135]	X							
Chen et al. [27]						X		
Ge et al. [136]	X			X		X		

Table 6.15: Summary of the different articles related to audio forensics techniques *based on attention mechanism.*

Article	Year	# datasets	Methodology	Code available	Task	Contribution
Zhang et al. [134]	2021	2	TE-ResNet, encoder, attention mechanism, ensemble	X	Binary classification	Novel ensemble-based model, where each model is composed of transformer encoder and residual neural network
Jung et al. [135]	2022	1	Heterogeneous stacking graph attention layer, maximum network operation	✓	Binary classification	Novel system, AASIST, with novel heterogeneous stacking graph attention layer, which processes temporal and spatial information
Chen et al. [27]	2022	1	Self-supervised, attention mechanism, genric model	✓	Learn patterns	New pretrained self-supervised model, learns universal speech representations from unlabelled data
Ge et al. [136]	2023	3	ResNet-34, excitation blocks, ResNetSE-34	✓	Binary classification	Ensemble-based system, composed of automatic speaker verification and spoofing countmeasure, using a combined optimisation

6.6 Techiques for multimodal forensics

In the final section of our analysis, manipulation detection is examined through the combined use of video and audio modalities. A key distinction from the previous sections lies in the comprehensive integration of all available information. In most cases, the focus is on the identification of differences between data extracted from video frames and audio content. Although all these models share the common objective of detecting inconsistencies, two distinct trends are observable, as depicted in Figure 6.5: *audiovisual inconsistencies*, This approach directly employs extracted features from both audio and visual sources to identify discrepancies; emotional

inconsistencies, This approach examines the limitations of generative systems in their ability to replicate human emotions. Despite the considerable success of current generative systems in the field of manipulation, they have difficulties in authentically reproducing human emotional expressions.

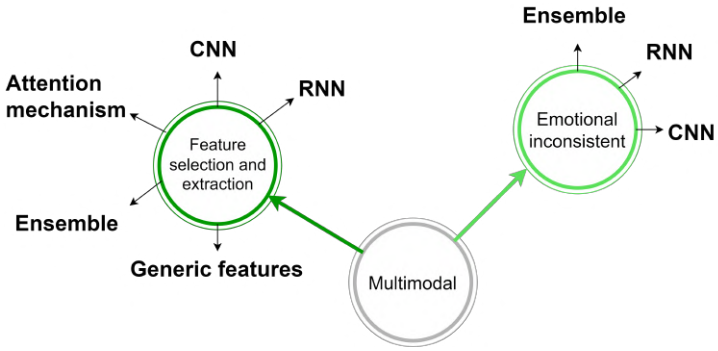


Figure 6.5: Representation of main multimodal forensics techniques.

6.6.1 Inconsistencies between modalities

An important part of the work in this domain uses information fusion techniques in which CNN and RNN are combined for video and audio processing, respectively. The information extracted by both architectures will be processed together to produce the final result. Another approach that is obtaining very good results in the search for inconsistencies is generic models, as they are able to learn patterns from real samples and easily detect inconsistencies, and they do not require manipulated samples for training.

As with audio and video analysed independently, attentional mechanisms and transformers are very present in this domain, due to their results in both modalities independently. However, in addition to CNNs and transformers, we can also find identity verification. Although these models are not designed for this specific task, they are nonetheless capable of detecting whether a voice is manipulated. Although this technique gives very good results, it has a main limitation: It has to be trained with the voices that are going to be analysed, or at least with very similar voices. Finally, if we analyse the tasks that are attempted to be solved, most of the models attempt to solve binary classification problems, and it can be observed that some works carry out localisation of the voices to be analysed, or at least with very similar voices.

Table 6.16: Comparative table of multimodal forensics works based on *audiovisual inncoherencies* and the datasets used. [1]: code available.

	UADFV	DF-TIMIT	HOHA-based	FF++	DFDCp	DFDC	Celeb-DF	DeeperForensics	WildDeepFake	ForgeryNet	KoDF	FFIW 10K	FakeAVCeleb	VoxCeleb	VideoSham	LAV-DF	Other dataset
Lewis et al. [137]						X											
Shang et al. [138]																	X
Cheng et al. [139]						X							X				
Wang et al. [140]																	X
Cozzolino et al. [141]		X			X								X				X
Cai et al. [88]																X	
Ilyas et al. [142]													X				
Yang et al. [143]						X							X				X

Table 6.17: Summary table of the different articles related to multimodal forensics techniques *based on inconsistencies between modalities.*

Article	Year	# datasets	Methodology	Code available	Task	Contribution
Lewis et al. [137]	2020	1	NOLANet, Xception, LipNet, DeepSpeech, fusion learning, LSTM	✓	Binary classification	Novel hybrid deep learning approach, NOLANet, combines spatial, spectral and temporal information.
Shang et al. [138]	2021	1	CNN, RNN, , co-attentive information fusion module	X	Binary classification	Novel framework, TikTec, that used text, audio and video, combining the information with co-attentive information fusion module
Cheng et al. [139]	2022	2	Voice-face matching, XAI, Real-fake loss function	X	Binary classification	New approach to voice-face matching, training with real videos and retraining for the task of forensics.
Wang et al. [140]	2022	1	FTFDNet, audio-visual attention mechanism, fussion learning	X	Binary classification	Novel framework, FTFDNet, based on attention mechanism (AVAM) ,applicable to any CNN.
Cozzolino et al. [141]	2022	4	Identity verification, contrastive loss, similarity matrix	✓	Binary classification	New approach, POI-Forensics which learn a person-of-interest based on multimodality consistencies
Cai et al. [88]	2022	1	3D-CNN, 2D-CNN, Bounday matching layer, fusion learning	✓	Spatial localisation	New multimodal method, Boundary aware temporal forgery detection
Ilyas et al. [142]	2023	1	AVFakeNet, Dense swin transformer Net, fusion learning	X	Binary classification	Novel unified framework, AVFakeNet, composed of two DST-Net to compute dense hierarchical features maps
Yang et al. [143]	2023	3	Multi-Head Self-Attention (MSA) sub-layer, cross-attention block, TSE, MMD, bidirectional cross attention	✓	Binary classification	New audio-visual joint learning, AVoiD-D, composed of temporal spatial encoder, multimodal joint decoder and cross modal classifier

6.6.2 Emotional inconsistencies

In this section, attention will be directed towards articles employing the utilisation of "emotions" as a means to identify "inconsistencies" between audio and video data. Despite the considerable success of contemporary generative systems, their inability to authentically replicate human emotions persists. This limitation is primarily attributed to the subjective and intricate nature of emotions, among other contributing factors. Consequently, the detection of such disparities in manipulated videos can serve as a valuable indicator.

In this domain we find a limited number of approaches, we can find authors who look for inconsistencies between video and audio, but we can also find articles where they compare with real video samples. The main problem of the latter approach is that real samples have to be available to train the model and compare with the sample of interest. Among the works related to this field we can find authors who have decided not to focus on binary classification tasks, but instead focus on multicalse classification problems, between manipulated, synthetic, and real. Due to the nature of the task to be solved we will find a large presence of ensembles and information fusion techniques, since in many cases video and audio are processed independently and then combined.

If we compare the number of works related to the field of multimodal forensics, Table 6.18, we can see that it is lower than the other two modalities and the works are from recent years, so it is expected that this approach will increase in the coming years.

Table 6.18: Summary table of the different articles related to multimodal forensics techniques *based on emotional inconsistencies.*

Article	Year	# datasets	Methodology	Code available	Task	Contribution
Lomnitz et al. [144]	2020	DFDC	Xception, Bi-LSTM, sincNet, MLP,	X	Multiclass classification	Novel systemm, based on fusion learning, using MLP and CNN for frame analysis and SincNet forr audio analysis
Mittal et al. [145]	2020	VideoSham	siamese network architecture, triplet loss, CNN, Memory fusion network	X	Binary classification	Novel system using an architecture based on a Siamese network, which uses the triplet loss function.
Hosler et al. [146]	2021	DFDC	LSTM, Low-Level Descriptor,	X	Binary classification	New classification system based on arousal and valence, using LSTM

6.7 Available tools for end-users

As demonstrated in this survey, several systems have been developed to identify manipulation in video, audio, and multimodal content. However, a substantial limitation is that many of these solutions remain inaccessible to the average social media user or the Internet user. To effectively fight misinformation and empower users to validate information sourced from the Internet, it is imperative that we develop user-friendly tools that cater to the needs of a diverse audience. These tools should prioritise simplicity, boost credibility, ensure content integrity, and enhance reliable dissemination of information.

InVID Verification [147] is a web-based toolkit designed to authenticate user-generated videos and provide comprehensive insights into their contextual background. It encompasses various components dedicated to the analysis and management of user-generated content, presenting the analysis results through an intuitive user interface. This versatile tool enables end users to examine multiple aspects of a video, including its previous usage, origin, rights, contextual information, and even allows for forensic analysis. The application offers full support for videos obtained from popular platforms such as YouTube, Facebook, and Twitter. However, for videos from other platforms, it offers only partial analysis capabilities. The InVID Verification Application is designed for a wide range of users, including researchers, journalists, fact-checkers, and law enforcement agencies, all of whom require a means to verify the authenticity and contextual details of user-generated videos.

Real-Time Deepfake Detector [148] is a detection platform, this system leverages advanced AI models for face and landmark detection algorithms. These algorithms meticulously analyse facial features, such as the eyes, nose, and mouth, while also evaluating blood flow signals within real-time videos to discern their authenticity. Designed to operate without problems on a server, this platform offers a user-friendly web-based interface. Remarkably, it has the capability to simultaneously process up to 72 distinct detection streams, all powered by 3rd Gen Intel Xeon Scalable processors. This innovative solution has immense potential for several applications. Social media platforms can use it to prevent the upload of harmful deepfake videos, protecting their user communities. International news organisations can use it to steer clear of inadvertently amplifying manipulated videos. International news organisations can rely on it to avoid inadvertently amplifying manipulated videos, thereby preserving their credibility.

Deepfake-o-meter [149] an open web platform that integrates more than ten state-of-the-art deepfake image and video detection methods, with plans to continually expand its capabilities. This platform serves as a valuable resource for benchmarking the effectiveness of multiple deepfake detection algorithms within a single interface. The platform is structured around three components: a user-friendly front-end, a robust back-end, and seamless data synchronisation. Users can easily upload a video, select their preferred detection methods, and provide their email address. Subsequently, after the software backend completes the analysis, the user will receive an email containing a comprehensive detection report. This platform not only empowers researchers to assess the performance of their own detection algorithms against the latest methods but also offers a practical utility to users seeking to verify the authenticity of a given video or image.

Reality Defender [150, 151] is an innovative deepfake detection and generative AI platform that's dedicated to combatting fake news and propaganda. This versatile system offers a comprehensive approach by swiftly analysing various forms of media, including audio, images, videos,

and documents, with the ability to detect deepfakes in a matter of milliseconds. End-users are provided with essential tools, such as real-time risk assessment, email alerts, and detailed forensic review reports for all types of media. Additionally, the platform ensures users stay updated with weekly reports on the latest deepfake and generative AI threats, keeping them informed. Reality Defender harnesses a blend of advanced technologies to uncover deepfakes and other forms of disinformation. When a media file is uploaded to the platform, a suite of algorithms examines it to identify any signs of manipulation. Furthermore, the platform cross-references the media file against an extensive dataset containing known deepfakes and manipulated media to enhance its accuracy. Users can access the results of the analysis through an easy-to-use dashboard, facilitating efficient verification of media content.

DeepDetector [152] is an advanced AI-powered deepfake detection software crafted and trained to identify AI-generated or AI-manipulated faces. This cutting-edge tool operates by identifying visible faces within media content and subjecting them to in-depth analysis to uncover traces of deepfakes. It not only provides a probability assessment that indicates the likelihood that the input is a deepfake, but also supplies an activation map to elucidate the classification process. This versatile software boasts the capability to detect a wide array of deepfake manipulations and AI-generated content, making it a comprehensive solution for identifying synthetic media. In addition, it operates as a cloud-based service, ensuring scalability and accessibility for users. Importantly, it adheres to European data protection laws, guaranteeing the privacy and security of user data, while combating the proliferation of deepfake content.

DeepfakeProof [153] A plugin examines all images within the webpages you browse, identifying any deepfakes or altered media. It offers real-time notifications to users and delivers precise, reliable deepfake detection, against the dissemination of deceptive and dangerous content. Installation is straightforward, ensuring a smooth browsing experience for everyone.

Currently, the selection of tools accessible to end users is rather restricted. Furthermore, these tools lack integration with any social networks, placing the onus on users to discover and learn about them, which could potentially restrict their adoption among less tech-savvy individuals. Nonetheless, it's highly probable that in the upcoming years, we'll witness a proliferation of tools and applications, marked by enhanced functionality and performance.

Table 6.19: Comparative Analysis of Deepfake Detection Applications and Tools.

Applications/Tools	Supported Media Types	Detection Methods	Integration Options	User interface Ease of Use	Cost & Pricing
InVID Verification	Videos, Images	AI-based	Plugin, Application	✓	Free
Real-Time Deepfake Detector	Videos	AI-based on PPG signals	-	-	-
Deep-fake-o-meter	Videos	10+ SOTA DeepFake Detectors	Application	✓	Open-source
Reality Defender	Audios, Images, Videos,Documents	AI-based	Web-based App, API	✓	-
DeepDetector	Videos, Images	AI-based	API	-	-
DeepfakeProof	Images	AI-based	Plugin	✓	Free

6.8 Answer to research questions

In Section 6.1, four research questions related to the current status and trends of forensics in multimedia data are formulated. The methodology and literature analysis conducted in this review have been driven by these questions. Based on the knowledge gained from the review process

conducted in the previous sections, the different research questions can now be answered. A summary of the main conclusions drawn from this in-depth analysis is shown in Figure 6.6. These conclusions highlight the results of the research questions, and a comprehensive explanation of all these responses is provided below.

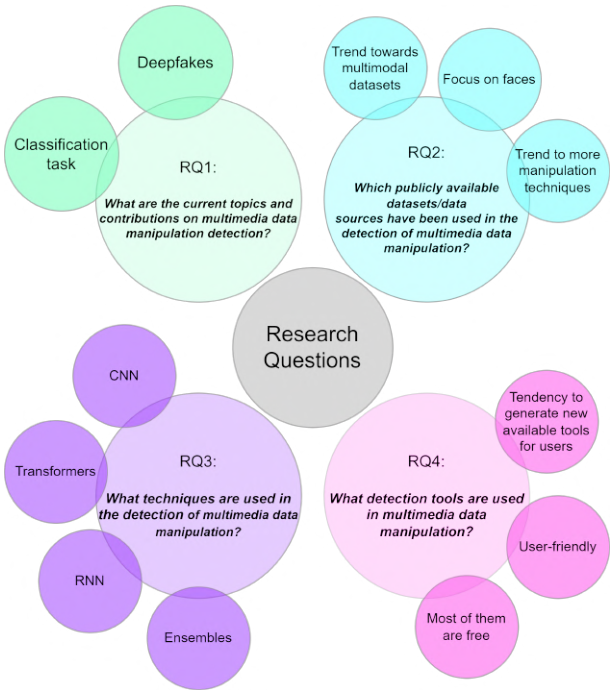


Figure 6.6: Key items of the answers to the proposed research questions.

RQ1. What are the current topics and contributions on multimedia data forensics?

Throughout the review, it is evident that in recent years both datasets and forensics techniques have placed their primary emphasis on "Deepfakes" and "face forensics," which constitutes a limitation and an issue. Alternative ways are now being explored by some current datasets, such as VideoSham [87]. This dataset diverges from conventional focus on different objectives, including background modification, object/people addition or removal, and audio signal replacement/addition.

Within the domain of multimedia data manipulation, several types of tasks can be addressed. While the state of the art has concentrated on binary classification, distinguishing between authentic and manipulated data, other unexplored options exist. One such uncharted territory is temporal localisation, which delineates the specific segments within videos that have been altered. This not only discerns the authenticity of the samples, but also pinpoints the precise portions subjected to manipulation.

RQ2. Which publicly available datasets/data sources have been used in the detection of multimedia data manipulation?

In this domain of multimedia data manipulation, public datasets are primarily employed by most authors, although some still opt to create their own datasets. These datasets can be used independently or in conjunction with other state-of-the-art datasets. The availability of information varies between different modalities. Figure 6.2 illustrates that the video modality

has a larger number of datasets, whereas multimodal datasets are comparatively scarce within the field.

There is a noticeable trend in these datasets towards increasing complexity, encompassing a broader array of manipulation techniques, a larger sample size, and a more realistic approach. These advances bring the datasets closer to real-world scenarios, enabling the detection of a wider range of manipulation techniques. A clear evolution can be observed, starting with the initial focus on detecting manipulation in video using visual features alone. Subsequently, the researchers expanded their scope to include the domain of audio. In recent years, there has been a growing emphasis on multimodal datasets, which not only offer more extensive information, but also present more diverse scenarios. This emphasis supports a more comprehensive analysis of multimedia data manipulation.

RQ3: What techniques are used in the detection of multimedia data manipulation?

The analysis of forensic techniques has revealed that the architecture most commonly used for multimedia data forensics is CNN. This preference arises because many authors process different modalities as images, even including audio signals. The other two architectures that also enjoy significant representation in this domain are RNN and transformers.

Another noteworthy consideration is that most research efforts focus on classifying samples as real or manipulated, without exploring more informative alternatives. Furthermore, there exists a pervasive lack of explainability in this field, with very few authors employing explainable AI techniques. When analysing the number of articles devoted to each modality, a consistent pattern is observed, reflecting the distribution of the dataset. Video forensics systems receive the greatest attention, while multimodal forensics garners the fewest research papers. However, it appears that in the coming years this relationship will undergo a reversal, with increasing importance placed on multimodality.

RQ4: What detection tools are used in multimedia data manipulation?

Disinformation has become a pressing issue in modern times, with an increasing prevalence of manipulated multimedia content circulating on social networks and websites. This proliferation can have far-reaching consequences for numerous users. Consequently, there is a growing demand for tools that enable the verification of multimedia content encountered on the Internet. In recent years, a multitude of such tools have emerged, many of them freely accessible, thus extending verification capabilities to a wider user base.

Similarly to the developments observed in other modalities, such as text, the prevailing trend suggests that the coming years will witness the creation of more precise and user-friendly verification tools. It is conceivable that some of these tools may even become integrated features in several social networks. Given that scientific works tend to be highly technical and less accessible to the general public, tools of this nature become essential components in the fight against information manipulation and consequently disinformation.

6.9 Future trends and challenges

The preceding research questions and their corresponding answers have offered a comprehensive and in-depth overview of the existing developments in techniques and systems for detecting manipulations in multimedia data. However, the essential literature review conducted to address

these questions has also unveiled several ways of future exploration within this research domain. This section delineates the coming trends that the literature is ready to receive, building upon its current state. Moreover, it identifies the challenges that the research community is likely to face and proposes potential strategies for their effective resolution. Figure 6.7 provides a visual summary of these emerging trends and challenges.

Throughout this review of the state of the art, several trends have been observed, and it has also been possible to see which ways will establish the ***future trends*** in this field. We will highlight the most promising research lines for the near future.

1. ***Diffusion models*** have demonstrated exceptional performance in the field of image and video generation in recent years (as evidenced by studies like Blattmann et al. [154] or Ho et al. [155])), and are expected to gain prominence in the field of multimedia data manipulation in the near future. These techniques have already shown promise in image manipulation, as exemplified by [156]. In the context of generating manipulated multimedia data, diffusion models offer the capability to manipulate existing videos and create modified versions of them. They can be applied to a range of tasks, including style transformation, object removal or addition, facial attribute alteration, and object appearance manipulation, among others. The diffusion process in these models involves iterative refinement of the input data to match a target distribution or achieve the desired effect. During each iteration step, small perturbations are introduced into the initial data to generate a new version, which is refined further as the process unfolds. Diffusion models can be trained using different techniques, such as supervised learning or reinforcement learning, depending on the specific task. Moreover, they can be synergistically employed with other techniques, such as Generative Adversarial Networks (GAN), to enhance the quality and diversity of manipulated multimedia generations. In scenarios where audio manipulation is also required, additional audio signal manipulation techniques must be applied and meticulously synchronised to avoid inconsistencies between both modalities, namely video and audio. This synchronisation ensures a seamless and coherent multimedia manipulation process.

2. In the ***development of systems for the detection of manipulated data***, it is anticipated that the observed trend throughout this survey will persist, with an increasing emphasis on ***multimodal classification systems***. This approach offers a more comprehensive and nuanced perspective due to the combination of audio and video modalities, resulting in a richer and more complex information source. Combining audio and video data provides a representation of the content. Audio contributes details such as speech, music, and sound effects, while video captures visual elements such as facial expressions, movements, and object interactions. The use of both modalities enables the capture of a wider range of features and contextual information, resulting in a more detailed representation of the data. By focussing on both audio and video modalities, users can identify subtle signs of manipulation, including inconsistencies and temporal discontinuities, among others. Furthermore, unlike approaches that only manipulate visual information, multimodal systems introduce new scenarios where one modality remains constant while the other is manipulated or where both modalities are manipulated. Training detection systems to recognise this diversity brings them closer to real-world scenarios and enhances their generalisability.

3. In most datasets, both public and private, the labelling convention typically operates at

the level of the entire sample, whether it is a video, an audio clip, or multimodal content. However, there exists another possibility of labelling, although less common, observed in certain articles, *temporal localisation*. This approach involves ***enhancing the granularity of labelling*** by specifying the precise time segments within the content that exhibit manipulation. Adopting temporal localisation not only enhances the quality of datasets but also augments the available information, thereby contributing to improved explainability of detection systems. Although this approach may initially appear to be a challenging task, the development of labelling tools has facilitated the way for segment-level labelling of both visual and acoustic information as real or manipulated. An example of such a tool with several opportunities is Label Studio [157].

4. The absence of ***XAI techniques*** in most articles is indeed quite striking. Considering the end-users of these systems, who often include non-expert users, the application of explainable AI techniques becomes paramount. This is especially crucial for users to trust the systems that operate as "black box" algorithms, lacking transparency and interpretability. It is a well-established fact that the incorporation of explainable AI techniques has the potential to significantly enhance the reliability of classification or localisation systems, especially when dealing with dynamic samples. However, it is worth noting that the choice of explainability techniques should be made judiciously. Inappropriate visualisations or explanations could potentially have the opposite effect, confusing users rather than providing clarity, thus compromising the trustworthiness of the system. Therefore, careful consideration and evaluation of the selected explainability techniques are essential to ensure that they effectively convey the system's decision-making processes and enhance user understanding without introducing additional confusion. This approach will be crucial to fostering trust and acceptance among a wider user base.

5. Manipulation techniques still exhibit errors, such as issues with synchronisation between different modalities, such as video and audio, and discrepancies between manipulated and genuine segments. These inconsistencies have become focal points in multimodal forensic techniques. In the future, dataset development will need to place particular ***emphasis*** on achieving video ***fluency***, striving to ***minimise incosistencies*** between real and manipulated segments as well as among different modalities. Synchronisation challenges should not be confined solely to different modalities like video and audio; they should also encompass alignment between manipulated and authentic segments. These vulnerabilities provide opportunities for forensic systems to operate effectively. Consequently, future generative and manipulation models should prioritise addressing these issues and try to improve the overall quality of datasets. This concerted effort will contribute to more robust and reliable forensic systems in the future.

Secondly, the different ***challenges*** facing the field of multimedia forensics will be examined. Throughout this review, we have identified several weaknesses that need to be overcome and present a challenge for future work in the area.

1. In the field of audio and multimodal, there tends to be a predominant focus on a single language, often English, with occasional other languages like Japanese and Chinese. This language-centric limitation in the datasets raises concerns about potential bias, as detection systems can make the manipulation signals difficult to identify in languages in which they were not trained. Given the lack of in-depth analysis in this regard, comprehensive studies are urgently needed to determine whether bias is indeed being introduced. If such bias is

identified, future data sets should consider strongly adopting a ***multilingual approach***. Furthermore, acoustic information can offer valuable insights into different situations and contexts, which can vary significantly depending on culture. While incorporating multiple languages into future datasets is undeniably complex, that is invaluable. Such an approach would mitigate biases and provide a broader spectrum of diverse cultural contexts and situations, enriching the dataset's representativeness and encompassing a wider range of population groups. This effort would contribute to more robust and inclusive multimedia content forensics.

2. The ***quality*** of current ***datasets*** is limited, they predominantly centre around specific subjects and often originate from controlled environments. These datasets exhibit signs of manipulation and are restricted in terms of the diversity of techniques used. Consequently, forensic systems have difficulties in generalising effectively to real-life situations. Therefore, the creation of more intricate and novel datasets that encompass a diverse spectrum of manipulation techniques becomes crucial. Such datasets will facilitate the evolution of advanced quality detection techniques applicable to emerging challenges.

3. As demonstrated in this survey, a multitude of manipulation techniques are evident, encompassing alterations to facial features, adjustments to colour or backgrounds, and the addition or removal of elements within samples, among others. These techniques can be produced using diverse models and architectures, each leaving ***representative traces*** in the generated samples. Such distinct features can introduce biases into forensic systems. When a system is trained only with samples generated by a single model, it may be difficult to effectively generalise to samples created by other models. It is therefore imperative that datasets cover a wide spectrum of methodologies and models, even for the same manipulation technique. This approach allows the model to improve its generalisability across unknown methodologies and improves its ability to identify samples that closely approximate real-world scenarios.

4. This research in this field unfolds along two parallel tracks: manipulation techniques and forensic techniques. Malicious actors are continually refining their skills to create manipulated samples that are increasingly precise and challenging to detect. Simultaneously, researchers are dedicated to designing novel detection systems capable of identifying these evolving manipulation techniques. This domain operates as an ongoing race between the creation of new manipulation techniques and the development of innovative detection systems. Hence, it is crucial to remain ***adaptable and consistently*** generate datasets that are very close to reality and exhibit higher quality. These datasets serve as the foundation for the creation of new and more effective detection systems. Such advancements are essential to combat the growing threat of disinformation, which poses an escalating problem within social networks.

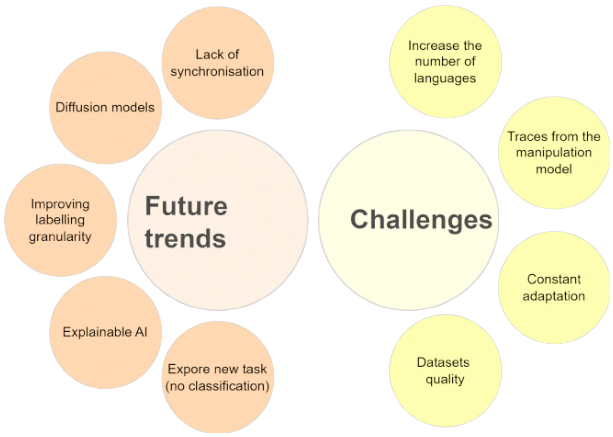


Figure 6.7: Future trends and challenges of forensics techniques on multimedia data.

Chapter 7

Conclusions and Future Work

Even the smallest person can change
the course of the future.

— R. R. Tolkien

In this last chapter, the various conclusions drawn from the research presented in this dissertation will be presented. The different Research Questions, which were raised in Chapter 1, will be answered with the aim of giving a general idea and the necessary details for future work within image processing to overcome the problems present in image datasets. Finally, potential future lines of work within this domain that would be interesting to explore are proposed.

7.1 Conclusions

Throughout the different chapters of this dissertation, different problems arising from image datasets have been solved. These datasets in most cases present a series of problems or limitations that cannot be solved by improving the quality of the datasets by adding new samples. Two main approaches have been used to solve these difficulties: information fusion techniques and ensembles. Due to the application domains addressed throughout this dissertation, another pillar of the dissertation has been the development of visualisation techniques based on heatmaps to facilitate their application in the different domains.

In chapter 2, the problems related to small size and low quality datasets in binary classification problems were first analysed using a dataset of paediatric chest radiographs consisting of 950 samples. For this purpose, an ensemble-based system composed of five models created from scratch was created, applying basic combination techniques, in particular, using the mean of the predictions. It was also demonstrated that when the available data do not contain enough information, simple models perform better than more complex models, such as CheXNet. The more complex models require a minimum quality dataset, which is not possible in many situations.

After solving binary classification problems, more complex problems related to dataset quality were explored further. In Chapter 3, a multilabel classification problem was solved, where the number of classes was very high (more than 30 different classes) within the medical field. Datasets with such a large number of classes tend to have imbalance problems, i.e. some classes

do not contain enough samples. Within the medical field, this problem is very common because not all diseases have the same incidence in the population. To solve this unbalance related problem, different ensemble techniques focused on multi-label classification problems were explored. Different ensemble-based systems were created using pre-trained state-of-the-art models, which demonstrated that ensemble techniques allow better performance and improved generalisation capability.

In chapter 4, an Early fusion-based methodology was proposed to solve regression problems in the forest fire management domain. To solve this problem, three sources of information were available, two of them were used to create the samples, GI and atmospheric variables, and another source of information was used to create the labels, the fire information. By combining these data sources we were able to design a regression system capable of predicting the resources needed in the event of a fire in a specific location in Castilla y León.

As has been shown throughout this dissertation, this research is very much focused on the application of DL or DL-related techniques in other domains far removed from DL, such as medicine or disinformation. The application of these techniques in domains outside the domains of DL is extremely difficult due to the opacity of these systems and models. Therefore, a pillar of this dissertation, more specifically Chapter 5, has been the design of heatmap visualisation techniques that allow us to show the features that the models or systems have used to generate the final result. Demonstrating that within the visualisations obtained using the grad-CAM algorithm, a wide variety of visualisations can be obtained depending on the problems to be solved and the preferences of the end users. Demonstrating that these techniques not only facilitate their application in external domains but also allow us to detect errors and biases in our systems due largely to biases and limitations due to the datasets used.

Finally, Chapter 6 has been the starting point for the extension of this dissertation to other data modalities, specifically video and audio. Due to their nature, they are particularly susceptible to be analyzed by means of ensemble and information fusion techniques. In order to properly understand the topic and discover the possible gaps that have not yet been exploited, it was considered that a review of the publications of the last few years would allow us to get a general idea of the state of the domain and discover in which areas we could focus and apply the techniques covered by this research. This research has allowed us to see that as in other domains, in the field of multimedia and multimodal information manipulation, datasets still present many weaknesses and limitations that can be overcome with the techniques analyzed throughout this dissertation.

7.1.1 Response to Research Questions

This section will focus on answering the different research questions posed in Chapter 1, thanks to the experimental analyses carried out throughout Chapters that compose this dissertation.

RQ1: Can ensemble techniques improve the performance of binary classification systems with datasets limited by sample size and quality in medicine field?

In Chapter 2, the performance of an ensemble approach using the mean to combine the probabilities obtained from the different models created from scratch was analysed, compared with the individual models and a state-of-the-art model widely used in similar tasks, CheXNet. In addition, the methodology presented in another similar state-of-the-art dataset was also applied.

The experiments performed showed an improvement in AUC of 11% compared to the best individual model proposed and an improvement of 6% in TPR for the dataset proposed in this research. If we compare the results obtained by the ensemble (AUC = 0.92 and TPR = 0.73) with the model used as the baseline of the state of the art, the improvement in AUC is 16% and an improvement in TPR of 29%. In other words, although the state of the art model presented an adequate AUC, although slightly lower than expected, it was not able to correctly detect the class of interest, consolidation. This showed that the proposed ensemble technique is able to obtain better and more robust results than the individual models and the baseline established in this research. Secondly, it can be observed that models created from scratch can present remarkable results when the dataset is not large enough and does not present enough features to be analysed with more complex models, such as CheXNet. Second, another state-of-the-art dataset was used to test the robustness of the approach proposed in this research. The dataset presented two different classification problems: normal vs. pneumonia and bacterial vs. viral pneumonia. In the first classification problem, the ensemble-based system showed a 2% and 6.8% improvement in AUC and TPR, respectively, improving the results obtained in the original article. In the second classification problem, although the TPR obtained in the original article was not improved, an improvement in the AUC of 2.4% was obtained.

The results obtained throughout this research show first of all that the proposed ensemble-based system is able to overcome binary classification problems with datasets that have two weaknesses, small size and limited quality of the samples. Second, it shows robust results that hold for other similar datasets in the state of the art. This confirms our initial hypothesis that a system based on ensembles with models created from scratch, less complex than the usual state-of-the-art models, is able to solve binary classification problems, although the datasets do not present the expected features.

RQ2: Can enesembles improve the performance of multilabel classification problems with data sets with an extreme imbalance within the medical field?

In Chapter 3, a deep learning methodology is proposed in this research for classification tasks with multilabel unbalanced datasets. An ensemble of five state-of-the-art architectures, namely DenseNet-201, EfficientNet B0, Inception, InceptionResNet, and Xception, has been constructed using this methodology. Weighted crossentropy with logit loss was employed to mitigate data imbalance. Additionally, a novel technique for generating heatmaps in multilabel classification problems was developed. Regarding the combination techniques, we have chosen to use three techniques specific to multilabel problems, CTP, PTC-mode and PTC-lw.

The results of the experiments are promising and have exceeded expectations. First, unlike existing state-of-the-art articles, we have established a methodologically robust baseline for future research, regardless of whether specific or general labels are used. This approach enables us to assess the performance of these models across varying numbers of labels. Our system achieves high AUC values for the classes utilised. Specifically, for specific labels, the system demonstrates outstanding performance with an AUC of 0.84. For general labels, we obtain an AUC of 0.819. This lower value may be attributed to the extensive classes and diverse radiological signs within the general classification. Consequently, the variability is greater, making classification more challenging.

If the results obtained with the different combination techniques are analysed, the CTP

approach performs better than the PTC approaches, as it is more informative and in classes where the probabilities are intermediate the PTC will not allow minimising individual errors. Moreover, this technique has been able to improve the individual results of the different models. The results of this research show how the use of ensembles, with the appropriate combinatorial technique, can overcome problems of multilabel datasets such as extreme unbalance.

RQ3: Can data fusion techniques help overcome regression problems within the wildfires domain, when the size of the dataset is extremely limited?

The architecture of the WAM model is proposed in Chapter 4 to predict wildfire resources using atmospheric variables and the GI. Due to the restricted number of labelled samples, an autoencoder was used to train on unlabelled samples, enabling it to discern trends and patterns within the variables of a specific geographic area. This acquired knowledge was then applied to a regression task, which predicts the required resources, control and extinguishing time, and the potential burnt area in the event of a fire. In conclusion, its capacity to generalise to different areas with varying meteorological conditions was examined, using a dataset from Andaluca. However, the applicability of the model to a specific location is constrained. Therefore, prediction maps were generated for each label, illustrating the essential resources needed at different locations within the autonomous region of Castilla y León for a particular date.

The promising results highlight the innovative nature of the methodology in the wildfire domain. In contrast to most published articles that exclusively employ ML algorithms, this study explores diverse DL approaches and techniques. Specifically, the residual autoencoder architecture surpasses the sequential one in performance. This is primarily attributed to skip connections, which offer alternative routes for gradients during the backpropagation process, allowing deeper layers to extract information from the initial layers. Significantly, better outcomes were achieved when the residual autoencoder architecture was applied after retraining the encoder on the regression task. This is likely because the introduction of new samples facilitated the regression model in capturing patterns that were originally overlooked by the initial autoencoder.

It is important to note that our model was initially trained in a specific region, Castilla y León, which is an autonomous region in Spain. Consequently, we conducted an analysis of its adaptability to other areas with different meteorological conditions, such as Andalucía, another autonomous region within the same country. The results indicated that the WAM model effectively predicted two labels, control and extinguishing time. However, it did not correctly predict the remaining labels. Specifically, the WAM model could not adapt to other regions without undergoing retraining. As mentioned above, the model displayed enhanced performance when the encoder was retrained in the context of the regression task, even when the samples were derived from the same autonomous region. The model's challenge in generalising to areas with distinct meteorological conditions can be attributed to its original pre-training on a specific region.

This research has shown that the application of early fusion techniques can increase the information of the samples, even if the dataset contains a limited number of samples. The combination of information fusion with a previous training with unlabelled samples has allowed us to solve the regression task in Castilla y León with a dataset composed of 445 samples. However, it has not been able to generalise to other areas such as Andalu-

sia, possibly because the pre-training has been performed in a limited area with specific conditions.

RQ4: How can XAI techniques be adapted to different binary and multilabel classification problems to facilitate the use of DL models within the medical domain?

Chapter 5 shows different visualisation techniques based on heatmaps generated using the grad-CAM algorithm. The different techniques have been generated for different classification problems, binary and multilabel, and for different end-user preferences, from very simple visualisations that only show the areas used by the model to reach the final result to more complex techniques where all the information obtained from the model is represented, i.e. the visualisation of each different class, the probability obtained by the model(s) and in case of application of ensemble techniques the agreement between the different models (both numerically and visually).

The visualization techniques presented in this chapter are derived from the classification problems in chapters 2 and 3, where two different classification tasks were solved: binary and multilabel. Several visualisations were also made for each visualisation problem from the simplest option to the most complex visualisation possible, containing as much information as possible. This wide variety of representations was done with the aim of representing all possible visualisations within the allowed spectrum, because each end-user may have different preferences. the course of this research it has been possible to validate these visualisations by a group of doctors. Throughout this research it has been possible to validate these visualisations by a group of doctors, both with seniors and junios doctors. Domínguez-Rodríguez et al. [1] analysed the agreement between the doctors and the visualisations. In the case of the senior doctors, they agreed with the visualisations obtained by 70%, and in the case of the residents, in all the cases analysed, their accuracy in reading the radiographs improved when the visualisations were used. This shows that the application of these systems in medicine can improve diagnosis.

However, it is important to note that the quality of the visualisations is entirely dependent on the performance of the classification system, so it is very important that the systems have high performance in order to obtain classification visualisations.

RQ5: How can the techniques of information fusion and ensembles analysed in this dissertation be applied to other multimedia information modalities, within the domain of information disorders?

Throughout this chapter, a review of the state of the art of manipulation detection techniques in video, audio and multimodal data has been carried out. The aim is to apply the techniques and knowledge of this dissertation to other multimedia information modalities, such as video and audio.

Regarding the datasets in this field, a series of limitations have been detected that can be solved with the techniques and knowledge of this dissertation, such as the imbalance between classes, many datasets present a higher number of false samples, especially in video and multimodal datasets. Secondly, not all datasets are of adequate size considering the complexity of the problem to be solved and the wide variety of manipulation techniques that can be found.

In terms of detection techniques, there is a strong presence of ensembles, not only applied

to multimodal multimedia information, but also in video treated as image sequences. On the other hand, information fusion, although present in numerous articles, has been less explored in this domain. Taking into account the nature of the information, multimodality, it is susceptible to be analysed by means of information assembly and fusion techniques, which would allow to use all the available information, improve the performance of the systems and improve the generalisation capacity.

7.2 Future works

Throughout this dissertation, we have tried to solve several problems arising from weaknesses of the datasets and to facilitate their applicability in different domains; however, we have also observed some points where this research can be extended:

- Within chest X-ray analysis we would like to explore the application of ensembles and fusion techniques to combine different views of the same X-ray. It is very common in this domain to perform different views from different angles in order to obtain more information about the lungs. This approach would allow us to increase the information available for the models without increasing the number of samples and with commonly available information. This would enable the performance of any classification task, binary or multi-label.

- As demonstrated in chapter 3, forcing the model to focus on the areas relevant to the problem improves performance, improves generalisability and reduces errors, so we want to explore new techniques such as attention mechanisms or other segmentation and segmentation-based cropping techniques. The use of these techniques is expected to improve the performance of classification tasks.

- In Chapter 4 we have observed problems of generalisation to other areas with different atmospheric conditions, in order to improve the performance of the model we want to create an autoencoder based on self-supervised learning with data from a larger area than the autonomous community of Castilla y León. In addition, new atmospheric variables can be included to allow the model to be applied to other ecological problems, such as droughts.

- To improve the performance of the autoencoder in Chapter 4, new options for encoder architectures are to be explored, e.g. transformers [2] or ConvNeXT [3]. The creation of new models capable of more precise reconstruction of representations will be able to understand patterns and trends in greater detail and thus create more accurate regression models.

- Another research line within the field of information fusion that we wish to explore or develop are techniques that allow us to know the relevance of each source of information, in order to select those sources of information that are really relevant to the task and avoid including information that may confuse or not provide relevant information. An example would be the analysis of the different variables and their influence on the problem in order to select or weight the different sources of information.

- Within the field of explainability, we want to explore other algorithms and XAI techniques to improve the quality of the visualisations. We also want to explore the possibility of combining different XAI algorithms to better visualise the features used in the different

models. This would facilitate the applicability of the systems in external fields and also a tool to detect possible errors during the generation of the systems.

- We want to develop a methodology that allows the application of the XAI techniques developed throughout this technique in the development process of DL models that allows an early detection of errors and inadequate selection of the image areas throughout the training. This methodology would provide information about the weaknesses of models and datasets quickly and more effectively than the analysis of numerical evaluations alone.

- Finally, we want to extend the application of the methodologies of this thesis to other multimodal data such as video, due to its visual and acoustic component, continuing the research conducted in Chapter 6, especially focusing on fine-grained manipulation detection systems, more precise and informative than most published works in the area, which focus on classification at the sample level.

Part II

Publications

PUBLICATION 1

Ensembles of Convolutional Neural Network models for pediatric pneumonia diagnosis

J1: Liz, H., Sánchez-Montañés, M., Tagarro, A., Domínguez-Rodríguez, S., Dagan, R., & Camacho, D. (2021). Ensembles of Convolutional Neural Network models for pediatric pneumonia diagnosis. Future Generation Computer Systems, 122, 220-233.
DOI: 10.1016/j.future.2021.04.007
Impact factor: 7.187 (JCR, 2021) [Q1, 10/110 CS, Artificial Intelligence]

- **Overall contribution**: This article presents a new approach to diagnosing paediatric pneumonia using ensembles of Convolutional Neural Network (CNN) models and explainable AI XAI techniques. Ensembles of models are used to improve classification performance, as presented in Chapter 2 while XAI techniques are used to overcome the lack of interpretability in CNN "black-box" algorithms. Specifically, the authors propose a new XAI technique based on combining individual heatmaps obtained from each model in the ensemble, which highlights the areas of the image that are most relevant to generate the classification, as presented in Chapter 5.

- **Contribution of the PhD candidate**:
  - First author of the article.
  - Contribution to the conception of the presented idea.
  - Design and execution of the experiments.
  - Co-author of the interpretation and discussion of the results provided.
  - Elaboration of the manuscript and visualisations.

Future Generation Computer Systems 122 (2021) 220–233

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Ensembles of Convolutional Neural Network models for pediatric pneumonia diagnosis

Helena Liz [a,f], Manuel Sánchez-Montañés [b,*], Alfredo Tagarro [c,d], Sara Domínguez-Rodríguez [d], Ron Dagan [e], David Camacho [a,*]

[a] Computer Systems Engineering Department,Universidad Politécnica de Madrid, Spain
[b] Computer Science Department Universidad Autónoma de Madrid, Spain
[c] Pediatrics Department. Hospital Universitario Infanta Sofía. Pediatrics Research Group. Universidad Europea de Madrid. Madrid, Spain
[d] Pediatric Research and Clinical Trials Unit (UPIC). Instituto de Investigación Sanitaria Hospital 12 de Octubre (IMAS12), Fundación para la Investigación Biomédica del Hospital 12 de Octubre, Madrid, Spain
[e] Faculty of Health Sciences, Ben-Gurion University of the Negev, Beer-Sheva, Israel
[f] Dept. Computer Sciences, Universidad Rey Juan Carlos, Spain

ARTICLE INFO

Article history:
Received 16 October 2020
Received in revised form 23 March 2021
Accepted 10 April 2021
Available online 17 April 2021

Keywords:
Ensembles of Convolutional Neural Networks
eXplainable Artificial Intelligence
Heatmaps
Pneumonia
Pediatrics

ABSTRACT

Pneumonia is a lung infection that causes 15% of childhood mortality (under 5 years old), over 800,000 children under five every year, around 2,200 every day, all over the world. This pathology is mainly caused by viruses or bacteria. X-rays imaging analysis is one of the most used methods for pneumonia diagnosis. These clinical images can be analyzed using machine learning methods such as convolutional neural networks (CNN), which learn to extract critical features for the classification. However, the usability of these systems is limited in medicine due to the lack of *interpretability*, because of these models cannot be used to generate an understandable explanation (from a human-based perspective), about how they have reached those results. Another problem that difficults the impact of this technology is the limited amount of labeled data in many medicine domains. The main contributions of this work are two fold: the first one is the design of a new explainable artificial intelligence (XAI) technique based on combining the individual *heatmaps* obtained from each model in the ensemble. This allows to overcome the explainability and interpretability problems of the CNN "black boxes", highlighting those areas of the image which are more relevant to generate the classification. The second one is the development of new ensemble deep learning models to classify chest X-rays that allow highly competitive results using small datasets for training. We tested our ensemble model using a small dataset of *pediatric X-rays* (950 samples of children between one month and 16 years old) with low quality and anatomical variability (which represents one of the biggest challenges addressed in this work). We also tested other strategies such as single CNNs trained from scratch and transfer learning using CheXNet. Our results show that our ensemble model clearly outperforms these strategies obtaining highly competitive results. Finally we confirmed the robustness of our approach using another pneumonia diagnosis dataset (Kermany et al., 2018).

1. Introduction

Pneumonia is the most common infectious disease in humans and the leading cause of childhood morbidity (refers to having a disease or a symptom of disease, or to the amount of disease within a population) and mortality in the world [1] that causes inflammation of the alveoli [2]. It especially affects children under 2 years old and elderly above 65. Globally, 15% of childhood mortality is caused by this disease, over 808,694 children in 2017 [3]. This mortality still remains around 800,000 children in 2018 under five every year, or 2200 children every day, resulting, as UNICEF shows in [4], the dramatic number of one child dead every 39 s, which includes over 153,000 newborns. Pneumonia is mainly caused by viruses or bacteria. Most frequent associated viruses are respiratory syncytial virus (RSV), influenza virus and human parainfluenza virus (HPIV) [5]. Most frequent bacteria are *Streptococcus pneumoniae*, *Haemophilus influenzae*, *Streptococcus pyogenes* and *Staphylococcus aureus*, and

* Corresponding authors.
E-mail addresses: helena.liz@alumnos.upm.es (H. Liz), manuel.smontanes@uam.es (M. Sánchez-Montañés), alfredo.tagarro@salud.madrid.org (A. Tagarro), sara.dominguez.r@gmail.com (S. Domínguez-Rodríguez), rdagan@bgu.ac.il (R. Dagan), david.camacho@upm.es (D. Camacho).

https://doi.org/10.1016/j.future.2021.04.007
0167-739X/

Mycoplasma pneumoniae [5]. The global prevalence of viral pneumonia origin in children is 14%–62%, being higher in children under two years old and decreasing with age [6,7].

Children with bacterial pneumonia should receive antibiotics as soon as possible, whereas children with viral pneumonia usually only need supportive care. However, antivirals may have a relevant role in the treatment of viral infections [8]. When in doubt, and given that bacterial pneumonia is more serious, the symptoms are usually resolved by using empirical antibiotic treatment, which in most cases is unnecessary since a virus is the most frequent cause of community-acquired pneumonia (CAP) [9].

Pneumonia diagnosis is fundamentally clinical, without reaching an etiological diagnosis most of the time. Due to the high economic cost, and the time it takes to obtain results (days, or even weeks), pediatricians often diagnose based on: laboratory tests, x-rays, and the examination of the patient. X-rays are one of the most important to diagnose CAP [10]. Furthermore, it is clear that a high proportion of all pneumonia cases are in fact vial-bacterial co-infections, complicating decisions regarding antibiotic administration [11]. Therefore, pediatricians have to decide empirically whether the child needs antibiotics, and choose the best treatment with their limited available tools. As a result, most children receive antibiotics. These results in what is considered over-treatment with antibiotics, leading to the need to narrow the indications by an appropriate discriminative diagnosis [12]. Proxies for typical bacterial pneumonia have been proposed, but the consensus is that no single biomarker alone is enough for diagnosing bacterial pneumonia [13,14]. The old paradigm that bacterial pneumonia is associated with a specific radiographic pattern different from the pattern of viral pneumonia is now often criticized, although the radiological pictures of alveolar pneumonia (also termed lobar pneumonia, or *pneumonia with consolidation*) appear to be bacterial in most of the cases [15]. The interpretation of the chest X-ray radiography (CXR) is usually performed following the standards of the "WHO Vaccine Trial Investigators Radiology Working Group" [16]. These standards establish two possible interpretations: "consolidation" (including consolidation and/or pleural effusion as per WHO standards) and "other infiltrates". However, the inter-observer agreement for these two categories is low [17].

Convolutional Neural Networks (CNNs) are well-known Deep Learning architectures. Recently great advances have undergone helping to solve several visual-related tasks [18]. This kind of neural networks is inspired by the biological neurons of the visual cortex [19], which allows them to solve problems such as image classification [20] and object recognition [21], among other image and video processing and recognition tasks. CNNs are also successful in other problems such as speech recognition [22], malware detection [23,24], natural language processing [25,26], among many others. These systems process the information in two main steps: *feature extractor*, where relevant features are detected; and *classification*, where these features obtained from the previous step are analyzed and different probabilities will be assigned to the detected structures to carry out the classification. In areas such as medicine, where the diagnosis is often based on the analysis of clinical images (e.g. x-rays), CNNs and Deep Learning methods have proven both their usefulness and effectiveness in the detection and classification of multiple diseases [19,27,28]. The classification performance in computer vision problems, such as those mentioned above, can be improved through ensembles of models [29,30]. An ensemble is a kind of method based on the combination of multiple learning algorithms, the ensemble is designed to improve the performance of the particular elements that builds up the new algorithm. This technique combine the individual predictions to produce a consensus prediction. Its advantages over individual models are the performance, because the combination of multiples models can improve their individual power and the final model can approximate better the optimal solution [31]; and robustness, because the ensemble models reduce the variance of prediction errors made but the contributing models by adding bias, avoiding overfitting of the final model [32]. For this reason this technique can be critical in small datasets like ours, where the information is particular limited in both size and quality.

However, CNNs, like many other Deep Learning and machine learning methods, are considered as "*black-box*" algorithms, where both the input and output can be easily analyzed and interpreted by the users, but where the inference process carried out by the algorithm is opaque hinders end-users' confidence in the results obtained, and therefore makes decision-making negatively affected. It makes this essential process ("how" and "why" the algorithm has obtained this outcome) uninterpretable for the human being [33]. This may limit its application in fields such as medicine, where the practitioners need to know how the algorithm has inferred the output for each specific patient [34] (e.g. why the algorithm is assigning a 90% probability for alveolar pneumonia?). This limitation can be overcome using automatic explanatory systems, called explainable AI (XAI) [35], which allows us to visualize which areas of the image (features) have been used to obtain the outcome generated by the algorithm, or at least alleviating, the aforementioned problem. These XAI-based systems will generate new images highlighting the areas of highest interest that the system uses to obtain the result (e.g. in our case to predict a particular kind of disease) [36].

The combination of Deep Learning models with medical knowledge allows the development of new clinical decision support systems (CDSS). These automatic systems can help in medical diagnosis reducing some typical clinical problems such as subjectivity in the interpretation of medical tests or human errors (fatigue, distraction, etc.) [37].

This combination of machine learning methods with human-based knowledge can improve the performance of the diagnosis process, as it was stated in [38]. In that project, leaded by Dr. Andrew Beck, it was demonstrated that the combination of pathologists and Deep Learning models provide a significant reduction of the error rate for breast cancer diagnosis. In the initial results pathologists obtained a 3.5% of error during classification of the pathology, whereas the Deep Learning algorithm obtained a slightly better result of 2.9% error. However, when both humans and AI model were combined this error decreased to an impressive 0.5% (so, the 99.5% of cases were correctly classified) [38].

The main contribution of this work can be briefly summarized as the design and development of a *novel Machine Learning system based on ensembles* for pneumonia diagnosis in childhood. Our system estimates the probability that the X-ray has a consolidation or other infiltrates, which would be a helpful tool for the unclear cases in case of disagreement among professionals, or in case of work overload. The result generated from the system should be user-friendly (from a health professional perspective), to achieve that, the system creates a graphical visualization using an explainable AI technique named heatmap, which highlights the areas of the image which are more relevant for the diagnosis according to the AI system [39].

An interesting point of this work is to understand how the neural network infers the pneumonia type (alveolar versus non alveolar) from the chest X-ray. This could help to expedite the treatment of patients who require medication. On the other hand, this could also help to avoid giving antibiotics to patients who do not need them. This is crucial since an incorrect use of antibiotics (that is, using them in patients who do not have a bacterial infection), or an excessive use of broad-spectrum antibiotics, can cause

antibiotic resistance. This can become a global problem making it more difficult to treat patients, the only solution being the development of new, more powerful antibiotics [12]. Therefore, in order to reduce the overuse of antibiotics in viral pneumonia, a correct diagnosis of bacterial pneumonia is crucial.

This article has been structured as follows: Section 2 provides a short description of some relevant works in the area of AI-based detection of lung diseases; Section 3 describes the methodology followed to design and train our CNN models; Section 4 shows the experimental results; Section 5 presents the conclusions and finally, future work, with some future lines of work.

2. Related work

AI techniques have been intensively applied in medicine, especially in diagnosis processes, forecasting and prediction of diseases evolution, global health impact, etc. We can find a very large number of examples, such as the diagnosis of brain abnormality with Magnetic resonance imaging (MRI) [40], cardiac disorders using clinical data [41], breast lesions [42], tuberculosis [43] or COVID-19 with chest X-ray [44]

Within the field of pneumonia there is also a multitude of systems with different kinds of data, such as: clinical data [45], ultrasounds [46], computed tomography [47] or X-rays [48,49], among many others. However, there are also numerous studies focused on other aspects, such as predicting the evolution of patients [50] or classification of patients according to severity using chest X-rays [51,52]. Due to the main contributions of this work, only some relevant works in the area of Deep Learning, and its application to the pathology considered (pneumonia), will be briefly described.

2.1. Deep learning methods for pneumonia diagnosis

Most of the previous works in this field use a pediatric dataset published by Kermany et al. [48], consisting of a total of 5856 radiographs divided in three classes (2780 bacterial, 1493 viral and 1583 normal) of patients between 1 and 5 years of age. Other works in the field [37] use other datasets and a classification hierarchy. First, radiographs are classified as "pneumonia" or "normal". Then, radiographs labeled as pneumonia are classified as "viral" or "bacterial". Kermany et al. obtained an AUC of 0.85 for the first classification (pneumonia versus normal) and 0.81 for the second (viral versus bacterial). We can also find the same classification system in [48], which will be explained later.

The first step that must be carried out is the pre-processing of the images. There are different algorithms, depending on the basic features, and pre-processing needs, to apply the target algorithms, of each dataset. The most commons are Histogram equalization, that enhances pathological signs [53,54]; Lung segmentation, which removes irrelevant data on X-rays and recovers useful information (because consolidation signs only appear inside lungs, therefore the rest of the image is irrelevant for the models). This one is applied in multiplicity of models [37,55,56].

Currently, one of the most successful AI methods for automatic image classification is Convolutional Neural Networks (CNN) [57]. Two main problems in CNN training for pneumonia diagnosis (and in most of the medical diagnosis problems) are low dataset availability and small dataset size. To solve these problems several works use *Transfer Learning*, a technique that takes advantage of the knowledge obtained from models used in other (similar) areas and trained with bigger datasets. The idea is that part of the CNN (the first layers) is "inherited" from a model previously trained in other dataset, while the rest of the CNN is trained with the penumonia dataset. This technique was used in the work by Kermany et al. [48], where part of the CNN was trained with ImageNet dataset and the rest of the CNN was trained with medical images including pediatric pneumonia. This resulted in an AUC of 0.96 for the classification task "pneumonia" versus "normal", and an AUC of 0.94 for the classification task "viral pneumonia" versus "bacterial pneumonia". This classification also shows an interesting classification distinction, unlike the rest of the networks, between three different categories (normal, bacterial and viral), achieving an AUC of 0.918. Other approaches, such as CheXNet [58], use a CNN called "DenseNet" which was also trained using ImageNet dataset and re-trained with a dataset of 14 different lung diseases (including pneumonia). This model has an AUC value for pneumonia of 0.768, however, the AUC of cardiomegaly and emphysema has higher values than the pneumonia AUC (0.925 and 0.937 respectively). Other technique to solve this problem is *Data augmentation*: a data-space solution to the problem of limited data, a quiet common problem in this field. It encompasses a suite of techniques that enhance the size and quality of training datasets such that better models can be built using them [59], increasing artificially, synthetically, the amount of training data using information only in it, so it helps to avoid overfitting. There are several techniques, such as geometric transformation (flippling, cropping, rotation, translation), color space transformation or noise injection [60].

Finally, CNN *ensembles* can improve the performance of a particular model based on an unique architecture, this technique is based on the idea of combining predictions from multiple statistical models (e.g. CNN) to form one final prediction (i.e. a new model made by a combination of selected models). There exist different ways to generate those ensembles [61]: *averaging*, ensemble averaging creates a group of models, each with low bias and high variance, then combines them into a new model with (hopefully) low bias and low variance. Some relevant examples can be seen in Shin et al. [62], which classifies two different datasets: the first one, Thoracoabdominal Lymph Node and Interstitial Lung Disease, comparing the performance of different architectures from the state of the art (CifarNet, Google-Net, AlexNet, etc.), and defining a new architecture that is made by the average value from the considered SoTA methods, or in the work by Christodoulidis et al. [63], which designs a system that classify between seven different Interstitial lung diseases using CT, and applying transfer learning technique; *majority voting*, in this approach the new model is built taking into account the outcomes from the different models considered, there are two approaches to the majority vote prediction for classification; *hard voting* and *soft voting*. Hard voting involves summing the predictions for each class label and predicting the class label with the most votes. Soft voting involves summing the predicted probabilities (or probability-like scores) for each class label and predicting the class label with the largest probability. Some examples of majority voting ensembles can be found in Yan et al. [64], which designs a CAD system for lung nodule malignancy risk classification from CT, using different CNNs with the objective of learning different levels of image spatial context, and improving detection performance; and finally, another popular ensemble method is the *weighted averaging*, which is based on a weighted combination of the former method, where a particular weight will be given to each model before the final model is generated. Some approaches that follows this method as is shown in Bermejo-Peláez et al. [65], which designs a model to classify computed tomography between 8 classes within Interstitial Lung Disease, another example of this approach can be found in Sirazitdinov et al. [28], which generates an object detection model for pneumonia detection and location from chest X-rays.

2.2. Explainable AI methods in medicine

As was previously mentioned, CNNs are considered *black-box* algorithms, which increases the difficulty of applying them in areas such as Medicine. The reason is that even these algorithms can be very precise classifying radiographs or making predictions, the end-users (medical staff in our case) will need to interpret and understand how and why the algorithm has reached the conclusion (the outcome provided to the end-user). For this reason, it is important to develop eXplainable AI (XAI) systems that allow end-users to understand how the system works (i.e. how and why classifies the radiography in our case).

We can distinguish between two main techniques, the first being *object detection* systems. These systems are based on CNN models that locate different objects in images. An example of this type of model is CoupleNet, which classifies and locates signs of pneumonia on a chest radiography and produces a visualization of the original image with bounding boxes in lung areas that show signs of disease [66]. These algorithms are used in different works to detect and locate pneumonia signs [28] combining two CNNs for the detection and location of pneumonia signs on lungs.

The second kind of methods uses additional techniques to *visualize how the model classifies*, such as *heatmap* generation. This method is used in a wide variety of problems including pneumonia diagnosis. For example in [37], where different lung areas can be seen with different color intensities according their relevance to the prediction made by the model. Other similar work is presented in Zech et al. [67].

A substantial difference between the two previous methods is that in the first one it is necessary to build a training dataset where the areas with signs of the disease have been marked by the experts. Whereas, in the second type of methods, it is only necessary to label each radiography of the training set with the classification given by the experts (consolidation/non consolidation in our case). After training, medical staff need to review the heatmaps generated by the model to validate the clinical sense.

3. Methodology

This section describes the methodology carried out to design and develop our model for pneumonia diagnosis.

The system has been designed following the three basic stages shown in Fig. 1.

All the processing was implemented in Python using well-known libraries such as numpy (mathematical computing) [68], matplotlib (visualization) [69], Keras (Deep Learning, [70]), and Keras-Vis (heatmaps calculation) [71].

The first step is the *data preprocessing* stage where the X-rays dataset is divided into training and test subsets. In this stage, individual X-rays images are also normalized. Data augmentation of training images was performed for a robust model construction [72]. The second stage is the *generation of a model* that classifies each X-rays into two classes (consolidation/non-consolidation). The accuracy of this model is validated using a k-fold cross-validation technique (where k, the number of folds, has been fixed to 5). Finally, the last step is the application of the explainable AI technique that we have selected to increase the interpretability of our system, *heatmap creation* [73]. To evaluate the quality of our system we follow two different strategies: (1) generating the heatmap using only a model, and (2) generating the heatmap using an ensemble of models with the same architecture, but trained with different data folds. The second strategy will allow us to compute an uncertainty level (given by the standard deviation) associated with each pixel, that allow us to analyze the robustness of the heatmap.

3.1. Datasets

In this work we use two different datasets. The first one is an X-ray pediatric-pneumonia (XrPP) dataset provided by Ben-Gurion University (Israel), and the second one is a public pediatric dataset of chest X-rays [48], which will be used to analyze the generalization capability of our model. X-rays have been labeled by experts using one of the following two mutually exclusive classes:

- **Consolidation**, denoting a chest x-ray image with signs of consolidation (alveolar pneumonia).
- **Non-consolidation**, denoting a chest x-ray image with other infiltrates signs that correspond with non-alveolar pneumonia.

The first dataset is formed by 950 labeled chest X-rays of children (between one month and 16 years), 403 cases of consolidation (42.42%) and 547 cases of non-consolidation (57.58%). These chest X-rays are posteroanterior (PA) radiographs showing the posterior view of the chest. Each case was classified by a panel of experts consisting of two senior pediatricians and a radiologist from Hospital 12 de Octubre Research Institute (Madrid, Spain). The experts also had access to lateral radiographs of the patients to increase the precision of the labeling of each case. During this classification, 50 samples were withdrawn due to lack of consensus from the expert panel on the diagnosis.[1] The authors have obtained the necessary permission from the ethical boards of Ben-Gurion University and Hospital 12 de Octubre Research Institute to work with these data.

Note that the distribution of classes in the dataset is relatively balanced. This is important, especially in small datasets like this, since an unbalanced distribution of classes can severely affect the model's performance (e.g. example accuracy, true positive rate: TPR, false positive rate: FPR) [74].

Another interesting feature of the dataset is the size of the images. The average size of the images is similar in both classes (approximately 200,000 pixels), a very low value for X-ray images, which implies that image resolution (and therefore quality) is limited. Therefore, another contribution of our work is the study of the reliability of CNNs when trained with small and low-resolution datasets.

The second dataset studied in our work is composed by 5856 X-rays of children between one and five years of age. There are 2780 cases of bacterial (47.5%), 1493 cases of viral (25.5%) and 1583 cases of normal (27.0%) [48], so the class distribution is not balanced. As described in that paper, all chest radiographs were screened for quality control, removing low quality or unreadable scans. Then, diagnoses for the images were graded by two expert physicians. Since the sizes of the images are between 1,000,000 and 2,000,000 pixels, both the images resolution (quality) and the number of images are clearly superior to the first dataset.

3.2. Data preprocessing

The X-ray images were provided by medical centers in jpg format. This format codes the color at each pixel using three values, the "RGB" components. In our dataset these values are redundant since RGB components are identical in grayscale images. Therefore we keep only the first one. The original images do not have the same size, so we normalize their shapes to 150 × 150 pixels. On the other hand, the pixel values of each image are normalized by dividing them by the average pixel value of the image.

[1] The dataset is not public and cannot be shared with the community.

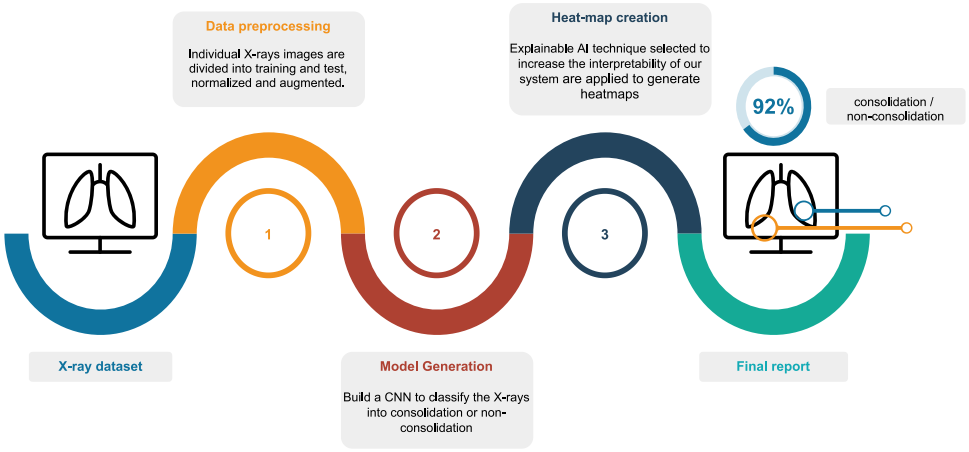


Fig. 1. Data flow diagram.

As we mentioned before, CNN usually needs a very large number of training images to avoid overfitting, however, our available dataset is particularly small. To overcome this problem we used a popular technique named *Data Augmentation*, which allows us to increase the size of our dataset [72]. It generates batches of images with real-time data augmentation. During each epoch, a different set of variations of the original training images is generated using different types of transformations [70]. In this work shearing (0.2), zoom (0.05), rotation (0.2), horizontal shift (0.1), vertical shift (0.1) and horizontal flip transformations have been used with a batch size of 32.

3.3. Data partitioning

In order to robustly evaluate the different architectures for our models, we generated a pool of different training/validation/test partitions of the dataset. For each of those partitions, we first divided the dataset randomly into construction (70%) and test (30%) subsets using stratified partitioning. Then the construction subset was randomly divided into training (80%) and validation (20%) subsets using stratified partitioning. Therefore each partition is a division of the original dataset into training (56%), validation (14%) and test (30%) subsets. These are mutually exclusive, so each image in each partition is only included in one subset (training, validation or test). Training subset will be used to learn the CNN's weights, whereas validation subset will be used for monitoring CNN's metrics throughout model's learning and avoid overfitting. Finally, test subset have been used for estimating the model's generalization capabilities (performance when new radiographs are considered).

3.4. Convolutional Neural Network model

We considered different architectures for our CNN model. The number of convolutional layers in them was in a 3–4 range (see Table 1), because a low number of convolutional layers may not be able to extract all the information from the images and a high number can lead to overfitting. Each convolutional layer has 32 kernels and a ReLU activation function. The output of the last convolutional layer is flattened and then perturbed by a Dropout with a rate of 70%. Then this information is processed by a dense layer ("FC layer") with a number of neurons depending on the architecture (Table 1) and ReLU activation function. Finally, the classification layer of our CNN consists in a dense layer of two neurons with Softmax activation. Therefore, we consider a total of six architectures (Table 1). Kernel L2 regularization with a strength of 0.01 was applied to FC dense layer. Each CNN was trained using Adam optimizer a learning rate of $1e^{-4}$.

Table 1
Hyperparameters of each of the six architectures considered.

	Architecture					
	Arch 1	Arch 2	Arch 3	Arch 4	Arch 5	Arch 6
Number of convolutional layers	4			3		
Number of neurons (FC layers)	64	128	256	64	128	256

3.5. Ensemble model

In order to increase the performance and robustness of our system, ensembles are considered. Each ensemble is composed by five different CNNs, each constructed using a different partition with different training/validation subsets but same test subset. This ensures models diversity, which is crucial for the ensemble performance [31].

The partitions were created as follows: first we randomly divided the dataset into construction (70%) and test (30%) subsets. Then we generated five different training/validation random partitions of the construction subset (80%/20%). For each of these five partitions a CNN was trained from scratch. The predictions for test subset of the ensemble formed by these five CNNs were then computed and the performance metrics evaluated. Ensemble's prediction for the probability of consolidation (and non consolidation) was computed as the average probability prediction across the five CNNs in the ensemble.

Summarizing, in order to compute in a robust and consistent way the metrics for the ensemble, the entire process described in this subsection was repeated using five different construction/test divisions. On the other hand, the total number of CNN models built for each division was 5. Therefore, a total of 25 models were constructed.

3.6. Reference model

In order to compare our system, CheXNet [58] has been selected as our reference model. CheXNet is a CNN model trained to classify chest X-rays over 14 different lung diseases, which is a very similar problem to ours (one of these lung diseases is precisely pneumonia). For this reason and the high performance with a similar problem was selected as the best model to compare the performance of ours. In order to make an adequate comparison between both models, the number of neurons in CheXNet's last layer was changed from 14 (number of classes in the original

paper) to two (number of classes in our work - consolidation/non-consolidation). Once the final layer is modified, the rest of layers in CheXNet's are freezing except the last two, which we will re-trained using the target dataset. Taking a previously trained model in another similar domain and retraining it in the current dataset is a popular strategy in Deep Learning called *transfer learning*.

3.7. Performance metrics

To analyze the performance of the different models, we used different metrics obtained from the ROC (Receiver Operating Characteristic Curve), a graphical representation that illustrates how the diagnostic capacity of a binary classifier system changes as its discrimination threshold is varied. More specifically, this curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold values [75].

A standard metric derived from the ROC is the *Area Under the Curve* (AUC), which measures the overall quality and accuracy of the classifier. Finally, we also measured the TPR, which in our problem corresponds to the fraction of positive (consolidation) cases that are correctly detected by the model. Accordingly to that, the fraction of patients that need antibiotics immediately that are correctly detected by the model.

3.8. Visual explanation using heatmaps

As it was previously mentioned, in order to generate an interpretable output for medical staff, we decided to generate a set of heatmaps. A heatmap is a matrix with the same size as the input image, where the value of each pixel is proportional to the importance of that pixel in the classification performed by the model has. Heatmaps were generated using the "Keras Vis" package [71]. These are shown overlaid with the original X-ray to make them more interpretable by medical staff. To overlay the images, the original X-ray and the obtained heatmap have a transparency degree of 50%. This allows us to see under the heatmap.

Typically, binary classification problems use a single output neuron. However, we used two in order to obtain a separate heatmap and generate the desired visualization for each of the two classes. Therefore our classifier layer has two output neurons, each one estimating the probability of the corresponding class (consolidation/non consolidation). The first step in heatmap generation is changing "the activation function of the last layer" of the network (the "output layer") from Softmax to Linear [71]. These two heatmaps were generated to be able to compare both classifications, allowing the end-user to comparing the areas that the model considers relevant when determining if the patient presents consolidation, or non-consolidation.

As explained above, ensembles are formed by five models. We generated ensemble heatmaps by averaging the individual heatmaps generated by those models. We also computed for each pixel the uncertainty of the heatmap by calculating the standard deviation of the individual heatmaps.

4. Experimental results

4.1. CNNs trained from scratch versus transfer learning with CheXNet

As described in 3.4, we considered six architectures for the CNN. These architectures were compared using CheX-Net, which was retrained in our dataset using transfer learning (see 3.6). Table 2 shows the performance metrics of the different models.

Table 2
AUC and TPR values of our six architectures and CheXNet. Possible significant differences are analyzed using Anova One Way test ($\alpha = 0.05$).

Arch	AUC		TPR	
	Value	p-value	Value	pvalue
Arch 1	**0.80 ± 0.03**		**0.62 ± 0.04**	
Arch 2	0.77 ± 0.02		0.56 ± 0.06	
Arch 3	0.78 ± 0.02		0.55 ± 0.08	
Arch 4	0.76 ± 0.02	0.04	0.57 ± 0.01	0.80
Arch 5	0.75 ± 0.01		0.55 ± 0.07	
Arch 6	0.77 ± 0.02		0.55 ± 0.09	
CheXNet	0.76 ± 0.02		0.43 ± 0.08	

Table 3
AUC and TPR values for Arch1 across the five different construction/test partitions. For each of them, five different training/validation splits were generated. Total number of CNNs trained from scratch: 25.

Partition	AUC		TPR	
	Value	p-value	Value	p-value
1	0.78 ± 0.01		0.60 ± 0.05	
2	**0.81 ± 0.01**		**0.67 ± 0.06**	
3	**0.80 ± 0.02**	0.20	**0.62 ± 0.07**	0.44
4	0.79 ± 0.01		0.64 ± 0.10	
5	0.80 ± 0.02		0.71 ± 0.09	
Average	0.80 ± 0.01		0.65 ± 0.04	

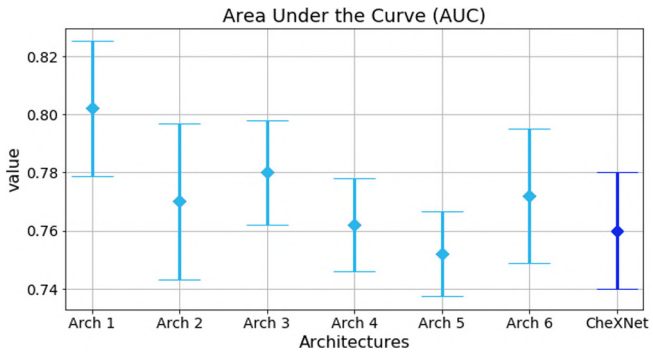


Fig. 2. AUC values for CheXNet-based model (dark blue) and the CNN models trained from scratch.

The statistics of AUC and TPR were calculated for all architectures using five different training/validation/test partitions.

Firstly, we can observe that our reference model, CheXNet, obtains a similar AUC value when is compared against our CNN architectures trained from scratch (see Fig. 2). However, the TPR obtained by CheXNet is 0.43 (Fig. 3). The reason might rely on the fact that CheXNet has been pre-trained using adult X-rays, and the differences between the original diseases are greater than the differences between alveolar and non-alveolar pneumonia. Furthermore, we can observe that the best architecture is Arch 1 (Architecture 1), which achieves the highest AUC and TPR (see Table 2). Since the differences in AUC are statistically significant (p-value = 0.04), Arch 1 was selected as the best architecture.

In order to analyze Arch 1 in depth, we generated five different divisions of the dataset into construction/test subsets, and for each construction subset we generated five different training/validation splits. For each training/validation/test partition a model was trained from scratch using Arch1, obtaining a total of 25 models. This allowed us to analyze the robustness and statistics of the architecture across different partitions (see Table 3).

We can observe in Table 3 and Fig. 4 that the performance of the architecture across different construction/test partitions is

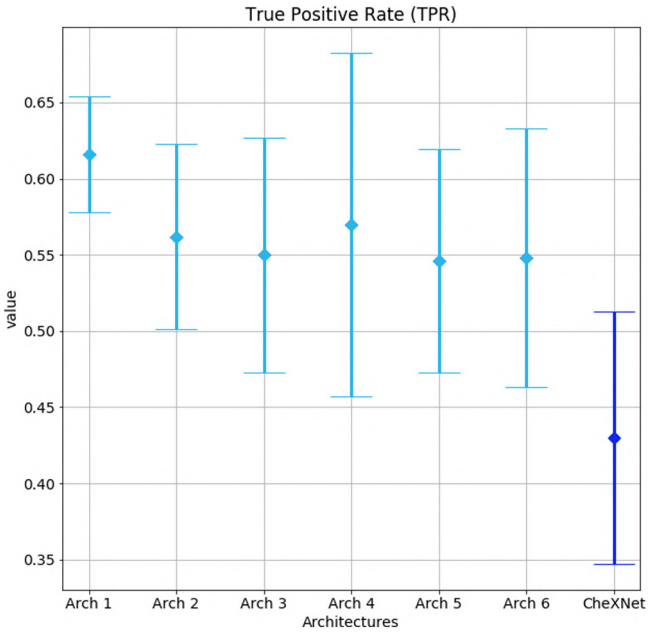


Fig. 3. TPR values for CheXNet-based model (dark blue) and the CNN models trained from scratch.

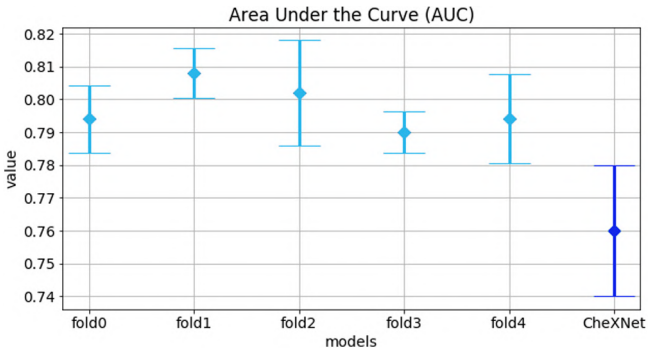


Fig. 4. AUC for CheXNet and our different models.

very similar, and the variance of the metrics is low. Therefore, we can conclude that this architecture is robust over our dataset.

Table 3 shows the performance of the individual Arch1 models. We will now analyze the performance of an ensemble consisting of five individual models. Taking in consideration that for each construction/test partition we generated 5 different training/validation partitions and trained a different CNN from scratch for each of them. Now we will analyze the performance of an ensemble formed by those five individual models. As it is shown in Table 4, a clear improvement in performance, both in AUC and TPR values, is now obtained. These values are higher than in the individual models with a difference of 9%/7% for AUC/TPR respectively. Therefore, it can be concluded that the ensemble is more robust against overfitting and provides better generalization capacity than the individual CNNs (see Fig. 5).

4.2. Visual explanation using heatmaps

The last step in our system is the heatmaps generation.

4.2.1. Individual heatmaps

Fig. 6 shows the prediction and heatmaps of an individual CNN for a non-consolidation test sample. The non-consolidation

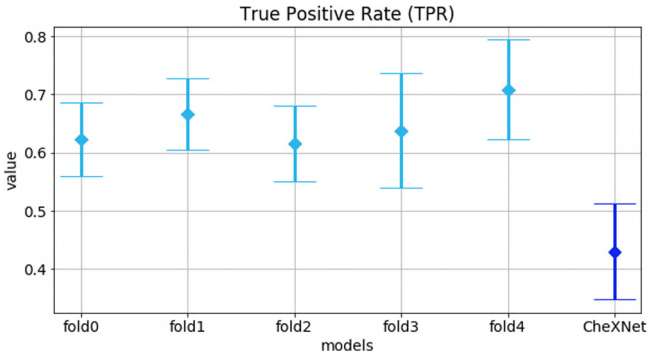


Fig. 5. TPR values for the five models and CheXNet.

Table 4
AUC and TPR values for Arch 1 ensembles.

Partition	AUC	TPR
1	0.89	0.71
2	0.92	0.73
3	0.88	0.65
4	0.88	0.73
5	0.87	0.79
Average	**0.89 ± 0.02**	**0.72 ± 0.04**

probability estimated by the model is 99.9%. If we analyze and compare the heatmaps, we can see that neuron 0 heatmap (non-consolidation) is clearly brighter than neuron 1 heatmap (consolidation), and in neuron 1 heatmaps there are different areas marked that should not be, corresponding to the clavicle and diaphragm. The conclusion is that the CNN has not found any signs of consolidation, but the previously referred areas of neuron 1 heatmap should not be lighted up.

In Fig. 8 we can see in neuron 1 heatmap that there is a consolidation sign in the upper left lung, but the consolidation probability estimated by the model is only 26.3%. The conclusions is that the CNN model incorrectly classified the X-ray as a non-consolidation sample but the heatmap is able to find the consolidation evidences.

Fig. 7 shows that there are signs of consolidation in the left lung. We can observe how both the probability estimated by the CNN model, 98.7%, and the heatmap corresponding to the consolidation class show that the system has correctly classified the radiography. However, we can see how the heatmap only marks the left side of the consolidations area, which does not cover all the consolidation signs. We can conclude that the CNN model has correctly classified the sample but it has not found all interesting area of the image.

4.2.2. Ensemble heatmaps

The ensembles are composed by five individual CNNs, therefore, for each radiography we have five different heatmaps for each class (consolidation/non-consolidation). This allows us to compute the average heatmap and the uncertainty at each pixel (standard deviation at each pixel) for each class.

In Figs. 9–11 we can see from left to right, and from top to bottom, the original X-ray, the average heatmap for the "non-consolidation" class, the average heatmap for the "consolidation" class, and the standard deviation heatmaps.

This visual representation provides to medical staff with relevant information. Firstly, the probabilities of each class predicted by the model are shown in the title of each average heatmap. Secondly, the average heatmaps show the areas of the X-rays that are most informative according to the ensemble. Finally, the standard deviation heatmaps provide information about the areas

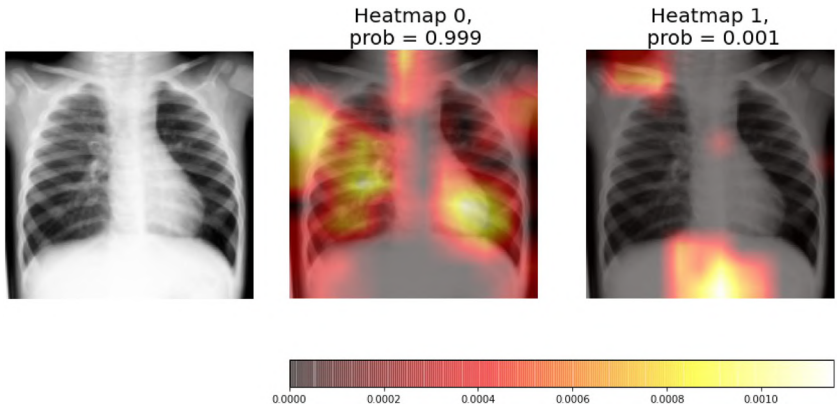


Fig. 6. Heatmaps generated with an individual CNN (Arch1) for a **non-consolidation** test X-ray (sample 1).

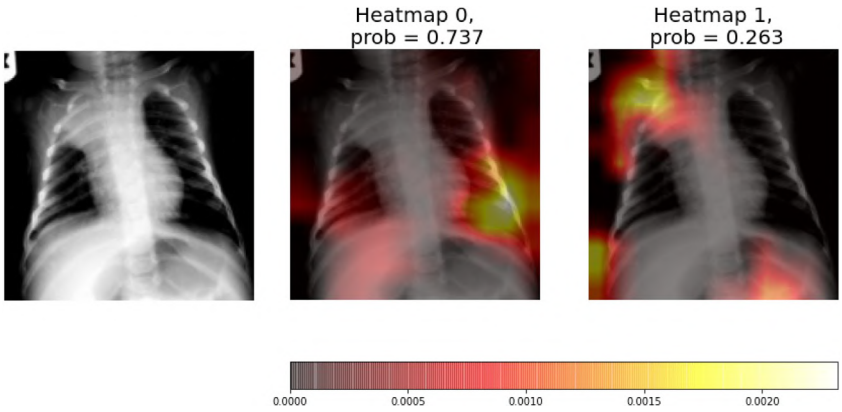


Fig. 7. Heatmaps generated with an individual CNN (Arch1) for a **consolidation** test radiography (sample 1).

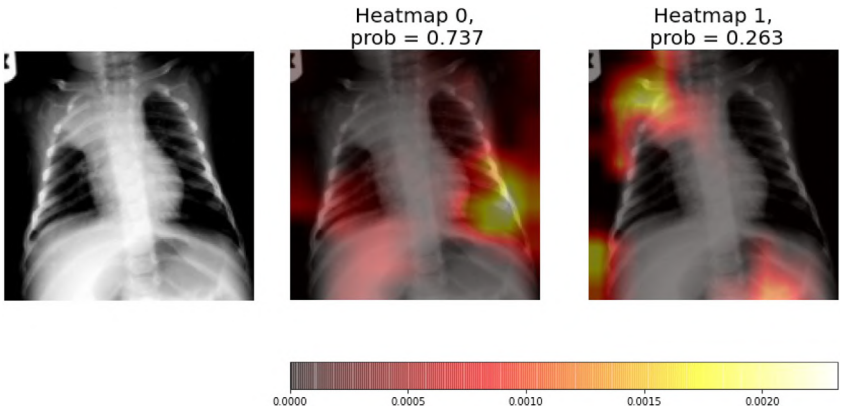


Fig. 8. Heatmaps generated with an individual CNN (Arch1) for a **consolidation** test radiography (sample 1).

of greatest disagreement among individual CNNs (that is, the areas with the greatest uncertainty). This suggests that medical staff should pay more attention to areas with higher values in the average heatmaps and in the standard deviation heatmaps.

In Fig. 9 we can see that the average heatmap for neuron 0 (non-consolidation class) is much brighter than the average heatmap for neuron 1 (consolidation). If we compare this visualization with the one obtained by an individual CNN (Fig. 6), we can see that it analyzes left lung in greater depth than the stand alone model, and focuses on the most relevant features to perform the X-ray classification. It also has a low standard deviation in this area and, unlike the visualization obtained by the individual CNN, it is hardly fixed on the heart area. Therefore, it can be concluded that the ensemble result is better than the one obtained by the individual CNN.

In Fig. 10 we can see how the average heatmaps of both neurons have lighted up, whereas only neuron 1 heatmap should light up because signs of consolidation appears in the upper area of the left lung. We can see how the standard deviation heatmap of neuron 0 presents higher values than neuron 1, which means that results of neuron 1 (consolidation class) are more robust than those of neuron 0 (non-consolidation class). If we look at the consolidation class average heatmap, the marked area corresponds to the pathology signs and it also marks a larger area than the same heatmap of the individual CNN (Fig. 7). Furthermore, if we look at the estimated probability of the models (81%) and we compare it with the one obtained with the individual CNN (26.7%), we can

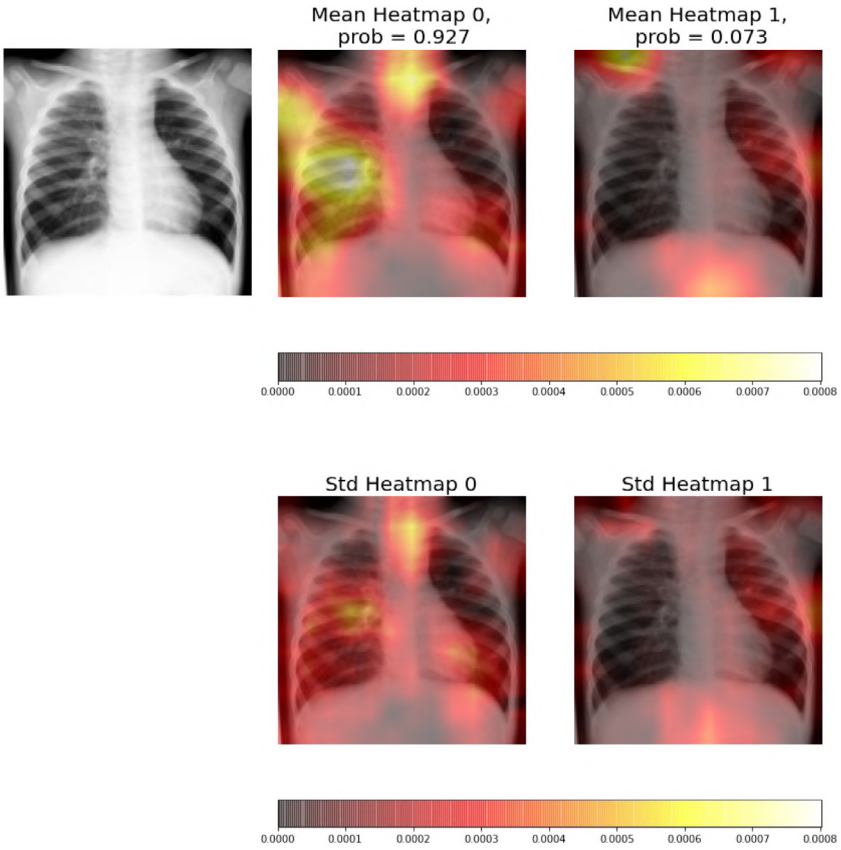


Fig. 9. Heatmaps of a **non-consolidation** sample 1. Left side, Neuron 0, shows the heatmaps for the non-consolidation class, whereas the right side, Neuron 1, shows the heatmaps for consolidation class for five individual CNNs.

see that ensembles give greater robustness to the system and avoid possible misclassifications obtained with the stand alone models.

Fig. 11 shows that the X-ray has consolidation signs in the left lung. The average heatmap marks practically all area affected by the pathology, unlike the visualization of an individual CNN that only marks a small part of the affected area. We can also observe how the standard deviation heatmap for neuron 1 indicates the adjacent areas to the one of interest. The probability of this class (consolidation) is 96.6%, lower than the obtained by individual CNN (98.7%), this is because the result is obtained from the average of the five models and not all of them have to present the optimal results, however by obtaining the average of all the models we increase the robustness of the model.

The average and standard deviation heatmaps (Figs. 9–11) provide very interesting and relevant information related to the ensemble output and how the system made a particular decision. However, the main problem related to this approach is the computational and time-consuming requirements. In our computer system, it took about 400 s to calculate the heatmaps for the five models in the ensemble.

4.3. Results with Kermany et al. dataset

Finally, we want to investigate the robustness of our approach now using the dataset of Kermany et al. [48]. Analogous to that paper, two classification problems were considered, normal versus pneumonia and bacterial versus viral pneumonia. We used the same training/test sets partition as in that work, and randomly subdivided training set into training (70%) + validation (30%) partitions. We trained 5 CNN models, each with a different training/validation partition. We considered Arch1 architecture for the individual models and followed a similar approach to previous experiments when this new dataset is considered.

Table 5
Kermany et al. dataset, normal versus pneumonia classification: comparison of AUC and TPR values originally reported by Kermany et al. to results obtained by our models.

	AUC	TPR
Kermany et al. model	0.968	0.932
Individual Arch 1 models	0.966 ± 0.007	0.91 ± 0.02
Arch 1 ensemble	0.976	1

Table 5 shows the results related to normal versus pneumonia classification problem. We observe that individual Arch1 models achieve AUCs similar to that reported in [48]. On the other hand, the TPR of the individual models is smaller (0.91 versus 0.932). However, the ensemble formed by our individual models achieves an AUC of 0.976 and a TPR of 1.0. Therefore, the ensemble shows better robustness and metrics than our individual models and the model presented in [48] based on DenseNet and transfer learning.

Note that the metrics are better than our previous dataset (see Section 4.1), which could be due Kermany et al. dataset is larger than ours (5856 X-rays versus 950) and the original images are higher quality (1–2 million pixels versus 200 K), so the training set contains much more information for constructing the models.

Table 6 indicates the results for the bacterial versus viral pneumonia classification problem. Again, it can be observed that individual Arch1 models achieve AUCs values similar to that reported in [48]. On the other hand, the ensemble formed by the individual models achieves an AUC value of 0.964 [48].

These results are remarkable as we obtained better AUCs using simple CNN ensembles than very deep CNN transfer learning

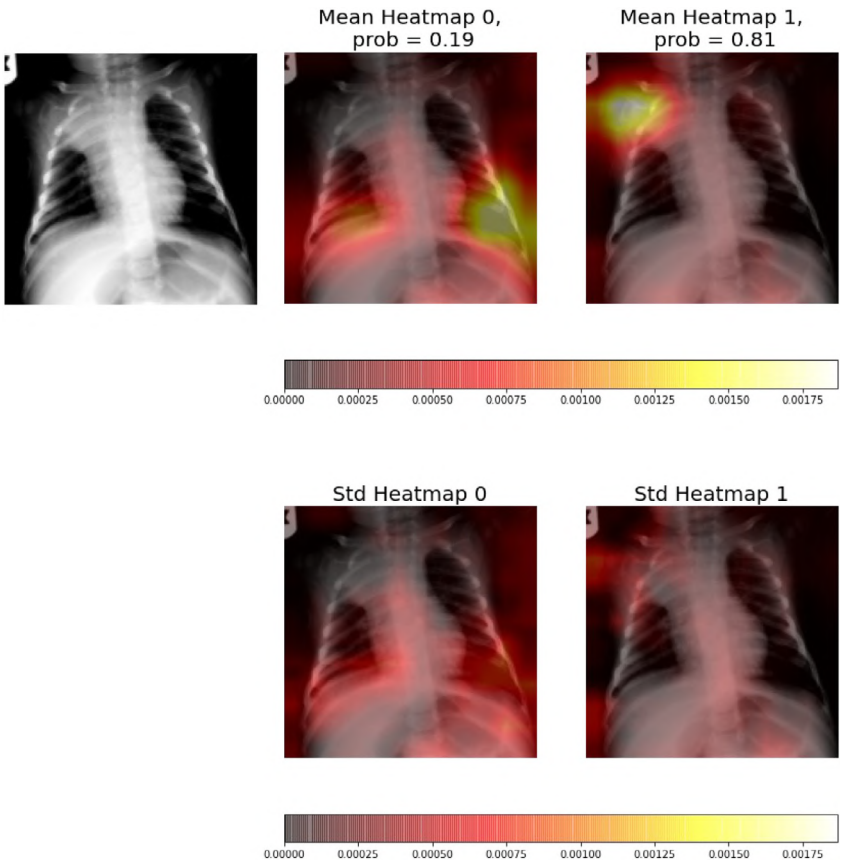


Fig. 10. Heatmaps of a **consolidation** sample 1. Left side, Neuron 0, shows the heatmaps for the non-consolidation class, whereas the right side, Neuron 1, shows the heatmaps for consolidation class for five individual CNNs.

Table 6
Kermany et al. dataset, classification bacterial versus virical: comparison of AUC and TPR values originally reported by Kermany et al. to results obtained by our models.

	AUC	TPR
Kermany et al. model	0.940	0.886
Individual Arch 1 models	0.94 ± 0.02	0.78 ± 0.05
Arch 1 ensemble	0.964	0.791

techniques such as those shown in [48], where transfer learning with a DenseNet architecture with 121 convolutional layers was used.

We can see that the application of ensembles in the Kermany et al. dataset also presents an improvement in the result for the bacterial and viral pneumonia classification obtained in AUC, with a difference of 2%. However, it can be observed that the improvement obtained is lower than that obtained in our dataset. The difference between the datasets is probably due to their quality. The dataset provided by Kermany et al. presents greater size and quality, so the models can achieve better values using only individual models and they can adequately generalize, while our dataset, as it is more limited, the individual models are not able to obtain the best possible results.

5. Conclusions

A large number of Convolutional Neural Network architectures have been recently proposed to help and support medical staff in the pneumonia diagnosis task. In this work, a new Machine Learning system based on ensembles, which combines XAI techniques and CNN models, has been designed for the childhood pneumonia diagnosis. When a simple model, over the target dataset, is used an AUC of 0.81 and a TPR of 0.67 values are obtained. However, applying ensemble techniques the performance of the model is improved to an AUC of 0.92 and a TPR of 0.73 values for this ensemble model. These results are in the line of the current state of the art results, although in some cases (as was described in Section 2.1) are lower. However, and opposite to other previous approaches from the state of the art, our CNN models classify between the presence or the absence of consolidation. The objective of this work was to analyze and study the applicability of XAI techniques in this domain, and applying our approach in a particular dataset (950 X-rays). Three main problems have been overcome, the small size of samples in our dataset (less than 1000), the low quality of images, and the anatomical variability existing within the age range of our dataset (between one month and 16 years old) [76,77].

Related to the visualizations provided (heatmaps) as the main XAI technique used, they present adequate results and provide much more information to medical staff than other methods. However, as it was described before, the quality of the model will depends on the quality of data, for this reason in some cases, as shown in Fig. 8, the model does not work correctly highlighting areas outside the lungs. Therefore, the model could be improved by using other techniques, such as segmentation methods to focus only on the lungs, and increasing the quality of the dataset. Finally we compare the results of models trained with our dataset and the dataset published by Kermany et al. and the results show that we can obtain similar values without very deep convolutional neural networks or transfer learning techniques.

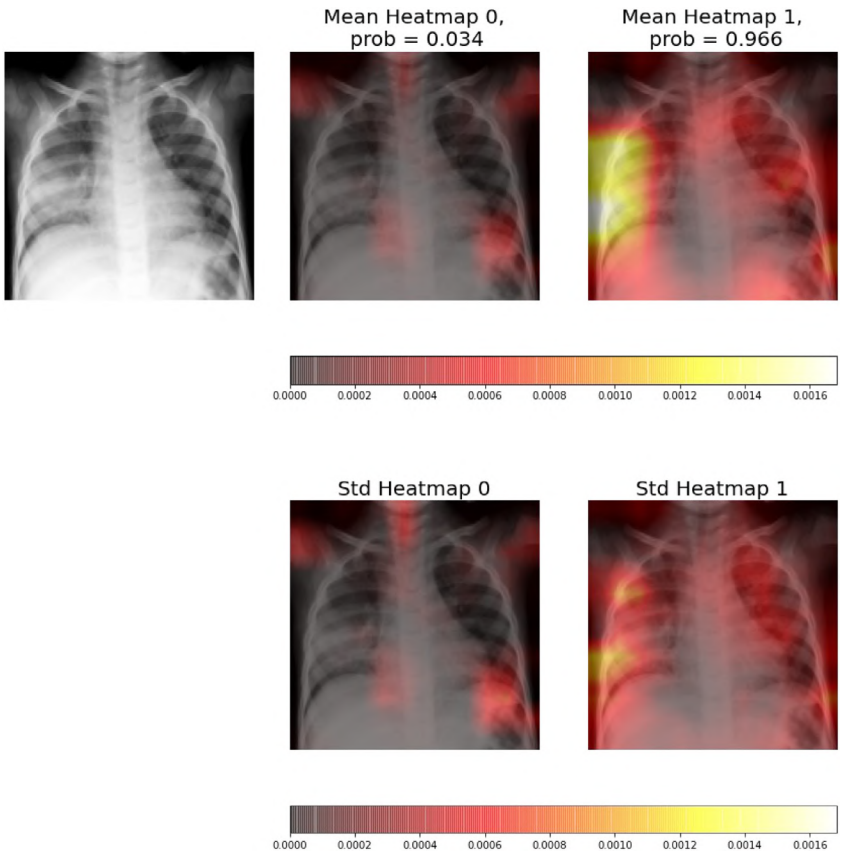


Fig. 11. Heatmaps of a **consolidation** sample 2. Left side, Neuron 0, shows the heatmaps for the non-consolidation class, whereas the right side, Neuron 1, shows the heatmaps for consolidation class for five individual CNNs.

6. Future work

Firstly, we would like to explore different preprocessing techniques, such as lung segmentation to force the model to focus only on the lungs, or different data augmentation techniques to improve the quality of the dataset. These techniques would improve the performance of our model. Secondly, we will study the performance of this technique with other datasets: for example with different thoracic diseases, such as COVID-19, Covid Data Save Lives (from HM hospitals) [78], that contains different kinds of clinical images and clinical data, BIMCV-COVID19 [79], which it is composed of different kinds of medical imaging and testing COVID-19; or multi-label datasets, for example such as PadChest [80], composed by different clinical images of lung and heart diseases or COVID-19 Image Data Collection [81], that contains X-rays by five lung diseases. The architecture could be improved with the application of a statistically-driven Coral Reef Optimisation algorithm for automatic selection of hyperparameter and the design of CNN [82]. Also we would to compare our system with other architectures, such as: COVID-Net [83], the model used by kermany et al. [48], Multichannel Convolutional Neural Network [84] among others, to better understand and study the robustness of the approach presented.

Other complementary XAI techniques based on visualizations will be considered with the aim to facilitate the work of medical staff. Finally, and to test the generality of the method proposed in this work, other domains, such as those related to industrial domains [85], where the combinations of CNN ensemble models and XAI methods, could increase the capabilities (analysis, prediction, explainability) of current used methods, will be explored and tested in the near future.

CRediT authorship contribution statement

Helena Liz: Implementation of the system, Analysis and/or interpretation of data, Writing - original draft, Writing - review & editing. **Manuel Sánchez-Montañés:** Implementation of the system, Analysis and/or interpretation of data, Writing - original draft, Writing - review & editing. **Alfredo Tagarro:** Conception and design of study, Acquisition of data, Analysis and/or interpretation of data, Writing - review & editing. **Sara Domínguez-Rodríguez:** Conception and design of study, Acquisition of data, Analysis and/or interpretation of data, Writing - review & editing. **Ron Dagan:** Analysis and/or interpretation of data, Writing - review & editing. **David Camacho:** Analysis and/or interpretation of data, Writing - original draft, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has been supported by Spanish Ministry of Science and Education under TIN2017-85727-C4-3-P (DeepBio) grant, by CHIST-ERA 2017 BDSI PACMEL project, Spain (PCI2019-103623, UPM-Spain), by Agencia Estatal de Investigación AEI/FEDER Spain, Project PGC2018-095895-B-I00, and Comunidad Autó-noma de Madrid, Spain under S2018/TCS-4566 (CYNAMON), S2017/BMD-3688 grants. We gratefully acknowledge the support of NVIDIA Corporation, Spain with the donation of the Titan V GPU used for this research.

All authors approved the version of the manuscript to be published.

References

[1] E. Prina, O.T. Ranzani, A. Torres, Community-acquired pneumonia, Lancet 386 (9998) (2015) 1097–1108.
[2] A. Thompson, Pneumonia, J. Amer. Manual Med. Assoc. 315 (6) (2016) 626.
[3] WHO, World health organization, Pneumonia (2020) https://www.who.int/news-room/fact-sheets/detail/pneumonia (last accessed: 23rd january 2021).
[4] UNICEF, Unicef data: Monitoring the situation of children and women, Pneumonia (2020) https://data.unicef.org/topic/child-health/pneumonia/ (last accessed: 23rd january 2021).
[5] M.U. Bhuiyan, T.L. Snelling, R. West, J. Lang, T. Rahman, M.L. Borland, R. Thornton, L.-A. Kirkham, C. Sikazwe, A.C. Martin, et al., Role of viral and bacterial pathogens in causing pneumonia among western Australian children: a case–control study protocol, Br. Med. J. open 8 (3) (2018) e020646.
[6] M.R. van den Bergh, G. Biesbroek, J.W. Rossen, W.A. de Steenhuijsen Piters, A.A. Bosch, E.J. van Gils, X. Wang, C.W. Boonacker, R.H. Veenhoven, J.P. Bruin, et al., Associations between pathogens in the upper respiratory tract of young children: interplay between viruses and bacteria, PLoS One 7 (10) (2012) e47711.
[7] C. Rodrigues, H. Groves, Community-acquired pneumonia in children: the challenges of microbiological diagnosis, J. Clin. Microbiol. 56 (3) (2018) e01318–17.
[8] L. Dong, S. Hu, J. Gao, Discovering drugs to treat coronavirus disease 2019 (COVID-19), Drug Discov. Ther. 14 (1) (2020) 58–60.
[9] WHO, World health organization, Antibiot. resist. (2020) https://www.who.int/news-room/fact-sheets/detail/antibiotic-resistance (last accessed: 23rd january 2021).
[10] W.G. Boersma, J.M. Daniels, A. Löwenberg, W.-J. Boeve, E.J. van de Jagt, Reliability of radiographic findings and the relation to etiologic agents in community-acquired pneumonia, Respir. Med. 100 (5) (2006) 926–932.
[11] O. Ruuskanen, E. Lahti, L.C. Jennings, D.R. Murdoch, Viral pneumonia, Lancet 377 (9773) (2011) 1264–1275.
[12] G. Tomson, I. Vlad, The need to look at antibiotic resistance from a health systems perspective, Upsala J. Med. Sci. 119 (2) (2014) 117–124.
[13] M.U. Bhuiyan, C.C. Blyth, R. West, J. Lang, T. Rahman, C. Granland, C. de Gier, M.L. Borland, R.B. Thornton, L.-A.S. Kirkham, et al., Combination of clinical symptoms and blood biomarkers can improve discrimination between bacterial or viral community-acquired pneumonia in children, BMC Pulm. Med. 19 (1) (2019) 71.
[14] M.M. Higdon, T. Le, K.L. O'Brien, D.R. Murdoch, C. Prosperi, H.C. Baggett, W.A. Brooks, D.R. Feikin, L.L. Hammitt, S.R. Howie, et al., Association of C-reactive protein with bacterial and respiratory syncytial virus–associated pneumonia among children aged< 5 years in the PERCH study, Clin. Infect. Dis. 64 (suppl_3) (2017) S378–S386.
[15] W.H. Organization, et al., Standardization of Interpretation of Chest Radiographs for the Diagnosis of Pneumonia in Children, Technical Report, World Health Organization, 2001.
[16] T. Cherian, E.K. Mulholland, J.B. Carlin, H. Ostensen, R. Amin, M.d. Campo, D. Greenberg, R. Lagos, M. Lucero, S.A. Madhi, et al., Standardized interpretation of paediatric chest radiographs for the diagnosis of pneumonia in epidemiological studies, Bull. World Health Organ. 83 (2005) 353–359.
[17] G. Mackenzie, The definition and classification of pneumonia, Pneumonia 8 (1) (2016) 14.
[18] A. Martín, V.M. Vargas, P.A. Gutiérrez, D. Camacho, C. Hervás-Martínez, Optimising convolutional neural networks using a hybrid statistically-driven coral reef optimisation algorithm, Appl. Soft Comput. 90 (2020) 106144.
[19] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.
[20] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[21] S. Ji, W. Xu, M. Yang, K. Yu, 3D Convolutional neural networks for human action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 35 (1) (2012) 221–231.
[22] G. Hinton, L. Deng, D. Yu, G.E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, et al., Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, IEEE Signal Process. Mag. 29 (6) (2012) 82–97.
[23] A. Martín, F. Fuentes-Hurtado, V. Naranjo, D. Camacho, Evolving deep neural networks architectures for android malware classification, in: 2017 IEEE Congress on Evolutionary Computation (CEC), IEEE, 2017, pp. 1659–1666.
[24] A. Martín, R. Lara-Cabrera, F. Fuentes-Hurtado, V. Naranjo, D. Camacho, Evodeep: a new evolutionary approach for automatic deep neural networks parametrisation, J. Parallel Distrib. Comput. 117 (2018) 180–191.
[25] E. Cambria, B. White, Jumping NLP curves: A review of natural language processing research, IEEE Comput. Intell. Mag. 9 (2) (2014) 48–57.
[26] T. Young, D. Hazarika, S. Poria, E. Cambria, Recent trends in deep learning based natural language processing, IEEE Comput. Intell. Mag. 13 (3) (2018) 55–75.
[27] G. Litjens, C.I. Sánchez, N. Timofeeva, M. Hermsen, I. Nagtegaal, I. Kovacs, C. Hulsbergen-Van De Kaa, P. Bult, B. Van Ginneken, J. Van Der Laak, Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis, Sci. Rep. 6 (2016) 26286.
[28] I. Sirazitdinov, M. Kholiavchenko, T. Mustafaev, Y. Yixuan, R. Kuleev, B. Ibragimov, Deep neural network ensemble for pneumonia localization from a large-scale chest x-ray database, Comput. Electr. Eng. 78 (2019) 388–399.
[29] H. Li, Z. Lin, X. Shen, J. Brandt, G. Hua, A convolutional neural network cascade for face detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5325–5334.
[30] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.
[31] T.G. Dietterich, Ensemble methods in machine learning, in: International Workshop on Multiple Classifier Systems, Springer, 2000, pp. 1–15.
[32] R. Minetto, M.P. Segundo, S. Sarkar, Hydra: An ensemble of convolutional neural networks for geospatial land classification, IEEE Trans. Geosci. Remote Sens. 57 (9) (2019) 6530–6541.
[33] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Comput. Surv. 51 (5) (2018) 1–42.
[34] A. Holzinger, C. Biemann, C.S. Pattichis, D.B. Kell, What do we need to build explainable AI systems for the medical domain?, 2017, arXiv:1712.09923.
[35] A.B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115.
[36] R.C. Fong, A. Vedaldi, Interpretable explanations of black boxes by meaningful perturbation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3429–3437.
[37] N. Mahomed, B. van Ginneken, R.H. Philipsen, J. Melendez, D.P. Moore, H. Moodley, T. Sewchuran, D. Mathew, S.A. Madhi, Computer-aided diagnosis for world health organization-defined chest radiograph primary-endpoint pneumonia in children, Pediatr. Radiol. (2020) 1–10.
[38] T. Kontzer, Deep learning drops error rate for breast cancer diagnoses by 85%, 2016, https://blogs.nvidia.com/blog/2016/09/19/deep-learning-breast-cancer-diagnosis/.
[39] W. Samek, T. Wiegand, K.-R. Müller, Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models, 2017, arXiv:1708.08296.
[40] M. Talo, U.B. Baloglu, Ö. Yıldırım, U.R. Acharya, Application of deep transfer learning for automated brain abnormality classification using MR images, Cogn. Syst. Res. 54 (2019) 176–188.
[41] K.-C. Chang, P.-H. Hsieh, M.-Y. Wu, Y.-C. Wang, J.-Y. Chen, F.-J. Tsai, E.S. Shih, M.-J. Hwang, T.-C. Huang, Usefulness of machine learning-based detection and classification of cardiac arrhythmias with 12-lead electrocardiograms, Canad. J. Cardiol. 37 (1) (2021) 94–104.
[42] M.A. Al-Antari, S.-M. Han, T.-S. Kim, Evaluation of deep learning detection and classification towards computer-aided diagnosis of breast lesions in digital X-ray mammograms, Comput. Methods Programs Biomed. 196 (2020) 105584.
[43] A. Hernández, Á. Panizo, D. Camacho, An ensemble algorithm based on deep learning for tuberculosis classification, in: International Conference on Intelligent Data Engineering and Automated Learning, Springer, 2019, pp. 145–154.
[44] A.M. Ismael, A. Şengür, Deep learning approaches for COVID-19 detection based on chest X-ray images, Expert Syst. Appl. 164 (2021) 114054.
[45] W. Dai, P.-F. Ke, Z.-Z. Li, Q.-Z. Zhuang, W. Huang, Y. Wang, Y. Xiong, X.-Z. Huang, Establishing classifiers with clinical laboratory indicators to distinguish COVID-19 from community-acquired pneumonia: Retrospective cohort study, J. Med. Internet Res. 23 (2) (2021) e23390.
[46] A. Singh, S. Shalini, R. Garg, Classification of pediatric pneumonia prediction approaches, in: 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), IEEE, 2021, pp. 709–712.
[47] L. Li, L. Qin, Z. Xu, Y. Yin, X. Wang, B. Kong, J. Bai, Y. Lu, Z. Fang, Q. Song, et al., Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT, Radiology (2020) 200905.
[48] D.S. Kermany, M. Goldbaum, W. Cai, C.C. Valentim, H. Liang, S.L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, et al., Identifying medical diagnoses and treatable diseases by image-based deep learning, Cell 172 (5) (2018) 1122–1131.
[49] Z. Yue, L. Ma, R. Zhang, Comparison and validation of deep learning models for the diagnosis of pneumonia, Comput. Intell. Neurosci. 2020 (2020).
[50] J. Mushtaq, R. Pennella, S. Lavalle, A. Colarieti, S. Steidler, C.M. Martinenghi, D. Palumbo, A. Esposito, P. Rovere-Querini, M. Tresoldi, et al., Initial chest radiographs and artificial intelligence (AI) predict clinical outcomes in COVID-19 patients: analysis of 697 Italian patients, Eur. Radiol. (2020) 1–10.
[51] S. Tabik, A. Gómez-Ríos, J. Martín-Rodríguez, I. Sevillano-García, M. Rey-Area, D. Charte, E. Guirado, J. Suárez, J. Luengo, M. Valero-González, et al., Covidgr dataset and COVID-sdnet methodology for predicting COVID-19 based on chest X-Ray images, 2020, arXiv preprint arXiv:2006.01409.
[52] J. Zhu, B. Shen, A. Abbasi, M. Hoshmand-Kochi, H. Li, T.Q. Duong, Deep transfer learning artificial intelligence accurately stages COVID-19 lung disease severity on portable chest radiographs, PLoS One 15 (7) (2020) e0236621.

[53] A. Sharma, D. Raju, S. Ranjan, Detection of pneumonia clouds in chest X-ray using image processing approach, in: 2017 Nirma University International Conference on Engineering (NUiCONE), IEEE, 2017, pp. 1–4.
[54] S. Khobragade, A. Tiwari, C. Patil, V. Narke, Automatic detection of major lung diseases using chest radiographs and classification by feed-forward artificial neural network, in: 2016 IEEE 1st International Conference on Power Electronics, Intelligent Control and Energy Systems, IEEE, 2016, pp. 1–5.
[55] B. Li, G. Kang, K. Cheng, N. Zhang, Attention-guided convolutional neural network for detecting pneumonia on chest X-Rays, in: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2019, pp. 4851–4854.
[56] M. Kim, B.-D. Lee, Automatic lung segmentation on chest X-rays using self-attention deep neural network, Sensors 21 (2) (2021) 369.
[57] O. Stephen, M. Sain, U.J. Maduh, D.-U. Jeong, An efficient deep learning approach to pneumonia classification in healthcare, J. Healthc. Eng. 2019 (2019).
[58] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, et al., Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning, 2017, arXiv preprint arXiv:1711.05225.
[59] M. Elgendi, M.U. Nasir, Q. Tang, D. Smith, J.-P. Grenier, C. Batte, B. Spieler, W.D. Leslie, C. Menon, R.R. Fletcher, et al., The effectiveness of image augmentation in deep learning networks for detecting COVID-19: A geometric transformation perspective, Front. Med. 8 (2021).
[60] C. Shorten, T.M. Khoshgoftaar, A survey on image data augmentation for deep learning, J. Big Data 6 (1) (2019) 60.
[61] F. Schwenker, Ensemble methods: Foundations and algorithms [book review], IEEE Comput. Intell. Mag. 8 (1) (2013) 77–79.
[62] H. Shin, H.R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, R.M. Summers, Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning, IEEE Trans. Med. Imaging 35 (5) (2016) 1285–1298.
[63] S. Christodoulidis, M. Anthimopoulos, L. Ebner, A. Christe, S. Mougiakakou, Multisource transfer learning with convolutional neural networks for lung pattern analysis, IEEE J. Biomed. Health Inform. 21 (1) (2016) 76–84.
[64] X. Yan, J. Pang, H. Qi, Y. Zhu, C. Bai, X. Geng, M. Liu, D. Terzopoulos, X. Ding, Classification of lung nodule malignancy risk on computed tomography images using convolutional neural network: A comparison between 2d and 3d strategies, in: Asian Conference on Computer Vision, Springer, 2016, pp. 91–101.
[65] D. Bermejo-Peláez, S.Y. Ash, G.R. Washko, R.S.J. Estépar, M.J. Ledesma-Carbayo, Classification of interstitial lung abnormality patterns with an ensemble of deep convolutional neural networks, Sci. Rep. 10 (1) (2020) 1–15.
[66] T.D. Team, Pneumonia detection in chest radiographs, 2018, arXiv:1811.08939.
[67] J.R. Zech, M.A. Badgeley, M. Liu, A.B. Costa, J.J. Titano, E.K. Oermann, Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study, PLoS Med. 15 (11) (2018).
[68] T.E. Oliphant, A Guide to Numpy, Vol. 1, Trelgol Publishing USA, 2006.
[69] J.D. Hunter, Matplotlib: A 2D graphics environment, Comput. Sci. Eng. 9 (3) (2007) 90–95.
[70] F. Chollet, et al., Keras, 2015, https://keras.io (Accessed: 2020-02-28).
[71] R. Kotikalapudi, contributors, Keras-vis, 2017, https://github.com/raghakot/keras-vis.
[72] L. Taylor, G. Nitschke, Improving deep learning using generic data augmentation, 2017, arXiv preprint arXiv:1708.06020.
[73] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (XAI), IEEE Access 6 (2018) 52138–52160.
[74] Q. Wei, R.L. Dunbrack Jr, The role of balanced training and testing data sets for binary classifiers in bioinformatics, PLoS One 8 (7) (2013).
[75] T. Fawcett, An introduction to ROC analysis, Pattern Recognit. Lett. 27 (8) (2006) 861–874.
[76] S. Andronikou, F.M. Vanhoenacker, A.I. De Backer, Advances in imaging chest tuberculosis: blurring of differences between children and adults, Clin. Chest Med. 30 (4) (2009) 717–744.
[77] A. Lander, J. Newman, Paediatric anatomy, Surgery (Oxford) 31 (3) (2013) 101–105.
[78] H. Hospitales, Covid data save lives - HM hospitales, 2021, https://www.hmhospitales.com/coronavirus/covid-data-save-lives (last accessed: 20th January 2021).
[79] M.d.l.I. Vayá, J.M. Saborit, J.A. Montell, A. Pertusa, A. Bustos, M. Cazorla, J. Galant, X. Barber, D. Orozco-Beltrán, F. Garcia, et al., BIMCV Covid-19+: a large annotated dataset of RX and CT images from COVID-19 patients, 2020, arXiv preprint arXiv:2006.01174.
[80] A. Bustos, A. Pertusa, J.-M. Salinas, M. de la Iglesia-Vayá, Padchest: A large chest x-ray image dataset with multi-label annotated reports, Med. Image Anal. 66 (2020) 101797.
[81] J.P. Cohen, P. Morrison, L. Dao, K. Roth, T.Q. Duong, M. Ghassemi, Covid-19 Image data collection: Prospective predictions are the future, 2020, arXiv preprint arXiv:2006.11988.
[82] A. Martin, R. Lara-Cabrera, V.M. Vargas, P.A. Gutiérrez, C. Hervás-Martínez, D. Camacho, Statistically-driven coral reef metaheuristic for automatic hyperparameter setting and architecture design of convolutional neural networks, in: 2020 IEEE Congress on Evolutionary Computation (CEC), IEEE, 2020, pp. 1–8.
[83] L. Wang, Z.Q. Lin, A. Wong, Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images, Sci. Rep. 10 (1) (2020) 1–12.
[84] A.-A. Nahid, N. Sikder, A.K. Bairagi, M. Razzaque, M. Masud, A. Z Kouzani, M. Mahmud, et al., A novel method to identify pneumonia through analyzing chest radiographs employing a multichannel convolutional neural network, Sensors 20 (12) (2020) 3482.
[85] V. Rodriguez-Fernandez, A. Trzcionkowska, A. Gonzalez-Pardo, E. Brzychczy, G.J. Nalepa, D. Camacho, Conformance checking for time-series-aware processes, IEEE Trans. Ind. Inf. 17 (2) (2020) 871–881.

Helena Liz is currently working as bioinformatician for the Fundación para la investigación Biomédica Hospital Infantil Universitario Niño Jesús (Madrid, Spain). Se has a Bachelor on Biology and a Master's degree in Bioinformatics and Computational Biology at the Universidad Autónoma de Madrid. Her research interests include Deep Learning, AI and Machien Learning applications in medicine, among others. Contact her at: helena.lizlopez@gmail.com

Manuel A. Sánchez-Montañés received his B.Sc. degree (with honors) in Physics from the Universidad Complutense de Madrid, Spain, 1997, and Ph.D. degree (cum laude) in Computer Science from the Universidad Autónoma de Madrid, Spain, 2003. He is currently part of the permanent faculty of the Computer Science Department, Universidad Autónoma de Madrid. His research activity is focused in Artificial Intelligence and Advanced Data Analysis, carrying out theoretical developments and applications.

Alfredo Tagarro is a Pediatrician and Clinical Researcher in Pediatrics since 2005. Dr. Tagarro obtained his Ph.D. in 2007. He has worked as a care provider in pediatric intensive care units, and since 2008 in Hospitalization, Emergency and Consultation, with a special interest in Infectious Diseases. At Infanta Sofía Hospital, where he has been working since 2008, doing all the Pediatric Infectious Disease protocols, and he is part of the Infectious Diseases Commission. He is an Associate Professor of Pediatrics at the European University of Madrid, and he is accredited by National Agency for Quality Assessment and Accreditation (ANECA) for Associate Professor. Currently, he is also PI of a project granted by the competitive National Call for Research (AES) Carlos III Health Institute, and co-researcher of several projects, including Co-researcher of Small Grant Award of European Society of Pediatric Infectious Diseases, 2018. Also, Dr. Tagarro is a co-researcher and Trial Coordinator of a clinical trial funded by EDCTP for trials in Africa (EMPIRICAL), funded with 7,6 million euros.

Sara Domínguez is currently the principal biostatistician in the Group of Translational Research on Pediatric infectious diseases (GTIP) in Hospital 12 Octubre in Madrid (Spain). She graduated from the Universidad Complutense of Madrid, with bachelor's in biology science in 2013. She is MsC in Biostatistics and Bioinformatics graduated from the Universidad Autónoma de Barcelona (Spain) and currently finishing her Ph.D. program in Medicine. She performs or supervises statistical analysis from the unit and assist the PI, study coordinators, and data managers in defining and maintaining SOPs and statistical analysis plans from clinical trials (phase II/phase III) and observational studies. Her research interests include Bayesian approaches, mixed-models, machine learning applied to the clinical field, and epidemiology. She has experience on HIV-1 epidemiology and cancer research. She is member of the Spanish Society for Epidemiology (SEE), Grupo de Estudio del SIDA-SEIMC, and young-member from the European Society for Pediatric Infectious Diseases and International AIDS Society. Contact her at: sara.dominguez.r@gmail.com

Ron Dagan is Distinguished Professor of Pediatrics and Infectious Diseases at the Ben-Gurion University of the Negev, Beer-Sheva, Israel. He founded the Pediatric Infectious Disease Unit at the Department of Pediatrics, Soroka University Medical Center in Beer-Sheva and served as its director from 1987 to 2014. He still remains part of the staff. His previous appointments include Adjunct Associate Professor of Pediatrics, University of Rochester, New York (1993-1998). He obtained his MD degree in 1974 (Hadassah School of Medicine, Hebrew University, Jerusalem). In 1982, he embarked on a 3-year Fellowship in Pediatric Infectious Diseases at the University of Rochester, NY.

David Camacho is currently working as Full Professor with the Departamento de Sistemas Informáticos at Universidad Politécnica de Madrid (Spain) and leads the Applied Intelligence and Data Analysis group (AIDA). He received a Ph.D. in Computer Science (2001) from Universidad Carlos III de Madrid, and a B.S. in Physics (1994) from Universidad Complutense de Madrid. His research interests include Data Mining, Evolutionary Computation, Social Network Analysis, and Swarm Intelligence, among others. He is on the editorial board of several journals, such as Information Fusion, International Journal of Bioinspired Computation, Journal of Ambient Intelligence and Humanized Computing and Expert systems. Contact him at: david.camacho@upm.es.

PUBLICATION 2

Deep learning for understanding multilabel imbalanced Chest X-ray datasets

J2: Liz, H., Huertas-Tato, J., Sánchez-Montañés, M., Del Ser, J., & Camacho, D. (2023). Deep learning for understanding multilabel imbalanced Chest X-ray datasets. Future Generation Computer Systems, 144, 291-306.
DOI: `10.1016/j.future.2023.03.005`
Impact factor: 7.5 (JCR, 2022) [Q1, 10/110 CS, Artificial Intelligence]

- **Overall contribution**: This article proposes a methodology for improving the performance imblalance classification tasks. The authors address the challenges associated with imbalanced multilabel datasets by applying ensemble techniques of models and using a specific loss function for imbalanced data, as presented in Chpater 3. The ensemble technique involves training multiple pretrained models and combining their predictions to obtain a final result. The authors also introduce a heatmap-based visualisation technique to highlight the most important areas for detecting each disease represented in the dataset. This technique generates a report that includes the visualisation of the heatmap, the probability produced by the system, and the agreement between the ensemble models, which is described in Chapter 5.

- **Contribution of the PhD candidate**:
  - First author of the article.
  - Contribution to the conception of the presented idea.
  - Implementation of the experiments.
  - Co-author of the interpretation and discussion of the results provided.
  - Elaboration of the manuscript and visualisations.

Future Generation Computer Systems 144 (2023) 291–306

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Deep learning for understanding multilabel imbalanced Chest X-ray datasets

Helena Liz [a,b,*], Javier Huertas-Tato [a], Manuel Sánchez-Montañés [c], Javier Del Ser [d,e], David Camacho [a]

[a] Computer Systems Engineering Department, Universidad Politécnica de Madrid, Alan Turing s/n, Madrid, 28031, Spain
[b] Department of Computer Sciences, Universidad Rey Juan Carlos, Tulipán s/n, Móstoles, 28933, Spain
[c] Computer Science Department, Universidad Autónoma de Madrid, Madrid, 28049, Spain
[d] TECNALIA Basque Research & Technology Alliance (BRTA), P. Tecnologico 700, Derio, Bizkaia, 48160, Spain
[e] University of the Basque Country (UPV/EHU), Bilbao, 48013, Spain

ARTICLE INFO

Article history:
Received 29 July 2022
Received in revised form 28 December 2022
Accepted 4 March 2023
Available online 6 March 2023

Keywords:
Convolutional neural networks
Chest X-rays
Explainable AI
Ensemble Methodology

ABSTRACT

Over the last few years, convolutional neural networks (CNNs) have dominated the field of computer vision thanks to their ability to extract features and their outstanding performance in classification problems, for example in the automatic analysis of X-rays. Unfortunately, these neural networks are considered black-box algorithms, i.e. it is impossible to understand how the algorithm has achieved the final result. To apply these algorithms in different fields and test how the methodology works, we need to use eXplainable AI techniques. Most of the work in the medical field focuses on binary or multiclass classification problems. However, in many real-life situations, such as chest X-rays, radiological signs of different diseases can appear at the same time. This gives rise to what is known as "multilabel classification problems". A disadvantage of these tasks is class imbalance, i.e. different labels do not have the same number of samples. The main contribution of this paper is a Deep Learning methodology for imbalanced, multilabel chest X-ray datasets. It establishes a baseline for the currently underutilised PadChest dataset and a new eXplainable AI technique based on heatmaps. This technique also includes probabilities and inter-model matching. The results of our system are promising, especially considering the number of labels used. Furthermore, the heatmaps match the expected areas, i.e. they mark the areas that an expert would use to make a decision.

1. Introduction

In recent years, the field of medicine has faced two relevant problems that hinder patient care: staff workload and subjectivity in the interpretation of tests [1,2]. These problems have no easy solution, which is especially dangerous in medicine because procedural errors can lead to serious health complications. Firstly, overwork in medicine, aggravated in recent times by the global COVID-19 pandemic, can lead to errors and delays in diagnosis and treatment. As mentioned above, there is also subjectivity in the interpretation of some medical tests. The expert analysing these tests, for example X-rays, may arrive at an erroneous diagnosis due to, for example, the existence of signs of different diseases to different degrees [3]. This type of imaging test is one of the most common in various diagnoses due to its low cost, speed of acquisition and the fact that it does not require much preparation [4]. Chest X-rays are useful for detecting a variety of diseases of the chest related to different organs such as the heart, lungs or bones. The features of X-rays make them suitable for analysis with convolutional neural networks (CNN) [5]. The combination of AI algorithms and medical knowledge can improve the performance of medical staff [6] and could also reduce patient waiting times by speeding up the diagnostic process and reducing the workload of doctors.

CNNs have been a breakthrough in computer vision due to their ability to extract features from images. These architectures are composed of different layers. The first has convolutional layers that are inspired by the notion of cells in visual neuroscience. The architectures are based on the visual cortex of animals. The main reason why these architectures have stood out is their great capacity to extract patterns from data, improving the performance of previous systems based on Machine Learning models. This advantage has made them a benchmark in Deep Learning due to their high performance in a wide range of tasks, such as speech recognition, computer vision or text analysis [7].

* Corresponding author at: Computer Systems Engineering Department, Universidad Politécnica de Madrid, Alan Turing s/n, Madrid, 28031, Spain.
E-mail addresses: helena.liz@urjc.es (H. Liz), javier.huertas.tato@upm.es (J. Huertas-Tato), manuel.smontanes@uam.es (M. Sánchez-Montañés), javier.delser@tecnalia.com (J. Del Ser), david.camacho@upm.es (D. Camacho).

https://doi.org/10.1016/j.future.2023.03.005

The properties of chest X-rays make them susceptible to be analysed by this type of algorithms. Some of the main advantages of CNNs over traditional techniques are that it is not necessary to manually extract image features or perform segmentation, and that by being able to learn from large volumes of data they can identify patterns that are difficult for the human eye to detect. Although in this article we focus on classification problems, other problems can be solved, such as X-ray segmentation [8], localisation, regression (such as predicting drug dosage), among others. CNNs are a potential tool for the analysis of chest radiographs. However, most of the work in this field focuses on binary and multiclass classification problems. Actual problems are usually more complex than the above; they tend to be multilabel classification problems, i.e. the different labels are not mutually exclusive, whereas in binary and multiclass classification problems there is only one label per radiograph [9]. To solve multilabel problems, we need to explore new strategies. Adapting algorithms can interpret this kind of problem by transforming them into simpler problems that can be solved by traditional algorithms, i.e., transforming them into binary problems [10]. In the field of chest X-rays we can find samples without labels, healthy patients and samples with radiological signs of several diseases at the same time. On the other hand, there are a large number of different radiological signs in chest X-rays, so if we want to build and validate a system that approximates realistic conditions, we have to use a dataset with a large number of mutually non-exclusive labels. This is the case of the PadChest database [11], which has 174 different radiological signs, substantially increasing the degree of realism and the complexity of the problem.

Many machine learning algorithms, including CNNs, work best when the classes in the dataset are balanced. However, in real life it is common to find datasets where this condition is not met; they are imbalanced datasets, where one or more classes have substantially more examples than the rest. As a consequence, with such datasets, machine learning algorithms learn a bias towards the majority class, even though the minority class is often more relevant. Therefore, it is necessary to apply different methods to improve the recognition rate [12]. There are several options to overcome this difficulty: (a) modify the dataset, reducing the samples from the majority class or increasing the number of samples from the minority class; (b) modify the algorithms to alleviate their bias towards the majority class, e.g. weighted learners [13]. The problem of unbalanced databases is exacerbated in multilabel classification problems, where multiple minority classes may appear, making this challenge more difficult to solve. In medicine, it is widespread because each disease has a different incidence in the population. Heart disorders top the list of the deadliest diseases, followed by chronic obstructive pulmonary disease, which causes more than 6 million deaths a year. In contrast, other diseases such as lung cancer are the sixth leading cause of death with less than 2 million deaths, according to the World Health Organization.[1] As a result, most radiographic datasets are imbalanced; a clear example is PadChest, the dataset used in this article, where the number of samples in each class approximates the incidence published by the World Health Organization.

These algorithms, like many other Deep Learning and Machine Learning methods, are considered "black box" algorithms because end users can only analyse the input and output, but the inference process is opaque, which reduces confidence in these algorithms. To alleviate this problem, explainable AI techniques have been developed, such as saliency maps, which produce heatmaps that highlight the pixels with the greatest influence on the final prediction [14]. This problem is serious in medicine, where errors can be dangerous for patients [15]. For this reason, explainable AI techniques are essential, as they allow users to understand how the system has arrived at the final result and use it to help diagnose [16]. However, the combination of medical knowledge and AI has many advantages, such as helping to reduce medical errors and speeding up diagnostic processes, leading to improved patient care, as doctors would have more time to attend patients.

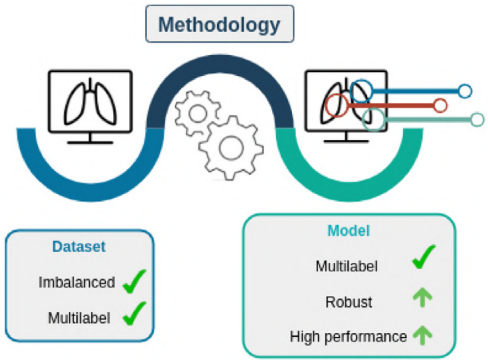


Fig. 1. Visual representation of the problem and the objective of the methodology.

The contribution of this manuscript is a methodology, see Fig. 1, for classifying imbalanced multilabel datasets with many classes The aim of this methodology is to generate robust and quality models; in this case, it has been applied to a highly imbalanced multilabel chest X-ray dataset with 174 classes. We selected this dataset for two reasons: (i) the number of classes, which is higher than in other state-of-the-art datasets; and (ii) the high imbalance between these classes. This methodology will allow to establish a suitable benchmark for this dataset against which future works can be compared, as there are currently very few published contributions using this dataset and they do not provide a detailed analysis of the problem.

We can summarise the main contributions of this work as follows:

- A methodology for imbalanced multilabel classification problems.
- A discussion about the experimental results obtained using a dataset with a large number of classes (more than 30) and a severe imbalance between them.
- An explainability interface using Grad-CAM for multilabel datasets.
- A suitable benchmark for this dataset serving as a reference against which to compare future proposals from the scientific community.

Finally, this manuscript is organised as follows. Section 2 summarises the most relevant work in the literature, with a special focus on chest X-ray classification problems for imbalanced multilabel datasets; Section 3 describes the methodology proposed for this type of problem, consisting of training a model and generating a visualisation based on heatmaps; Section 4 presents the CNN architectures, the hyperparameters used for training, details of the execution environment, and a link to the repository where the code used in the experimentation can be found; Section 5 presents the experimental results, and Section 6 presents the main conclusions and possible lines of future work.

2. Related work

Since the first application of AI techniques in medicine in the 1980s, the use of these algorithms has grown exponentially,

[1] https://www.who.int/es/news-room/fact-sheets/detail/the-top-10-causes-of-death

especially in recent years. Deep learning algorithms are applied to all kinds of clinical data [17]: biosignals, which include electrical [18,19], mechanical [20,21] and thermal signals [22,23]; biomedicine, which studies molecules of biological processes [24–26]; electronic health records (EHR), focused on optimising diagnosis [27–30]; and clinical imaging, widely used in the diagnosis of many diseases [31–34], as is the case with our problem. The practice of healthcare has evolved from observation-based medicine to evidence-based medicine. This makes deep learning and big data algorithms especially useful in this field as they can identify some radiological signs that medical staff cannot detect [35]. Although in this manuscript we focus on classification problems, there are papers where these algorithms are used in regression problems, such as estimating the dose of a drug [36]; generating medical reports from clinical tests [37]; support healthcare management [38]; or image processing, such as image segmentation [39] and image reconstruction [40].

The COVID-19 pandemic has had a strong impact on research into the application of machine learning and deep learning in medical image analysis. As expected, many of the classification systems investigated have focused on detecting signs of bilateral COVID-19-associated pneumonia. In Ahmed et al. [4] they use two different pre-trained architectures to classify chest X-rays, VGG16 and ResNet, and optimise the hyperparameters. In Pham [41] they train three different pre-trained architectures, AlexNet, GoogleNet and SqueezNet, with six datasets independently, testing different percentages of train set samples (50 and 80%), achieving an accuracy of 99.85% with SqueezeNet. However Ahmad et al. [42] develops an ensemble system based on MobileNet and InceptionV3 that achieves 96.49% accuracy. Soon, binary classification was extended to multiclass problems, making it possible to discern whether pneumonia is caused by COVID-19 or another virus/bacteria or whether the patient is healthy. As with binary classification problems, many works, such as Avola et al. [43], use state-of-the-art architectures to find the best performing ones, such as AlexNet, GoogleNet, ResNet and ShuffleNet, among others. MobilNet_v3 achieves the best result with a precision of 84.92% on a dataset composed of 6330 samples. In Zebin and Rezvy [44], in addition to training a pre-trained state-of-the-art architecture, a heatmap-based visualisation is generated that shows two images for each sample. The first is the original X-ray, and the second is the class activation map, i.e. the most important area for the CNN. However, the images do not overlap, making interpretation difficult. Other works, such as Teixeira et al. [45], apply segmentation techniques to remove all irrelevant areas of the system, which should improve performance and visualisation. Their dataset consists of three different classes: COVID-19, normal and lung opacity.

As we have discussed in Section 1, the explainability of deep learning models is a fundamental factor to be taken into account in their application. These models are black-box algorithms and need explainable AI techniques to make them more trustworthy [46]. There are two main ways to produce the final visualisation, (1) generate a heatmap per label, or (2) generate a single visualisation for all classes. The first one is more commonly used, [43,45,47] however, this technique has one main limitation: it is not feasible for a large number of labels, and it makes a global view difficult. The second one (e.g. Teixeira et al. [48]), shows the different signs as areas with higher colour intensity, but only one colour scale was used, which makes it difficult to identify which pathological sign indicates which area of interest. We propose a new technique where each visualisation shows a radiological sign including the probability and agreement between models.

Although most medical datasets have two classes (samples of a particular pathology and healthy samples), in chest X-rays it is common to find signs of more than one pathology. For this reason, in the last five years different authors have published multilabel radiological datasets. These datasets are closer to real situations than binary ones, with the additional challenge of imbalance of different classes. The size of each class in a realistic dataset should depend on the incidence of pathology in society, i.e. some classes are more represented than others. These characteristics of these datasets are interesting and need to be analysed in detail in order to address the problem adequately.

Table 1
Summary table of multilabel datasets in the field of chest radiography.

	# samples	# patients	Labels	Views	Reference
ChestX-ray 14	112120	32717	14	frontal	[54]
CheXpert	224316	65240	14	frontal/lateral	[55]
PadChest	160000	67000	174	frontal/lateral	[11]

2.1. Multilabel classification problems

As we have seen, much of the work generated in recent years has focused on binary and multiclass classification problems. In these problems, the labels are mutually exclusive, while multilabel classification problems have multiple classes that are not mutually exclusive, which increases the difficulty of the problem. There are two ways to solve these problems: (a) transform the multilabel problem into simple binary problems, or (b) adapt the algorithms to solve the multilabel problem directly, i.e. attack the problem globally [49].

Binary and multiclass classification systems are very restrictive, as they only serve to detect one type of radiological finding. However, patients can often present signs of multiple diseases at the same time. There are very few multilabel datasets that take into account a large number of signs, as they require a large number of samples and most of them have only a few labels. Three datasets are worth highlighting for their quality and relevance to the state of the art (see Table 1). The first two have been used extensively in image classification problems, but the third has been used mainly in medical report generation [37,50–53]. However, this third dataset has two advantages that make it very suitable also for classification problems: (1) the number of labels is larger; (2) it has the largest number of different patients, which implies a smaller number of similar samples from the same patient. Given the lack of application of algorithms for this task to increase the potential and interest of this dataset mentioned above, it was selected as a case study for this article.

ChestX-ray 14 dataset is one of the most widely used datasets in the field of chest X-ray classification since its publication in 2019 [54]. For example, Wang et al. [56] uses DenseNet-121 optimising its hyperparameters, obtaining an average AUC of 0.82. The AUC achieved for the pneumonia class was 0.662 (the lowest), while for the hernia class it was 0.923 (the highest). Other researchers use different architectures such as Inception-ResNet_v2 and ResNet152_v2 to achieve an AUC for pneumonia of 0.73 [57]. Much of the work on this dataset retrains state-of-the-art architectures, but there are other strategies for improving classification performance; for example, Almezhghwi et al. [58] switches the classifier from AlexNet and VGG16 to SVM with the intention of improving the results of previous manuscripts, achieving an AUC for pneumonia of 0.98 with both architectures. Having different types of radiographs of the patient can also improve the classification results, for example a frontal and a lateral X-ray. Finally, the main disadvantage of ChestX-ray 14 dataset is that it only contains radiographs with a frontal view, while CheXpert and PadChest datasets also contain X-rays with a lateral view.

The second dataset, **CheXpert** has 14 different labels and reports on all images. In terms of published classification work, we find a situation similar to ChestX-ray 14, with many works retraining state-of-the-art architectures, such as Seyyed-Kalantari et al. [59], where they adjust the hyperparameters of DenseNet-121 to optimise its performance. Other authors look for different strategies, such as Cohen et al. [60], where they make two modifications to DenseNet to improve its performance. First, they modify the loss function by assigning weights to the different labels, alleviating the imbalance problem. Second, they modify the threshold for discerning between the presence or not of each label, i.e. the probability at which the class is considered present. However, the CheXpert dataset has the same limitation as ChestX-ray 14: they contain only 14 possible diseases, which represents only a small subset of all possible diseases that may be present in the chest.

Finally, **PadChest** is the most interesting dataset of the three in our opinion because it has many more labels than the others. It is a massive multilabel classification problem, much closer to reality than the other datasets. The number of patients used is also larger than in the others, leading to more variability in the dataset, and the imbalance of the classes is larger too. One of the papers using this dataset, [61], combines the PA and lateral views to predict labels in four different ways: (a) the lateral view is stacked in the second channel of PA X-ray; (b) both views are processed by two CNNs and the combination of them is processed by a fully connected layer; (c) the model input is processed through two separate CNNs, the output is concatenated and passed by two dense layers with an average pooling layer between them; (d) a modification of (c) where two dense layers are added. A major limitation of that paper is that it shows overall results without performing a detailed analysis per label, which prevents comparison with other works in the area. On the other hand, in Pooch et al. [62] CheXNet is retrained, which is a state-of-the-art architecture previously trained with a multilabel chest X-ray dataset. In that paper, different models are trained with four datasets, and each model is tested with each dataset separately. The main limitation of that manuscript is the reorganisation of the labels of PadChest dataset: the label "Lesion" is generated to unify the samples of the atelectasis classes, using only 8 classes out of the 174 available. Given the limitations we have found in all classification works using the PadChest dataset and that some most of them are not replicable, we propose to create a benchmark that future works can use to compare results, with a methodology adapted for the two main problems: the high number of different labels, and the imbalance between them. As we have explained in this section, the PadChest dataset has several advantages over other multilabel datasets: (i) it has the most labels, which makes it closer to real-world scenarios; (ii) the number and diversity of patients is greater; and (iii) it contains lateral and frontal radiographs. We propose two ways of organising the dataset based on the term tree provided by its authors, which allows us to group radiological signs into higher classes. The first one uses the specific labels for a finer-grained classification. The second one works with more general labels, which indicate more general radiological signs.

2.2. Class imbalance in deep learning

As explained above, most machine learning algorithms work best when the number of samples for each class is similar. When there is a significant difference between the classes, the system will boost the majority class while the minority class(es) will have less relevance, even though the minority class is often the most relevant. There are several classification tasks with this problem, such as Cohen et al. [63], where the majority class is COVID-19 over the rest of the pneumonia classes. As expected, because the incidence of COVID-19 has been extremely high, the dataset contains more than 400 samples of COVID-19 followed by the class *Pneumocystis spp* with fewer than 30 samples. This phenomenon appears in many classification tasks, especially those with more than two classes, both multiclass and multilabel. For example, in Wang et al. [54] there are 15 classes and the class "No findings"/"Normal" exceeds 50,000 samples, while the other labels have less than 20,000 samples, of which only three exceed 10,000 samples.

As mentioned in the Introduction section, there are different strategies to alleviate the class imbalance problem. *Modify the dataset*, for example with oversampling techniques, which increase the number of samples from minority classes by applying data augmentation and histogram equalisation techniques [64]. Charte et al. [65] develops a new algorithm, Multilabel Synthetic Instance Generation, for multilabel problems. For each sample, a nearest-neighbour search is performed, the features are extrapolated and the label is generated from them. Another option for generating synthetic samples is to use generative adversarial networks (GANs), i.e. to use deep learning models to produce new samples from the original dataset. Salehinejad et al. [66] uses this method to generate new chest X-rays to balance the different classes. Another strategy for balancing the classes in the dataset is to reduce the samples of the majority of classes. This technique is called undersampling. Typically, random samples are removed from the majority classes, as in Qu et al. [67], where the maximum number of samples in each class is set to balance it. Undersampling is not as widespread as oversampling because Deep Learning systems need a large number of samples, so undersampling may not work.

Another strategy to alleviate class imbalance is to *modify the way the model learns* by increasing the weight of minority classes in learning, thus preventing the model from giving more importance to majority classes. One option is to apply class weights in the loss function that increase the relevance of the minority classes. One example is Rajpurkar et al. [68], which uses the chest X-ray14 dataset to classify the presence or absence of pneumonia. Another example is Monowar et al. [69], where the weighted binary cross-entropy loss function is applied. Ge et al. [70] developed a novel error function, Multilabel Softmax Loss, this method considers the relationship of multiple labels explicitly, the author computes the derivative of the error with respect to each class using the chain rule. In addition they applied it to a system composed of two CNNs combined by a bilinear pooling layer. Teixeira et al. [48] proposes a dual lesion attention network composed of two models, DenseNet-169 and ResNet-152, as feature extractors, after an attention module and average max pooling. The outputs are combined to generate three classifiers. Finally, all classifiers are merged to obtain the final prediction. In addition, they used a variant of the weighted binary cross-entropy loss. To tackle the class imbalance, we propose using weighted cross-entropy with logits using class weights.

2.3. The challenge of imbalance in multilabel classification problems

As we have explained, many real classification problems have two properties that make them difficult to solve: multilabeling and imbalance. Each of these two properties alone makes classification difficult, so together they can be very challenging. In medicine, multilabel and imbalance problems are common because medical staff can find different radiological signs on a chest X-ray, and different diseases do not have the same incidence in the population. All the datasets mentioned in Section 2.1 have both features; however, ChestX-ray 14 and CheXpert have a low number of classes, 14 labels, compared to PadChest [11], which is composed of 174 different labels with a large imbalance: the label

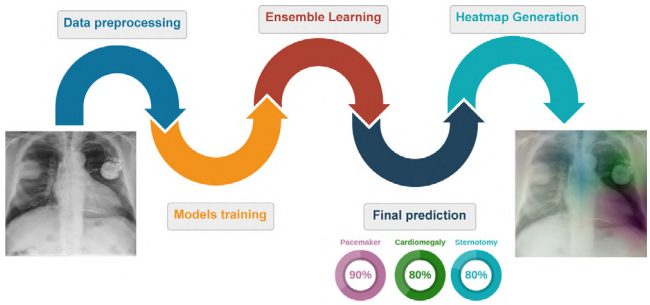


Fig. 2. Visual representation of the proposed ensemble system. We train each architecture with preprocessed images, and their outputs are combined to generate the ensemble output. Finally, the system produces the global prediction and heatmap visualisation.

"Normal" has more than 35000 samples, while other labels, such as round atelectasis, pleural mass or nephrostomy tube, have less than ten samples.

Most of the published work using these datasets modifies the architecture so that it can directly solve multilabel problems, but does not consider or apply any specific technique to solve the imbalance problem. However, other works explore different ways to overcome these difficulties and achieve better results. Such as Huang and Fu [71], which proposes a multi-attention convolutional neural network to reduce the performance difference between classes and, more interestingly, to extract discriminative features to classify similar classes, which is very common in this kind of dataset. Wang et al. [72] generates three images: the first one is the original chest X-ray, the second one is a segmentation-based cropping, where areas not interesting for the model are removed, and the last one is a cropping of the area where previous models have found pathological signs. The information extracted from the three images is fused and finally processed to obtain the final result. Another interesting strategy is the modification of the loss function to focus on the most interesting samples; for example, Qin et al. [73] proposes a loss function called "weight focal loss", which forces the model to pay more attention to the most difficult samples. This makes the model pay more attention to minority classes, avoiding false negatives.

These methods can help in class imbalance problems, but in extreme cases of multilabel and imbalance, such as the PadChest dataset, they may not be sufficient. Most of the published papers attempt to improve the performance of the architecture or solve these problems using a single strategy, which may not be sufficient for datasets such as PadChest.

In contrast to other works in the related literature, we have decided to address these problems by combining different strategies: (1) to avoid confounding the model with areas that do not present interesting radiological signs, we have applied segmentation-based cropping; (2) to make the system robust against the individual errors of the different architectures, we have created an ensemble whose hyperparameters have been adjusted in a validation split to obtain the best possible results; (3) we have applied a specific loss function for imbalanced data that weights each class by its inverse frequency. The combination of these techniques will allow us to substantially reduce the errors due to imbalance and the high number of labels. In addition, we have created a heatmap-based visualisation that highlights the most important areas for detecting each disease represented in the dataset, the estimated probability of that pathology, and the agreement between models (how many models have a probability higher than 50% for that disease), which facilitates interpretation and shows the degree of confidence in the result.

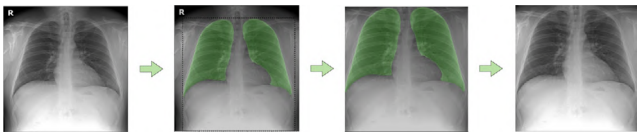


Fig. 3. A segmentation-based cropped sample. The first image corresponds to the original X-ray. The second shows the lung segmentation mask. The third one show the cropped image with lung mask, and finally the last image shows the input of our system, the preprocessing result.

3. Methodology

We can summarise the proposed methodology in Fig. 2, which has four sections. The first is the data pre-processing step, where we prepare the images for the model and apply data augmentation to alleviate class imbalance. In the second stage we build the model, training different state-of-the-art architectures. We then combine the results of each model to obtain the final probabilities. Finally, we developed a multilabel heatmap technique to areas of the image that are relevant in the classification. In this technique, the original X-ray is combined with one or more regions labelled with different colours to facilitate the application of these techniques in health centres or hospitals.

3.1. Label selection

As explained above, multilabel datasets are often imbalanced as they have classes with a low number of samples. For this reason, we must establish a criterion for choosing the labels to include in our classification system, especially in datasets where the number of classes is extremely high, as in our case. First, we set the minimum number of samples a label must have to be included in the classification problem, and we set the threshold at 200 X-rays. For a dataset of 90000 samples this is 0.22% of the total. The model cannot work correctly for under-represented labels as it is not a few-shot system. If a sample has only deleted minority labels we will remove it.

In this paper we consider two different experiments. First, we use the classes proposed by the authors of the dataset that correspond to the specific labels; this classification system has a smaller number of samples and labels due to the cleaning of under-represented labels explained in the previous paragraph, but is a more fine-grained classification system. In the second case, we use more general labels. We create these classes grouping the specific labels according to their characteristics. The number of samples and classes is larger at the cost of being less precise systems, but it allows us to cover a larger number of different classes.

3.2. Preprocessing

The raw images were preprocessed in order to train the model efficiently. First, we reduced the number of channels to one because although the original files are RGB images (three colour channels), the X-rays are grayscale images, so all three channels contain the same information. Next, we normalised their size to 512×512 pixels. The pixel values were then normalised between 0 and 1, Fig. 3 (first image).

Chest X-rays show an area larger than the area of interest (ROI). Areas such as arms or neck, among others, are irrelevant to the problem we want to solve, so a cropping based on segmentation masks was performed, forcing the system to focus on the relevant areas. This trimming is performed in three different steps: first, we generated the lung masks using a segmentation model based on the U-Net architecture [74], Fig. 3 (second image).

We also added the area underneath the lungs to the masks as it may contain radiological signs of interest. On many occasions, the segmentation models are not perfect; they generate more than two masks, leave gaps inside the masks, etc. Therefore, thirdly, we decided to use a mask post-processing system [75]. This system fills the possible gaps in the masks by applying the flood fill algorithm, which analyses the pixels neighbouring the one of interest and depending on whether or not they belong to the mask, it will decide to fill the gap or not. Then, if more than two masks have been generated (one per lung), those whose area is less than a predetermined value are removed. In addition, in case the lung masks are stuck together, they are separated. Finally, the image is cropped using the mask coordinates and the lower boundary of the sample, Fig. 3 (third image). As the images can have different sizes, we normalised their size to 224 × 224 pixels, because this is the normalised size of the samples in the state-of-the-art models, Fig. 3 (last image).

3.3. Image classification with CNNs

Five state-of-the-art architectures pre-trained with ImageNet were selected for their relevance:

EfficientNetB0. [76]: This architecture uses different scaling coefficients to scale width, depth and resolution. In the EfficientNet family, this architecture is the smallest. It is based on the idea that if the images are larger, the network needs more layers to extract the relevant information.

DenseNet-201. [77]: Instead of adding more layers to the architecture, the number of connections between units is increased by connecting each unit to the last, unlike ResNet50, which only connects one unit to the next output. This architecture has several advantages: it alleviates the vanishing gradient problem, enforces feature propagation and feature reuse, and reduces the number of parameters.

InceptionV3. [78]: This architecture is different from the previous ones. It factorises convolutions into smaller convolutions (which can be asymmetric) to reduce cost. In addition, this architecture has an auxiliary classifier between layers that acts as a regulariser.

InceptionResNetV2. [79]: This architecture combines ResNet and InceptionV3. It consists of several Inception units with shortcut connections between them; this enhances the capability of the architecture.

Xception. [80]: It consists of depth-wise separable convolutions involving two steps: depth-wise convolution, which differs from the standard convolution in that it only acts on one channel; and point-wise convolution, where a 1 × 1 convolution is applied to all channels. This architecture also includes shortcut connections, such as ResNet50.

We applied Transfer Learning on the above five architectures and retrained them with PadChest dataset, replacing the classifier in all cases with two dense layers. We froze the first 10% of the convolutional layers, as they detect basic patterns and do not need to be retrained. The remaining convolutional layers are retrained to learn patterns specific to our problem. The main relevant training parameters are summarised in Table 2. In addition, a checkpoint is used to save the best model using the validation loss. Finally, an early stopping algorithm was used to finish training when the validation loss did not improve over the last 25 epochs by more than a threshold of 0.001.

Table 2
Summary of the hyperparameters used in training: optimisation, data augmentation and training methodology.

Optimisation	
Optimiser	Adam
Learning rate	1e−4
Loss	weighted crossentropy with logits
Feed-forward classifier	
# Neurons	512
Activation	ReLu
Dropout	0.2
Data Augmentation	
Shear range	0.1
Zoom range	0.1
Rotation range	45
Width shift range	0.1
Height shift range	0.1
Horizontal flip	True
Fill mode	nearest
Brightness range	0.7–1.1
Channel shift range	0.05
Training methodology	
Maximum epochs	350
Early stopping patience	25
Early stopping threshold	0.001
Batch size	32
Image size	224 × 224

3.4. Ensemble technique

Ensemble learning is an effective way to improve the performance and robustness of deep learning algorithms. We combined the results of all trained models, obtaining a system composed of five different architectures with the same test set. We distinguish two approaches [81]: "Combine then predict" (CTP) and "Predict then combine" (PTC). In the CTP method, the label probabilities predicted by the individual models are first calculated, and then the average probability at each label is used to obtain the ensemble label prediction. The other method, PTC, combines the binary predictions to obtain the ensemble. We consider two versions of PTC: label-wise voting (PTC-lw), which calculates the number of positive and negative individual predictions for each label, adopting the majority. Thus, PTC-lw calculates the prediction of each label independently of the others. On the other hand, PTC-mode calculates the set of labels predicted by each individual model, and predicts the most frequent set.

3.5. Heatmap generation

As explained in Section 2, it is necessary to include XAI techniques for the medical staff to understand the output given by our system. For this reason, we developed a visualisation technique using heatmaps. A heatmap is a matrix of the same size as the input image. The value of each pixel is proportional to its importance for the classification of the model. A colour scale is used in the heatmap to highlight the most relevant pixels for the model.

The first step in generating the heatmaps is to change the activation function of the last layer (the classifier layer) from softmax to linear. Then, for each classifier neuron, we compute the weighted average of the last convolutional layer. Each channel is weighted by the gradient of the classifier neuron with respect to that channel. This is the so-called grad-CAM algorithm [82], which allows to compute a heatmap for each class.

As explained before, the ensemble consists of five models. We generate ensemble heatmaps by averaging the individual heatmaps generated by those models. Finally, we generate a of the average heatmap of each classifier neuron, which is overlaid

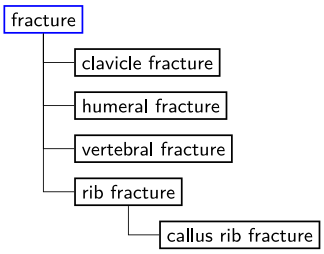


Fig. 4. Example of a section of the term tree of the dataset. The general label is boxed in blue, and the specific labels are marked in black.

Table 3
Summary table of the two types of experiments performed (general labels, and specific labels). The total number of labels and the total number of samples in each of the splits are shown.

	# classes	# samples	Train size	Val. size	Test size
General labels	54	90687	63475	9069	18143
Specific labels	35	85367	59753	8532	17082

on the original X-ray using a 10% of transparency to improve the information for the medical staff. We include in the title the estimated probability for this class and the inter-model agreement showing the confidence of the ensemble in that prediction, which facilitates the use of the system by medical staff.

4. Experimental setup

4.1. Dataset

In this article, we have used the PadChest dataset [11], an imbalanced and multilabel dataset. It was published in January 2019 by the University of Valencia together with BIMCV. The samples were collected at Hospital de San Juan (Spain) between 2009 and 2017. This dataset is composed of 160,868 clinical images from 67,625 patients, divided into 174 different labels, and corresponds to different signs of thoracic disease. This dataset contains chest X-rays with different projections: posteroanterior (PA), anteroposterior (AP) and lateral views; however, only PA X-rays were used for experimentation, corresponding to 91,728 clinical images from the original dataset. The authors of the dataset provided a term tree[2] in which all labels are grouped into more general labels, as can be seen in Fig. 4. In this example, the general label is fracture. The specific labels are clavicle fracture, humeral fracture, vertebral fracture, and rib and callus rib fractures. Therefore, we designed two experiments, the first using specific labels for classification and the second using more general labels, each grouping one or more specific labels. We then set the minimum number of samples that each class must have to be included in the classification system. The more general classification system has a larger number of classes that are more heterogeneous, while the more specific classification system has a smaller number of classes, but is more precise than the previous one.

Table 3 shows the details of the two classification systems, the number of samples, the classes and the size of the training, validation and test sets. In the train/test/validation split we stratify the samples according to classes and patient id, which avoids biases and problems between subsets. In addition, to facilitate the replicability and transparency of this article we will make the split available on the github in Section 4.2.

2 https://github.com/auriml/Rx-thorax-automatic-captioning

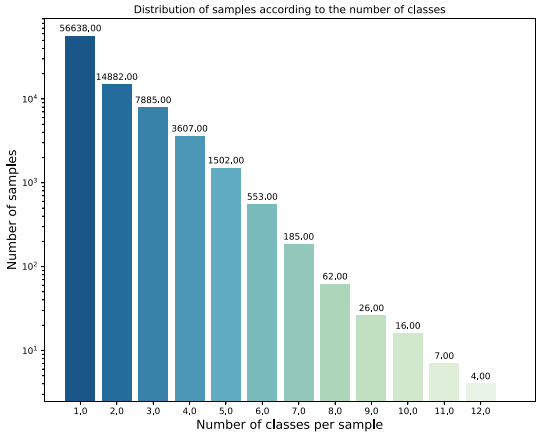


Fig. 5. Distribution of the number of labels per sample (specific labels experiment).

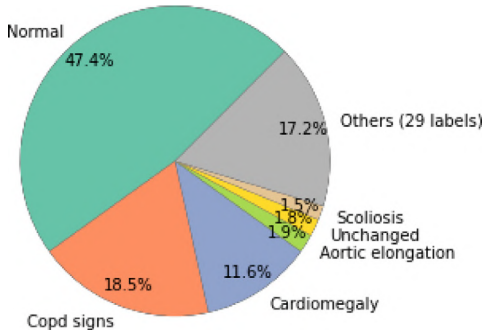


Fig. 6. Label distribution (specific labels experiment).

Label distribution: Specific labels. This experiment, as explained in Section 3.1, label selection, has a smaller number of samples and classes than the second case, but the radiological signs are more accurate. In this experiment we used a total of 85367 samples and 35 different classes. We can observe in Fig. 5, how more than half of the samples present a single class; however, we can observe that there are samples with a high number of classes, four of them presenting 12 different labels at the same time. This distribution of the samples is in line with expectations; the number of samples decreases as the number of labels per sample increases. In Fig. 6 we can see how the classes in this experiment are extremely imbalanced. Although there are 35 classes, the six majority classes account for 82.7% of the dataset. Only the normal class, which is the majority class, accounts for 47.4% of the total samples, while the supra aortic elongation class, which is the least represented class, accounts for only 0.28% of the total.

Label distribution: General labels. In this experiment, different classes were unified according to the tree of terms proposed by the authors. Therefore, the number of classes and samples is higher than in the first experiment. However, the radiological signs used in the classification are less precise, so in the end 54 classes and 90,687 samples were used. In Fig. 7 we can see how the number of classes per sample is distributed in a very similar way to the previous case. However, we can see that there are samples with 13 different labels, one more than in the previous case. If we look at Fig. 8, we can see that the six majority classes represent 51.3% of the total, while the other 48 classes do not reach 50%. The majority class, as in the previous case, is the normal class. This class accounts for 22.6% of the total while the minority class, vascular redistribution, accounts for only 0.13% of

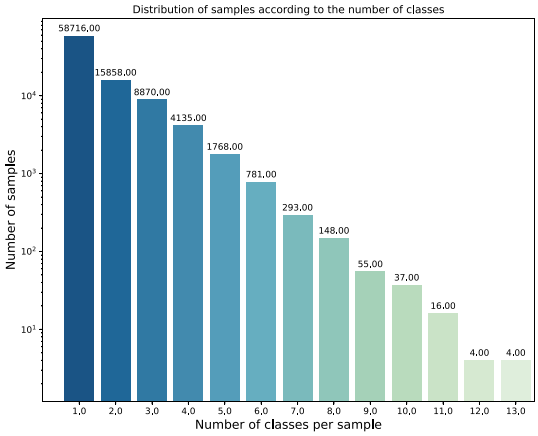


Fig. 7. Distribution of the number of labels per sample (general labels experiment).

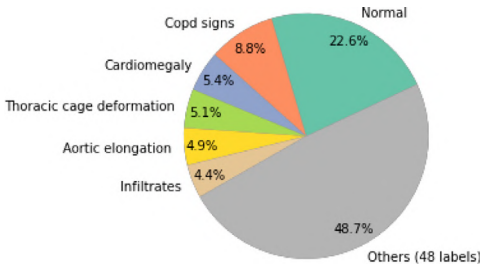


Fig. 8. Label distribution (general labels experiment).

the dataset. This shows that even if we group the radiological signs into higher classes, the dataset is very imbalanced.

4.2. Execution environment and Github repository

All experiments have been run on a 24 GB Nvidia GeForce RTX 3090. The main packages used in these experiments are the following: Tensorflow [83], Scikit-Learn [84] and openCV [85]. The code developed in our work is publicly available at GitHub.[3]

5. Experimental results

This section describes the results obtained with the proposed methodology and evaluates its performance on a multilabel and imbalanced problem, the PadChest dataset. We considered two strategies for the classes: directly using the labels proposed by the dataset creators, or grouping them into more generic classes that encompass similar radiological signs. First, we checked whether preprocessing improves the ensemble performance. Next, we checked the performance of both the individual models and the ensemble, and analyse the quality of the visualisations based on explainable AI techniques. To measure the performance of the different models, we have used three metrics suitable for multi-label problems: Area Under the Curve (AUC), Hamming Loss and F-measure [86].

5.1. Impact of preprocessing techniques

First, we trained the models with the images without segmentation-based cropping or data augmentation. The results obtained, Tables 4 and 5, show that only two individual models have been able to learn, EfficientNet and DenseNet, while the rest of the models were not able to learn and presented a flat training curve with an AUC of 0.5. As expected, the ensemble does not work correctly, and therefore the preprocessing step is necessary.

Tables 6 and 7 show the results training with segmentation-based cropping but without applying data augmentation techniques. At first, it is interesting that Inception does not learn, possibly because it is not able to generalise correctly without data augmentation techniques. InceptionResNet has the best results in most classes, but EfficientNet achieves the best overall result, achieving an AUC of 0.792 while InceptionResNet scores 0.779. Comparing Table 8 with these results shows that the application of data augmentation techniques improves the system performance. If we focus on the results for the different ensembles, we can see that for all labels, the CTP technique performs better than the two PTC methods. CTP also performs better than the individual models except in three cases: in one case it equals them, and in two cases it performs worse). We can conclude that data augmentation improves the performance of the system.

5.2. Performance analysis of CNN models

The first step is the comparison of the different architectures explained in Section 3.3. They are used as a baseline to compare the ensemble system. As explained in Section 3, we consider two types of classification problems: the first uses the original labels proposed by the authors of the dataset (“specific labels”), and the second uses general labels constructed by grouping specific labels. In the first problem, a finer-grained classification is performed, but it contains a small number of labels, 35, as many of the original 144 do not pass the filter of the minimum number of samples (200). In the second problem, general radiological patterns are classified, but there is a larger number of labels, 54, because when grouping labels there are a larger number of classes satisfying the minimum threshold of 200 samples.

Tables 8 and 9 show the results obtained by applying the proposed methodology for the first case study (classification using specific labels). The model with the best global AUC value is DenseNet, followed by EfficientNet, with 0.818 and 0.804 respectively. The other models (Inception, InceptionResNet and Xception) do not achieve an AUC $= 0.8$. These results are broken down by class. First of all, we can observe that the labels with fewer samples do not show worse results on average than the classes with more samples, which means that we have managed to overcome the data imbalance problems of. It can also be seen in the table that some models perform better with majority classes, such as Inception; others achieve the best results for minority classes, such as EfficientNet and Xception. However, DenseNet 201 and InceptionResNet perform well in both cases.

Secondly, we have analysed the results obtained with the ensemble techniques, using the individual models as baselines. Interestingly, only the CTP technique improves the individual models, as is also the case in Table 6. If we focus on this ensemble technique, we can see that there are two classes, Pleural effusion and pacemaker, where the results of the individual models are not improved. These two classes have 658 and 336 samples respectively, i.e. they are not majority classes, so one hypothesis would be that the ensemble performs worse in minority classes. However, the number of labels for which the ensemble does not outperform the individual models is very small compared to the total. Furthermore, the ensemble achieves an AUC above 0.85 for more than 40% of the labels, which is higher than expected. Since we can observe that the ensemble achieves an AUC higher than 0.9 for classes such as hemidiaphragm elevation, hiatal hernia, or sternotomy, all of them with less than 300 samples, we conclude that class imbalance does not affect our system significantly. Considering that the model is trained for 35 different classes, reaching

[3] https://github.com/helenalizlopez/multilabelimbalancedchestxraydataset

Table 4
Specific labels experiment: results obtained by training the models without segmentation-based cropping or data augmentation. For each label, the individual models with the best performance and the ensembles that outperform all individual models are marked in bold. The best ensemble result is marked in italics unless it ties the random classifier.

	# Samples	DenseNet		EfficientNet		Inception		InceptionResNet		Xception		PTC-mode		PTC-lw		CTP	
		AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1
Normal	34327	**0.589**	0.470	0.500	0.374	0.500	0.374	0.500	0.374	0.500	0.374	0.500	0.374	0.500	0.374	0.589	0.374
Copd signs	13419	0.500	0.457	0.500	0.457	0.500	0.457	0.500	0.457	0.500	0.457	0.500	0.457	0.500	0.457	0.500	0.457
Cardiomegaly	8412	**0.620**	0.551	0.611	0.563	0.500	0.475	0.500	0.475	0.500	0.475	0.500	0.475	0.500	0.475	**0.633**	0.475
Aorticelongation	1399	0.538	0.509	**0.553**	0.526	0.500	0.479	0.500	0.479	0.500	0.479	0.500	0.479	0.500	0.479	**0.558**	0.479
Unchanged	1311	**0.535**	0.483	0.526	0.504	0.500	0.480	0.500	0.480	0.500	0.480	0.500	0.480	0.500	0.480	**0.543**	0.480
Scoliosis	1073	0.500	0.484	**0.550**	0.522	0.500	0.484	0.500	0.484	0.500	0.484	0.500	0.484	0.500	0.484	**0.550**	0.484
Chronic changes	873	**0.581**	0.481	0.578	0.451	0.500	0.487	0.500	0.487	0.500	0.487	0.500	0.487	0.500	0.487	**0.585**	0.487
Costophrenic angle blunting	703	**0.556**	0.525	0.541	0.532	0.500	0.490	0.500	0.490	0.500	0.490	0.500	0.490	0.500	0.490	0.545	0.490
Air trapping	663	0.500	0.490	0.498	0.510	0.500	0.490	0.500	0.490	0.500	0.490	0.500	0.490	0.500	0.490	0.498	0.490
Pleural effusion	658	0.655	0.573	**0.656**	0.567	0.500	0.490	0.500	0.490	0.500	0.490	0.500	0.490	0.500	0.490	**0.676**	0.490
Pneumonia	651	0.626	0.556	**0.629**	0.566	0.500	0.490	0.500	0.490	0.500	0.490	0.500	0.490	0.500	0.490	**0.645**	0.490
Interstitial pattern	594	**0.597**	0.544	0.582	0.547	0.500	0.491	0.500	0.491	0.500	0.491	0.500	0.491	0.500	0.491	0.594	0.491
Infiltrates	591	**0.615**	0.540	0.594	0.542	0.500	0.491	0.500	0.491	0.500	0.491	0.500	0.491	0.500	0.491	0.612	0.491
Laminar atelectasis	578	0.500	0.491	**0.508**	0.491	0.500	0.491	0.500	0.491	0.500	0.491	0.500	0.491	0.500	0.491	0.508	0.491
Vertebral degenerative	575	0.500	0.491	**0.573**	0.485	0.500	0.491	0.500	0.491	0.500	0.491	0.500	0.491	0.500	0.491	**0.573**	0.491
Kyphosis	526	**0.602**	0.558	0.538	0.520	0.500	0.492	0.500	0.492	0.500	0.492	0.500	0.492	0.500	0.492	**0.606**	0.492
Apical pleural thickening	469	0.500	0.493	0.499	0.488	0.500	0.493	0.500	0.493	0.500	0.493	0.500	0.493	0.500	0.493	0.499	0.493
Vascular hilar enlargement	463	0.584	0.510	**0.587**	0.475	0.500	0.493	0.500	0.493	0.500	0.493	0.500	0.493	0.500	0.493	**0.602**	0.493
Fibrotic band	449	0.500	0.493	0.489	0.484	0.500	0.493	0.500	0.493	0.500	0.493	0.500	0.493	0.500	0.493	0.489	0.493
Nodule	449	0.500	0.493	0.500	0.493	0.500	0.493	0.500	0.493	0.500	0.493	0.500	0.493	0.500	0.493	0.500	0.493
Calcified granuloma	388	0.500	0.494	0.499	0.494	0.500	0.494	0.500	0.494	0.500	0.494	0.500	0.494	0.500	0.494	0.499	0.494
Callus rib fracture	360	0.500	0.495	0.500	0.495	0.500	0.495	0.500	0.495	0.500	0.495	0.500	0.495	0.500	0.495	0.500	0.495
Pacemaker	336	0.627	0.543	**0.646**	0.523	0.500	0.495	0.500	0.495	0.500	0.495	0.500	0.495	0.500	0.495	***0.663***	0.495
Aortic atheromatosis	318	0.500	0.495	**0.616**	0.457	0.500	0.495	0.500	0.495	0.500	0.495	0.500	0.495	0.500	0.495	*0.616*	0.495
Volume loss	294	0.500	0.496	**0.512**	0.496	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	*0.512*	0.496
Sternotomy	292	0.530	0.517	**0.539**	0.506	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	***0.545***	0.496
Bronchiectasis	290	0.500	0.496	0.480	0.496	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	0.480	0.496
Hiatal hernia	287	0.500	0.496	**0.533**	0.506	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	***0.533***	0.496
Pseudonodule	275	0.500	0.496	0.498	0.500	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	0.498	0.496
Hemidiaphragm elevation	254	0.515	0.496	**0.531**	0.496	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	***0.544***	0.496
Alveolar pattern	248	**0.664**	0.531	0.663	0.503	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	***0.695***	0.496
Increased density	239	0.528	0.513	**0.536**	0.502	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	0.500	0.496	***0.547***	0.496
Vertebral anterior compression	214	0.546	0.510	**0.548**	0.487	0.500	0.497	0.500	0.497	0.500	0.497	0.500	0.497	0.500	0.497	***0.559***	0.497
Suture material	210	0.500	0.497	**0.542**	0.509	0.500	0.497	0.500	0.497	0.500	0.497	0.500	0.497	0.500	0.497	*0.542*	0.497
Supra aortic elongation	200	0.500	0.497	**0.503**	0.497	0.500	0.497	0.500	0.497	0.500	0.497	0.500	0.497	0.500	0.497	***0.504***	0.497
Global		0.543	0.508	**0.547**	0.502	0.500	0.488	0.500	0.488	0.500	0.488	0.500	0.488	0.500	0.488	***0.558***	0.488

Table 5
Specific labels experiment: global results obtained by the individual models and the ensemble without using segmentation-based cropping or data augmentation techniques.

	DenseNet	EfficientNet	Inception	InceptionResNet	Xception	PTC-mode	PTC-lw	CTP
Hamming Loss	0.067	0.107	0.046	0.046	0.046	0.046	0.046	0.046
AUC	0.543	0.547	0.500	0.500	0.500	0.500	0.500	0.558
F1	0.508	0.502	0.488	0.488	0.488	0.488	0.488	0.488

an imbalance between majority and minority classes of 1:172, we can say that the performance of the system is sufficiently high, considering its characteristics.

In the second case study used to validate the proposed methodology, we have grouped the different radiological signs into higher level classes that are more general, as shown in the example of fracture types, Fig. 4. After this grouping, the number of labels passing the minimum 200-sample filter rises to 54 (in the specific labels experiment only 35 labels passed this threshold). Therefore, we now train the system with a larger number of labels, which is closer to the reality of health centres. Regarding the individual models, we can see that the best model is EfficientNet B0 followed by DenseNet, with an AUC of 0.767 and 0.761, respectively. The rest of the models have a value lower than 0.75. Regarding the performance per class of each model, we observe that Xception, EfficientNet and DenseNet perform better in majority classes, while Inception and ResNet perform better in minority classes.

If we look at the results obtained by the ensemble technique, as in the previous case, CTP is the best performer with an AUC of 0.819, which is an improvement of 0.052 over EfficientNet. There are four classes where the ensemble performs as well as the best individual model, but there is no class where the individual models perform better than the ensemble. The number of labels where the ensemble achieves an AUC above 0.85 is slightly lower than in the previous case, 37%, but more than 50% of the classes have an AUC greater than 0.8. This is interesting considering the number of classes (54) and their imbalance. Although the ensemble performs well, it does not perform well for all classes. For example, with the class "Sclerotic bone lesion" it obtains an AUC close to 0.5.

We can observe that in this case the ensemble further improves the individual models as the improvement over the best individual model is now high. The combination of different architectures avoids overfitting and improves the generalisation capacity in a problem where classification is more difficult due to the specificities of the dataset (high number of classes, multilabel, class imbalance). These results demonstrate that this methodology works well on highly imbalanced and multilabel datasets (see Tables 10 and 11).

Table 6
Specific labels experiment: results obtained by training the models with segmentation-based cropping, but without data augmentation. For each label, the individual models with the best performance and the ensembles that outperform all individual models are marked in bold. The best ensemble result is marked in italics.

	# Samples	DenseNet		EfficientNet		Inception		InceptionResnet		Xception		PTC-mode		PTC-lw		CTP	
		AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1
Normal	34327	0.5	0.374	0.802	0.725	0.453	0.374	**0.819**	0.723	0.5	0.374	0.528	0.444	0.500	0.374	*0.806*	0.374
Copd signs	13419	0.777	0.682	0.785	0.672	0.500	0.457	**0.799**	0.690	0.777	0.682	0.538	0.534	0.648	0.676	***0.825***	0.674
Cardiomegaly	8412	0.900	0.768	0.898	0.774	0.641	0.474	**0.918**	0.767	0.917	0.762	0.596	0.628	0.814	0.792	***0.938***	0.795
Aortic elongation	1399	0.863	0.700	0.874	0.686	0.500	0.479	**0.875**	0.719	0.837	0.705	0.594	0.623	0.767	0.724	***0.898***	0.724
Unchanged	1311	0.612	0.556	**0.625**	0.549	0.500	0.480	0.597	0.549	0.602	0.544	0.506	0.495	0.531	0.539	***0.642***	0.537
Scoliosis	1073	0.823	0.678	0.808	0.690	0.500	0.484	**0.830**	0.702	0.500	0.484	0.591	0.628	0.674	0.711	***0.863***	0.708
Chronic changes	873	0.707	0.537	**0.731**	0.553	0.500	0.487	0.696	0.547	0.695	0.538	0.515	0.518	0.625	0.568	***0.738***	0.568
Costophrenic angle blunting	703	0.810	0.698	0.837	0.691	0.500	0.489	**0.842**	0.655	0.810	0.704	0.558	0.587	0.729	0.713	***0.884***	0.712
Air trapping	663	0.500	0.490	0.671	0.568	0.500	0.490	0.500	0.490	**0.688**	0.553	0.508	0.506	0.500	0.490	***0.705***	0.490
Pleural effusion	658	0.925	0.839	0.942	0.818	0.479	0.046	**0.943**	0.770	0.927	0.838	0.818	0.542	0.901	0.823	*0.942*	0.825
Pneumonia	651	0.759	0.675	0.803	0.671	0.500	0.490	**0.808**	0.655	0.806	0.657	0.572	0.603	0.704	0.691	***0.851***	0.692
Interstitial pattern	594	0.799	0.638	0.795	0.650	0.500	0.491	**0.813**	0.637	0.812	0.615	0.562	0.576	0.714	0.678	***0.858***	0.680
Infiltrates	591	0.733	0.620	0.776	0.635	0.500	0.491	**0.802**	0.597	0.771	0.627	0.563	0.583	0.668	0.639	***0.831***	0.639
Laminar atelectasis	578	0.500	0.491	**0.806**	0.639	0.500	0.491	0.754	0.630	0.745	0.646	0.560	0.587	0.572	0.607	***0.837***	0.607
Vertebral degenerative	575	**0.730**	0.544	0.721	0.540	0.500	0.491	0.725	0.564	0.718	0.533	0.571	0.560	0.620	0.568	***0.771***	0.568
Kyphosis	526	0.796	0.611	**0.813**	0.644	0.500	0.492	0.794	0.628	**0.813**	0.615	0.569	0.585	0.683	0.664	***0.860***	0.664
Apical pleural thickening	469	**0.798**	0.591	0.787	0.573	0.500	0.493	0.775	0.569	0.758	0.575	0.574	0.567	0.701	0.619	***0.838***	0.619
Vascular hilar enlargement	463	0.679	0.562	**0.741**	0.506	0.500	0.493	0.717	0.559	0.715	0.547	0.531	0.533	0.596	0.578	***0.771***	0.582
Fibrotic band	449	0.756	0.568	**0.772**	0.599	0.500	0.493	0.767	0.608	0.758	0.593	0.573	0.585	0.688	0.636	***0.813***	0.638
Nodule	449	0.616	0.557	0.677	0.567	0.500	0.493	**0.688**	0.547	0.626	0.558	0.535	0.545	0.566	0.574	***0.719***	0.572
Calcified granuloma	388	0.741	0.651	0.752	0.641	0.500	0.494	**0.757**	0.622	0.689	0.611	0.578	0.601	0.645	0.654	***0.819***	0.656
Callus rib fracture	360	0.682	0.600	**0.773**	0.594	0.500	0.495	0.500	0.495	0.497	0.495	0.529	0.543	0.500	0.495	***0.799***	0.495
Pacemaker	336	**0.996**	0.948	**0.996**	0.945	0.500	0.495	**0.996**	0.946	**0.996**	0.949	0.741	0.799	0.992	0.951	*0.996*	0.951
Aortic atheromatosis	318	**0.812**	0.538	0.810	0.542	0.500	0.495	0.791	0.567	0.742	0.577	0.544	0.545	0.672	0.605	***0.852***	0.607
Volume loss	294	0.855	0.687	0.862	0.717	0.500	0.496	**0.882**	0.677	0.830	0.691	0.560	0.581	0.762	0.729	***0.910***	0.731
Sternotomy	292	0.991	0.945	0.991	0.939	0.500	0.496	**0.993**	0.872	**0.993**	0.918	0.756	0.814	0.983	0.948	***0.996***	0.948
Bronchiectasis	290	0.673	0.593	**0.726**	0.597	0.500	0.496	0.719	0.587	0.725	0.576	0.541	0.563	0.594	0.613	***0.784***	0.614
Hiatal hernia	287	0.912	0.852	0.920	0.826	0.500	0.496	0.939	0.843	**0.945**	0.726	0.747	0.784	0.877	0.870	***0.962***	0.872
Pseudonodule	275	0.632	0.524	**0.705**	0.547	0.500	0.496	0.536	0.514	0.639	0.540	0.530	0.536	0.540	0.544	***0.718***	0.545
Hemidiaphragm elevation	254	0.902	0.706	0.879	0.697	0.500	0.496	**0.911**	0.696	0.891	0.687	0.749	0.709	0.816	0.751	***0.951***	0.751
Alveolar pattern	248	0.791	0.626	0.834	0.603	0.500	0.496	**0.853**	0.580	0.810	0.604	0.568	0.579	0.715	0.621	***0.887***	0.622
Increased density	239	0.580	0.551	0.586	0.521	0.500	0.496	**0.619**	0.521	0.569	0.526	0.501	0.500	0.533	0.537	***0.634***	0.539
Vertebral anterior compression	214	0.640	0.536	**0.645**	0.530	0.500	0.497	0.644	0.537	0.623	0.517	0.516	0.522	0.524	0.524	***0.702***	0.525
Suture material	210	0.798	0.663	0.791	0.649	0.500	0.497	**0.824**	0.628	0.786	0.665	0.622	0.622	0.742	0.679	***0.833***	0.680
Supra aortic elongation	200	0.697	0.569	0.778	0.564	0.500	0.497	**0.832**	0.561	0.738	0.554	0.579	0.574	0.613	0.577	***0.861***	0.578
Global		0.751	0.633	**0.792**	0.648	0.502	0.475	0.779	0.636	0.750	0.622	0.584	0.586	0.677	0.650	***0.831***	0.651

Table 7
Specific labels experiment: global results obtained by the individual models and the ensemble with preprocessing (segmentation-based cropping) but without data augmentation.

	Densenet201	EfficientNet	Inception	InceptionResnet	Xception	PTC-mode	PTC-lw	CTP
Hamming Loss	0.077	0.079	0.072	0.070	0.077	0.056	0.057	0.057
AUC	0.751	**0.792**	0.502	0.779	0.750	0.584	0.677	**0.831**
F1-score	0.633	0.648	0.475	0.636	0.622	0.586	0.650	0.651

5.3. Visual explanation using heatmaps

As explained in Section 2, the visualisation of multilabel problems is an essential element for this methodology, but it is not a simple problem. Most of the work in this field has deficiencies. Therefore, we have developed a technique that for each label generates a heatmap, an estimated probability, and the ensemble agreement. In Fig. 9, we can see the original X-ray and the heatmaps of the different classes. The areas marked on the radiographs match the radiological signs, and the probabilities are high, with three of the four cases showing agreement between all models.

In the second example, Fig. 10, we can see that the class probabilities are lower than before. The class Atelectasis has an agreement of three models and a low probability (0.583), which means that the physician should be careful with this label. The last example, Fig. 11, belongs to the normal class. In this case, the heat map marks approximately the entire radiograph, as it scans the whole image for radiological signs. The performance of the visualisations is highly dependent on the performance of the model: if the model is better, the visualisations will be more accurate, and the probability and agreement between models will be higher. An advantage of this technique over the state of the art is that we generate a grad-CAM map for each sign that includes the probability generated by the system and the agreement between the models of the ensemble.

6. Discussion

As mentioned throughout the article, the PadChest dataset has a high quality and is really interesting due to the number of classes, which is higher than other multilabel datasets, and the challenge of class imbalance. Although we can find numerous papers using this dataset for medical report generation, it is underutilised in chest X-ray classification problems, which makes the available works for comparison scarce. Moreover, those articles present several problems that complicate an adequate comparison of our work. Therefore, one of our aims is to generate a methodologically correct baseline that allows comparison for future work. For this purpose, we have conducted two experiments: in the first one we have used the specific radiological signs, i.e. the original ones from the dataset, while in the second we have used more generic radiological signs from a tree of terms provided by the authors of the dataset. In Table 12 we can find a summary of

Table 8
Specific labels experiment: results obtained with by training the models with segmentation-based cropping and data augmentation. For each label, the individual models with the best performance and the ensembles that outperform all individual models are marked in bold. The best ensemble result is marked in italics.

	# Samples	Densenet201		EfficientNet		Inception		InceptionResnet		Xception		PTC-mode		PTC-lw		CTP	
		AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1
Normal	34327	0.820	0.722	0.811	0.716	**0.832**	0.727	0.820	0.709	0.827	0.732	0.725	0.731	0.725	0.731	***0.837***	0.730
Copd signs	13419	**0.823**	0.681	0.785	0.644	0.816	0.678	0.815	0.675	0.800	0.666	0.588	0.610	0.647	0.675	***0.833***	0.672
Cardiomegaly	8412	**0.927**	0.773	0.907	0.749	0.926	0.767	**0.927**	0.777	0.922	0.779	0.746	0.765	0.825	0.789	***0.937***	0.791
Aortic elongation	1399	0.885	0.690	0.846	0.655	0.882	0.676	**0.888**	0.698	**0.888**	0.702	0.690	0.682	0.777	0.705	***0.894***	0.707
Unchanged	1311	0.636	0.553	0.614	0.543	0.638	0.549	**0.642**	0.544	0.636	0.547	0.531	0.537	0.545	0.551	*0.641*	0.551
Scoliosis	1073	0.759	0.636	0.712	0.598	0.745	0.605	0.732	0.602	**0.774**	0.661	0.630	0.637	0.671	0.659	***0.793***	0.664
Chronic changes	873	0.759	0.518	0.720	0.549	**0.768**	0.546	0.762	0.519	0.752	0.533	0.621	0.556	0.684	0.545	***0.772***	0.547
Costophrenic angle blunting	703	**0.862**	0.674	0.845	0.674	0.832	0.662	0.831	0.665	0.855	0.663	0.685	0.676	0.739	0.693	***0.877***	0.693
Air trapping	663	**0.692**	0.557	0.687	0.560	0.515	0.490	0.469	0.490	0.506	0.490	0.525	0.532	0.500	0.490	***0.704***	0.490
Pleural effusion	658	**0.959**	0.827	0.951	0.811	0.956	0.823	0.955	0.830	0.945	0.816	0.862	0.822	0.886	0.839	***0.967***	0.840
Pneumonia	651	0.815	0.671	**0.821**	0.660	0.810	0.663	**0.821**	0.672	**0.821**	0.668	0.681	0.668	0.703	0.687	***0.850***	0.687
Interstitial pattern	594	0.834	0.625	0.828	0.636	**0.846**	0.651	0.843	0.616	0.830	0.613	0.727	0.651	0.743	0.650	***0.858***	0.651
Infiltrates	591	0.812	0.629	0.803	0.617	0.803	0.626	**0.815**	0.633	0.808	0.639	0.649	0.635	0.662	0.644	***0.840***	0.646
Laminar atelectasis	578	**0.843**	0.670	0.812	0.637	0.827	0.643	0.827	0.654	0.833	0.654	0.666	0.658	0.690	0.677	***0.858***	0.678
Vertebral degenerative changes	575	0.779	0.545	0.730	0.544	0.774	0.546	**0.785**	0.518	0.779	0.547	0.627	0.560	0.670	0.557	***0.797***	0.556
Kyphosis	526	**0.867**	0.640	0.834	0.589	0.845	0.587	0.849	0.609	0.839	0.625	0.691	0.617	0.736	0.639	***0.870***	0.640
Apical pleural thickening	469	**0.808**	0.553	0.801	0.568	0.789	0.573	0.509	0.493	0.500	0.493	0.592	0.574	0.661	0.619	***0.830***	0.621
Vascular hilar enlargement	463	**0.746**	0.549	0.742	0.568	0.769	0.522	0.755	0.544	0.745	0.515	0.618	0.556	0.651	0.562	***0.783***	0.563
Fibrotic band	449	**0.831**	0.583	0.809	0.600	0.813	0.614	0.575	0.493	0.806	0.611	0.641	0.604	0.716	0.658	***0.848***	0.659
Nodule	449	**0.706**	0.578	0.675	0.554	0.561	0.493	0.574	0.493	0.551	0.493	0.518	0.526	0.500	0.493	*0.704*	0.493
Calcified granuloma	388	**0.808**	0.653	0.802	0.649	0.542	0.494	0.554	0.494	0.496	0.494	0.572	0.593	0.500	0.494	***0.833***	0.494
Callus rib fracture	360	0.717	0.606	**0.765**	0.557	0.614	0.495	0.609	0.495	0.571	0.495	0.550	0.549	0.500	0.495	***0.787***	0.495
Pacemaker	336	0.993	0.927	**0.997**	0.942	0.996	0.919	0.996	0.931	0.984	0.926	0.984	0.930	0.993	0.946	***0.997***	0.946
Aortic atheromatosis	318	0.856	0.521	0.847	0.516	0.862	0.550	0.852	0.541	**0.871**	0.559	0.739	0.556	0.786	0.558	***0.885***	0.557
Volume loss	294	**0.917**	0.693	0.902	0.657	0.904	0.636	0.896	0.672	0.902	0.640	0.789	0.670	0.809	0.693	***0.928***	0.697
Sternotomy	292	0.992	0.898	0.987	0.920	0.990	0.926	**0.995**	0.936	0.992	0.865	0.961	0.912	0.984	0.941	***0.997***	0.941
Bronchiectasis	290	0.801	0.549	0.775	0.561	0.796	0.578	**0.805**	0.562	0.794	0.573	0.682	0.580	0.690	0.587	***0.820***	0.588
Hiatal hernia	287	0.939	0.856	0.941	0.697	0.947	0.824	**0.964**	0.801	0.947	0.851	0.871	0.786	0.876	0.867	***0.967***	0.867
Pseudonodule	275	0.612	0.496	**0.670**	0.550	0.598	0.496	0.589	0.496	0.545	0.496	0.536	0.543	0.500	0.496	***0.672***	0.496
Hemidiaphragm elevation	254	0.893	0.667	0.882	0.679	**0.915**	0.670	0.894	0.702	0.898	0.684	0.759	0.692	0.784	0.725	***0.934***	0.724
Alveolar pattern	248	0.876	0.627	0.877	0.589	0.885	0.607	**0.895**	0.606	0.871	0.594	0.748	0.612	0.774	0.623	***0.911***	0.622
Increased density	239	0.643	0.541	0.641	0.509	0.651	0.534	0.633	0.526	**0.668**	0.535	0.545	0.531	0.547	0.544	***0.673***	0.547
Vertebral anterior compression	214	**0.749**	0.532	0.696	0.524	0.736	0.517	0.743	0.527	0.734	0.535	0.573	0.530	0.597	0.540	***0.752***	0.539
Suture material	210	0.819	0.652	0.820	0.663	0.818	0.639	0.811	0.662	**0.822**	0.612	0.759	0.663	0.768	0.685	***0.847***	0.684
Supra aortic elongation	200	0.865	0.576	0.821	0.541	0.880	0.546	**0.882**	0.563	0.857	0.562	0.628	0.558	0.681	0.576	***0.894***	0.575
Global		**0.818**	0.642	0.804	0.629	0.797	0.625	0.780	0.621	0.782	0.625	0.677	0.637	0.701	0.647	***0.840***	0.647

Table 9
Specific labels experiment: global results obtained from the individual models and the ensembles.

	Densenet201	EfficientNet	Inception	InceptionResnet	Xception	PTC-mode	PTC-lw	CTP
Hamming Loss	0.082	0.078	0.081	0.077	0.074	0.063	0.065	0.065
AUC	0.818	0.804	0.797	0.780	0.782	0.677	0.701	0.840
F1	0.642	0.629	0.625	0.621	0.625	0.637	0.647	0.647

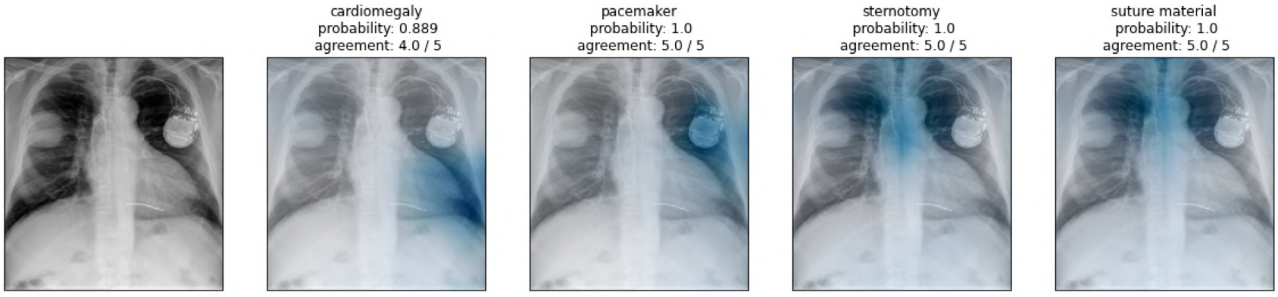


Fig. 9. First visualisation example. The heatmaps of four radiological signs detected (cardiomegaly, pacemaker, sternotomy and suture material) are shown. The title shows the label, the probability estimated by the ensemble, and the agreement between the models of the ensemble. The areas of interest for classification are marked in blue.

the different published systems and their global and class specific performance. First of all, it is interesting to note how most of the papers have selected different labels to perform the classification, and all papers, except [61], select a low number of total classes compared to the number of classes available.

In the case of Rimeika et al. [87], the publication does not show how the two models have been built; it does not provide information on the architecture, the other dataset used, or the criteria for selecting the classes from PadChest dataset, so there is no possibility to replicate these models, and therefore we cannot use it for comparison. In Pooch et al. [62], the PadChest classes have been adapted to match the classes of other multilabel datasets such as ChestX-ray14 and CheXpert. For example, regardless of the fact that the class "Lesion" does not exist in two of the datasets, they generated this class using the ChestX-ray14 labels "Nodules" and "Masses". However, PadChest was processed in

Table 10
General labels experiment: results obtained by training the models with segmentation-based cropping and data augmentation. For each label, the individual models with the best performance and the ensembles that outperform all individual models are marked in bold. The best ensemble result is marked in italics.

	# Samples	Densenet201		EfficientNet		Inception		InceptionResnet		Xception		PTC-mode		PTC-lw		CTP	
		AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1
Normal	34327	0.735	0.685	0.707	0.652	**0.750**	0.691	0.723	0.658	0.732	0.674	0.702	0.693	0.690	0.691	***0.770***	0.691
Copd signs	13419	0.771	0.629	0.761	0.649	0.771	0.618	**0.793**	0.665	0.779	0.645	0.596	0.621	0.615	0.645	***0.816***	0.640
Cardiomegaly	8120	0.899	0.736	**0.904**	0.746	0.890	0.723	0.892	0.751	0.898	0.741	0.766	0.747	0.816	0.760	***0.923***	0.762
Thoracic cage deformation	7778	0.706	0.603	**0.728**	0.627	0.500	0.478	0.675	0.595	0.708	0.609	0.577	0.586	0.601	0.612	***0.745***	0.614
Aortic elongation	7436	0.858	0.691	0.853	0.690	0.842	0.661	**0.866**	0.683	**0.866**	0.687	0.701	0.690	0.758	0.697	***0.886***	0.700
Infiltrates	6706	0.794	0.686	**0.802**	0.664	0.791	0.663	0.797	0.668	0.794	0.663	0.703	0.676	0.721	0.690	***0.827***	0.692
Unchanged	6487	0.630	0.538	**0.636**	0.552	0.618	0.543	0.633	0.545	0.631	0.552	0.536	0.543	0.536	0.546	***0.652***	0.547
Chronic changes	4312	**0.759**	0.548	0.754	0.542	0.752	0.525	0.734	0.520	0.740	0.522	0.656	0.567	0.685	0.543	***0.768***	0.545
Surgery	3928	0.813	0.730	0.815	0.766	0.750	0.722	0.766	0.713	**0.829**	0.724	0.726	0.739	0.739	0.762	***0.845***	0.765
Atelectasis	3565	**0.798**	0.628	0.756	0.636	0.698	0.570	0.729	0.596	0.759	0.628	0.587	0.598	0.647	0.632	***0.804***	0.630
Costophrenic angle blunting	3306	**0.845**	0.655	0.807	0.638	0.758	0.604	0.784	0.638	0.828	0.652	0.660	0.634	0.700	0.656	***0.864***	0.658
Calcified densities	3253	0.719	0.638	**0.751**	0.639	0.500	0.491	0.500	0.491	0.500	0.491	0.520	0.527	0.500	0.491	***0.764***	0.491
Vertebral degenerative changes	3203	**0.744**	0.502	0.726	0.528	0.676	0.497	0.733	0.487	0.730	0.512	0.643	0.532	0.664	0.514	***0.751***	0.514
Hilar enlargement	3162	**0.755**	0.549	0.732	0.544	0.699	0.551	0.738	0.533	0.731	0.538	0.618	0.562	0.649	0.560	***0.765***	0.565
Pleural thickening	3010	0.753	0.586	**0.773**	0.585	0.737	0.572	0.743	0.525	0.763	0.562	0.651	0.587	0.671	0.586	***0.790***	0.587
Mediastinal enlargement	2813	0.795	0.643	0.798	0.668	0.774	0.688	0.778	0.675	**0.822**	0.657	0.705	0.689	0.710	0.697	***0.841***	0.700
Air trapping	2765	0.654	0.528	0.669	0.536	0.500	0.492	0.665	0.495	**0.672**	0.536	0.520	0.523	0.602	0.544	***0.692***	0.546
Fracture	2529	**0.749**	0.663	0.725	0.599	0.5	0.493	0.640	0.507	0.732	0.611	0.574	0.590	0.579	0.615	***0.792***	0.617
Pleural effusion	2436	**0.942**	0.738	0.927	0.782	0.930	0.720	0.935	0.735	0.937	0.771	0.878	0.762	0.900	0.775	***0.956***	0.775
Granuloma	2306	0.500	0.493	**0.777**	0.652	0.500	0.493	0.500	0.493	0.500	0.493	0.513	0.519	0.500	0.493	***0.777***	0.493
Nodule	1936	0.653	0.583	**0.679**	0.571	0.617	0.542	0.643	0.547	0.622	0.575	0.560	0.569	0.558	0.575	***0.707***	0.573
Fibrotic band	1781	0.738	0.530	**0.747**	0.531	0.712	0.522	0.500	0.495	0.727	0.519	0.606	0.546	0.654	0.556	***0.770***	0.556
Electrical device	1772	0.992	0.959	0.992	0.913	0.992	0.889	**0.994**	0.871	0.992	0.929	0.990	0.935	0.992	0.942	***0.997***	0.942
Pneumonia	1652	0.804	0.594	0.790	0.567	0.804	0.549	**0.813**	0.577	0.799	0.599	0.712	0.602	0.728	0.606	***0.854***	0.607
Aortic atheromatosis	1581	0.834	0.502	0.830	0.540	0.813	0.477	0.840	0.524	**0.843**	0.519	0.713	0.554	0.769	0.522	***0.866***	0.523
Pseudonodule	1451	0.693	0.561	**0.727**	0.553	0.500	0.496	0.500	0.496	0.708	0.562	0.557	0.555	0.576	0.589	***0.759***	0.593
Bronchiectasis	1430	0.795	0.544	0.776	0.571	0.789	0.548	**0.814**	0.539	0.779	0.561	0.639	0.575	0.696	0.573	***0.833***	0.574
Hiatal hernia	1362	0.916	0.813	0.892	0.796	0.906	0.773	0.918	0.804	**0.927**	0.788	0.857	0.810	0.889	0.852	***0.959***	0.852
Hemidiaphragm elevation	1231	0.814	0.651	**0.841**	0.680	**0.841**	0.596	0.823	0.649	0.811	0.645	0.733	0.670	0.746	0.683	***0.890***	0.683
Increased density	1133	0.633	0.497	**0.640**	0.524	0.596	0.492	0.609	0.509	0.606	0.511	0.539	0.516	0.533	0.514	***0.661***	0.515
Diaphragmatic eventration	757	0.500	0.498	**0.775**	0.586	0.500	0.498	0.500	0.498	0.500	0.498	0.525	0.534	0.500	0.498	***0.775***	0.498
Volume loss	684	0.802	0.542	0.776	0.580	0.809	0.531	**0.814**	0.513	0.776	0.560	0.728	0.561	0.761	0.564	***0.865***	0.564
Adenopathy	659	0.500	0.498	**0.697**	0.538	0.500	0.498	0.548	0.520	0.583	0.543	0.500	0.498	0.521	0.528	***0.715***	0.528
Bronchovascular markings	602	0.712	0.570	0.738	0.537	**0.777**	0.576	0.765	0.545	0.704	0.585	0.685	0.592	0.703	0.585	***0.802***	0.584
Mass	574	0.707	0.621	0.715	0.608	0.744	0.570	0.732	0.574	**0.746**	0.616	0.700	0.615	0.707	0.641	***0.806***	0.641
Artificial heart valve	562	0.969	0.658	0.953	0.730	0.975	0.696	**0.977**	0.727	0.972	0.713	0.941	0.736	0.968	0.730	***0.981***	0.731
Catheter	545	0.871	0.740	0.874	0.721	0.866	0.673	**0.878**	0.639	0.861	0.717	0.799	0.724	0.848	0.773	***0.905***	0.773
Suboptimal study	544	0.743	0.524	0.693	0.510	**0.754**	0.522	0.697	0.531	0.727	0.506	0.666	0.526	0.681	0.539	***0.784***	0.540
Pulmonary fibrosis	523	0.850	0.584	0.834	0.587	0.837	0.551	**0.864**	0.577	0.862	0.565	0.795	0.577	0.810	0.591	***0.892***	0.591
Heart insufficiency	520	0.875	0.541	0.877	0.555	0.896	0.546	**0.884**	0.538	0.870	0.547	0.819	0.553	0.856	0.551	***0.920***	0.551
Hypoexpansion	476	0.838	0.541	0.745	0.545	**0.846**	0.534	0.768	0.571	0.500	0.499	0.651	0.556	0.677	0.571	***0.900***	0.573
Gynecomastia	437	0.852	0.527	0.810	0.552	0.852	0.501	**0.858**	0.507	0.806	0.550	0.772	0.540	0.825	0.554	***0.917***	0.555
Emphysema	410	0.780	0.508	0.715	0.520	0.801	0.512	**0.809**	0.506	0.724	0.521	0.684	0.529	0.732	0.524	***0.862***	0.525
Sclerotic bone lesion	352	**0.506**	0.511	0.500	0.499	0.500	0.499	0.500	0.499	0.500	0.499	0.500	0.499	0.500	0.499	***0.506***	0.499
Fissure thickening	336	0.816	0.533	0.802	0.573	0.819	0.518	**0.842**	0.526	0.798	0.539	0.746	0.547	0.806	0.557	***0.891***	0.558
Hilar congestion	318	0.785	0.503	0.798	0.519	0.790	0.514	**0.827**	0.519	0.808	0.520	0.734	0.526	0.756	0.523	***0.896***	0.522
Osteopenia	318	0.659	0.508	0.688	0.507	0.659	0.483	0.695	0.466	**0.701**	0.497	0.611	0.500	0.647	0.500	***0.752***	0.500
Tuberculosis	299	0.852	0.534	0.861	0.567	**0.869**	0.561	0.824	0.559	0.805	0.597	0.760	0.577	0.848	0.592	***0.909***	0.592
Bullas	290	**0.746**	0.520	0.685	0.532	0.739	0.524	0.715	0.512	0.651	0.549	0.667	0.543	0.714	0.547	***0.777***	0.547
Hyperinflated lung	272	0.715	0.506	0.630	0.502	**0.719**	0.504	0.645	0.485	0.659	0.501	0.649	0.513	0.658	0.512	***0.728***	0.512
Cavitation	243	0.780	0.556	0.834	0.575	**0.856**	0.546	0.789	0.539	0.823	0.590	0.679	0.555	0.823	0.585	***0.934***	0.585
Mediastinic lipomatosis	212	0.648	0.499	**0.654**	0.551	0.5	0.499	0.500	0.499	0.500	0.499	0.520	0.514	0.500	0.499	***0.681***	0.499
Pneumothorax	210	0.705	0.572	0.717	0.530	0.717	0.540	**0.721**	0.518	0.620	0.592	0.596	0.546	0.630	0.572	***0.847***	0.573
Vascular redistribution	204	**0.774**	0.499	0.752	0.526	0.705	0.508	0.694	0.516	0.667	0.507	0.635	0.514	0.676	0.519	***0.837***	0.519
Global		0.761	0.589	**0.767**	0.600	0.732	0.566	0.739	0.572	0.739	0.589	0.669	0.594	0.696	0.601	***0.819***	0.602

Table 11
General labels experiment: global results obtained by the individual models and the ensembles.

	Densenet201	EfficientNet	Inception	InceptionResnet	Xception	PTC-mode	PTC-lw	CTP
Hamming Loss	0.070	0.065	0.070	0.075	0.065	0.052	0.057	0.056
AUC	0.761	0.767	0.732	0.739	0.739	0.669	0.696	0.819
F1-score	0.589	0.600	0.566	0.572	0.589	0.594	0.601	0.602

that paper by unifying all classes related to Atelectasis, without providing any medical explanation for this decision, so the labels do not match ours, containing only 8 out of 174 classes available. Therefore, it cannot be compared with our methodology. In the case of Hashir et al. [61], they first select a single sample from each patient and the authors have used 32 different labels, which is not in line with expectations, since using a lower threshold than ours there should have a larger number of labels. In this case, we can compare the overall AUC of the system. These authors achieve an AUC of 0.800 using 32 labels while we obtain 0.8397 using 35 specific labels. Therefore, we have achieved a better AUC than in Hashir et al. [61]. Because of the above reasons, comparing our methodology with the state of the art is really difficult, and therefore one of the aims of this paper is to create a baseline

Fig. 10. Second visualisation example. The heatmaps of three radiological signs detected (pleural effusion, laminar atelectasis and suture material) are shown. The title shows the label, the probability estimated by the ensemble, and the agreement between the models of the ensemble. The areas of interest for classification are marked in blue.

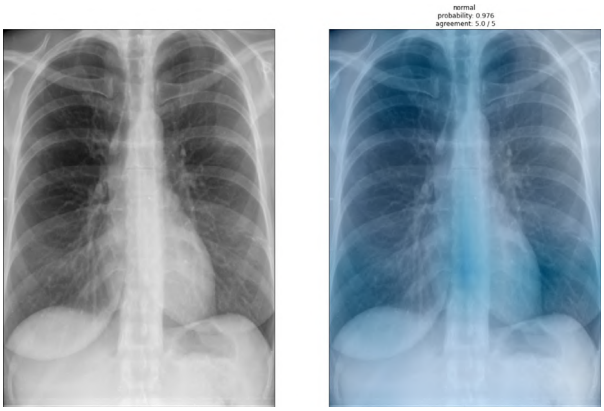


Fig. 11. Third visualisation example. The sample belongs to the normal class. The title shows the label, the probability estimated by the ensemble, and the agreement between the models of the ensemble. The areas of interest for classification are marked in blue.

Table 12
Comparative table of the different state-of-the-art models, their global and class performance.

	Rimeika G. et al. [87]		Pooch, E. H. [62]
	model1	model2	
cardiomegaly	90.36%	91.94%	90.75%
nodule	74.97%	71.42%	–
normal	–	–	87.10%
pleural effusion	95.42%	94.93%	–
pneumonia	–	–	79.90%
lobar collapse	88.86%	86.39%	–
edema	95.35%	96.05%	91.07%
subcutaneous emphysema	98.52%	93.79%	–
consolidation	87.39%	85.50%	86.07%
pneumothorax	89.95%	88.19%	82.76%
tuberculosis	92.62%	92.40%	–
Lymphadenopathy	77.11%	75.81%	–
linear atelectasis	84.16%	78.26%	76.41%
lymph node calficiation	82.64%	72.69%	–
congestion	85.39%	87.29%	–
Widened mediastinum	75.02%	77.50%	–
mass	86.90%	82.29%	–
lesion	–	–	69.75%
Global	*86.98%*	*84.97%*	*82.98%*

to facilitate the comparison of future work with this dataset. If we look at the overall AUC of the published models trained with PadChest and compare them with ours, we see that we only outperform two models, but we use a much higher number of classes. Therefore, we can see that our system performs well and that although it works with a much larger number of labels, it outperforms some of the published models.

7. Conclusions and future work

This paper proposes a Deep Learning methodology for classification tasks with imbalanced multilabel datasets. We have built with this methodology an ensemble of five state-of-the-art architectures: DenseNet-201, EfficientNet B0, Inception, InceptionResNet and Xception. We have used weighted crossentropy with logit loss to alleviate data imbalance and developed a new technique for generating heatmaps in multilabel classification problems.

The results of our experiments are promising. First, in contrast to state-of-the-art papers, we have established a methodologically sound baseline for future work, regardless of whether specific or general labels are used. It will also allow us to analyse the performance of these models when the number of labels varies. Our system obtains high AUC values for the number of classes used. In the case of specific labels, high performance is achieved with an AUC of 0.84. In the case of general labels, we obtain an AUC of 0.819. This value may be due to the fact that the general classification has more classes and each of them is composed of different radiological signs. Thus, the variability is high and it is more difficult to classify. The results of the visualisation technique show a great potential, as it allows a view of the whole radiograph that differentiates the different pathological signs. This technique generates a report that includes the visualisation of the heatmap, the probability produced by the system and the agreement between the ensemble models.

There are several ways to improve our methodology. First, other strategies can be used to alleviate data imbalance, such as adding new samples to the dataset. This can be done either by obtaining new images from other datasets such as CheXpert, ChestX-ray14, or other single disease datasets, or by creating them with generative adversarial networks (GANs). Another way to improve the performance of the proposed system is to use different X-ray views of each sample.

In our proposal, we used segmentation techniques to force the model to pay more attention to the most relevant areas. However, different techniques have recently been developed for this same purpose. For example, [88] used soft and hard attention mechanisms to prevent the model from focusing on areas that are not relevant to the problem. Another way to remove non-interesting areas is the application of semi-supervised learning methods to locate and distinguish different anatomical regions [89]. Based on

these recent advances, it would be interesting to study whether including them in our model improves its results.

To improve the visualisation technique, we can extend the displayed information by including a heatmap that shows the standard deviation of the visualisation of the ensemble [90]. This would help medical staff to know in which areas of the heatmap there is more uncertainty. Another line of work we would like to explore is the generation of a system that returns general and specific labels. In addition to a combination with report generation techniques, doctors would receive a report explaining the different radiological signs and a visual interpretation of these signs. This could be done using cascade models, which first classify the most general labels and later classify the subcategories. This would allow to include minority classes, or at least part of them.

Another possible improvement would be to retrain the system using feedback from experts in the field on the system's predictions and heatmaps. This poses multiple challenges in practice, mainly due to the need to implement close collaboration between specialised diagnostic models and medical staff, who may lack background and expertise to rely on the models' results. On the positive side, semi-supervised Deep Learning techniques are emerging lately and are yielding results that are unprecedented in the state of the art. For instance, Avilés-Rivero et al. [91] have developed a semi-supervised graph-based framework for classifying lung diseases (COVID-19, pneumonia and healthy). Such frameworks have been identified as promising for supporting the construction of human-in-the-loop models in medical applications [30], hence future efforts will be devoted in this direction.

CRediT authorship contribution statement

Helena Liz: Conceptualization, Visualization, Writing, Methodology. **Javier Huertas-Tato:** Conceptualization, Visualization, Validation, Methodology, Writing. **Manuel Sánchez-Montañés:** Conceptualization, Writing – review & editing. **Javier Del Ser:** Conceptualization, Writing – review & editing. **David Camacho:** Conceptualization, Writing, Resources, Funding acquisition, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgements

This work has been funded by Grant PLEC2021-007681 (XAI-DisInfodemics) and PID2020-117263GB-100 (FightDIS) funded by MCIN/AEI/ 10.13039/501100011033 and, as appropriate, by "ERDF A way of making Europe", by the "European Union NextGenerationEU/PRTR", by the research project CIVIC: Intelligent characterisation of the veracity of the information related to COVID-19, granted by BBVA FOUNDATION GRANTS FOR SCIENTIFIC RESEARCH TEAMS SARS-CoV-2 and COVID-19, by European Comission under IBERIFIER - Iberian Digital Media Research and Fact-Checking Hub (2020-EU-IA-0252), by "Convenio Plurianual with the Universidad Politécnica de Madrid in the actuation line of Programa de Excelencia para el Profesorado Universitario", and by Comunidad Autónoma de Madrid under S2018/TCS-4566 (CYNAMON) grant. M. Sánchez-Montañés has been supported by grants PID2021-127946OB-I00 and PID2021-122347NB-I00 (funded by MCIN/AEI/ 10.13039/501100011033 and ERDF - "A way of making Europe") and Comunidad Autónoma de Madrid, Spain (S2017/BMD-3688 MULTI-TARGET&VIEW-CM grant). J. Del Ser thanks the financial support of the Spanish Centro para el Desarrollo Tecnológico Industrial (CDTI, Ministry of Science and Innovation) through the "Red Cervera" Programme (AI4ES project), as well as the support of the Basque Government (consolidated research group MATHMODE, ref. IT1456-22)

References

[1] E. Moustaka, T.C. Constantinidis, Sources and effects of work-related stress in nursing, Health Sci. J. 4 (4) (2010) 210.
[2] S. Domínguez-Rodríguez, H. Liz, A. Panizo, Á. Ballesteros, R. Dagan, D. Greenberg, L. Gutiérrez, P. Rojo, E. Otheo, J.C. Galán, et al., Testing the performance, adequacy, and applicability of an artificial intelligent model for pediatric pneumonia diagnosis, 2022.
[3] N. Shaw, M. Hendry, O. Eden, Inter-observer variation in interpretation of chest X-rays, Scott. Med. J. 35 (5) (1990) 140–141.
[4] K.B. Ahmed, G.M. Goldgof, R. Paul, D.B. Goldgof, L.O. Hall, Discovery of a generalization gap of Convolutional Neural Networks on COVID-19 X-rays classification, Ieee Access 9 (2021) 72970–72979.
[5] Y. LeCun, Y. Bengio, et al., Convolutional networks for images, speech, and time series, in: The Handbook of Brain Theory and Neural Networks, Vol. 3361, no. 10, 1995, p. 1995.
[6] T. Kontzer, Deep learning drops error rate for breast cancer diagnoses by 85%, 2016, URL: https://blogs.nvidia.com/blog/2016/09/19/deep-learning-breast-cancer-diagnosis/.
[7] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.
[8] T. Agrawal, P. Choudhary, EfficientUNet: Modified encoder-decoder architecture for the lung segmentation in chest X-ray images, Expert Syst. (2022) e13012.
[9] I.M. Baltruschat, H. Nickisch, M. Grass, T. Knopp, A. Saalbach, Comparison of deep learning approaches for multi-label chest X-ray classification, Sci. Rep. 9 (1) (2019) 1–10.
[10] J.-Y. Park, Y. Hwang, D. Lee, J.-H. Kim, MarsNet: Multi-label classification network for images of various sizes, IEEE Access 8 (2020) 21832–21846.
[11] A. Bustos, A. Pertusa, J.-M. Salinas, M. de la Iglesia-Vayá, Padchest: A large chest X-ray image dataset with multi-label annotated reports, Med. Image Anal. 66 (2020) 101797.
[12] I. Al-Badarneh, M. Habib, I. Aljarah, H. Faris, Neuro-evolutionary models for imbalanced classification problems, J. King Saud Univ.-Comput. Inf. Sci. (2020).
[13] E. Lin, Q. Chen, X. Qi, Deep reinforcement learning for imbalanced classification, Appl. Intell. 50 (8) (2020) 2488–2502.
[14] T.N. Mundhenk, B.Y. Chen, G. Friedland, Efficient saliency maps for explainable AI, 2019, arXiv preprint arXiv:1911.11293.
[15] A. Holzinger, M. Dehmer, F. Emmert-Streib, R. Cucchiara, I. Augenstein, J. Del Ser, W. Samek, I. Jurisica, N. Díaz-Rodríguez, Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence, Inf. Fusion 79 (2022) 263–278.
[16] R.K. Singh, R. Pandey, R.N. Babu, Covidscreen: explainable deep learning framework for differential diagnosis of COVID-19 using chest X-rays, Neural Comput. Appl. 33 (14) (2021) 8871–8892.
[17] F. Piccialli, V. Di Somma, F. Giampaolo, S. Cuomo, G. Fortino, A survey on deep learning in medicine: Why, how and when? Inf. Fusion 66 (2021) 111–137.
[18] S.W. Baalman, F.E. Schroevers, A.J. Oakley, T.F. Brouwer, W. van der Stuijt, H. Bleijendaal, L.A. Ramos, R.R. Lopes, H.A. Marquering, R.E. Knops, et al., A morphology based deep learning model for atrial fibrillation detection using single cycle electrocardiographic samples, Int. J. Cardiol. 316 (2020) 130–136.
[19] E.A. Chung, M.E. Benalcázar, Real-time hand gesture recognition model using deep learning techniques and EMG signals, in: 2019 27th European Signal Processing Conference, EUSIPCO, IEEE, 2019, pp. 1–5.
[20] H. Li, X. Wang, C. Liu, Q. Zeng, Y. Zheng, X. Chu, L. Yao, J. Wang, Y. Jiao, C. Karmakar, A fusion framework based on multi-domain features and deep learning features of phonocardiogram for coronary artery disease detection, Comput. Biol. Med. 120 (2020) 103733.
[21] J. Pan, Y. Zi, J. Chen, Z. Zhou, B. Wang, LiftingNet: A novel deep learning network with layerwise feature learning from noisy mechanical data for fault classification, IEEE Trans. Ind. Electron. 65 (6) (2017) 4973–4982.
[22] Y.-S. Su, T.-J. Ding, M.-Y. Chen, Deep learning methods in internet of medical things for valvular heart disease screening system, IEEE Internet Things J. 8 (23) (2021) 16921–16932.

[23] F.N. Abdullah, M.N. Fauzan, N. Riza, Multiple linear regression and deep learning in body temperature detection and mask detection, IT J. Res. Dev. (2022) 109–121.
[24] M. Jost, D.A. Santos, R.A. Saunders, M.A. Horlbeck, J.S. Hawkins, S.M. Scaria, T.M. Norman, J.A. Hussmann, C.R. Liem, C.A. Gross, et al., Titrating gene expression using libraries of systematically attenuated CRISPR guide RNAs, Nature Biotechnol. 38 (3) (2020) 355–364.
[25] S. Li, K. Yu, D. Wang, Q. Zhang, Z.-X. Liu, L. Zhao, H. Cheng, Deep learning based prediction of species-specific protein S-glutathionylation sites, Biochim. Biophys. Acta (BBA)-Proteins Proteomics 1868 (7) (2020) 140422.
[26] G. Zampieri, S. Vijayakumar, E. Yaneske, C. Angione, Machine and deep learning meet genome-scale metabolic modeling, PLoS Comput. Biol. 15 (7) (2019) e1007084.
[27] M.L. Welch, C. McIntosh, A. McNiven, S.H. Huang, B.-B. Zhang, L. Wee, A. Traverso, B. O'Sullivan, F. Hoebers, A. Dekker, et al., User-controlled pipelines for feature integration and head and neck radiation therapy outcome predictions, Phys. Med. 70 (2020) 145–152.
[28] Z. Xu, J. Chou, X.S. Zhang, Y. Luo, T. Isakova, P. Adekkanattu, J.S. Ancker, G. Jiang, R.C. Kiefer, J.A. Pacheco, et al., Identifying sub-phenotypes of acute kidney injury using structured and unstructured electronic health record data with memory networks, J. Biomed. Inform. 102 (2020) 103361.
[29] K. Rough, A.M. Dai, K. Zhang, Y. Xue, L.M. Vardoulakis, C. Cui, A.J. Butte, M.D. Howell, A. Rajkomar, Predicting inpatient medication orders from electronic health record data, Clin. Pharmacol. Therapeutics 108 (1) (2020) 145–154.
[30] I. Ahmed, D. Camacho, G. Jeon, F. Piccialli, Internet of health things driven deep learning-based system for non-invasive patient discomfort detection using time frame rules and pairwise keypoints distance feature, Sustainable Cities Soc. 79 (2022) 103672.
[31] D. Arefan, A.A. Mohamed, W.A. Berg, M.L. Zuley, J.H. Sumkin, S. Wu, Deep learning modeling using normal mammograms for predicting breast cancer risk, Med. Phys. 47 (1) (2020) 110–118.
[32] M. Byra, M. Wu, X. Zhang, H. Jang, Y.-J. Ma, E.Y. Chang, S. Shah, J. Du, Knee menisci segmentation and relaxometry of 3D ultrashort echo time cones MR imaging using attention U-Net with transfer learning, Magn. Reson. Med. 83 (3) (2020) 1109–1122.
[33] S. Saha, A. Pagnozzi, P. Bourgeat, J.M. George, D. Bradford, P.B. Colditz, R.N. Boyd, S.E. Rose, J. Fripp, K. Pannek, Predicting motor outcome in preterm infants from very early brain diffusion MRI using a deep learning Convolutional Neural Network (CNN) model, Neuroimage 215 (2020) 116807.
[34] M. Saminathan, M. Ramachandran, A. Kumar, K. Rajkumar, A. Khanna, P. Singh, A study on specific learning algorithms pertaining to classify lung cancer disease, Expert Syst. 39 (3) (2022) e12797.
[35] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016.
[36] S. Kazemifar, A.M. Barragán Montero, K. Souris, S.T. Rivas, R. Timmerman, Y.K. Park, S. Jiang, X. Geets, E. Sterpin, A. Owrangi, Dosimetric evaluation of synthetic CT generated with GANs for MRI-only proton therapy treatment planning of brain tumors, J. Appl. Clin. Med. Phys. 21 (5) (2020) 76–86.
[37] G. Liu, T.-M.H. Hsu, M. McDermott, W. Boag, W.-H. Weng, P. Szolovits, M. Ghassemi, Clinically accurate chest X-ray report generation, in: Machine Learning for Healthcare Conference, PMLR, 2019, pp. 249–269.
[38] F. Piccialli, F. Giampaolo, E. Prezioso, D. Camacho, G. Acampora, Artificial intelligence and healthcare: Forecasting of medical bookings through multi-source time-series fusion, Inf. Fusion 74 (2021) 1–16.
[39] T. Nemoto, N. Futakami, M. Yagi, A. Kumabe, A. Takeda, E. Kunieda, N. Shigematsu, Efficacy evaluation of 2D, 3D U-Net semantic segmentation and atlas-based segmentation of normal lungs excluding the trachea and main bronchi, J. Radiat. Res. 61 (2) (2020) 257–264.
[40] D.C. Benz, G. Benetos, G. Rampidis, E. Von Felten, A. Bakula, A. Sustar, K. Kudura, M. Messerli, T.A. Fuchs, C. Gebhard, et al., Validation of deep-learning image reconstruction for coronary computed tomography angiography: Impact on noise, image quality and diagnostic accuracy, J. Cardiovasc. Comput. Tomography 14 (5) (2020) 444–451.
[41] T.D. Pham, Classification of COVID-19 chest X-rays with deep learning: New models or fine tuning? Health Inf. Sci. Syst. 9 (1) (2021) 1–11.
[42] F. Ahmad, A. Farooq, M.U. Ghani, Deep ensemble model for classification of novel coronavirus in chest X-ray images, Comput. Intell. Neurosci. 2021 (2021).
[43] D. Avola, A. Bacciu, L. Cinque, A. Fagioli, M.R. Marini, R. Taiello, Study on transfer learning capabilities for pneumonia classification in chest-X-rays image, 2021, arXiv preprint arXiv:2110.02780.
[44] T. Zebin, S. Rezvy, COVID-19 detection and disease progression visualization: Deep learning on chest X-rays for classification and coarse localization, Appl. Intell. 51 (2) (2021) 1010–1021.
[45] L.O. Teixeira, R.M. Pereira, D. Bertolini, L.S. Oliveira, L. Nanni, G.D. Cavalcanti, Y.M. Costa, Impact of lung segmentation on the diagnosis and explanation of COVID-19 in chest X-ray images, Sensors 21 (21) (2021) 7116.
[46] A.B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115.
[47] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R.M. Summers, ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2097–2106.
[48] V. Teixeira, L. Braz, H. Pedrini, Z. Dias, Dualanet: Dual lesion attention network for thoracic disease classification in chest X-rays, in: 2020 International Conference on Systems, Signals and Image Processing, IWSSIP, IEEE, 2020, pp. 69–74.
[49] I. Allaouzi, M.B. Ahmed, A novel approach for multi-label chest X-ray classification of common thorax diseases, IEEE Access 7 (2019) 64279–64288.
[50] M.M.A. Monshi, J. Poon, V. Chung, F.M. Monshi, Labeling chest X-Ray reports using deep learning, in: International Conference on Artificial Neural Networks, Springer, 2021, pp. 684–694.
[51] A. Smit, S. Jain, P. Rajpurkar, A. Pareek, A.Y. Ng, M.P. Lungren, CheXbert: combining automatic labelers and expert annotations for accurate radiology report labeling using BERT, 2020, arXiv preprint arXiv:2004.09167.
[52] W. Boag, T.-M.H. Hsu, M. McDermott, G. Berner, E. Alesentzer, P. Szolovits, Baselines for chest X-ray report generation, in: Machine Learning for Health Workshop, PMLR, 2020, pp. 126–140.
[53] S. Jain, A. Smit, S.Q. Truong, C.D. Nguyen, M.-T. Huynh, M. Jain, V.A. Young, A.Y. Ng, M.P. Lungren, P. Rajpurkar, VisualCheXbert: Addressing the discrepancy between radiology report labels and image labels, in: Proceedings of the Conference on Health, Inference, and Learning, 2021, pp. 105–115.
[54] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R.M. Summers, ChestX-ray: Hospital-scale chest X-ray database and benchmarks on weakly supervised classification and localization of common thorax diseases, in: Deep Learning and Convolutional Neural Networks for Medical Imaging and Clinical Informatics, Springer, 2019, p. 369.
[55] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, et al., Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, no. 01, 2019, pp. 590–597.
[56] H. Wang, Y.-Y. Yang, Y. Pan, P. Han, Z.-X. Li, H.-G. Huang, S.-Z. Zhu, Detecting thoracic diseases via representation learning with adaptive sampling, Neurocomputing 406 (2020) 354–360.
[57] S. Albahli, H.T. Rauf, A. Algosaibi, V.E. Balas, AI-driven deep CNN approach for multi-label pathology classification using chest X-rays, PeerJ Comput. Sci. 7 (2021) e495.
[58] K. Almezhghwi, S. Serte, F. Al-Turjman, Convolutional neural networks for the classification of chest X-rays in the IoT era, Multimedia Tools Appl. 80 (19) (2021) 29051–29065.
[59] L. Seyyed-Kalantari, G. Liu, M. McDermott, I.Y. Chen, M. Ghassemi, CheXclusion: Fairness gaps in deep chest X-ray classifiers, in: BIOCOMPUTING 2021: Proceedings of the Pacific Symposium, World Scientific, 2020, pp. 232–243.
[60] J.P. Cohen, M. Hashir, R. Brooks, H. Bertrand, On the limits of cross-domain generalization in automated X-ray prediction, in: Medical Imaging with Deep Learning, PMLR, 2020, pp. 136–155.
[61] M. Hashir, H. Bertrand, J.P. Cohen, Quantifying the value of lateral views in deep learning for chest X-rays, in: Medical Imaging with Deep Learning, PMLR, 2020, pp. 288–303.
[62] E.H. Pooch, P. Ballester, R.C. Barros, Can we trust deep learning based diagnosis? the impact of domain shift in chest radiograph classification, in: International Workshop on Thoracic Image Analysis, Springer, 2020, pp. 74–83.
[63] J.P. Cohen, P. Morrison, L. Dao, K. Roth, T.Q. Duong, M. Ghassemi, COVID-19 image data collection: Prospective predictions are the future, 2020, arXiv preprint arXiv:2006.11988.
[64] Y. Wang, L. Sun, Q. Jin, Enhanced diagnosis of pneumothorax with an improved real-time augmentation for imbalanced chest X-rays data based on DCNN, IEEE/ACM Trans. Comput. Biol. Bioinform. 18 (3) (2019) 951–962.
[65] F. Charte, A.J. Rivera, M.J. del Jesus, F. Herrera, MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation, Knowl.-Based Syst. 89 (2015) 385–397.
[66] H. Salehinejad, E. Colak, T. Dowdell, J. Barfett, S. Valaee, Synthesizing chest X-ray pathology for training deep convolutional neural networks, IEEE Trans. Med. Imaging 38 (5) (2018) 1197–1206.
[67] W. Qu, I. Balki, M. Mendez, J. Valen, J. Levman, P.N. Tyrrell, Assessing and mitigating the effects of class imbalance in machine learning with application to X-ray imaging, Int. J. Comput. Assist. Radiol. Surg. 15 (12) (2020) 2041–2048.

[68] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, et al., Chexnet: Radiologist-level pneumonia detection on chest X-rays with deep learning, 2017, arXiv preprint arXiv:1711.05225.
[69] K.F. Monowar, M.A.M. Hasan, J. Shin, Lung opacity classification with Convolutional Neural Networks using chest X-rays, in: 2020 11th International Conference on Electrical and Computer Engineering, ICECE, IEEE, 2020, pp. 169–172.
[70] Z. Ge, D. Mahapatra, S. Sedai, R. Garnavi, R. Chakravorty, Chest X-rays classification: A multi-label and fine-grained problem, 2018, arXiv preprint arXiv:1807.07247.
[71] Z. Huang, D. Fu, Diagnose chest pathology in X-ray images by learning multi-attention Convolutional Neural Network, in: 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference, ITAIC, IEEE, 2019, pp. 294–299.
[72] K. Wang, X. Zhang, S. Huang, KGZNet: Knowledge-guided deep zoom neural networks for thoracic disease classification, in: 2019 IEEE International Conference on Bioinformatics and Biomedicine, BIBM, IEEE, 2019, pp. 1396–1401.
[73] R. Qin, K. Qiao, L. Wang, L. Zeng, J. Chen, B. Yan, Weighted focal loss: An effective loss function to overcome unbalance problem of chest X-ray14, IOP Conf. Ser.: Mater. Sci. Eng. 428 (1) (2018) 012022.
[74] J. Islam, Y. Zhang, Towards robust lung segmentation in chest radiographs with deep learning, 2018, arXiv preprint arXiv:1811.12638.
[75] S. Reza, O.B. Amin, M. Hashem, TransResUNet: Improving U-net architecture for robust lungs segmentation in chest X-rays, in: 2020 IEEE Region 10 Symposium, TENSYMP, IEEE, 2020, pp. 1592–1595.
[76] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning, PMLR, 2019, pp. 6105–6114.
[77] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.
[78] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.
[79] C. Szegedy, S. Ioffe, V. Vanhoucke, A.A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Thirty-First AAAI Conference on Artificial Intelligence, 2017.
[80] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1251–1258.
[81] V.-L. Nguyen, E. Hüllermeier, M. Rapp, E. Loza Mencía, J. Fürnkranz, On aggregation in ensembles of multilabel classifiers, in: International Conference on Discovery Science, Springer, 2020, pp. 533–547.
[82] F. Chollet, et al., Keras, 2015, GitHub, URL: https://github.com/fchollet/keras.
[83] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, M. Devin, et al., Tensorflow: Large-scale machine learning on heterogeneous distributed systems, 2016, arXiv preprint arXiv:1603.04467.
[84] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[85] G. Bradski, The OpenCV Library, Dr. Dobb's J. Softw. Tools (2000).
[86] F. Charte, A.J. Rivera, M.J. del Jesus, F. Herrera, REMEDIAL-HwR: Tackling multilabel imbalance through label decoupling and data resampling hybridization, Neurocomputing 326 (2019) 110–122.
[87] G. Rimeika, E. Mockiene, et al., Deep learning model for chest X-ray pathology classification performance on an independent spanish dataset, in: European Congress of Radiology-ECR 2020, 2020.
[88] T. Zhang, X. Li, Z. Qu, Lesion attentive thoracic disease diagnosis with large decision margin loss, Biomed. Signal Process. Control 71 (2022) 103202.
[89] U. Kamal, M. Zunaed, N.B. Nizam, T. Hasan, Anatomy-xnet: An anatomy aware Convolutional Neural Network for thoracic disease classification in chest X-rays, IEEE J. Biomed. Health Inf. 26 (11) (2022) 5518–5528.
[90] H. Liz, M. Sánchez-Montañés, A. Tagarro, S. Domínguez-Rodríguez, R. Dagan, D. Camacho, Ensembles of Convolutional Neural Network models for pediatric pneumonia diagnosis, Future Gener. Comput. Syst. 122 (2021) 220–233.
[91] A.I. Aviles-Rivero, P. Sellars, C.-B. Schönlieb, N. Papadakis, GraphXCOVID: Explainable deep graph diffusion pseudo-labelling for identifying COVID-19 on chest X-rays, Pattern Recognit. 122 (2022) 108274.

Helena Liz is an Associate Researcher at Universidad Rey Juan Carlos (URJC) Computer Science Department under project CYNAMON. She did her undergraduate studies in Biology at Universidad Autónoma de Madrid and she has an M.Sc in Bioinformatics and Computational Biology at Universidad Autónoma de Madrid. Currently, she is Ph.D candidate at Universidad Politécnica de Madrid. Her research interests include Deep Learning, AI and Machine Learning applications in medicine, among others. Contact her at: helena.liz@urjc.es.

Dr. Javier Huertas-Tato obtained his PhD in Computer Science at Universidad Carlos III de Madrid under a FPI research grant. Currently, he is working as a Ph.D. assistant lecturer at Universidad Politecnica de Madrid and collaborating with national and international research projects such as CIVIC, FightDIS, and IBERIFIER. His current research topics are disinformation detection, tracking, and countering; machine learning applied to environmental issues; and deep learning techniques such as convolutional networks and transformers.

Manuel A. Sánchez-Montañés received his B.Sc. degree (with honors) in Physics from the Universidad Complutense de Madrid, Spain, 1997, and Ph.D. degree (cum laude) in Computer Science from the Universidad Autónoma de Madrid, Spain, 2003. He is currently working with the Computer Science Department, Universidad Autónoma de Madrid. His research activity is focused in Artificial Intelligence and Advanced Data Analysis, carrying out theoretical developments and applications.

Javier Del Ser received his first PhD in Telecommunication Engineering from the University of Navarra (2006), and a second PhD in Computational Intelligence from the University of Alcala (2013). Currently he is a Research Professor in Artificial Intelligence at TECNALIA (Spain) and an adjunct professor at the University of the Basque Country (UPV/EHU). His research interests gravitate on Artificial Intelligence for data modelling and optimisation tasks in a diverse range of application fields (Energy, Transport, Telecommunications, Health and Industry, etc.). He also serves as an associate editor in a number of indexed journals, including Information Fusion, Swarm and Evolutionary Computation and IEEE Transactions on Intelligent Transportation Systems. He is a IEEE Senior Member, and has been included in the list of the 2% most influential authors in Artificial Intelligence in 2021 and 2022 elaborated by Stanford University.

David Camacho is currently working as Full Professor with the Departamento de Sistemas Informáticos at Universidad Politécnica de Madrid (Spain) and leads the Applied Intelligence and Data Analysis group (AIDA). He received a PhD in Computer Science (2001) from Universidad Carlos III de Madrid. His research interests include Data Mining, Evolutionary Computation, Social Network Analysis, and Swarm Intelligence, among others. Contact him at: david.camacho@upm.es.

PUBLICATION 3

Generation and detection of manipulated multimodal audiovisual content: advances, trends and open challenges

J3: Generation and detection of manipulated multimodal audiovisual content: advances, trends and open challenges". Helena Liz-López, Mamadou Keita, Abdelmalik Taleb-Ahmed, Abdenour Hadid, Javier Huertas, David Camacho. Information Fusion. ISSN: 1566-2535, eISSN: 1872-6305. Vol. xx, pp. 1–44, 2023. Submitted, 26th July 2023. DOI: 10.1016/j.inffus.2023.102103
Impact factor: 18.6 (JCR, 2022) [Q1, 4/145 CS, Artificial Intelligence]

- **Overall contribution**: The main contribution of this article is a comprehensive review of the current state of the art in multimodal multimedia data forensics, specifically focuses on audio and video data (visual and multimodal approach), as described in Chapter 6. It includes manipulation techniques, available datasets, and detection tools. The article also presents a summary of the main challenges and future trends in the field.

- **Contribution of the PhD candidate**:
  - First author of the article.
  - The conception of the idea presented.
  - Elaboration of the methodology.
  - Co-author of the manuscript, figures, and tables.
  - Co-author of the interpretation and discussion of the results provided.

Highlights

Generation and detection of manipulated multimodal audiovisual content: advances, trends and open challenges

Helena Liz-López,Mamadou Keita,Abdelmalik Taleb-Ahmed,Abdenour Hadid,Javier Huertas-Tato,David Camacho

- We provide a comprehensive review of the manipulation techniques, and the available datasets, for multimedia content.
- We present the current state-of-the-art forensics techniques for identifying manipulation in multimodal multimedia content.
- We listed a catalogue of prominent disinformation detection tools that have been implemented for utilization by end-users.
- The main open challenges and trends in the field of forensics for multimedia content are summarised.

Generation and detection of manipulated multimodal audiovisual content: advances, trends and open challenges

Helena Liz-López[a,*], Mamadou Keita[b], Abdelmalik Taleb-Ahmed[b], Abdenour Hadid[c], Javier Huertas-Tato[a] and David Camacho[a]

[a]Computer Systems Department, Universidad Politécnica de Madrid, Calle Alan Turing s/n, Madrid, 28031, Spain
[b]Institut d'Electronique de Microélectronique et de Nanotechnologie (IEMN) Université Polytechnique Hauts de France, Valenciennes, 59313, France
[c]Sorbonne Center for Artificial Intelligence Sorbonne University Abu Dhabi, Abu Dhabi, UAE

ARTICLE INFO

Keywords:
Multimedia data manipulation generation
Multimedia data forensics
Deep Learning
Video
Audio
Multimodal

ABSTRACT

Generative deep learning techniques have invaded the public discourse recently. Despite the advantages, the applications to disinformation are concerning as the counter-measures advance slowly. As the manipulation of multimedia content becomes easier, faster, and more credible, developing effective forensics becomes invaluable. Other works have identified this need but neglect that disinformation is inherently multimodal. Overall in this survey, we exhaustively describe modern manipulation and forensic techniques from the lens of video, audio and their multimodal fusion. For manipulation techniques, we give a classification of the most commonly applied manipulations. Generative techniques can be exploited to generate datasets; we provide a list of current datasets useful for forensics. We have reviewed forensic techniques from 2018 to 2023, examined the usage of datasets, and given a comparative analysis of each modality. Finally, we give another comparison of end-to-end forensics tools for end-users. From our analysis clear trends are found with diffusion models, dataset granularity, explainability techniques, synchronisation improvements, and learning task diversity. We find a roadmap of deep challenges ahead, including multilinguality, multimodality, improving data quality (and variety), all in an adversarial ever-changing environment.

1. Introduction

In recent years, the global use of social networks has increased, enabling individuals to express their emotions and opinions freely. While this proliferation enriches online content, it also facilitates the spread of false information [1]. Information can be classified into two categories based on intent: ***misinformation***, denoting the unintentional spread of false or misleading information; and ***disinformation***, which refers to deliberately disseminated false information intended to deceive and harm others. These terms are relatively modern, having been introduced in 2018 by Ireton and Posetti [2]. Although initially disinformation only affects textual content, however, social networks allow us to use other modalities, such as images, audios or videos, that are becoming very popular on social networks and therefore a greater dissemination [3] of false information for two main reasons: first, multimedia content information is more attractive and impactful, it means that this information is easier to consume and attracts much more attention than textual information [4]; and second, even if the information is false, it gives a greater sense of reliability than news that only includes text because visual content is a direct representation of the reality while words are an abstract form of communication, making information with visual content more credible [5].

The detection of manipulated information is an important part of factuality prediction, a fundamental tool to prevent the spread of misleading or false information. There are different approaches to generate false multimedia information, such as manipulation of multimedia content [6, 7, 8, 9, 10], computer generated content [11, 12, 13], or multimedia content outside the context, that is, multimedia content that does not correspond to text [14, 15, 16, 17, 18]. The dissemination of false information is dangerous regardless of the technique by which it is generated, but information manipulation is one of the most dangerous. Manipulation of multimedia information is not only used in a negative way, we also have examples of a beneficial use of this information [19], such as for education, like videos of historical characters [20]; entertainment, such as in movies where we can find special effects or cartoon characters [21]; or to help people communication, people with diseases like ALS, who could communicate with the help of these fake videos, to express themselves or in speech

*This work has been funded by the project PCI2022-134990-2 (MARTINI) of the CHISTERA IV Cofund 2021 program, funded by MCIN/AEI/10.13039/501100011033 and by the "European Union NextGenerationEU/PRTR"; by the research project DisTrack: Tracking disinformation in Online Social Networks through Deep Natural Language Processing, granted by Mobile World Capital Foundation; by the Spanish Ministry of Science and Innovation under FightDIS (PID2020-117263GB-I00); by MCIN/AEI/10.13039/501100011033/ and European Union NextGenerationEU/PRTR for XAI-Disinfodemics (PLEC2021-007681) grant, by Comunidad Autónoma de Madrid under S2018/TCS-4566 grant, by European Comission under IBERIFIER - Iberian Digital Media Research and Fact-Checking Hub (2020-EU-IA-0252); and by "Convenio Plurianual with the Universidad Politécnica de Madrid in the actuation line of Programa de Excelencia para el Profesorado Universitario".

*Corresponding author
helena.liz@alumnos.upm.es (H. Liz-López); Mamadou.Keita@uphf.fr (M. Keita); Abdelmalik.Taleb-Ahmed@uphf.fr (A. Taleb-Ahmed); abdenour.hadid@ieee.org (A. Hadid); javier.huertas.tato@upm.es (J. Huertas-Tato); david.camacho@upm.es (D. Camacho)
ORCID(s): 0000-0003-4962-6314 (H. Liz-López); 0000-0001-7218-3799 (A. Taleb-Ahmed); 0000-0001-9092-735X (A. Hadid); 0000-0003-4127-5505 (J. Huertas-Tato); 0000-0002-5051-3475 (D. Camacho)

to reach an audience that understands a different dialect from the speaker [22].

While information manipulation can be beneficial in specific contexts, many instances are harmful and negative in the real world. For example, in 2017, numerous pornographic videos were uploaded to Reddit under the pseudonym "deepfake", having a huge impact and threatening the security and privacy of society. The author used a synthetic technique to replace the face of a person in an image or video with the features of another [23]. These techniques can be used to sabotage opponents, threaten or extort people [24], and even damage the reputation of a person, group, or institution [25]. There are two main factors that have enabled the creation and dissemination of false multimedia information [26]: firstly, the rise of social media; and secondly, the development and improvements in Deep Learning algorithms to generate this false information. Advances in deep learning, such as generative adversarial networks (GANs) [27], autoencoders (AEs) [28] and Diffusion models [29], have enabled the manipulation of multimedia data. Manipulation techniques in themselves have neither positive nor negative connotations, depending on how they are applied they will have one objective or another, however throughout the survey we will focus on the application of multimedia data ***manipulation techniques*** with the *aim of misinforming* or harming people or collectives.

We can find several literature overviews related to disinformation [30, 31], which can be analysed from two perspectives [32]: the factuality of information or content, where facutality refers to the truthfulness or accuracy of the information presented, while content refers to the information or message itself. We have decided to focus on the *factuality* of information because of the social and technological moment in which we live. As we have already explained, information with multimedia content tends to reach more people and spread faster. There are several surveys that analyse the factuality of text and image [33, 34], but video has been less studied, and audio is the least analysed modality. Therefore, we decided to focus on *video, audio and multimodality*. Finally, within the analysis of factuality, we have decided to focus on manipulation detection techniques. Typically, ***forensics*** pertains to scientific methods applied to investigate computer systems, networks, and electronic devices, with the aim of uncovering, preserving, analysing, and presenting valid data. However, in this context, we use it to describe techniques for ***detecting multimedia content manipulation***.

The generation of manipulated content can be found in many areas of our lives, such as rumours, media and internet sources. However, due to the increasing use of social networks, these are a great source of manipulated information and also a widely used way to disseminate information from other sources, for all these reasons we decided to focus on them as a source of manipulated multimedia content.

Comparing this survey with others in the field (see Table 1), several key differences emerge. Firstly, while most surveys concentrate on deepfakes, our review is not limited to this topic; instead, it encompasses various manipulation targets. On the other hand, we can observe that state-of-the-art works usually do not include tools already implemented for users outside the field of computer science, so we decided to explore the available tools, which are very useful to limit the spread of disinformation.

Four research questions are formulated to organise the contributions of this review.

- **RQ1**: What are the main topics within the field of multimedia data manipulation detection?
- **RQ2**: Which publicly available datasets are currently used in multimedia data manipulation detection?
- **RQ3**: What DL techniques techniques are used in the multimedia data manipulation detection?
- **RQ4**: What multimedia manipulation detection tools are available for non-expert users?

The main contribution of this article, drawn from these research questions and the process of answering them, can be summarised as follows.

1. It presents an updated picture of video and audio manipulation techniques in social media.
2. It presents an updated picture of available datasets, current techniques for video and audio manipulation, and Online Social Networks (OSNs) with academic APIs for data mining.
3. Presents an updated picture of the audio and video forensics techniques and available trained models.
4. The paper presents the implemented tools for end-users that would be of help to authors interested in conducting experiments or making future advances in this field.
5. It describes the trends, challenges, and research directions that can be pursued in the field of forensics and supports them with the conclusions of the analysis.

Finally, this manuscript is organised as follows: Section 2 explained the methodology used to conduct the survey and the criteria for the selection of articles. Section 3 is a general descriptive analysis of the articles included in this survey. Section 5 describes the main Deep Learning techniques for manipulating multimedia data and open datasets in the state of the art. Section 6 describes and compares the different forensics techniques used in the literature. Section 7 focusses on the applications of these techniques in social media. Section 8 answers the research questions posed, shows the future trends and challenges of this topic, and the final conclusions.

2. Methodology

In this section, we will describe the systematic process used to conduct a survey on articles, based on PRISMA protocol [43] related to forensics techniques in video, audio, or both domains. The process involved retrieving articles from

Table 1
Comparison of similar surveys available.

	Year	Modalities			Manipulation techniques	Manipulation detection	Manipulation detection tools	Methodology protocol	References included	Main Contribution	Limitations
		Video	Audio	Multimodal							
Tolosana et al. [35]	2020	✓	X	X	X	✓	X	-	200	It provides a comprehensive analysis of facial manipulation techniques and the performance of manipulation and detection techniques.	Focus on deepfake only and manipulation techniques.
Ju [36]	2020	✓	X	X	✓	✓	X	-	53	Performance of manipulation detection work is compared.	Focus on deepfake.
Pashine et al. [37]	2021	✓	X	X	✓	X	X	-	22	Comparison of the performance of state-of-the-art neural networks in the detection of manipulation.	Extremely limited analysis with a low number of manipulation techniques and datasets and not very novel algorithms.
Yu et al. [38]	2021	✓	X	X	X	✓	X	-	88	They analyse in detail the temporal inconsistencies and modified generic elements due to manipulation.	Focus on deepfake.
Weerawardana and Fernando [39]	2021	✓	✓	✓	X	✓	X	-	48	Includes traditional Deepfake detection methods.	Focuses on Deepfakes and detection. techniques.
Alam et al. [32]	2021	✓	✓	X	X	✓			152	Includes image, text and network information.	It does not focus only on manipulation and insufficiently deepen the domain.
Malik et al. [40]	2022	✓	X	X	✓	✓	X	-	130	Provides an introduction to the different typical architectures and a list of tools implemented to manipulate samples.	Focus on deepfake.
Zhang [28]	2022	✓	X	X	✓	✓	X	-	99	Includes an overview of the main architectures.	Focus on deepfake.
Masood et al. [41]	2022	✓	✓	X	✓	✓	X	✓	339	It includes a section describing possible malicious agents and an analysis of the evolution of deepfakes.	Focus on deepfake and it does not give sufficient relevance to multimodal analysis.
Comito et al. [42]	2023	✓	✓	X	✓	✓	X	-	101	Includes text, image and social network information modalities.	Focuses on deepfakes and gives more attention to text and images.

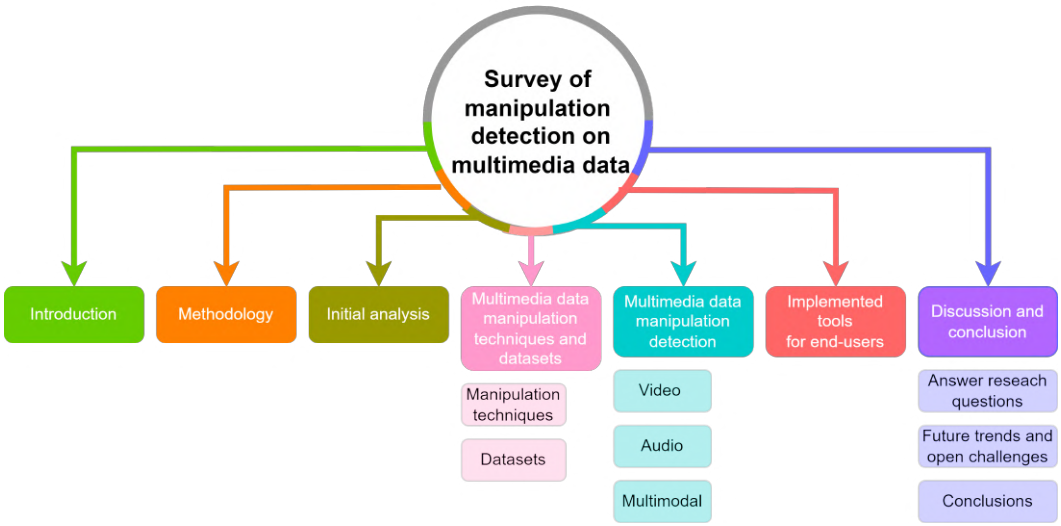


Figure 1: Schematic representation of the structure of the survey.

four prominent scientific databases: Scopus, ScienceDirect, IEEE Xplore, and arXiv.

The steps followed to carry out the manual selection and review process of the articles were as follows:

1. *Search in databases*: We are going to focus on the detection of multimedia content manipulation, but since the number of articles is extremely large and most of the surveys on the subject deal mainly with the detection of image manipulation, we decided to focus on *video*, *audio* and *multimodal(video and audio content)*. Also, thesaurus "social media" was included because they are the main sources of this type of content. Therefore, the thesaurus finally included in the research is: ("Video manipulation detection" OR "audio manipulation detection" OR "multimodal manipulation detection") AND ("social media"). However, other filters were used in this first search, excluding surveys or reviews. In terms of research fields, only Computer Science was selected, as we are interested in studying Deep Learning techniques applied in this field. Finally, we select the date range between 2018 and 2023 (up to May), to analise the most recent and relevant work in this topic.
2. *First screening*: title, abstract, and methodology. We checked the title, abstract, keywords, and methodology to determine whether the articles met the review inclusion criteria. The summary criteria are as follows:
   - The articles must be written in English.
   - The methodology must be clear, including all the necessary details for its implementation, and the results must be clearly explained.
   - The analysis must be quantitative.

Table 2
Articles extracted from different databases about forensics in video, audio and multimodal.

Data source	# articles
ScienceDirect	688
Scopus	51
arXiv	30
IEEE Xplore	26
Total	**795**

- The articles have to apply forensics techniques for extracted multimedia information.

3. *Second screening*: article's content. This was a second more exhaustive review of these articles that carefully read the content of each document and excluded those that confirmed that they did not satisfy the above-mentioned criteria. 729 items were discarded after this step, and we maintain *66 articles* related to the forensics in multimedia data, video, audio, and multimodal.
4. *Analysis of the selected articles and extraction of information*: The final step was the analysis and comparison of the information from the articles of the detection of multimedia data manipulation.

3. Initial Analysis

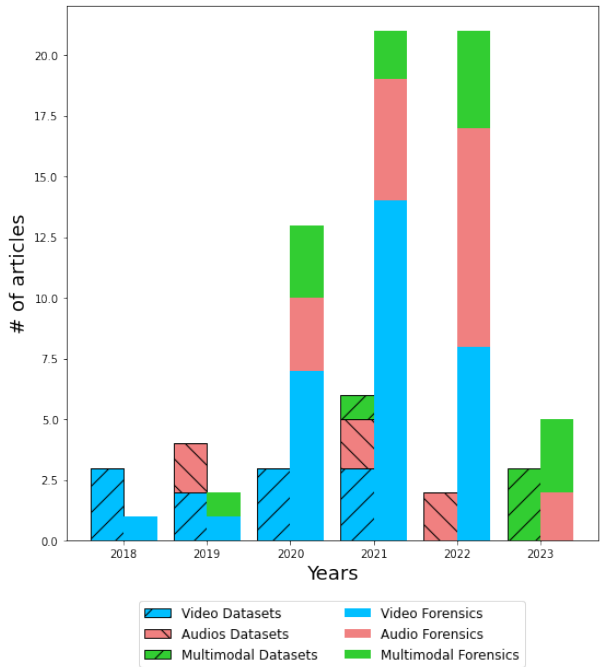


Figure 2: Evolution over the years of articles on manipulation and forensics techniques, divided into the three categories of the study: video, audio, and multimodal.

In this section, we are going to describe the articles to be analysed throughout the survey and the evolution of the field of forensics over the period defined in the methodology, from 2018 to 2023. The aim of this section is to create an overview of the state of the art and thus facilitate the understanding of the next sections.

First, we will compare the number of datasets and manipulation detection works published throughout the study period, which will allow us to analyse the trends of the different modalities analysed in the survey (video, audio, and multimodal); see Figure 2. We can observe how the first years, from 2018 to 2021, are dominated by papers related to video, more specifically with the visual information of the video without including audio. Subsequently, audio and multimodal samples are explored in greater detail, including visual and audio information. However, it is observed that the study of forensics in multimodal data starts to have more relevance in 2021, where new datasets focused on this problem are also being created. In summary, there has been an increase of attention in this area in recent years and an evolution that indicates that multimodality will become more important in this area of study in the upcoming years.

If we focus on Figure 3, where the evolution of the datasets is studied over time, only the datasets that indicate the number of samples in the dataset have been included. Two variables have been used for this analysis: the number of manipulation techniques used, using the y-axis; and the number of dataset samples, using the size of the circle. It can be seen that the datasets tend to use a greater number of manipulation techniques and the datasets are becoming larger and larger, i.e. the datasets show a tendency to increase in complexity. On the other hand, regarding the differences between video, audio, and multimodal, first of all, the same trend as in Figure 2 is observed and it can also be seen that multimodal datasets are more limited than the rest, the number of manipulation techniques used, and the total number of samples is smaller, compared to video and audio datasets. Interestingly, the size of audio datasets are similar, even larger, than video datasets but this has not been reflected in the number of publications of audio forensics techniques, although it is possible that the number of works on audio forensics will increase in the upcoming years.

4. Manipulation techniques

In this section, we present a comprehensive analysis of current state of the art techniques for manipulating audio and videos. We provide an overview of each modality and examine its manipulation methods to gain a deeper understanding of the different approaches used.

4.1. Video manipulation techniques

Video manipulation techniques have become effective instruments for changing and modifying visual content in the last years, opening up both creative opportunities and ethical concerns. These methods, sometimes associated with the term "deepfake", involve modifying videos to obtain incredibly realistic and often false representations.

Through Deep Learning and Machine Learning algorithms, video manipulation techniques can convincingly

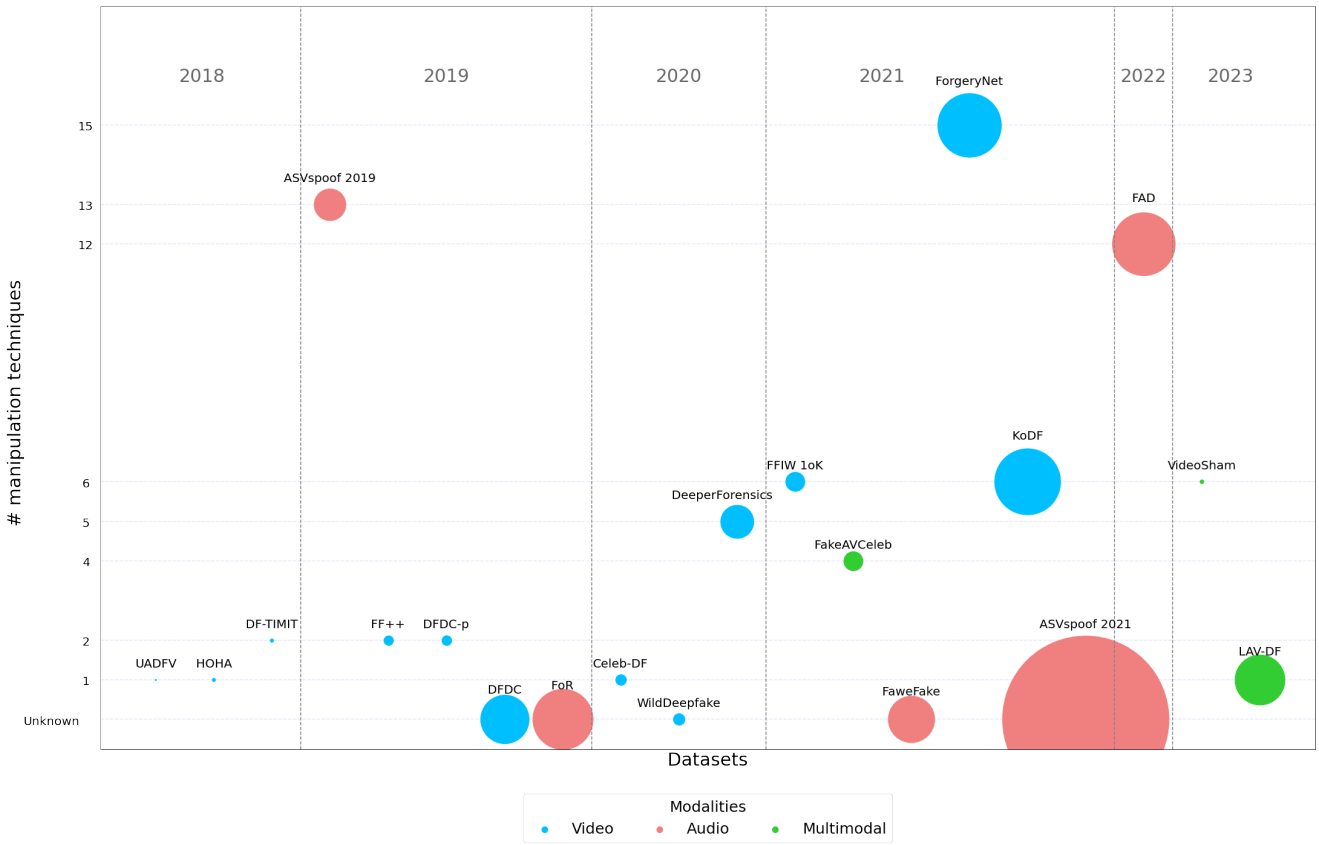


Figure 3: Evolution of the datasets, ordered chronologically. Two variables are compared: the number of manipulation techniques used, marked on the y-axis; and the total number of samples in the dataset, represented by the size of the circles.

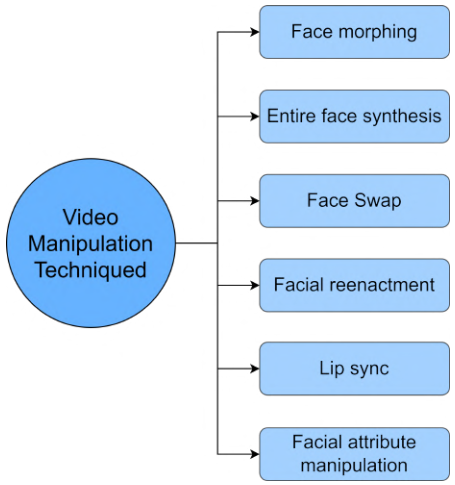


Figure 4: Representation of main video manipulation techniques.

swap faces, change expressions, and even create entirely fabricated scenes, blurring the boundary between reality and fiction. While such techniques offer great possibilities in terms of entertainment, art, and visual effects, they also pose significant problems in terms of privacy protection, disinformation, and the erosion of trust in digital media.

In this section, we will discuss various video manipulation techniques, see Figure 4, including image-level manipulation, face morphing, full-face synthesis, face swapping, facial re-enactment and lip synchronization. These techniques involve altering and modifying videos to manipulate visual content, create realistic faces, swap faces between individuals, generate full-face animations and lip-sync with audio. We will explore the latest technologies, the methods used and give examples of applications.

4.1.1. Face morphing

This technique integrates two or more faces into a single face. Such a morphing results in resembling all of the input faces. Technically, it inherits features from each of the contributing faces. Popular morphological generation approaches use landmark-based techniques [44], where morphing is performed by blending images based on corresponding landmarks. More recent research overcomes landmark limitations by leveraging deep network architectures [45, 46, 47].

Several techniques are employed in the literature to perform face morphing, such as mesh wrapping, field morphing, and radial basis morphing. Nowadays, with the advancement of Deep Learning, significant progress has been made in the use of ***GANs*** for face morphing [45, 47, 49]. For example, Zhang et al. [45] proposed a standalone solution for morph creation by modifying StyleGAN [50] with

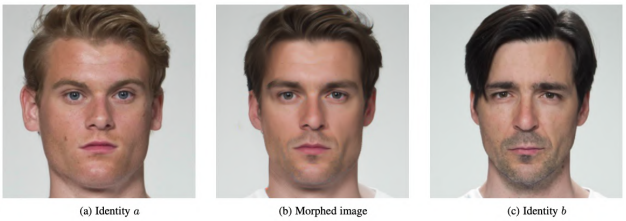


Figure 5: Examples of morphed image generated using Diffusion-based generative model [48].

a newly formulated loss function that takes into account perceptual quality and an added identity factor. Another approach to face morphing is presented by Zhang et al. [51], who introduced MorphGANFormer, a transformer-based method inspired by GANFormer [52]. MorphGANFormer consists of a generator that maps latent space samples to images and a discriminator that distinguishes between true and false images. It uses a bipartite transformer that iteratively propagates information between latent variables and evolving visual features for mutual refinement. The approach also incorporates a latent code to control styling globally or by region, and uses several loss functions to maximise similarity with target face images.

In addition, Blasingame and Liu [48] have proposed a ***diffusion model*** to generate high-quality morphed images. Their methodology uses stochastic and semantic representations of an image, leveraging a diffusion-based process. The architecture includes core DDIM [53] components and a semantic encoder, as well as additional features such as image space processing, interpolation functions and a latent space interpolation space. These elements enable the generation of realistic morphed images of high visual quality.

This technique is not only used to alter faces, with the aim of harming a person, it can also be used in numerous situations, without negatively affecting any individual, for example: graphic effects in animated films and entertainment videos, and it is important to note that it is also misused to create fraudulent identities. Face morphing integrates multiple faces into a single image, yielding realistic results and leveraging techniques such as landmark-based approaches and deep network architectures like GANs. However, articles on diffusion models are beginning to emerge with outstanding results.

4.1.2. Entire face synthesis

This generative technique creates non-existent faces on the basis of learnt high-level attributes from a face dataset. The most popular and widely used approach for face synthesis is StyleGAN [50].

Most of the authors use GANs and Variational AutoEncoders (VAEs). In addition, work using Neural Radiance Fields (NeRFs) and 3D models has also emerged [54, 55]. Sketch-based face synthesis has also been an active area of research. Recent publications have focused on developing methods to generate faces from hand-drawn or digital sketches [56, 57, 58].

Within the ***GAN architecture***, we can find different examples, Karras et al. [59] introduced modifications to StyleGAN, an image generation model, aimed at improving image quality and dealing with artefacts. They proposed a two-part architecture called StyleGAN2. The first part consists of a mapping network that converts a latent representation into an intermediate one. The second part is a generative network that uses convolutions and per-pixel noise to generate images. The intermediate code plays a role in controlling the modulation of the convolution kernel. In addition, StyleGAN2 incorporates skipped connections, blending regularisation and path-length regularisation. Building on these advances, Zhang et al. [60] presented a new GAN model called StyleSwin, which adopts a style-based architecture for high-resolution face synthesis. This model uses a Swin transformer and operates on the basis of a dual-attention mechanism to efficiently capture local and offset contexts. The StyleSwin model consists of a generator that takes a low-resolution image as input and gradually scales it to the desired high-resolution using a series of oversampling layers. In addition, it incorporates a spatial discriminator and a wavelet discriminator to combat blocking artefacts commonly found in generated images.

Figure 6: Face synthesized with StyleSwin [60].

Thanks to advancements in facial synthesis, researchers have ventured into the intricate domain of talking head generation, pushing the limits of computer-generated faces even further [61, 62]. Stypulkowski et al. [63] introduced a ***diffusion model*** to generate speech-driven talking faces by correlating video frames in an auto-regressive manner that restricts inference speed. The technique uses a pre-trained audio encoder to embed audio features during the denoising process. Furthermore, motion frames are incorporated to tackle problems related to unnatural sequences, along with a simple modification in the loss function to preserve lip-sync consistency.

These advances in talking head generation not only impact the realm of disinformation, but also offer diverse *positive applications*, such as the generation of virtual characters for video games, the support of the 3D face modelling industry, and the assistance of face recognition applications by generating additional training data. However, it is crucial to acknowledge that face synthesis can also be used maliciously, such as generating fake photorealistic faces to propagate disinformation on social media.

In the field of synthetic face generation, we can see that the predominant architecture is GAN, but other architectures such as VAEs are also used, and recently the field of diffusion

models is beginning to be explored. The main limitation of this technique is that it generates good results in samples that only include faces, since, being synthetic samples, synchronisation with the original bodies can be complex.

4.1.3. Face Swap

FaceSwap is a technique involving automatically replacing the face of a person in the source video with the face of another person in a target video while keeping the attributes (e.g., expression, posture, lighting, etc.) of the target face unchanged.

Figure 7: Face swapping by TransFS [64].

A general process for face swapping, outlined by Perov et al. [65], includes three primary stages: extraction, training, and conversion. During the extraction phase, various algorithms and steps such as face detection, alignment, and segmentation are used to extract faces from both the source and target data. In the training phase, a model is trained to transfer the source person's face onto the target person's head. Finally, in the conversion step, the original face is placed onto the target head and refined for a seamless fit.

Earlier face-swapping techniques used operations on ***3D models*** [66, 67] or 2D image composition [68]. However, with advances in deep learning, contemporary methods rely predominantly on neural network architectures such as ***GANs, VAEs, and CNNs***. Various publications have explored different neural network structures and automation levels for face swapping, as demonstrated by the work of [69, 70]. Moreover, recent research has explored the possibilities of disentangling latent spaces to swap faces [71, 72]. Additionally, modifications to GAN architectures have been proposed for face swapping, as seen in [59, 73]

In recent years, there has been a surge in the development of ***zero-shot*** and ***few-shot face-swapping*** methods, eliminating the need for initial neural network training. These advances have resulted in the publication of numerous articles [74, 72, 75, 76]. These innovative methods have made face swapping more accessible to the general public. Chen et al. [77] introduced SimSwap, which employs a spatial transformation technique and comprises two fundamental concepts. The first concept is the *identity injection module*, which transfers identity information from a source face to a target face at the feature level. The second concept is the *weak feature matching loss*, which implicitly preserves the attributes of the face.

In an effort to solve the problems encountered in face swapping, Cao et al. [64] proposed a ***transformer model*** that overcomes the challenges related to pose, expression preservation, and high-resolution image quality. Their approach incorporates a Swin transformer-based cross-window face encoder for learning facial features, a transformer-based identity generator to reconstruct high-resolution faces, a patch discriminator to enhance image details, a face conversion module for swapping faces, and feature embedding for improved feature learning and facial reconstruction.

Face swapping technology has evolved beyond its initial use and is now used for entertainment purposes, for example to create humorous or satirical videos by swapping the faces of politicians or celebrities. On the other hand, face swapping can also be used as an alternative to blurring or pixelation of faces in graphic documents.

We can see that face moprphing and face swap are very similar, however they are not the same. The former focuses on the gradual transposition between two different faces, while the latter focuses on replacing the face while maintaining the attributes of the original subject. The main advantage of these techniques is that by modifying the existing face in the video sample, it does not have the synchronisation problems of face synthesis.

4.1.4. Facial Reenactment

Facial reenactment refers to the process of creating a photo-realistic animation of a target video that mimics the facial expressions of a source person. Compared to face-swapping, facial reenactment generates a new visual representation of a person while preserving their identity, while the first modifies an existing visual representation by replacing the person's face with someone else's. Various deep learning techniques are used in face reenactment technology, such as GANs, CNNs, recurrent neural networks (RNNs), AEs and VAEs [78, 79, 80, 81].

Figure 8: A facial reenactment illustration [82].

From the three categories of facial reenactment, we will be interested only in Video-Driven Facial Reenactment, which mainly relies on reconstructing a source and target face with a parametric face model. Bounareli et al. [82] performs face reenactment by disentangling identity features from head pose and expression. Generate two sets of face images and obtains style vectors from them in the pretrained ***StyleGAN2 models*** style space. It then mixes and masks the style vectors to create a new set with target identity and

source pose/expression, which is used to generate new face images with desired pose/expression.

Beyond faces, some recent work has expanded from head to ***full-body synthesis*** [83, 84], where the target's expression along with mannerism is manipulated to create realistic deepfakes. The generated video from the above-mentioned technique can then be combined with fake speeches [85]. In the field of full-body reenactment, Liu et al. [86] introduced a technique to transfer poses from a source video to the target person. A ***3D model*** of the person is reconstructed and a generative model is trained to produce photorealistic frames from images rendered with that 3D model. This method seamlessly transfers poses between different people, creating visually compelling and lifelike results. Furthermore, Chan et al. [83] reenacted full-body movements by learning the mapping directly from the detected poses to generate a video of the target person, applying pix2pixHD [87] with temporal smoothing and GAN structures for faces. This approach performs well in a constrained video domain, where the model is trained on laboratory videos of a single person with a range of poses.

Facial reenactment has a variety of applications, such as editing a participant's facial expression and mouth movement in an online multilingual video conference, dubbing or editing an actor's head and facial expressions in entertainment industry, creating photorealistic animations for games, or virtual reality, etc. Facial reenactment involves creating a photorealistic animation of a target video that mimics the facial expressions of a source person. It uses deep learning techniques like GANs, CNNs, RNNs, AEs, and VAEs for this purpose. Additionally, there have been advances in full-body reenactment, expanding from head to full-body movements.

4.1.5. Lip sync

Lipsync is a manipulation technique that allows us to improve the lip synchronisation of the sample subjects. This technique can have two different functions: to synchronise the movement of the lips in a sequence of frames or to synchronise the video with the audio. Lip-sync is not a generative technique, but a post-processing technique, and it is of great importance in the field of multimedia information manipulation [88]. It allows to reduce the signals or manipulation traces.

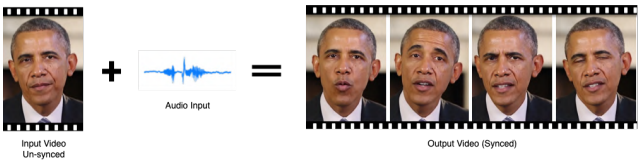


Figure 9: Representation of lip-syncing a video with arbitrary audio.

Several techniques for lip synchronisation have been proposed. A popular approach is to use ***GANs*** [89, 61], which have been applied to synthesise realistic lip-sync videos with high precision and visual quality [90, 91]. Researchers have proposed different models to improve lip-sync accuracy and visual quality. LipGAN, introduced by KR et al. [91], is a face-to-face translation model that generates realistic talking face videos regardless of language while preserving facial pose and expression by taking as input a single image and its corresponding lower half masked as pose prior. It incorporates a generator that generates faces conditioned by audio inputs and a discriminator that checks if the generated face is synchronised with the given audio. The discriminator was trained in conjunction with the generator to discriminate in-sync and out-of-sync audio-video pairs.

In addition, ***RNNs*** were used to capture the temporal dynamics of audio and video data for lip-sync [92] ***AE*** and ***VAEs*** have been applied to encode and decode lip movements, allowing the generation of novel lip-sync videos [93]. In recent years, there has been a significant increase in the development of 3D models for lip-sync [94, 95]. Prajwal et al. [90] proposed Wav2Lip, which improves lip-sync by incorporating a buffer of contiguous frames to efficiently exploit temporal context information and a pre-trained lip-sync discriminator named lip-sync expert for adversarial training. However, it was not successful in generating speech-synchronised head movements. Another modification of Wav2Lip, called AttnWav2Lip, was introduced by Wang et al. [96]. It incorporates spatial and channel attention. This allows the network to learn where to emphasise and suppress in feature maps across channels and spatial axes, resulting in the accurate synchronisation of lip movements with arbitrary speech in the wild.

Unlike the other techniques explained in this section, lip-sync is a post-processing technique that allows us to reduce the modification traces and improve the quality of the samples, making detection tasks more difficult. Therefore, although it is not directly a manipulation technique, it requires careful attention.

4.1.6. Facial attribute manipulation

A technique used to change the specific attributes of the face of a target in a video while maintaining its identity. This includes changing a person's point of view, hair colour, adding/removing glasses, or changing/transferring the head's expressions, age or pose.

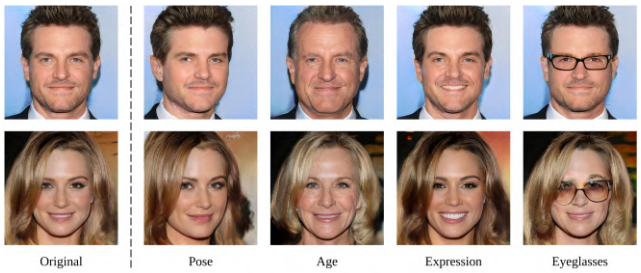


Figure 10: Manipulating various facial attributes. The first column shows the original images, while each of the other columns shows the results of manipulating a specific attribute [97].

GANs or ***VAEs*** are the technologies most commonly used, which undergo modifications in different aspects [97, 98, 99]. In particular, Fard et al. [99] introduced GANalyzer, a technique to analyse and manipulate the latent space of GANs to modify facial attributes in images. The GANalyzer uses a transformation function that operates in the latent space of the GANs to modify attributes such as point of view, hair colour, glasses, and facial expressions. This transformation function is created by analysing a large set of GAN-generated facial images to identify commonalities and using them to modify attributes in a given image. However, a drawback of this approach is the use of standard classifiers to label synthesised images.

As we have seen in other manipulation techniques, this technique can have several positive applications in our lives, including in the entertainment industry to create special effects in movies or in the beauty industry to virtually try on different hairstyles or makeup. However, in this article we focus on its negative use in the field of disinformation. Facial attribute manipulation is a technique used to change specific facial attributes of a target in a video while preserving its identity. The most commonly used technologies are GANs and VAEs, allowing for analysis and manipulation of the latent space to modify attributes. Although this technique has positive applications in the entertainment and beauty industries, its potential for misuse in disinformation is a concern.

In conclusion, manipulation technologies, including Face Swap, Facial Reenactment, Lip Sync and Facial Attribute Manipulation, have undergone rapid evolution thanks to advances in deep learning and artificial intelligence. These techniques have moved from initial 3D approaches and 2D image compositing to relying heavily on deep learning algorithms such as GANs, RNNs and VAEs to achieve strikingly realistic results.

These innovations not only have applications in entertainment, advertising and education, but also raise ethical and privacy challenges, as they can be used to create misleading and manipulated content. As these technologies become more accessible, it is crucial to consider how they are used and to regulate their use to ensure an ethical and safe digital environment. Ultimately, these technologies continue to transform the way we interact with visual content, requiring a balance between innovation and responsibility.

4.2. Audio manipulation techniques

The field of audio manipulation encompasses a wide range of methodologies and approaches, from basic audio editing tools to advanced signal processing algorithms. Such techniques make it possible to modify and transform audio signals, giving researchers unprecedented control over various aspects of sound. By modifying parameters such as pitch, timbre, tempo, and spatial features, audio manipulation techniques facilitate the exploration of acoustic properties and open up new possibilities for audio forensics.

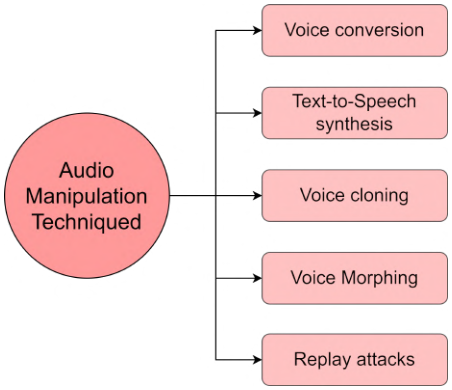


Figure 11: Representation of main audio manipulation techniques.

This section aims to deep dive into the field of audio manipulation, presenting an overview of the methodologies and algorithms used in the field (see Figure 11).

4.2.1. Voice conversion

Voice conversion (VC) is a technique used to modify the vocal features of a given speech of a source speaker to match those of a target speaker. VC methods can be divided into two categories: parallel and non-parallel. It can also be distinguished according to their dependence on text transcriptions. Text-dependent methods require word or phonetic transcriptions, while text-independent methods need no transcription and find similar speech segments themselves to build a conversion function. VC approaches can also be classified according to language: language-independent methods deal with different languages of source and target speakers, while language-dependent methods require the same language of both source and target speakers [100].

Parallel VC involves training models using a parallel dataset where both source and target utterances contain the same speech. This approach has been explored using statistical techniques [101, 102], linear and non-linear algebra techniques [103, 104], signal processing techniques [105], as well as cognitive techniques such as neural networks, classification and regression trees [106, 107], restricted Boltzmann machines [108], and transformer models [109].

In contrast, ***non-parallel VC*** does not require a parallel dataset and learns speaker correspondence without content alignment. Neural networks such as GANs, AEs, and VAEs are commonly used for this purpose [110, 111, 112, 113].

For ***GANs***, Nguyen and Cardinaux [114] introduces an end-to-end architecture that does not require a vocoder. Rather, VC is performed directly on the raw audio waveform, which greatly improves the final speech quality. Regarding AEs, Kim et al. [115] proposed a modification of the encoder and decoder architecture (end-to-end architecture) by changing the layer configuration. And lastly, about VAEs, Kameoka et al. [116] proposed to incorporate an auxiliary classifier into the conversion system in order to encourage the converted speech to be correctly classified as belonging to the source speaker.

One specific approach, proposed by Lin et al. [109] known as FragmentVC, utilises ***self-attention*** and Wav2Vec 2.0 to extract and merge fine-grained speech fragments from the target speaker's utterances based on the latent phonetic structure of the source speaker's utterance. FragmentVC consists of a source coder, a target coder, and a decoder, and it does not depend on disentangling speaker and content. This model has shown good generalisation performance on unknown speakers.

Voice conversion (VC) is a technique used to modify the vocal features of a source speaker's speech to match those of a target speaker. VC methods can be parallel or non-parallel, depending on whether a parallel dataset is used for training. Some of the most used techniques are neural networks like GANs, AEs, and VAEs and have shown promising results.

4.2.2. Text-to-speech synthesis

Text-to-speech is a process of generating speech on the basis of written text using the rules of linguistic description of the text. Such technique aims to generate a synthetic speech that is highly intelligible and indistinguishable from human speech.

Text-to-speech synthesis has evolved over time, starting with ***concatenative synthesis***, which involved assembling pre-recorded speech fragments to create synthetic speech. However, this approach suffered from naturalness issues that were easily detectable by the human ear. Neural network-based speech synthesis techniques have emerged as a more advanced alternative. These techniques have progressed from ***using CNN/RNN*** [117] to ***transformer models*** [118, 119, 120] and from autoregressive generative models [121] to more powerful generative models such as ***VAE*** and ***GAN*** [115, 122, 123]. Furthermore, there has been a shift in the architecture of these models, from cascaded acoustic models/vocoders [124, 125] to end-to-end models [126, 127], with a recent focus on non-autoregressive models [126, 128]. In recent research, Lei et al. [129] introduced Glow-WaveGAN 2, a novel architecture that uses a GAN backbone to extract the latent distribution of speech and regenerate the waveform. They trained a flow-based acoustic model to learn the same latent space from text input, reducing discrepancies between the acoustic model and the vocoder. This approach enables the generation of high-quality synthesised speech without requiring model fine-tuning.

Another notable advancement is the progressive and fast ***diffusion model*** proposed by Huang et al. [130]. ProDiff parameterizes the model by predicting clean data and employs a teacher-synthesised mel spectrogram as the target, reducing data discrepancies and enabling sharp predictions. Unlike previous models that required hundreds of iterations, ProDiff achieves high-quality text-to-speech synthesis with fewer iterations. Furthermore, Jeong et al. [131] introduced Diff-TTS, a non-autoregressive speech synthesis model based on a probabilistic framework. Diff-TTS factorises the model into several transitions using the Markov chain property. It aims to restore a mel-spectrogram from Gaussian noise, with the learning objective focused on minimising likelihood-based optimisation for TTS. The architecture consists of a text encoder, pitch encoder, duration predictor, and decoder, allowing for the extraction of contextual information from phoneme sequences and integrating it with duration and diffusion steps for Gaussian noise prediction.

Text-to-speech synthesis has evolved from concatenative synthesis to more advanced neural network-based techniques such as transformer models, VAEs, and GANs. The focus has shifted to end-to-end models and non-autoregressive approaches, leading to significant improvements in the naturalness and quality of synthetic speech. Recent advancements in diffusion models achieve high-quality results with fewer iterations.

4.2.3. Voice cloning

Voice cloning, an advanced text-to-speech technology, aims to convert a given paragraph of text into speech using the desired voice from a reference audio. While voice cloning is essentially text-to-speech technology to copy the target speaker's voice, two main approaches are being used to this end: speaker adaptation and speaker encoding. Speaker adaptation involves fine-tuning a multispeaker generative model, while speaker encoding trains a separate model to deduce a new speaker embedding, which is then applied to a multispeaker generative model [132, 133, 134].

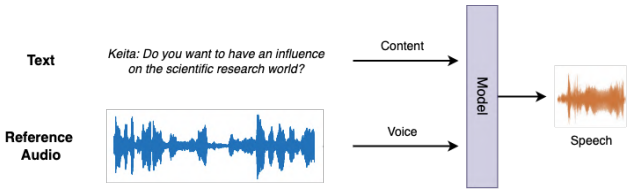


Figure 12: Voice cloning scheme.

In addition to these approaches, ***voice conversion*** is another method used to achieve voice cloning. In this approach, the linguistic information from the source utterance is extracted directly from speech, rather than relying on text [132].

A model proposed by Chen et al. [135], V2C-Net, builds on the widely used ***FastSpeech2 TTS framework***. It involves converting text into speech with the desired voice and emotion derived from reference audio and video. V2C-Net consists of a multimodal encoder, a synthesiser, and a vocoder. The synthesiser generates a mel-spectrogram by extracting features from the multimodal encoder, while the vocoder (HiFi-GAN [125]) converts the mel-spectrogram into waveform. Sadekova et al. [132] proposed a voice cloning and conversion model using a single diffusion model with two encoders and a shared decoder. The model includes a mel encoder, a text encoder, and a decoder. The mel encoder is trained to minimise the mean square error between the generated mel-spectrograms and average voice mel-spectrograms. The text encoder allows voice cloning using a variant of the popular Tacotron2 [120] acoustic feature generator. The

decoder maximises the weighted variational lower bound on data log-likelihood. During speaker adaptation, the decoder is fine-tuned with the target speaker's data, while both encoders remain speaker-independent. The model can perform voice cloning and voice conversion, producing speech of comparable quality to recently proposed algorithms for each task.

In summary, voice cloning focuses on replicating the voice of a specific individual, while text-to-speech is a more general technique for synthesising speech from written text with a focus on naturalness and intelligibility. It allows us to create synthetic audios with the qualities of the voice of a subject we are interested in, which will allow better synchronisation with real fragments of the sample, so it has advantages if the objective is to modify only some fragments of the sample and not to generate it from scratch.

4.2.4. Voice morphing

As the literature review shows, voice morphing, or speech morphing, is often confused with voice conversion, which refers to modifying the speech of a source speaker to match the voice of a target speaker (see Section 4.2.3). In the scope of this review, voice morphing is defined as a technique for algorithmically transforming the voice of one person into that of another, or, in other terms, a technique for smoothly transforming one signal into another. It enables a user to transform a person's voice model into another model with distinct features, giving it a new identity while preserving the original content. Various methods of voice morphing rely on speech parameter interpolation and modelling techniques, such as harmonic-plus-noise model parameters, mixed time and frequency domain methods to alter the pitch, duration, and spectral features, and techniques like Short-time Fourier Transform (STFT), Linear Predictive Coding (LPC), or Sinusoidal Models [136, 137].

4.2.5. Replay Attacks

Replay attack is a type of spoofing in which an attacker targets an automatic speaker verification (ASV) system by using a prerecorded speech sample collected from a genuine speaker they intend to imitate. These attacks, despite their simplicity, have received limited attention in research. They can be classified into two subtypes: far-field detection attacks and copy-and-paste detection attacks. Far-field detection attacks involve replaying a recording made with the victim's remote microphone through a loudspeaker-equipped telephone handset. On the other hand, copy-and-paste detection attacks involve pasting together short recordings to mimic the required sentence for a text-based system [138].

In summary, the field of audio manipulation has undergone a significant transformation, from early concatenative synthesis techniques to advanced neural network-based approaches such as GANs, AEs and VAEs, and powerful generative models such as VAE and GAN. These techniques have revolutionised the way we generate and modify speech, enabling everything from voice conversion to adapt the vocal characteristics of one source speaker to another, to voice cloning to replicate the voice of a specific person. In addition, voice morphology techniques allow us to transform one person's voice into that of another, creating an entirely new vocal identity while preserving the original content.

However, these advances have also presented challenges, especially in the context of security, such as in replay attacks, where attackers use pre-recorded recordings to fool automatic speaker verification systems. Audio manipulation has become an interdisciplinary field that combines linguistics, artificial intelligence and data science to open up a world of creative and applied possibilities. As these technologies continue to evolve, it is crucial to address ethical and security concerns to ensure their responsible and beneficial use in a variety of applications, from the entertainment industry to cyber security.

4.3. Multimodal manipulation techniques

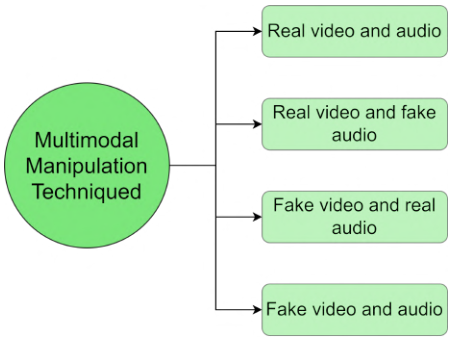


Figure 13: Representation of main multimodal manipulation techniques.

In the previous two sections, we have focused on manipulation techniques for video, visual features, and audio. However, in this section we are going to focus on the combination of the techniques explained above, i.e. *multimodal manipulation techniques*. There are four possible combinations:

- **Real video and real audio**: Both the audio and visual information in the video are real and correspond to unmanipulated samples. Although these samples are real, they are used in manipulated contexts, as discussed by Abdelnabi et al. [139]. However, the focus of this article is on techniques used to alter visual or acoustic information from video and audio samples.
- **Real video and fake audio**: The visual content of the video samples remains unchanged while the audio is modified using one of the techniques explained above. The properties of the video, such as tone, can be modified [140], or the content, that is, the message of the audio sample, can be altered [141]. When one mode is kept constant, the final duration of the sample is fixed, ensuring compatibility between the audio and video modalities [142].
- **Fake video and real audio**: Audio signals are kept constant and visual information is manipulated using

techniques explained in Section 4.1. Similarly to the previous combination, audio remains constant, allowing only manipulation techniques that affect frames to be applied. Techniques such as adding or removing frames [143] cannot be used, as they would cause incompatibility between the modalities [144].

- **Fake video and fake audio**: In this case, both the video and audio modalities are manipulated within the same sample. This allows for a wider range of possible sample alterations and a variety of manipulation techniques. The duration of the sample is not fixed, allowing the increase or decrease of the total duration. These manipulation techniques are the most complex, requiring careful attention to the synchronisation between both modalities [145].

We can see that the possible manipulations using these combinations are almost infinite, but they are not simple. When manipulating one or both modalities, video or audio, inconsistencies are generated between them. These inconsistencies are usually quite significant, to the point that they are used by most detection work in the field. Therefore, the correct *synchronization* between the two modalities is crucial [146]. The most commonly used technique for the synchronisation of both modalities is lip-sync, which we have also explained in the video section. This technique can be used to synchronize both lip movements between different frames and audio with the lips. In addition to lip sync, some authors use diffusion models to resynchronize modalities, video and audio, improving the quality of the samples, such as Bigioi et al. [147], demonstrated that lip and jaw movements can be resynchronized without relying on intermediate structural representations, such as facial landmarks or a 3D face model.

Diffusion models are now being used in some multimedia content manipulation tasks. Some authors use audio to create a video of a person that incorporates the input audio, as shown by Stypulkowski et al. [63]. Other authors have developed diffusion models capable of generating multimodal information, audio, and video simultaneously. Tang et al. [148] developed a diffusion model called CoDi, which can produce various combinations of output formats, including language, images, video, or audio, from different input formats. Unlike current generative AI systems, CoDi can simultaneously generate multiple modalities and is not limited to a specific subset, such as text or images. Although most of the works on multimodal manipulation use combinations of the previously explained techniques, we are starting to find works that take advantage of diffusion models to manipulate or create multimodal content, with outstanding results. As we can see, multimodal manipulation has not been explored as much as audio and video manipulation, since it is an emerging field within multimedia manipulation. It is expected that in the coming years the number of works related to multimodal manipulation will increase, since it provides more information and allows us to analyse a greater number of manipulation techniques.

In this section, the focus shifts to multimodal manipulation techniques, combining both audio and video to create modified samples. Four possible combinations are explored: real video and real audio (unmanipulated samples), real video and fake audio (modifying only the audio), fake video and real audio (modifying only the video), and fake video and fake audio (manipulating both modalities). Synchronisation between audio and video is crucial to maintain consistency, and lip-sync is commonly used for this purpose. Diffusion models are emerging as a promising technique for synchronising multimodal content and generating high-quality manipulations. Although multimodal manipulation is an emerging field, it has great potential for diverse applications, and further research in this area is expected to increase in the future.

5. Dataset of manipulated multimedia content

The creation of high-quality and realistic datasets is extremely important in the field of multimedia data forensics. These datasets play a critical role in providing representative samples of genuine and manipulated content, enabling the development and the assessment of effective detection algorithms and techniques.

The creation of high-quality and realistic datasets is of utmost importance in the field of multimedia data forensics. These datasets play a pivotal role in furnishing representative samples of authentic and manipulated content, facilitating the development and assessment of effective detection algorithms and techniques.

Realistic datasets enable us to simulate diverse and authentic scenarios in which manipulations can occur in audiovisual media. This is crucial, since manipulation techniques are continually evolving and growing more sophisticated. By having access to realistic data, we can train and evaluate the ability of detection systems to handle a wide range of manipulations, thereby enhancing their effectiveness in detecting manipulated content. Furthermore, the quality of the dataset is crucial to prevent bias and ensure fair and impartial outcomes. A high-quality dataset should encompass diversity, incorporating samples from different contexts, geographic regions, cultures, and demographic groups. This approach mitigates inherent biases and ensures that detection algorithms perform uniformly well in identifying manipulations across all scenarios and user profiles.

In summary, constructing high-quality and realistic datasets is critical to advancing video and audio forensics. These datasets form the foundational base for developing more resilient and efficient algorithms capable of addressing increasingly sophisticated manipulations.

All datasets described in this section are public and have a download link (see Table 21).

5.1. Video

In this section, we will focus on video manipulation datasets, that is, we will focus on *visual content*. Table 3 presents a concise overview of the video datasets used in the literature. Next, a short description of each dataset is given:

UADFV [149] is the first dataset for depfake detection. It was provided by the University of Albany in 2018. It contains 49 real and 49 fake Youtube videos. To generate fake videos, the authors used the FakeApp [1] mobile application, where the faces are swapped with Nicolas Cage's face, and each sample has only one person. The videos have a resolution of 294x500 pixels and 11.14 seconds on average. The objective is for fake faces detection using physiological signals, such as eye-blinking. However, this dataset is of very low quality and modifications are easily detectable.

HOHA-based dataset [150] composed of 300 real videos randomly taken from the HOHA dataset and 300 fake videos from different websites, the authors select the HOHA dataset because it contains realistic samples from known movies. Samples usually have a resolution of 360x240 and 24 frames per second.

Deepfake-TIMIT [151] was published in 2018 and has 640 fake videos from 16 pairs of similar people from the VidTIMIT database [152], which has 10 videos per person. For each individual, there are videos of two quality levels: DeepFake-TIMIT-LQ (low quality) with a resolution of 64x64 and DeepFake-TIMIT-HQ (high quality) with a resolution of 128x128. To generate this dataset, the authors used the open source GAN-based approach [2], which is created from the original deepfake algorithm based on AE methods [3]. However, the audio was not modified. The length of each generated video is 4 seconds, and the audio channel is kept unchanged. Another interesting feature of this dataset is that the samples are blurred and the actors appear frontally with a monochrome colour background.

FaceForensics++ [153] was published in 2019 and is currently one of the most famous and widely used datasets in the field. This dataset is an extension of the FaceForensics dataset (2018). It is composed of four subsets: FaceSwap [4], DeepFake [3], Face2Face [154] and NeuralTexture [155]. This dataset has 1000 original videos from Youtube and 3000 manipulated videos using deep learning and deepfake approaches specified in [156]. Like the Deepfake-TIMIT dataset, it has two quality levels: compressed and H264 compressed format, allowing performance to be evaluated for different types of video. A curious observation about this dataset is that it does not have good lip-sync, which is visible in the samples.

Deepfake Detection Challenge (DFDC) [157] Facebook community launched a challenge, called Deepfake Detection Challenge (DFDC)-preview, in 2019 to accelerate the development of new tools to detect deepfakes. The dataset is composed of 1,131 original videos and 4119 manipulated ones, with two unspecified techniques. This dataset was extended to the final version [158] consisting of 100,000 manipulated videos and 19,000 real ones, using various techniques based on face-swap and data augmentation (geometric, color transformation, varying frame rate, distortion, among others).

Celeb-DF [159] was published in 2020 and like FaceForensics is another famous and widely used disinformation dataset. This dataset attempts to overcome a limitation of most of the previous datasets, the visible artefacts. It has 408 original videos and 795 fake ones. The original samples were collected from Youtube and can be divided into two groups: The first has more than a subject per sample, and the second has only one subject per sample. To generate manipulated videos, two deepfake algorithms were refined [160, 161].

DeeperForensics (DF) [162] was introduced in 2020 and contains 50,000 original videos and 10,000 manipulated ones. The authors developed a new VAE, called DF-VAE; in addition, they applied different augmentation techniques, like blur, noise, among others. One of the advantages of this dataset is the variety of appearances in the set of actors and real-world scenarios. All the features explained above have allowed the generated samples to improve the quality of most of the previous datasets.

WildDeepfake [163] was published in 2020 and is considered one of the most complex datasets because it is a real-world dataset composed of 7,314 face sequences from 707 videos recollected from different internet sources. The complexity of this dataset lies in the variety of video types, in terms of activities, resolution, scenarios, manipulation techniques, among others.

ForgeryNet [164] was presented in the ForgeryNet Challenge 2021. This dataset has 2.9 million images and 221,247 videos. He et al. [164] used seven image manipulation techniques and eight for video and added 36 perturbation attacks to increase the complexity of the dataset and make it more similar to a real situation. This dataset contains annotations for four different tasks: image classification, spatial forgery localisation, video forgery classification, and temporal forgery localisation.

KoDF [165] is a Korean dataset published in 2021 with 175,776 fake videos from 403 different participants. Six different techniques have been used to generate the manipulated samples, such as FaceSwap [83], DeepFaceLab [166], FSGAN [167] or First Order Motion Model (FOMM) [168]. After the generation of the fake samples, Kwon et al. [165] applied the Talking Face Head Pose (ATFHP) [134] and WavLip [90] to reduce artefacts and improve fidelity.

FFIW 10K [169] is a large-scale dataset used in the detection of facial manipulation in multi-person videos. This dataset consists of 10,000 videos, including both real and

[1] https://fakeapp.es/
[2] https://github.com/shaoanlu/faceswap-GAN
[3] https://github.com/deepfakes/faceswap
[4] https://github.com/MarekKowalski/FaceSwap/

Table 3
Comparison of available video datasets. Download links for the following datasets can be found in Table 21.

Name	Year	# manipulation methods	# samples			Source	Average duration	Resolution	Visual quality	Reference
			Total	Fake	Real					
UADFV	2018	1	98	49	49	Youtube	11.4 sec	294x500	Low	[149]
HOHA-based dataset	2018	1	600	300	300	Internet	varied	360x240	Low	[150]
DeepFake TIMIT	2018	2	640	640	0	Actors	4 sec	64x64 (LQ) 128x128 (HQ)	Low	[151]
FaceForensics++	2019	2	5,000	4,000	1,000	Youtube	18 sec	480p, 7200p, 1080p	Low	[153]
Celeb-DF	2019	1	6,229	5,639	590	Youtube	13 sec	varied	High	[170]
DFDC-Preview	2020	2	5,250	4,119	1,131	Actors	30 sec	180p, 2160p	High	[157]
DFDC	2020	Unknow	128,154	104,500	23,654	Actors	30 sec	180p, 2160p	High	[158]
DeeperForensics	2020	5	60,000	50,000	10,000	Actors	varied	1920x1080	High	[162]
WildDeepFake	2020	Unknow	7,314	3,805	3,509	Internet	varied	varied	High	[163]
ForgeryNet	2021	15	221,247	121,617	99,630	Internet	varied	varied	High and low	[164]
KoDF	2021	6	237,942	175,776	62,166	Actors	90 sec	1920×1080	High	[165]
FFIW 10K	2021	6	20,000	10,000	10,000	Youtube	varied	varied	High	[169]

manipulated videos. Provides both face-level and video-level labels, with a total of 3.2 million real faces and 1.1 million fake faces from 3.6 thousand annotated persons. The number of unique fake videos in FFIW 10K is ten orders of magnitude higher than other datasets, such as DeeperForensics-1.0, which treats a manipulated video and its distorted versions as different videos. The dataset uses high-fidelity facial manipulation techniques, guaranteed by a model-independent quality assessment network called Q-Net.

5.2. Audio

In this section, our focus will be on audio manipulation datasets, which constitute an underexplored domain within the field of multimedia data manipulation. Thus, it is crucial to create high-quality datasets that enable the development of new systems and models for detecting manipulation signals in audio samples, thus advancing this domain. Table 4 provides a concise overview of the audio datasets used in the literature. Following that, a brief description of each dataset is provided:

ASV spoof 2019 [171] is one of the most used dataset related to audio manipulation, created for the automatic verification of speakers (ASV). This dataset was created from the VCTK base corpus [172], which is composed of 107 different speakers. ASV spoof 2019 has two parts: Logical access (PA), that contains text-to-text and voice conversion attacks; and physical access (PA), where the attackers acquire a recording of the speaker, which is replayed to the ASV system. Both parts are divided into three different sets: training, development, and evaluation.

Fake-or-Real (FOR) datasets [173] was published in 2019 and was composed of more than 198,000, more than 111,000 real English sequences and more than 87,000 fake utterances. Reimao and Tzerpos [173] use the last text-to-speech synthesis methods, such as Deep Voice 3 [174] and Google Wavenet [121]. The real samples were taken from different datasets: Arctic [175], LJSpeech [176], VoxForce datasets [5]. FoR dataset is divided into four versions: for-original, that contains unbalanced voices without any modification or class/gender balancing (195541 samples); for-norm, contains the samples but normalised by gender, class and volume; F2S, includes samples of for-norm but trimmed to 2 seconds; and FR, is a re-recorded version of the previous one, to simulate an attack in which the attacker sends a utterance over a voice channel.

ASV spoof 2021 [177] is the fourth edition in a series of biannual challenges aimed at promoting the study of spoofing and the design of countermeasures to protect automatic speaker verification systems from manipulation. It consists of speech recordings collected from multiple speakers. Training partitions include recordings from 20 speakers (8 men, 12 women), while development partitions include recordings from 10 speakers (4 men, 6 women). The evaluation partitions contain recordings from the same 48 speakers (21 male, 27 female) as the ASVspoof 2019 evaluation partition. Additionally, the dataset for the deepfake (DF) task includes data from the VCTK base corpus and undisclosed corpora.

WaveFake [178] consists of 117,985 generated audio clips (16-bit PCM wav). To generate the samples Frank and Schönherr [178] use two audio datasets: LJSpeech [176] and Japanese Speech Corpus (JSUT) [179] and six different methods: MelGAN [180], Parallel WaveGAN (PWG) [121], Multiband MelGAN (MB-MelGAN) [181], Full-band MelGAN (FB-MelGAN) , HiFi-GAN (HiFi-GAN) [125] and WaveGlow [182]. The authors use two different languages. Although the synthetic samples are of good quality and resemble real samples, each sample includes only one speaker, so the diversity is limited and is not close to reality.

In-The-Wild Audio Deepfake (IWA) [183] is a dataset of politicians and public figures, unlike most datasets that

[5] http://www.voxforge.org/

Table 4
Comparison of available audio manipulation datasets. * The WaveFake dataset contains only manipulated samples. ** The number of samples is not given, but the total time of each class (real or fake) is presented. The download links for the datasets mentioned above are available in Table 21.

Dataset	Language	Condition	# manipulation techniques	# samples			# Speakers	Reference
				Total	Fake	Real		
ASVspoof 2019	English	Clean	13	55,200	15,600	39,000	107	[171]
ASVspoof 2021	English	Clean, noisy	Unknow	1,513,852	130,032	1,383,820	149	[177]
FoR	English	Clean	Unknow	+ 198,000	+ 87,000	+110,000	140 real 33 fake	[173]
WaveFake*	English, Japanese	Clean	Unknow	117,985	117,985	-	+100	[178]
FAD	Chinese	Clean, noisy	12	173,800	-	-	1,024 real, 279 fake	[184]
IWA**	English	Clean, noisy	Unknow	38 h	20,8 h	17,2 h	58	[183]

generate fake samples, this dataset obtains them from different sources such as social networks and strimming platforms. A total of 58 celebrities were used and a total of 38 hours were collected, 20.8 hours of real audios, and 17.2 hours of fake ones. The average of the samples is 23 and 18 minutes, respectively, per celebrity.

Chinese fake audio detection dataset (FAD) [184] is the first public chinense dataset for fake audio detection, together with ASV spoof, are the only datasets where the samples have background noise, which is interesting as it is closer to reality. The dataset can be divided into two versions: clean and noise. The samples belonging to the clean version were collected from five sources from OpenSLR[6]. To generate clean and false, Ma et al. [184] used 11 representative synthetic audio generation methods divided into traditional vocoder-based systems, such as Griffin-Lim [185] or STRAIGHT [186] and neural vocoder-based systems, such as WaveNet [121] or HifiGan [125]. For the samples with noise, five different signal noise ratios between 0 and 20dB were applied to generate them.

5.3. Multimodal video and audio

Lastly, we will outline the published multimodal datasets, incorporating both audio and video components. In contrast to the previous sections, these datasets examine acoustic and visual features jointly rather than analyzing them independently. Such datasets offer richer information and facilitate generalization to a broader range of modified samples compared to their predecessors. Table 5 provides a succinct overview of the multimodal datasets employed in the literature. Subsequently, brief descriptions of these datasets are presented:

FakeAVCeleb [142] was published in 2021. It is a multimodal dataset that involve manipulation in audio and video with accurate lip-syncing, resulting in four combinations: real audio and video, real audio and fake video, fake audio and real video, and fake audio and video. The videos were collected from youtube and Khalid et al. [142] used different algorithms to manipulate the videos, audios or both. To manipulate the videos they used Faceswap [187] and FSGAN [167], for audio they used a transfer learning-based real-time voice cloning tool (SV2TTS) [188] and finally after generating fake videos and audios, they applied Wav2Lip [90] to a perfect lip-synced.

VideoSham *[144]* consists of 826 videos, divided into 413 real videos and 413 manipulated videos. Unlike other existing deepfake datasets, VideoSham focusses on manipulations beyond faces, including changes in background context, text, audio, aesthetic edits, entity addition/removal, and temporal edits. It uses a combination of six different attacks to manipulate videos. These attacks are classified into spatial, temporal and geometric manipulation techniques. Examples of spatial attacks include copying and moving objects in the video, while temporal attacks involve changing the temporal sequence of events in the video. Geometric attacks refer to modifications in the geometry of objects in the video.

Localized Audio Visual DeepFake (LAV-DF) [145] is an audiovisual dataset specifically designed for the detection and localisation of temporary deepfakes. First, Speech-to-Text techniques are used to modify the audio information, and then voice and face reenactment is applied. The dataset contains more than 130,000 samples, of which more than 90,000 are fake. The LAV-DF dataset is composed of videos containing real segments and fake segments generated by audio and video manipulation techniques. The fake videos were created by modifying the real segments, allowing a comparative analysis of the differences between the two. The dataset includes key features such as sentiment score, sentiment changes, length of fake segments, length of videos, and proportion of fake segments. These features provide important information for analysis and deepfake detection. Unlike most audio manipulation datasets, this one does not focus solely on the classification of the samples but on the location of manipulation signals within the samples.

6. Multimedia data forensics

In this section, we will focus on one of the main aspects of disinformation, the factuality of multimedia, video, and audio content. Within the analysis of the veracity of multimedia content, we will focus on the manipulation traces. The spread of manipulated information has become a major problem in social networks. As a result, more and more

[6] http://www.openslr.org/12/

Table 5
Comparison of available multimodal manipulation datasets. Table 21 shows the download links for the different datasets.

Dataset	Year	# samples			Subject	# manipulation techniques	Average duration	Source	Reference
		Total	Fake	Real					
FakeAVCeleb	2021	20,000	19,500	500	500	4	7,8 sec	VoxCeleb2 dataset	[142]
VideoSham	2023	826	413	413		6	8 sec	Vimeo	[144]
LAV-DF	2023	136,304	99,873	36,431	153	1	0,64 sec	VocCeleb2 dataset	[145]

researchers are developing new systems, architectures, and frameworks to detect this content and ultimately fight against disinformation. Although this may seem surprising as a result of the relationship between the two modalities, most authors still work with these modalities independently. All download links for models and datasets not explained in this article appear in Tables 22 and 23.

6.1. Techniques for video forensics

The first modality that we are going to analyse is video. This type of data can be analysed from different approaches, Figure 14: **visual techniques**, a set of techniques that are based on the information of the frames independently without using the temporal information of the samples; **visio-temporal techniques**, unlike the previous technique, it takes advantage of the temporal information of the videos looking for, for example, inconsistencies between frames; and finally **metadata techniques**, unlike the previous techniques, it will not use the information of the video content but the metadata of the files, such as compression, which will show traces that the file has been altered.

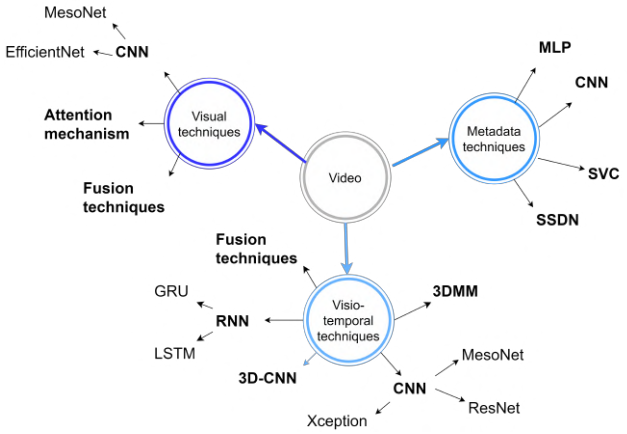


Figure 14: Representation of main video forensics techniques.

6.1.1. Techniques for detecting manipulated metadata

Manipulation often leaves traces in the samples, such as inconsistencies in the texture or compressed video information. In this section, we will focus on techniques developed to detect traces of manipulation in metadata. High-definition videos need compression for efficient storage and transmission. *High-Efficiency Video Coding (HEVC)* is the latest video codec for high-definition video compression, reducing spatial and temporal redundancy. When high-quality videos are manipulated, a second HEVC compression is inevitable, and these traces left by manipulation techniques become the focus of several projects. Uddin et al. [189] proposed using double compression detection as an efficient way to validate video forensically, especially in high-efficiency video. They introduced a novel methodology based on frame partition information to distinguish between single and double compression. This model extracts several statistical and convolutional features from the neural network. The extracted features are then combined and classified with an SVM. Hong et al. [190] developed a system to detect frame modifications in HEVC-coded video within the compressed domain. The proposed methodology consists of two modules, as illustrated in Figure 15: the first module extracts features from the compressed information, and the second module classifies the integrity of the video based on these extracted features. The first module focuses on specific changes in the coding pattern. The model used in the classification step is an MLP, composed of three fully connected layers. Zhang et al. [191] proposed a self-supervised decoupling network (SSDN), that incorporates compression irrelevance. This system learns two separate feature representations for the suspect video: authenticity and compression. Then a joint self-supervised strategy is used for the decoupling of the features, to decouple features by similarity, and similarity learning of authentic features is used. On the other hand, the model performs adversarial decoupling, training the SSDN model in an adversarial way for robust learning. The key of this system is to decouple the relevance between facial authenticity and compression to learn an intrinsic feature representation for robust facial forensics. Finally, the compression ratio and authenticity will be classified independently. Huamán et al. [192] presented a new technique to detect attacks against various video format files, affecting both integrity and authenticity. The system analyzes the video structure generated by mobile devices. At first, they used an atom extraction algorithm, and the duplicity of atoms and the existence of child atoms were verified. Finally, your behaviour is analysed when it is shared on social networks.

This approach, which emphasizes metadata information over traces of manipulation in the content, is underrepresented in the field of study, but it should not be overlooked. Metadata analysis can be one way to control the dissemination of manipulated multimedia content. This section focusses on a completely different approach; instead of focussing on the features extracted from the video content, it focusses on the properties of the video. Another striking detail is that there are no papers that combine both sources of information. We can find several examples that combine

Table 6
Summary table of the different articles related to video forensics techniques based on *metadata information*. Download links to the datasets not included in Section 5 but used in these papers are listed in Table 22.

	Year	Methodology	Dataset available	Code available	Task	Main contribution
Hong et al. [190]	2019	CNN, SVM, statistical and visual features, Fusion Learning	✓	X	Binary classification	Classification system based on coding patterns using MLP
Huamán et al. [192]	2020	Atom extraction algorithm, video structure	✓	X	Binary classification	Novel system, for different video formats, based on the structure of video containers and behaviour in social networks.
Zhang et al. [191]	2021	Self-supervised, Ensemble, EfficientNet-B2,	✓	X	Binary classification	Ensemble-based system, SSDN, focus on compression ratio and authenticity
Uddin et al. [189]	2022	CNN, SVM, statistical and visual features, Fusion Learning, Picture partitioning	X	X	Binary classification	Compression level classification system, between single and double compression, focussing on statistical and visual features from frame partitioning infromation

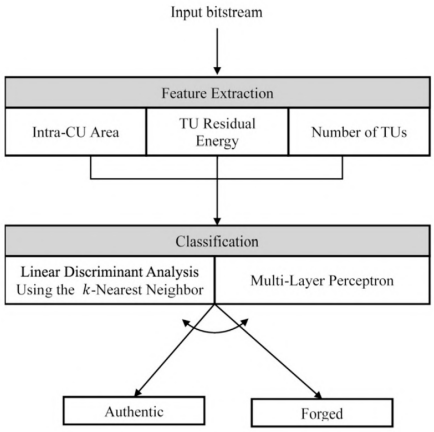


Figure 15: Overall structure of the proposed forgery detection system Hong et al. [190].

frame and video level information as well as metadata and video or frame information. A concise summary of video manipulation techniques based on metadata information is presented in Table 6.

6.1.2. Techniques for detecting manipulated video

In this section, we will present a comprehensive overview of video forensics techniques that emphasize frame features. These authors analyze *distinct frames* of the video *independently*. However, this approach to video forensics has a drawback: it doesn't leverage the available temporal information within video files. Nevertheless, it enables the analysis of samples using image detection systems that have shown success in recent years. Certain authors try to classify videos as real or fake solely by retraining widely used state-of-the-art CNNs. For example, Pokroy and Egorov [193] retrained various EfficientNet architectures (B0-B7) to differentiate frames extracted from the samples as real or fake. Mitra et al. [9] proposed a CNN-based method for detecting manipulated videos. The system comprises three steps: *preprocessing*, which involves: 1) dataset preparation, combining different datasets; 2) extraction of key video frames, where start and end frames of sequences within the original sample are selected to reduce computational cost; and 3) data preprocessing, involving detection, cropping, normalization, and resizing of faces in various frames for input into the CNN. Face samples are then analyzed using Xception [194], excluding the classifier. Finally, the samples are classified using the last module, which consists of a classifier composed of a 2D Global Average Pooling layer, two dropout layers, a dense layer, and a softmax activation function. Although it is a very simple approach, it still achieves competitive performance.

One of the traces that manipulation techniques can leave in a video is the *inconsistency of the texture*. Some researchers explore additional sources of information beyond the image itself, with texture being the most analysed. Kingra et al. [195] developed a novel approach to detect deepfakes by examining inconsistencies in texture information. Specifically, they analysed the local binary pattern (LBP) of faces using a CNN. LBP is notable for its tolerance to illumination variations and is based on comparing each pixel with its surrounding ones. The classification model comprises three steps: preprocessing, where the face is detected and extracted; feature extraction using the LBP technique; and the third step involves CNN training from scratch. A significant advantage of this model is its tolerance to different compression levels and types of manipulation.

Guo et al. [196] introduced a novel architecture, AMTEN-net, which integrates an adaptive manipulation traces extraction network (AMTEN) and a CNN. The preprocessing module, AMTEN, is based on a CNN and learns manipulation traces during back propagation. The output of the AMTEN module is processed by a CNN built from scratch, consisting of three blocks and two dense layers with a softmax activation function.

In contrast, Kim and Cho [197] proposed an innovative approach based on convolutional neural networks that combines frame information with trace features. The model employs two feature extractors to focus on content features and trace features from faces. The first extractor is a pretrained model applied with fine-tuning, while the second one relies on the local relationship between neighbouring pixels. Initially, a multichannel constrained convolution is used, causing colour and contrast to disappear, leaving outlines and traces. Subsequently, the information from both extractors is fused with an average pooling layer, and a fully connected layer produces the final probabilities. An advantage over most state-of-the-art systems is the application of

explainable AI techniques to display the features used in the classification system.

As we can see, other techniques used for video forensics or classification are *ensemble learning* or *fusion techniques*, which combine information obtained from different sources to arrive at the final result. Xu et al. [198] propose a new architecture based on ensemble techniques, called Set Convolutional Neural Network (SCNN). This architecture has three path function paths: the backbone path, the set reduction path, and the discriminative path. They used two common architectures for the backbone path: MesoNet [199] and XceptionNet [194]. The reduction path merges the information from frame-level to set-level and from low-level to high-level. The discriminative path concatenates the information from the two previous paths and generates discriminative features that are the basis of the classifier. Yu et al. [200] developed a new deep forgery detector, called Patch-DFD. This system applied a patch-based solution in facial patch mapping (FPM) with the objective of obtaining several part-based feature maps, keeping the original details of each facial patch to the greatest extent. In addition, the BM-pooling module fixes the size of the feature maps and reduces the quantisation errors. Finally, the local voting strategy merges the results of the part detectors. The advantage of this scheme is efficiency due to the absence of repeated input. Mazaheri and Roy-Chowdhury [201] proposed the facial expression forensics (EMD) framework. This system is composed of two different modules. The first one, facial expression recognition (FER), extracts the relevant information from the facial expressions, and the feature maps provide information about facial regions that encode the expression. The second module is an encoder-decoder architecture used in forensics. The encoder generates a lower dimensional space, where the features from the FER module are combined, and the decoder gives the final result. Another strength of this work is generating heatmaps to know which facial regions the decision is based on. Chen et al. [202] developed an end-to-end framework for detecting and localising manipulated faces, named DLFMNet. This framework combines face detection and face forensics into one model; unlike most models, it does not apply pre-processing techniques to crop faces or re-extract features. This framework is composed of an RGB feature extractor, noise feature, Face Detection Branch (FDB), Manipulation Classification Branch (MCB) and Manipulation Localisation Branch (MLB). In the FDB module, the authors used RetinaNet [203] in each feature map, and then a face prediction head is used in each scale feature map to classify and bound with the box of the manipulated face. In the MCB module, given the boxes of the proposal, they extract feature maps using ROIAlign [204], the information is combined by compact bilinear pooling to fuse both modalities. Then, a fully connected layer is applied to predict the final probabilities.

Other authors have focused on another approach to detect manipulation in videos, the *attention mechanisms*, based on the idea of using only the relevant features of the input. Dang et al. [205] developed a system capable of discriminating between manipulated and real regions within the same sample, at the local level. Initially, the samples were processed using two convolutional networks, Manipulation Appearance Model (MAM) and regression-based methods. The output of both models was combined and served as input for an attention layer. This layer improved classification performance and generated attention maps showing the modified areas of the image. The authors also introduced a new metric, Inverse Intersection Noncontainment (IINC), and a new dataset. Qian et al. [206] presented a novel frequency in the face counterfeiting network, F^3 -Net. This system is composed of two frequency-aware branches: the first one tries to learn subtle forgery patterns through Frequency-aware Image Decomposition (FAD); the second one extracts high-level semantics from Local Frequency Statistics (LFS), which describe the frequency-aware statistical discrepancy between real and fake samples. Both branches are gradually fused using a cross-attention module. The features obtained by both architectures were analysed using two different CNNs, Slowfast-R101 [207] with modifications to suit the problem, the information from both branches was combined using a pooling layer, and the final result was obtained. Zhao et al. [208] proposed a novel architecture, Figure 16, with three main components: multiple spatial attention heads to make the network pay attention to local regions; textural feature enhancement block, to detect artefacts in shallow features; and aggregate low-level textural feature and high-level features guided by attention feature maps. It can be seen that in this work they have combined different approaches: the use of texture information, attention mechanisms, and ensemble techniques to take advantage of different sources of information.

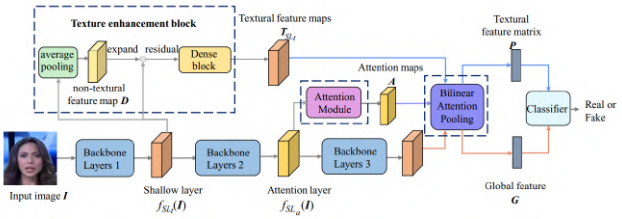


Figure 16: Framework of Zhao et al. [208].

In visual techniques for detecting manipulation in videos, using information extracted from frames without considering the video's temporal aspects, authors often explore textural features to identify inconsistencies. Regarding the most prevalent techniques and architectures, CNNs and attention mechanisms consistently deliver strong performances. Additionally, ensembles remain a viable option for improving system performance. A concise overview of video manipulation techniques based on visual information is provided in Tables 7 and 8.

Table 7
Comparative table of video forensics works based on *visual features* analysis and the datasets used. [1]: Data will be made available on request. [2]: the dataset is available in the article or in a reference given in the article. Table 22 lists the download links for datasets used in these papers, which are not included in Section 5.

	UADFV	DF-TIMIT	HOHA-based	FF+*	DFDCp	DFDC	Celeb-DF	DeeperForensics	WildDeepFake	ForgeryNet	FakeAVCeleb	KoDF	Dessa [7]	VoxCeleb	AffectNet	Other dataset
Dang et al. [205]																X[2]
Qian et al. [206]				X												
Pokroy and Egorov [193]						X										
Chen et al. [202]				X												
Guo et al. [196]																X[2]
Kim and Cho [197]				X												X
Mitra et al. [9]				X		X										
Xu et al. [198]		X		X	X											
Zhao et al. [208]	X			X			X									
Yu et al. [200]		X		X			X									
Mazaheri and Roy-Chowdhury [201]				X											X	
Kingra et al. [195]				X		X	X	X								X[1]

Table 8
Summary table of the different articles related to video forensics techniques based on *visual information*. Download links to the codes used in the following papers can be found in Table 23.

Article	Year	# datasets	Methodology	Code available	Task	Main contribution
Dang et al. [205]	2020	1	CNN, Attention mechanism	✓	Binary classification	Verification system by comparison with real subjects Novel metric, Inverse Intersection Noncontainment
Qian et al. [206]	2020	1	CNN, Colaborative Learning, F^3−Net	X	Binary classification	System based on frequency domain based on two own modules
Pokroy and Egorov [193]	2021	1	CNN, EfficientNet	X	Binary classification	Establish a baseline with a single pretrained CNN
Chen et al. [202]	2021	1	CNN, Ensemble,	✓	Binary classification, Localisation	Manipulation localisation system based ensemble of frame and noise from frame
Guo et al. [196]	2021	1	CNN, AMTENnet	✓	Binary classification	Novel preprocessing model, Adaptative manipulation trace extraction network
Kim and Cho [197]	2021	1	CNN, Ensemble, XAI	X	Binary classification	Ensemble combining frames with trace features
Mitra et al. [9]	2021	2	CNN, Xception	X	Binary classification	Lower computational requirements
Xu et al. [198]	2021	3	CNN, Ensemble, MesoNet, Xception, SCNN	X	Binary classification	Novel system, Set Convolutional Neural Network, based on ensembles
Zhao et al. [208]	2021	3	Multiattention Network, Texture enhancement	✓	Binary classification	Novel Multiattention Network based in local information and texture
Yu et al. [200]	2022	3	CNN, Facial Patch Mapping, Patch-DFD, XAI	X	Binary classification	Ensemble combining patches and frames
Mazaheri and Roy-Chowdhury [201]	2022	2	CNN, Xception,XAI	X	Binary classification, Localisation	Manipulation localisation system with XAI
Kingra et al. [195]	2022	5	CNN, Local Binary Pattern, LBPNet	X	Binary classification	Novel model, LBPNet, based on facial texture irregularities

6.1.3. Techniques for detecting manipulated video-temporal continuity features

Videos have a unique feature that differentiates them from images: *temporal continuity*. The video is composed of multiple frames, these frames present continuity and a correlation with the adjacent ones. Manipulation can alter these two properties of the video, causing temporal incoherence in the samples. The current techniques used to manipulate the video have improved their performance and many may not leave visual traces to the naked eye. However, they can generate temporal inconsistencies, such as the displacement of objects and eye blinks, among others. This temporal information can be extremely important for detecting manipulated signals.

As we have seen in the previous section, *CNNs* are one of the most widely used architectures and continue to perform very well in different computer vision problems, such as manipulation detection. CNNs are capable of obtaining good results in the analysis of individual video frames, so Tran et al. [209] developed 3D-CNNs that are capable of factoring 3D convolutional filters into separate spatial and temporal components, obtaining a significant improvement in accuracy. Many authors have used this architecture to detect video manipulation. Although it is a rather basic approach, like CNNs in frame processing, it gives very good results. Nguyen et al. [210] proposed a 3D convolutional neural network (3D-CNN) capable of extracting spatiotemporal features from short video clips, composed of 16 frames.

The 3D convolution multiplies a 3D kernel with successive stacked frames. To capture the spatio-temporal information from consecutive frame sequences, the feature maps in a convolutional layer are connected with the previous one's successive frames. Das et al. [211] tries to learn discriminative features based on the evolution of faces to detect manipulation in videos. For this, they have used 3D ResNet adding after each convolutional block an attention layer, the CNN was pre-trained with the Kinetics-400 [212] dataset and fine-tuned with FF++. To create the samples with which they train CNN, they detect and crop faces based on easy landmarks, using the method of Bulat and Tzimiropoulos [213], when trimming the sample, they leave a frame of pixels external to the face to avoid losing relevant information for this task. Most works directly use the information from the videos; however, other authors extract the low-level features. For example, Zi et al. [163] proposed two new deepfake detection networks based on Attention mechanism, ADDNets (ADDNet-2D, ADDNet-3D). These networks exploit the facial landmarks extracted by a facial landmark detector to generate an attention mask to process the low-level features of a face. Finally, the low-level features are classified with 2D CNN for image-level detection and with a 3D CNN for video-level, and this network concatenates and reshapes the masks obtained from the attention module before processing them with the 3D CNN. Remarkably, in the comparison of both networks, when both networks are compared, ADDNet-2D showed better results, although both performed better than the state-of-the-art in most datasets.

There are several successful works, such as those mentioned above, that show the outstanding performance of 3D-CNNs, but many authors are in favour of continuing to use *2D-CNNs* to extract information from individual frames and then process them. They then process them *with other architectures*, mainly Recurrent Neural Networks (RNN), such as Gated Recurrent Unit (GRU) or Long Short-Term Memory (LSTM). The main differences between the two architectures is their complexity, where GRU is a simpler architecture and is more focused on simpler tasks, while LSTM is more complex and performs better on complex tasks. Another difference is the internal memory, which LSTM has because it has output gates (it has three gates), while GRU has no internal memory and is composed of two gates. Chamot et al. [214] have developed a system to classify true and false videos made up of two modules. The first module involves a convolutional neural network, which analyses the samples at the frame level; and the second module comprises an RNN that assesses the output of the CNN at the video level. The output of this RNN is a sequence vector utilised by the classifier to reach the final decision. To get closer to reality, Chamot et al. [214] have performed experiments with clear samples and applied perturbations, making classification more difficult. The CNNs used were MesoNet [199] and EfficientNet [193] and the selected RNN was LSTM. Finally, they applied explainable AI techniques to check which areas of the image were used in the classification, which, in addition to eliminating the problem of black box algorithms, allows us to know if there is any bias in the system. Chintha et al. [215] presented a novel loss function that is more efficient in deepfake detection. This loss function was applied in the XcepTemporal model [216], which combines Xception [194] with a convolutional recurrent neural network [150, 217]. The architecture is composed of convolutional blocks from Xception, two bidirectional LSTM and finally a dense layer as the classifier. This methodology allows the model to focus on the low-level feature space by complementing the RGB channel information with a combination of edge and dense flow maps.

Montserrat et al. [218] presented a novel architecture that combines a CNN with a recurrent neural network (RNN). This architecture automatically detects the most important frames using a weighting mechanism with a GRU, which gives the final probability. At first, they used MTCNN to detect and crop faces, then EfficientNet-B5 [219] is used to extract features from faces, to improve feature extraction, this architecture was combined with additive angular margin loss, ArcFace [220], instead of a regular softmax and cross-entropy loss. ArcFace reduces the intraclass difference and enlarges the interclass differences. The extracted features from CNN were analysed with an automatic weighing mechanism to emphasise the most interesting regions. To combine the information from all regions and frames, they used an RNN on top of the automatic face weighting. Finally, the GRU combines the features, logits, and calculated weights of all face regions to obtain the final estimation. Fernando et al. [221] proposed a Hierarchical Attention Memory Network (HAMN) framework for deepfake detection. Initially, they utilize the pre-trained ResNet [222] to extract facial features. The extracted embeddings are then restructured into a sequence, serving as input to a bidirectional GRU [223] to map relationships. This bidirectional GRU is employed to extract informative embeddings, forming the vector used to query the memory. The memory module provides the output information regarding the authenticity of the face.

Although RNN is the architecture most widely used for combining information from individual frames, other authors use other techniques such as *channel-wise spatio-temporal aggregation*. Lu et al. [224] proposed the Channel-Wise Spatiotemporal Aggregation (CWSA) module to fuse the information of continuous frames without any recurrent units. Initially, the videos undergo pre-processing, during which the face is detected and cropped, maintaining a margin where the background is visible. Subsequently, the deep features of consecutive frames from various channels are fused using the CSWA module and analysed using EfficientNet B0. They also show how the skip connection preserves low-level features useful for detection.

Some models do not focus on frame information, but use other features extracted from the frames, such as *biometrical features*. These are based on measures and characteristics unique to each individual. Cozzolino et al. [225] they have developed a methodology, ID-Reveal, for deepfake detection based on biometric features. The facial replacement

techniques provoked a disconnection between visual identity and biometrical features, the facial reenactment keeps the visual identity but there are incoherences in biometrical features. The main advantage over other techniques is that it does not need manipulated videos for training, it only needs real videos. The goal is for the system to detect deepfakes independently of the techniques used for their generation, as many of the state-of-the-art techniques focus on one or a limited set of techniques. The system can be divided into three modules: extraction of compact representations of each frame using the 3D Morphable model (3DMM) [226]. These extracted features serve as the input for the Temporal ID network, which computes an embedding vector. A metric in the embedding space is used to compare the video with the biometric data recorded for a specific person. To prevent the system from focussing on visual features, they use another 3DMM generative network, which is trained adversarially using a Temporal ID network as a discriminator.

Finally, in this section, we can highlight a set of techniques that *independently process information from frame and video levels*. This approach allows us to take advantage of available information. Hu et al. [10] proposed a method to detect manipulation in compressed videos. This system analyses independently at the level of frame and temporality. The first branch uses a pre-trained CNN, MesoNet [199], with a pruning module, to prevent the model from locking into compression noise. The compressed videos contain I-frames and P-frames, the I-frames are used in the frame-level stream. However, the temporality level is trained on time-dependent features of videos, linked to P-frames. Firstly, Hu et al. [10] cleans the videos, keeping the data to keep the information from the frame-level stream. In this module, residual features are used to detect temporal inconsistency. Finally, ResNet-18 was used to classify the samples. The last step is the combination of the score generated in each stream with a Softmax function. Pu et al. [227] developed a collaborative framework to detect frame-level and video-level forgeries simultaneously, Figure 16. A joint loss function was used to optimise the AUC and the error. First, the faces of each video frame were extracted with Dlib and the features were extracted with the pretrained ResNet-50 [222], which allows the detection of high-level information. The output contains the spatial domain information of each face, which is fed into the temporal learning module, implemented with LSTM or gated recurrent unit (GRU) layers. The output of the temporal learning module will be the simultaneous input of a classifier at the frame level and a classifier at the video level to predict the integrity at both levels. It also includes an explainable AI module at the frame level that allows the analysis of the area of the image used by the system.

Agarwal et al. [228] developed a novel architecture, MD-CSDNetworwk, which combines features in the spatial and frequency domains to extract discriminative representations for classifying manipulated videos. MD-CSDNetwork is a novel cross-stitched network with two parallel streams that analyse spatial and temporal information. The combination of both domains helps the system achieve better

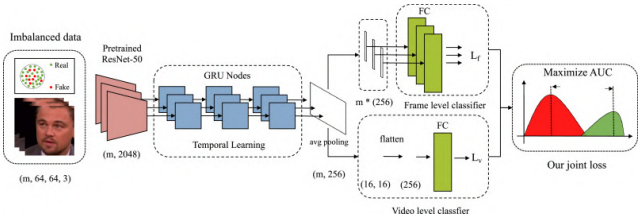


Figure 17: Framework of Pu et al. [227].

performance and generalisation. Another interesting feature of this architecture is the use of cross-stitch connections that are inserted between the two branches to automatically learn an optimal combination of domain-specific representations and shared representations of the other domain. In addition, the authors have generated heatmaps, an explainable AI technique, where the areas of a frame relevant to the classification are marked. Wang et al. [229] proposed a framework, multimodal contrastive classification by locally correlated representation (ML-LCR), the advantage of this model consists in the amplification of the implicit local discrepancies between manipulated and non-manipulated faces from the spatial and frequency domains. At first, a shallow-style representation block measures the pairwise correlation of shallow feature maps, endocing local information to extract relevant features in the spatial domain. A patch-wise amplitude and phase dual attention module capture local incoherences in the frequency domain. They used a combination of contrastive loss with cross-entropy loss to learn discriminative and generalised representations. Kolagati et al. [230] show a deep hybrid neural network model to detect deepfake videos. At first, the frames are extracted from the raw video. The system is based on the fusion of a CNN and an MLP. First, the raw video frames will be extracted and then analysed by both modules. The first module detects the facial landmarks (temporal analysis): eye blink, eye shape, lip shape, and nose shape, this information is fed to the MLP model. The second one is composed of a CNN from scratch that analyses the frames. Finally, the information from both models is concatenated and processed with two dense layers.

Within this set of techniques, combining information at the frame and video level, we can observe a wider variety of techniques. The other articles in this section use a rather limited number of architectures based on CNNs and RNNs. We can also observe how the information used in the techniques of this section is much more complete than those shown in the previous section, where temporal information is not exploited. A brief summary of video manipulation techniques based on visio-temporal information appears in Tables 9 and 10.

6.2. Techniques for detecting manipulated audio

The second modality that we are going to analyse is audio, which, as we will see below, is the least researched modality of multimedia information in the state of the art. As in the previous subsection, the authors have tackled the

Table 9
Comparative table of video forensics works based on *visio-temporal information* and the datasets used. [1]: code available.

	UADFV	DF-TIMIT	HOHA-based	FF+ +	DFDCp	DFDC	Celeb-DF	DeeperForensics	WildDeepFake	ForgeryNet	FakeAVCeleb	KoDF	Dessa [8]	VoxCeleb	Other dataset
Tran et al. [209]															X
Montserrat et al. [218]						X									
Fernando et al. [221]				X											
Chintha et al. [215]				X		X									
Zi et al. [163]		X		X	X	X	X		X						
Das et al. [211]				X											
Cozzolino et al. [225]														X	
Nguyen et al. [210]		X		X											
Hu et al. [10]				X			X								
Lu et al. [224]				X	X		X								
Agarwal et al. [228]				X		X	X								
Kolagati et al. [230]						X							X		
Chamot et al. [214]				X			X								X
Pu et al. [227]				X		X	X								
Wang et al. [229]				X		X	X	X							

problem of detection of manipulated audio signals from different perspectives, see Figure 18: *approach based on the selection and extraction of features*, where the authors try to use the most relevant features for this task; *CNN audio application*, encompasses all the works that have used CNNs or have relied on them in the generation of their models; and finally *attention layer approach*, includes those models that have used attention layers or Transformers models in their approach, architectures that are performing very well in similar tasks.

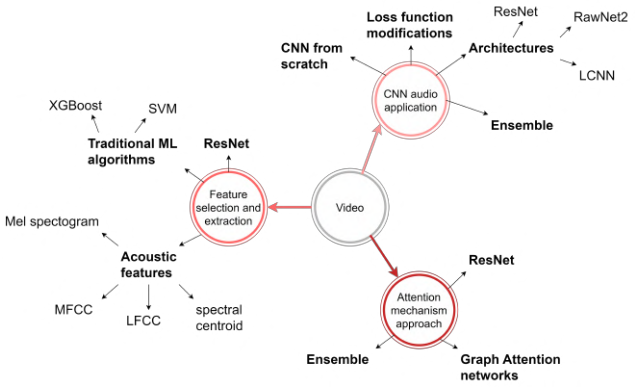


Figure 18: Representation of main forensics techniques in audio.

6.2.1. Audio forensics based on feature selection and extraction

In this section, we will provide a comprehensive overview of audio forensics systems that focus on features selection and extraction. The authors search for different techniques so that the input of the model or system contains the most relevant features to detect evidence of manipulation. Some authors are still using traditional *ML techniques* in feature selection models, for example Iqbal et al. [231] who focused on optimal feature engineering. To extract the relevant features, the signal is first converted into a readable format, mel spectrogram, and then five feature extraction techniques such as mel frequency cepstral coefficients (MFCC), a spectral_roll_off or spectral centroid were selected. The extracted information forms an array from which we calculate the means, medians, and standard deviation, generating a total of 270 features. To reduce the number of variables and keep the most relevant ones for this task, they applied principal component analysis (PCA), reducing to 65 unique features. Unlike most work in the field, they focus on extracting the most relevant features for the classification task and then processing them with traditional machine learning algorithms, such as support vector machine, decision tree, or XGBoost. Although this is a simple methodology, the results are promising, with an accuracy greater than 0.95 for SVM and XGB.

Many authors use *features* such as *MFCC* to detect manipulation in videos; another work based on them is Dongre et al. [232] which proposed an adaptive channel-wise feature recalibration technique to counter adversarial attacks. This technique is based on the analysis of the features of each audio channel separately and dynamically adjusts the weights assigned to each channel during the classification process. They proposed an attentional feature fusion technique to combine linear frequency cepstral coefficients (LFCC) and MFCC, creating enhanced input features that can aid simpler models in effectively generalising synthetic speech classification tasks. Other authors are in favour of other manual feature extraction techniques, such as the application of filters. Rahman et al. [233] proposed a model based on x-ResNet with a probabilistic linear discriminant analysis (PLDA) classifier. The system proposed in this work can be divided into three different modules. Firstly, the audio front-end module, where the acoustic features are extracted using the Linear Filter Banks (LFB) technique, which is less expensive than other techniques, such as cepstral features, that are widely used in this field. The extracted features

Table 10
Summary of different articles related to video forensics technique based on *visio-temporal information*. The links to the available codes of the following works can be found in Table 23

Article	Year	# datasets	Methodology	Code available	Task	Main contribution
Tran et al. [209]	2018	4	3DCNN, skip connection	X	Binary Classification	Comparison between different 3DCNNs, some with skip connections
Montserrat et al. [218]	2020	1	CNN, RNN, EfficientNet-B5, Facial feature extraction	X	Binary Classification	New system composed of ensemble CNN-based and RNN focuss on facial features
Fernando et al. [221]	2020	1	Attention mechanism, RNN, CNN, ResNet, GRU	X	Binary Classification	Novel Hierarchical Attention Memory Model (HAMN), using knowledge stored in neural memories
Chintha et al. [215]	2020	2	CNN, RNN, Xception, LSTM	X	Binary Classification	Novel system, XceptionTemporal, and loss function
Zi et al. [163]	2020	6	3DCNN, Attention mechanism, low-level facial features	X	Binary Classification	System based on 3D CNN that used attention masks
Das et al. [211]	2021	1	3DCNN, ResNet, Attention layers, Face	✓	Binary Classification	Novel system based on 3DCNN with attention layer in convolutional blocks, pretrained with real samples
Cozzolino et al. [225]	2021	1	Biometric features, Generalisability, 3D morphable model, Temporal ID network	✓	Binary Classification	Novel system, ID-Reveal, trained only with real samples and focused on the inconsistencies between the visual identity and biometric features
Nguyen et al. [210]	2021	2	3DCNN, 3D convolution kernels, temporal face extractor	X	Binary Classification	System based on 3D CNN with 2D convolutional kernels Novel method for constructing 3D images from the faces of consecutive frames
Hu et al. [10]	2021	2	CNN, MesoNet, XAI, Face Swap, Frame and temporal features, ResNet 18, Ensemble	X	Binary Classification	Ensemble-based system, combining frames and temporal information preocessed by CNN
Lu et al. [224]	2021	3	CNN, EfficientNet-B0, Face features, Channel-wise Spatiotemporal Aggregation module	✓	Binary Classification	A novel fusion module, CWSA, which combines the information from the frames and classifies using a CNN
Agarwal et al. [228]	2021	3	CNN, Xception, XAI, cross-stitched network	X	Binary Classification	Novel system focus on frame and temporal information using cross stitch connection
Kolagati et al. [230]	2022	2	CNN, MLP, Fusion Learning, Hybrid Neural Network, Facial landmarks	X	Binary Classification	Novel hybrid neural network, that combinea CNN, frames, and MLP, temporal analysis using facila landmarks
Chamot et al. [214]	2022	3	CNN, RNN, LSTM, MesoNet, EfficientNet, XAI	✓	Binary Classification	Combination of CNN and RNN, to process the frame information, taking advantage of temporal information. Apply XAI techniques for frames.
Pu et al. [227]	2022	3	CNN, RNN, Colaborative learning, ResNet, Ensemble, Joint loss function	✓	Binary Classification	Ensemble composed of frame and video classification system with joint loss function to maximise the performance
Wang et al. [229]	2022	4	CNN, Xception, MLP, Local correlation, Facial features, Local incoherences	X	Binary Classification	Novel system, MCLCR, based on frame and patches from frames information with contrastive loss with cross-entropy loss

are then processed by x-ResNet, after which a Squeeze-Excitation (SE) block is applied to adaptively recalibrate the independencies of the convolution channels into a global feature. Finally, they used probabilistic linear discriminant analysis (PLDA) to compute scores between embeddings.

One of the disadvantages of this type of feature is their capacity for generalisation, they learn the training data correctly but obtain low performance when analysing samples manipulated with techniques they have not seen, limiting their application in real situations. For this reason, other authors have decided to explore another very promising way: the analysis of *biometric features*, which are different for each person, regardless of the manipulation techniques that have been applied. Pianese et al. [234] proposed a new detection approach based on the biometric features of the speaker to improve generalisability. They explored the capabilities of the person-of-interest (PoI) approach in situations where only the audio track is accessible or usable. They assume prior knowledge of the speaker's identity in the analysed audio and possess a set of high-quality reference audio tracks associated with that identity. The verification process involves comparing appropriately extracted embedded vectors from the test and reference audio tracks to confirm the speaker's identity.

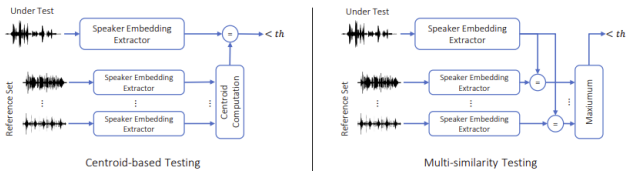


Figure 19: Approach of Pianese et al. [234].

Wang and Yamagishi [235] focused on *active learning* (AL) to select the most interesting information from the training set. They proposed a novel energy-score-based AL method. The aim of this methodology is that the model is responsible for selecting its own training set with the most relevant information from the set. The aim of this methodology is that the model is responsible for selecting its own training set with the most relevant information from the set and that the fine-tuning is iterative. The main advantage of this approach is that it significantly reduces

Table 11
Comparative table of audio forensics works, focus on *feature selection and extraction*, and the datasets used.

	ASVspoof	FoR	WaveFake	FAD	IWA	VoxCeleb	FakeAVCeleb	Other dataset
Iqbal et al. [231]		X						
Dongre et al. [232]	X	X						
Rahman et al. [233]	X	X						X
Pianese et al. [234]	X				X		X	
Wang and Yamagishi [235]			X					

the size of the training set, reducing the cost of the system. Finally, other authors have chosen approaches that are not related to manual feature selection, but rather the models themselves select the training samples, based on the most relevant features for forensics, without human intervention. This approach, although very different from the state-of-the-art work described above, offers many possibilities, and in addition to obtaining better results than other work in the field, it reduces the computational cost by eliminating irrelevant information that can also introduce noise into the system.

Although feature extraction is a well-studied field, many works still focus on it, as they still show promising results, however, many cases still extract features manually, which can be an important limitation, as it reduces the ability to generalise to new scenarios. A brief summary of the audio manipulation techniques based on feature extraction and selection appears in Tables 11 and 12.

6.2.2. Audio forensics based on CNN architectures

In this section, we focus on the articles that have used CNNs or based on them for the generation of their audio forensics models. CNNs have demonstrated outstanding performance on a number of tasks, such as classifying audio according to whether they have been tampered with. Although these architectures have been used for more than a decade, they are still a good choice for many tasks.

In order to be able to analyse audio signals with this type of architectures, it is necessary to transform these signals into a format that can be understood by the model; in this case, we will have to transform them into images. First of all, we need to represent the signal where the amplitude of the signal is represented with respect to time; then we apply the Fourier transform, which allows us to decompose a signal into its individual frequencies and the amplitude of the frequency; i.e., it transforms the signal from the time domain to the frequency domain, producing a spectrum. The spectrogram is the result of calculating the spectrum of a signal by time windows of the same. This already allows us to work efficiently with CNN. However, as we will see later, many authors transform the spectrogram into a Mel spectrogram, i.e. it is converted to the Mel scale.

Some authors continue to focus on building basic *CNNs from scratch*, although these networks do not have the same feature extraction capability as other pre-trained CNNs widely used in the state of the art, for simple tasks they still give very good results. Zhang et al. [236] propose two models, based on convolutional networks. First, they transform the audio signals into images, which are representations of the acoustic features. Then, two baselines are proposed for the audio forensics. Firstly, they propose ResNet to process the constant Q cepstral coefficient (CQCC), and secondly, they propose to use TSSDNet to process the speech waveform. It also introduces a new audio manipulation dataset in Mandarin.

To improve the performance of these architectures in the task of detecting the manipulation of audio signals is the *modification of loss functions*. Chen et al. [237] developed a manipulated audio detection model that incorporates two key enhancements over state-of-the-art models. Firstly, they applied data augmentation techniques by introducing public domain noise. Second, they employed a novel loss function, the large-margin cosine loss function (LMCL), designed to maximise the variance between classes and minimise intraclass variance. This model consists of four different modules: in the first one, features are extracted by linear filters, which is a compressed version of the short-time Fourier transforms (SFT); the main advantage is the reduction of the computational cost; second, some frequency channels are randomly masked during the training. Subsequently, they proposed to add FeqAugment, a layer that during training adjacent frequency channels are masked randomly, to improve generalisation ability. Finally, the information is the input of a fully connected layer and classifier with the LMCL loss function to produce the final result. Other authors also include data augmentation techniques, which improve the generalisability of the system, Wang and Yamagishi [238] focussing on loss functions and their effect on different models. They have developed a new loss function, *mean-square-error loss function with P2SGrad*. The main advantage of this function is that it does not require hyperparameter settings, unlike other loss functions such as margin-based softmax, which is widely used in state-of-the-art audio and image processing. The loss function is based on the mean square error (MSE) metric and the probability gradient to similarity (P2SGrad). The combination of the loss function proposed in this paper, together with a light convolution network (LCNN) with average pooling, obtaining very competitive results without the need for data augmentation. Modifications or the creation of new loss functions is only one of the ways to improve the performance of these architectures. For example, Tak et al. [239] present an architecture based on RawNet2 [240] with modifications. This architecture is capable of processing raw audio and detecting details that were impossible to detect with traditional techniques. The authors have modified the filter lengths and size of the kernel filters of the second residual block. As in the original architecture, they use FMS independently in each residual block. Therefore, they adopted a combined additive and multiplicative feature scaling approach as Jung et al. [240].

There are also numerous examples where authors do not use directly the pre-trained models but are inspired by them

Table 12
Summary of different articles related to audio forensics technique based on *feature selection and extraction*. Download links to the codes used in the following papers can be found in Table 23.

Article	Year	# datasets	Methodology	Code available	Task	Main contribution
Iqbal et al. [231]	2022	1	ML algorithms,PCA, mel spectrogram	X	Binary classification	Novel ML model focusing on feature selection through principal component analysis
Dongre et al. [232]	2022	2	Adaptative channel-wise feature recalibration technique, Attention feature fusion block	✓	Binary classification	The introduction of the attentional feature fusion block of image domian, to combine LFCC and MFCC
Rahman et al. [233]	2022	3	CNN, x-ResNet, Probabilistic linear, discriminant analysis, Squeeze excitation block	X	Binary classification	Combination on x-ResNet and probabilistic linear discriminant analysis
Pianese et al. [234]	2022	3	Person-of-Interest, Centroid-based and multi-similarity testing	X	Binary classification	New system based on speaaker approach, exploring capability of person-of-interest (POI)
Wang and Yamagishi [235]	2023	1	Active Learning, wav2vec model	✓	Binary classification	Novel spoofing countermeasure using active learning to remove useless samples from a pool

to create their own architectures. Hua et al. [241] propose a new model, called *Time-domain Synthetic Speech Detection Net* (TSSDNet), where they consider two advanced CNN structures, firstly skip connection, such as ResNet, called Res-TSSDNet; and secondly, parallel convolutions, such as Inception, called Inc-TSSDNet. The last architecture includes extended convolutions that allow for increased receptive field and control model complexity. Also, it is interesting to note that the authors use the mix-up regularisation, which improves the generalisation capacity, which is very important for detecting different manipulation techniques. Kawa et al. [242] proposed a novel architecture, called *SpecRNet*, characterised by a fast inference time and a low computational requirement. The backbone is inspired by RawNet-2 [240], the main difference is that this architecture processes two-dimensional spectrogram information, in particular linear frequency cepstral coefficients (LFCC). The model uses an LFCC representation of the audio signal as an input. The architecture is composed of a normalisation layer and three sets of residual blocks (ResBlock) and FMS attention layer, then a pre-recurrent normalisation, two bidirectional GRU layers, and two fully connected layers. Wang et al. [243] focus on the power of layered neural activation patterns, which are hypothesised to be able to capture subtle differences that can provide more information to classification systems than raw input. They proposed a new approach, called *DeepSonar*, based on the monitoring of neuronal behaviour in a DNN-based speaker recognition system with a binary classifier to recognise AI-synthesised false voices. A DNN-based SR system is used to capture the raw layer-wise neurone behaviours from the input and determine the most valuable neurones in subtle differences detection. Finally, the binary classifier, based on the activated layer-wise behaviours of the input, produces the final result.

As we have already seen in the section on techniques for detecting manipulated video, *ensembles* are still widely used techniques in the domain of forensics, and in the case of audio signals it is no different, with many authors opting for this option to improve the performance of individual models. Gomez-Alanis et al. [244] proposed an model based on fusion learning, a new early integration technique based on a deep learning model. They used two different systems: standalone presentation attack detection (PAD) and standalone automatic speaker verification (ASV). The LA and PA spoofing embeddings are obtained by extracting features from the short-time Fourier transform (STFT) of the i-th test utterance, while the x-vectors, extracted from Mel-frequency cepstral coefficients (MFCC) features, of the enrolment and test utterances are merged into a single ASV embedding. The different embeddings are concatenated and processed with three fully connected layers. Khochare et al. [245] compare two approaches: feature-based and image-based. In the first one, features are extracted from the audio samples and then processed with classical machine learning techniques, such as Support Vector Machine, Random Forest or XGBoost. In the image-based approach, audio samples are transformed into melspectrograms and then processed with Temporal Convolutional Networks (TCN), architectures that give good results on sequential data, and Spatial Transformer Network (STN). They have created an ensemble with both models and the outputs are combined to generate the final prediction. In this work, it is observed that deep learning techniques obtain better results than classical machine learning algorithms.

As we have observed throughout this section, most of the published work is focused on classifying audio samples into manipulated and real samples; however, this has some limitations, such as not knowing which parts of the sample have been manipulated. In addition to providing interesting information to the user, it also helps to understand how the model gives at the result, i.e. it adds explainability to the system. For example, Zhang et al. [246] focused on addressing the challenge of detecting short segments of fake speech embedded within a longer utterance. To facilitate research in this area, the authors introduce the *PartialSpoof* dataset, which is specifically designed for the detection of these misleading segments. The dataset consists of real voice recordings that include both genuine and fake speech segments embedded within utterances. The proposed model for forensics is capable of labelling samples at the segment level and at the sample level, which allows us to know not only whether the audio signal has been manipulated, but also which segments have been manipulated. A brief summary of audio manipulation techniques based on CNN appears in Tables 13 and 14.

Table 13
Comparison of audio forensics works *based on CNN* and the datasets used.

	ASVspoof	FoR	WaveFake	FAD	IWA	VoxCeleb	FakeAVCeleb	Other dataset
Gomez-Alanis et al. [244]	X							
Wang et al. [243]		X						
Chen et al. [237]	X							
Wang and Yamagishi [238]	X							
Tak et al. [239]	X							
Hua et al. [241]	X							
Khochare et al. [245]		X						
Zhang et al. [246]	X							X
Zhang et al. [236]								X
Kawa et al. [242]	X		X					X

6.2.3. Audio forensics based on attention layers

Transformer models were introduced in 2017, revolutionising the field of natural language processing (NLP). This architecture, based on *attention mechanisms*, has proven to be highly effective in capturing complex relationships in sequences and temporal data, showcasing a strong ability to learn universal patterns and representations. These advantages make Transformers an interesting option for audio forensics. However, despite their successes, Transformers are not as widely used as CNNs, which still hold two main advantages: greater computational efficiency and greater applicability to spatial data. Therefore, in this section, we focus on papers that have used this approach to detect audio manipulation. In some works, we can see how the authors have used transformer models, for example, Zhang et al. [247] proposed a novel model, called TE-ResNet, which combined a residual network scheme with a transformer encoder. First, the encoder is used to extract a contextual representation of the acoustic features, and the residual network processes them and provides the score. Like other authors, they apply data augmentation (five techniques) and ensemble techniques to improve the results of individual models and to merge the scores of the different models, and they use a logistic regression model to calculate the final score.

Other researchers have chosen to incorporate *attention layers into different architectures*. For example, Jung et al. [248] introduced a novel heterogeneous stacking graph attention layer. This layer enables artefact detection in diverse temporal and spectral intervals using a heterogeneous attention mechanism and a stacking node. By adding a new maximum network operation and a new reading scheme, the AASIST approach has managed to improve state-of-the-art models by a 20% relative. The authors process the audio samples with an encoder and then use two graph modules to process the temporal and spectral information in parallel. The output will be combined and the maximum graph operation technique is applied to two branches independently. Finally, the read-out scheme concatenates node-wise maximum and average, and the stack node, which is then followed by an output layer. Ge et al. [249] have developed an ensemble of two different models, automatic speaker verification (ASV) and spoofing countermeasure, where they focus on analysing how joint optimisation obtains better results than if they are optimised independently. Although this approach reduces the performance of individual models, it improves the complementarity and overall performance of the system. The ASV subsystem has used ResNet34 with excitation blocks [250], called the ResNetSE34 model [251]. The CM subsystem has used a RawNet2-based encoder to decompose audio signals [240, 239]. Features are integrated through a graph attention network [239], and finally the information is processed with the SASV 2022 AASIST model [248]. The results of the subsystems are then stacked and processed by the back-end classifier. This classifier is composed of three convolutional blocks, an average pooling layer, followed by a flatten layer, and finally three linear layers and an OC-softmax activation.

Finally, we will look at other types of system, which are not exactly models or forensics systems but have proven their advantages in these tasks. *self-supervised models* are able to achieve great knowledge of the samples, understand their patterns and trends and then apply this knowledge to more specific tasks, such as detection of audio manipulation or speaker verification. Chen et al. [252] presented a new pre-trained self-supervised model, WavLM, whose aim is to predict masked speech and eliminate speech noise, Figure 20. WavLM used gated relative position bias within the Transformer structure to effectively capture the sequential order of input speech. The model proposed by the authors learns universal speech representations from unlabelled data and is capable of performing different speech processing tasks; nine, in particular, have been evaluated in this task. Among the checked subtasks is speaker verification, which allows us to detect manipulated audio. The model uses a Transformer model as the backbone. First, this model has a convolutional encoder composed of seven blocks of temporal convolutional layers and a normalisation layer with a GELU activation layer. Then the information is processed by a Transformer encoder with a gated relative position bias. The encoder has a convolution-based relative position embedding layer with 128 kernel sizes and 16 groups at the bottom. The gated relative position bias is coded according to the differences between the key sample and the query.

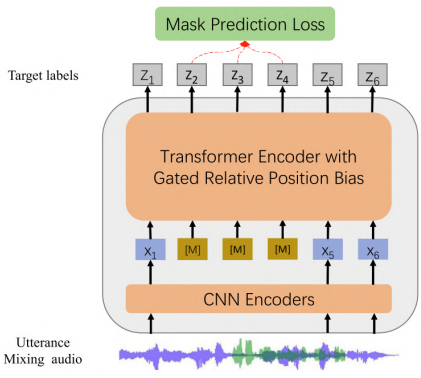


Figure 20: WavLM architecture [252].

Table 14
Summary of the different articles related to audio forensics techniques *based on CNN*. Table 23 contains the download links for the codes utilized in the papers mentioned above.

Article	Year	# datasets	Methodology	Code available	Task	Contribution
Gomez-Alanis et al. [244]	1	2020	Fusion learning, presentation attack detection, automatic speaker verification, DNN	X	Binary classification	Novel model that combine presentation attack detection and automatic speaker verification
Wang et al. [243]	1	2020	DeepSonar, layer-wise neurone activation patterns	X	Binary classification	New approach, DeepSonar, based on the monitoring of neuronal behaviour in a DNN-based speaker recognition system with a binary classifier
Chen et al. [237]	1	2020	CNN, ResNet, Data augmentation, LMCL	X	Binary classification	New loss function, XXX, which maximises the interclass variance and minimises the intraclass variance. Include FreqAugment layer to improve generalisability
Wang and Yamagishi [238]	1	2021	mean-square-error loss function with P2SGrad, gradient similarity, LCNN	✓	Binary classification	New loss functio, mean-square-error with P2SGrad, no hyperparameter setting needed, based on mean square error metric and probability gradient to similarity
Tak et al. [239]	1	2021	RawNet2, Combined adaptative and multiplicative feature scaling approach	✓	Binary classification	Modification of the parameters of the convolutional blocks and the inclusion of combined additive and multiplicative feature scaling approach
Hua et al. [241]	1	2021	Time-domain Synthetic Speech Detection Net, Res-TSSDNet, Inc-TSSDNet	✓	Binary classification	Novel architecture, Time-domain Synthetic Speech Detection Net, with skip conenection or parallel convolutions with mixup relugarisation
Khochare et al. [245]	1	2021	Traditional machine learning algorithms, TCN, STN, ensemble	X	Binary classification	New ensemble model, combining features spatial and temporal information and comparison with traditional machine learning approach
Zhang et al. [246]	2	2022	Self-supervised, labelling granularity	X	Temporal location, segment detection	Novel dataset for segments detection Novel model for labelling at segment label
Zhang et al. [236]	1	2022	CNN, ResNet, Res-TSSDNet	✓	Binary classification	Two baselines: ResNet for processing CQCC and Res-TSSDNet for processing the raw speech waveform New mandarin dataset
Kawa et al. [242]	3	2022	SpecRNet, RawNet-2, FMS attention layer, GRU	✓	Binary classification	New architecture, SpecRNet, composed of residual blocks, FMS attention layer and GRU layers

Table 15
Comparative table of audio forensics works based on *attention mechanism* and the datasets used.

	ASVspoof	FoR	WaveFake	FAD	IWA	VoxCeleb	FakeAVCeleb	Other dataset
Zhang et al. [247]	X	X						
Jung et al. [248]	X							
Chen et al. [252]						X		
Ge et al. [249]	X			X		X		

As we can see, the number of articles available using transformers for the task we are interested in is lower than that using CNNs. However, this does not mean that in the next few years the current situation will change and transformer models will become more relevant in this domain. A brief summary of audio manipulation techniques based on attention mechanism appears in Tables 15 and 16.

6.3. Techiques for detecting manipulated multimodal content

Finally, we will analyse works that have used both video and audio modalities to detect manipulation. The main difference between this subsection and the previous ones is that this approach takes advantage of all available information; most of the work focusses on finding inconsistencies between the information coming from the frames and the audio. Although all models are based on the search for inconsistencies, two different trends can be observed, Figure 21: *approach based on audiovisual inconsistencies*, which is based on directly using extracted features, or *inconsistent emotion-based approach*, which looks for failures of generative systems in emotion imitation. Although current generative systems obtain very good results in the manipulation domain, they are not capable of imitating people's emotions.

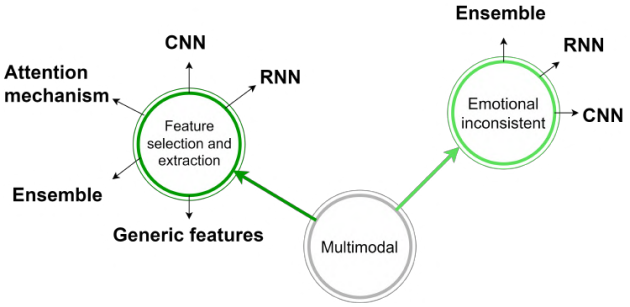


Figure 21: Representation of main multimodal forensics techniques.

6.3.1. Inconsistencies between modalities

In this section, we have focused on published articles that focus on comparing audio and visual features of video for inconsistencies that may be signs of manipulation. As we have seen throughout the article, the most commonly used architectures are CNNs and transformers, as they perform correctly in both tasks independently.

Some methodologies focus on finding inconsistencies between modalities using *CNN and combined RNN* for the processing of visual and acoustic features, respectively. For example, Lewis et al. [253] proposed a hybrid deep learning approach, NOLANet, that uses spatial, spectral, and temporal information. This multimodal network processes

Table 16
Summary of the different articles related to audio forensics techniques *based on attention mechanism*. Table 23 contains the download links for the codes used in the papers mentioned below.

Article	Year	# datasets	Methodology	Code available	Task	Contribution
Zhang et al. [247]	2021	2	TE-ResNet, encoder, attention mechanism, ensemble	X	Binary classification	Novel ensemble-based model, where each model is composed of transformer encoder and residual neural network
Jung et al. [248]	2022	1	Heterogeneous stacking graph attention layer, maximum network operation	✓	Binary classification	Novel system, AASIST, with novel heterogeneous stacking graph attention layer, which processes temporal and spatial information
Chen et al. [252]	2022	1	Self-supervised, attention mechanism, genric model	✓	Learn patterns	New pretrained self-supervised model, learns universal speech representations from unlabelled data
Ge et al. [249]	2023	3	ResNet-34, excitation blocks, ResNetSE-34	✓	Binary classification	Ensemble-based system, composed of automatic speaker verification and spoofing countmeasure, using a combined optimisation

visual neural, visual spectral and audio spectral information with three different pretrained models, Xception [194], LipNet [254] and DeepSpeech2 [255] respectively. They introduced a novel method of spectral feature analysis using a discrete cosine transform (DCT). Each network generates a feature map for each sample frame, the output of each network is processed by three LSTMs independently, and the information obtained is combined and processed by a multilayer perceptron (MLP) model to obtain the final prediction. Other works include other sources of information, such as text; Shang et al. [256] developed a novel framework, TikTok disinformation deTection (TikTec), a TikTok video forensics model that focusses on COVID-19-related videos. Unlike other works, this approach uses three different sources of information: audio, video, and text. In addition to the information provided in the videos, it also uses the subtitle information extracted from the audio samples to extract the relevant information. Initially, a CNN is employed to extract visual features from the frames, while an RNN processes features from the other two modalities, audio and text. Subsequently, a visual-speech co-attentive information fusion module is utilized, combining verbal and visual information from these diverse modalities.

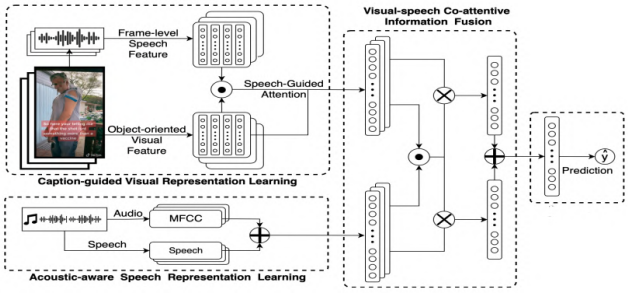


Figure 22: TikTok disinformation deTection (TikTec) architecture [256].

Another way to detect manipulation in multimodal samples is to train a more *generic model* that is able to *understand generic features* and relationships between modalities. Cheng et al. [257] focused on comparing the homogeneity between the voice and face of the video samples. They propose to pre-train a model with a real generic dataset so that the system learns its general features from the samples and subsequently retrains itself from the specific task, forensics. The main advantage of this approach is that the system can be adapted to any manipulation technique. First, in the pre-training stage, 3-second clips are extracted from the audio samples and plotted as a spectrogram, then two CNNs extract the features from the faces and audio samples independently. The extracted features are combined in a voice-face comparison network to measure the degree of match between voice and faces. In the fine-tuning stage, the model is adjusted to detect manipulation in video samples, jointly optimising both feature extractors. This approach allows the detection of a wide variety of manipulation techniques since the homogeneity of the manipulated samples is lower than the real samples with which the model was pre-trained.

The other main way in the field for forensics is the application of *attention mechanisms and transformer models*. Yang et al. [258] proposed an audio-visual joint learning to detect deepfakes, AVoiD-DF, based on the inconsistencies between the two modalities, audio and video. This methodology is divided into three modules: 1) a temporal spatial encoder (TSE); 2) a multimodal joint decoder (MMD); 3) and a cross-modal classifier. The video samples were split into individual frames, and the audio signals were transformed into Mel frequency spectral coefficients (MFCC) features, using the Kaldi toolkit. Then, both types of samples were split into patches. Then, the TSE extracts the audio-visual features independently by means of two modules, a temporal encoder that extracts the features from the patches and aggregates them to obtain a temporal representation. The spatial encoder extracts the features from spatially synchronised positions between modalities. Then the MMD module merges the information from the previous module. Regarding bidirectional cross-attention (BiCroAtt), it allows the exchange of information between the audio and video channels. Then the self-attention layers learn the relationships between both modalities, and finally, the features are transformed so that the classifier can give a final result. In the Cross-Modal classifier, the tokens of both modalities are concatenated and processed by a dense layer. Wang et al. [259] proposed a novel framework, called FTFDNet, based on the audio-visual attention mechanism (AVAM), which

allows the discovery of the most useful features of the task. This mechanism can be applied to any CNN architecture using modularisation. Like all multimodal models, the audio signal and the frames are processed independently using a stack of residual convolutional layers to encode the input into low-dimensional features. The AVAM mechanism fuses the features of both modalities to create the joint representation, which is processed with three dense layers to give the final result.

Ilyas et al. [260] proposed a novel multimodal framework, AVFakeNet, composed of the two-stream network, using Dense Swin Transformer Net (DST-Net) to analyse video and audio independently. First, the frames and the audio signal are extracted from the video; then the audio is transformed into a Mel spectrogram and a Multi-Task Cascaded Convolution Neural Network (MTCNN) algorithm is used to extract the faces of the frames. The two sources of information are then independently processed using DST-Net. This model has three steps: the input block, which has three dense layers; the feature extractor block has a swin transformer module and a 1D-global average; and the output block, composed of two dense layers followed by a dropout layer and the classifier. To make the final decision, max voting is applied to the predictions of the different frames and if the classification of the video or audio is false, the final decision of the system is false. Feng et al. [261] proposed an autoregressive model capable of capturing the temporal synchronisation between the frames and the sound of the sample. The main advantage of this approach is that it does not need manipulated samples to be trained because it uses features common to real samples. This enhances the generalisability of the system, making it applicable to various situations, and enabling the detection of different manipulation techniques. First, video and audio features are extracted independently, using ResNet-18 2D+3D [262] and VGG-M [263] respectively. Information from both modalities is merged using a transformer model. Subsequently, a decoder-only autoregressive transformer is used to learn the distribution of the features. Finally, from the information obtained over time of the sample, using the code's log likelihood, averaged over each video frame, for anomaly detection.

In addition to CNNs and transformers, there are other techniques that have been shown to be successful in detecting manipulation, such as *identity verification* systems. Cozzolino et al. [264] proposed a forensics system based on audiovisual identity verification, called POI-Forensics, using each person's own features to create the system. To do so, they have taken advantage of the contrastive learning paradigm to learn the manipulated fragments of the video. This system does not need samples of manipulated videos for training, but learns the patterns and features of the real videos and then searches for inconsistencies in the embedding space. While this approach enables the detection of manipulation signals regardless of the technique, enhancing the generalisability of the system, it has a notable limitation: it can only detect manipulation involving individuals it has been trained on. This restricts its applicability in real-world scenarios.

One of the most intriguing articles is Cai et al. [145], which introduced a new dataset and model known as boundary-aware temporal forgery detection (BA-TFD). This study focused on two tasks: classifying samples and *temporal localization* of modifications. The main advantage is the ability to locate the manipulated fragments, which not only provides more information but also serves as an explainable AI method to understand whether the model uses the right features for the task. First, the system will extract the video and audio signal features independently by means of two encoders, a 3DCNN and a 2DCNNN respectively. These features are then used to classify the audio and video frames independently. Subsequently, the video or audio boundary matching layer detects the boundaries between real and fake segments from the frame sequence. Finally, the fusion module combines video and audio features and produces a sequence of fused boundary maps that indicate the probability that each video frame and each audio frame are real or fake. From the fused boundary maps produced by the fusion module, it generates the video and audio segments using a threshold-based segmentation algorithm. A brief summary of multimodal manipulation techniques based on inconsistencies between modalities appears in Tables 17 and 18.

6.3.2. Emotional inconsistencies

In this section, we will focus on articles that use *emotions* to detect *inconsistencies* between audio and video. Although current generative systems achieve very good results, they are still unable to imitate human emotions. This is due to the subjective nature of emotions and their complexity, among other factors. Therefore, it can be a good indicator of inconsistencies in manipulated videos. Lomnitz et al. [265] concentrated on using emotions to detect manipulation in video samples, seeking disparities between emotions in fake videos and authentic emotions. The primary drawback of this approach is the requirement for real samples from the same actors to make comparisons; nevertheless, it enhances generalisability. Unlike most published work in the domain, this system is a *multiclass* problem, where the labels are real, synthetic and manipulated, rather than the binary (real or manipulated) classification systems of most articles. This system is composed of three different branches: the first branch processes frames individually using Xception and an MLP; the second branch analyses the frame sequence using Xception and two sequential bidirectional bidirectional long-short-term memory units; and lastly, it processes audio signals with SincNet [266]. Finally, an MLP is used to combine all the information and perform the classification.

Mittal et al. [267] focused on the extraction and analysis of two audio and visual modalities to classify the samples as real or manipulated. This model is inspired by the siamese network architecture and the triplet loss. The model is based on the comparison of the interest samples against real samples, first of all the face and audio features are extracted using

Table 17
Comparative table of multimodal forensics works based on *audiovisual inncoherencies* and the datasets used. [1]: code available.

	UADFV	DF-TIMIT	HOHA-based	FF+-	DFDCp	DFDC	Celeb-DF	DeeperForensics	WildDeepFake	ForgeryNet	KoDF	FFIW 10K	FakeAVCeleb	Vox-Celeb	VideoSham	LAV-DF	Other dataset
Lewis et al. [253]						X											
Shang et al. [256]																	X
Cheng et al. [257]						X							X				
Wang et al. [259]																	X
Cozzolino et al. [264]		X			X								X				X
Cai et al. [145]																X	
Ilyas et al. [260]													X				
Yang et al. [258]						X							X				X

Table 18
Summary table of the different articles related to multimodal forensics techniques *based on inconsistencies between modalities*. The download links for the codes used in the referenced articles can be found in Table 23.

Article	Year	# datasets	Methodology	Code available	Task	Contribution
Lewis et al. [253]	2020	1	NOLANet, Xception, LipNet, DeepSpeech, fusion learning, LSTM	✓	Binary classification	Novel hybrid deep learning approach, NOLANet, combines spatial, spectral and temporal information.
Shang et al. [256]	2021	1	CNN, RNN, , co-attentive information fusion module	X	Binary classification	Novel framework, TikTec, that used text, audio and video, combining the information with co-attentive information fusion module
Cheng et al. [257]	2022	2	Voice-face matching, XAI, Real-fake loss function	X	Binary classification	New approach to voice-face matching, training with real videos and retraining for the task of forensics.
Wang et al. [259]	2022	1	FTFDNet, audio-visual attention mechanism, fussion learning	X	Binary classification	Novel framework, FTFDNet, based on attention mechanism (AVAM) ,applicable to any CNN.
Cozzolino et al. [264]	2022	4	Identity verification, contrastive loss, similarity matrix	✓	Binary classification	New approach, POI-Forensics which learn a person-of-interest based on multimodality consistencies
Cai et al. [145]	2022	1	3D-CNN, 2D-CNN, Bounday matching layer, fusion learning	✓	Spatial localisation	New multimodal method, Boundary aware temporal forgery detection
Ilyas et al. [260]	2023	1	AVFakeNet, Dense swin transformer Net, fusion learning	X	Binary classification	Novel unified framework, AVFakeNet, composed of two DST-Net to compute dense hierarchical features maps
Yang et al. [258]	2023	3	Multi-Head Self-Attention (MSA) sub-layer, cross-attention block, TSE, MMD, bidirectional cross attention	✓	Binary classification	New audio-visual joint learning, AVoiD-D, composed of temporal spatial encoder, multimodal joint decoder and cross modal classifier

OpenFace and pyAudioAnalysis, respectively. The extracted features are the input of the neural networks: CNN and Memory Fusion Network (MFN) which are trained using a combination of two triplet loss functions. This approach is similar to a Siamese network, because the network weights are used to operate on two different inputs. Hosler et al. [268] proposed a video forensics system based on sentiment recognition from frames and audio signals, looking for inconsistent and non-natural emotions especially when they are not explicitly restricted to do so during creation. First, they extracted the low-level features (LLDs) from the video that describe the speaker's face and voice to reduce input dimensionality, and only tracked important features that are linked to emotion. Then they used a valence-arousal model of emotion where they measured two aspects of emotion: positivity and arousal from LLD audio and faces. They have used a model based on the LSTM architecture presented by Ringeval et al. [269], which is continuous, allowing subtle or ambiguous emotions to be captured in the samples. The last stage classifies the samples based on the information obtained from the emotions based on the inconsistencies between the faces and the voice. A brief summary of the multimodal manipulation techniques based on emotional inconsistencies appears in Table 19.

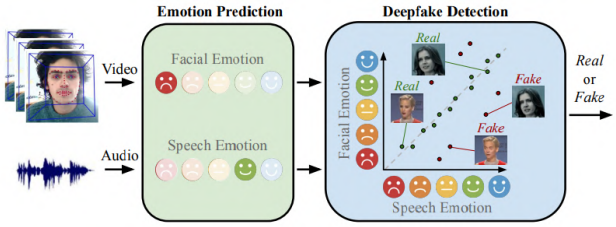


Figure 23: Proposed system of Hosler et al. [268].

7. Available tools for non-technical end-users

As we have seen throughout this survey, there are many systems that have been designed for detecting manipulation in video, audio, and multimodal samples; however, many of these works cannot be used by any user of social networks or the internet. So, if we want to fight misinformation and allow users to check or verify information coming from the Internet, we need to create simple tools that can be used by anyone who wants to check the veracity of multimedia samples they find on the Internet. Although the audience is very diverse, it is important that the application is adapted to their needs, providing greater credibility, content integrity, and more reliable information dissemination.

In this section, we explore practical applications and existing tools for detecting video and speech manipulation,

Table 19
Summary table of the different articles related to multimodal forensics techniques *based on emotional inconsistencies*.

Article	Year	# datasets	Methodology	Code available	Task	Contribution
Lomnitz et al. [265]	2020	DFDC	Xception, Bi-LSTM, sincNet, MLP,	X	Multiclass classification	Novel systemm, based on fusion learning, using MLP and CNN for frame analysis and SincNet forr audio analysis
Mittal et al. [267]	2020	VideoSham	siamese network architecture, triplet loss, CNN, Memory fusion network	X	Binary classification	Novel system using an architecture based on a Siamese network, which uses the triplet loss function.
Hosler et al. [268]	2021	DFDC	LSTM, Low-Level Descriptor,	X	Binary classification	New classification system based on arousal and valence, using LSTM

highlighting advances in this field. Table 20 provides a comparative analysis of the different tools described in this section.

InVID Verification [270] a web-based toolset specifically created to verify user-generated videos and provide insights into their context. It comprises various components dedicated to the analysis and management of user-generated content, and it presents the analysis results through a user-friendly interface.

With this tool, end-users can verify several aspects of a video, such as its previous use, origin, rights, contextual information, and even perform forensics analysis. The application offers full support for videos sourced from popular platforms like YouTube, Facebook, and Twitter. However, for videos from other platforms, only a partial analysis can be conducted. InVID Verification Application is designed for anyone who needs to verify the authenticity and context of user-generated videos, including researchers, journalists, fact-checkers, and law enforcement agencies.

Real-Time Deepfake Detector [271] a real-time deepfake detection platform, which uses AI models for face and landmark detection algorithms to identify key markers of a person's face, such as the shape of their eyes, nose, and mouth and assesses blood flow signals in real videos to detect whether they are genuine or fake.

It is designed to run on a server and interfaces through a web-based platform, and can run up to 72 different detection streams simultaneously on 3rd Gen Intel Xeon Scalable processors.

It could be used by social media platforms to prevent users from uploading harmful deepfake videos, by international news organisations to avoid inadvertently amplifying manipulated videos, and by nonprofit organisations to democratise deepfake detection for everyone.

Deepfake-o-meter [272] proposed an open web platfom that incorporate more than ten state-of-art deepfake image and video detection methods and continues adding more capacities. It can be used to benchmark the effectiveness of multiple deepfake detection algorithms on a single entry.

The tool is composed of three components: front-end, back-end, and data synchronisation. An end-user can upload a video, select the desired detection methods, and enter his email address. Upon completion of the process in the software back-end, the user will receive an email with the detailed detection result.

It enables researchers to compare their own detection algorithms with the latest methods, and users can identify whether a given video or image is genuine or not.

Poster [217] proposed a deepfake detection system for journalists. They design an intuitive application that journalists can use to determine whether a video is genuine or fake.

Reality Defender [273, 274] a deepfake detection and generative AI platform that detects fake news and propaganda. It can analyse audio, images, videos, and documents to detect deepfakes in milliseconds.

End-users can access real-time risk assessment, email alerts, and forensic review reports for all media types.The platform also provides users with weekly updates on the latest deepfake and generative AI threats.

Reality Defender uses a combination of advanced technologies to detect deepfakes and other types of disinformation. When a media file is uploaded to the platform, it is analysed by a set of algorithms to identify any signs of manipulation or falsity. The platform also compares the media file to a large dataset of known deepfakes and other types of manipulated media to improve its precision. Users can view the results of the analysis in real time through a user-friendly dashboard.

DeepDetector [275] an AI-based deepfake detection software designed and trained to recognise AI-generated or AI-manipulated faces. DeepDetector works by extracting visible faces in the media and analyses them to find deepfake traces, provides a probability of the input being a deepfake, and an activation map to explain the classification. It can detect various types of deepfake manipulations and AI-generated content, and it is cloud-based and compliant with European data protection laws.

DeepfakeProof [276] an AI-powered plug-in that scans every image on the webpages you visit to detect any deepfake or manipulated media. It alerts users in real-time and provides accurate and reliable deepfake detection to protect against the spread of misleading and harmful deepfakes online. It is easy to install and provides a seamless browsing experience for all users.

Table 20
Comparative Analysis of Deepfake Detection Applications and Tools.

Applications/Tools	Supported Media Types	Detection Methods	Integration Options	User interface Ease of Use	Cost & Pricing
InVID Verification	Videos, Images	AI-based	Plugin, Application	✓	Free
Real-Time Deepfake Detector	Videos	AI-based on PPG signals	-	-	-
Deep-fake-o-meter	Videos	10+ SOTA DeepFake Detectors	Application	✓	Open-source
Poster	Audios, Videos	AI-based	Web-based App	✓	-
Reality Defender	Audios, Images, Videos,Documents	AI-based	Web-based App, API	✓	-
DeepDetector	Videos, Images	AI-based	API	-	-
DeepfakeProof	Images	AI-based	Plugin	✓	Free

We can see that the number of tools available to end users, moreover, these tools are not integrated into any social network, but it is the users who have to know them in order to use them, which may limit their use by the inexpert public. However, it is very likely that in the coming years the number of tools and applications available will increase and their performance will improve.

8. Discussion and conclusion

This survey has reported in detail the contributions in the domain of multimedia data manipulation, more specifically, for video, audio, and their multimodal combination. Throughout this review, our research questions have been analysed in depth, both descriptively and comparatively. This final section aims to complete the overview of this domain by providing three objectives: the answers to the research questions, the challenges and future trends of this problem, and finally, a brief conclusion.

8.1. Answer to research questions

In Section 1, we formulate four research questions related to the current status and trends of forensics in multimedia data. These questions have driven the methodology and literature analysis conducted in this review. Following the analysis conducted throughout this article, the different research questions can now be answered based on the knowledge gained from the review process conducted in the previous sections. Figure 24 shows a summary of the main conclusions drawn from this in-depth analysis. These conclusions highlight the results of the research questions. A comprehensive explanation of all these responses is provided below.

RQ1. What are the main topics within the field of multimedia data manipulation detection? Throughout the review, it is clear that in recent years both datasets and forensics techniques have focused on *Deepfakes* and *face forensics*, which is a limitation and a problem. Most current datasets are beginning to explore other ways, such as VideoSham [144], which focuses on other objectives, such as modifying the background, adding/removing objects/people, or replacing/adding audio signals.

Within the domain of manipulation in multimedia data there are different types of task that we can solve, the state of the art has focused on the *binary classification* of data between authentic and manipulated, however other options have not been explored, such as temporal localisation, which shows which parts of the videos are modified, which not only shows if the samples are manipulated but also indicates the specific parts that have been manipulated.

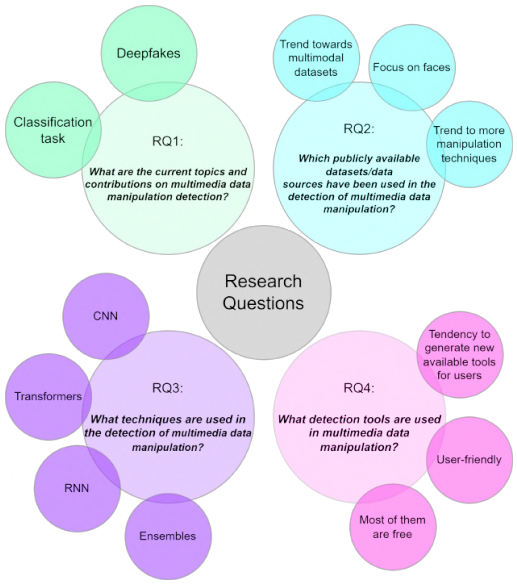


Figure 24: Key items of the answers to the proposed research questions.

RQ2. Which publicly available datasets are currently used in multimedia data manipulation detection?

In this topic, multimedia data manipulation, most authors primarily use public datasets, although some still create their own datasets. These datasets can be used independently or in combination with other state-of-the-art datasets. The availability of information varies between different modalities. Figure 3 shoes that the video modality has a larger number of datasets, while multimodal datasets are the most limited in the field. The trend in these datasets is towards increasing complexity, encompassing a larger variety of manipulation techniques, a greater number of samples, and a more realistic approach. These advances bring the datasets closer to real-world situations and enable the detection of a wider range of manipulation techniques.

A clear evolution can be observed where the initial focus was on detecting manipulation in video by utilising visual

features alone. Subsequently, the researchers expanded their study to include the field of audio. In recent years, there has been a growing emphasis on multimodal datasets, which not only provide a greater amount of information, but also present more diverse scenarios. This emphasis supports a more comprehensive analysis of multimedia data manipulation.

RQ3: What DL techniques techniques are used in the multimedia data manipulation detection?

Section 6 has revealed that the most widely used architecture for multimedia data forensics is CNNs, because many authors process the different modalities as images, including audio signals. The other two architectures that are also widely represented in this domain are RNNs and transformers. Another relevant aspect is that most works focus on classifying the samples between real and manipulated, without exploring other options that would be more informative. Furthermore, there is a generalised lack of explainability, with very few authors applying explainable AI techniques. If we analyse the number of articles of each modality, we continue to observe the same pattern as in the datasets, a greater number of video forensics systems, and the one with the lowest number of papers is multimodal forensics. However, it seems that in the coming years this relationship will be reversed and more and more importance will be given to multimodality.

RQ4: What multimedia manipulation detection tools are available for non-expert users?

Disinformation is a serious problem nowadays, and we find more and more manipulated multimedia information on social networks and websites, which can affect many users. This has led to a great need for tools that allow us to corroborate the multimedia content that we find on the Internet. Numerous tools have appeared in recent years, many of them free of charge, making verification available to all users.

As has been seen with other modalities, such as text, the trend seems to indicate that in the coming years much more accurate and user-friendly tools will be created, and some of them will probably even be integrated into many social networks. Scientific works are too technical and not accessible to the general public, so this type of tool will become an essential pillar in the fight against information manipulation and therefore disinformation.

8.2. Future trends and challenges

The preceding research questions and their corresponding answers provide an extensive and comprehensive overview of existing developments in in the field of techiques and sytems for the detection of manipulations in multimedia data. However, the essential literature review conducted to answer these questions has also revealed several ideas and prospects for future exploration in this field of research. This section describes the next trends that the literature is ready to receive, building on its current state. In addition, it identifies the challenges that the research community will face and suggests possible ways to address them effectively. Figure 25 presents a schematic summary of these emerging trends and challenges.

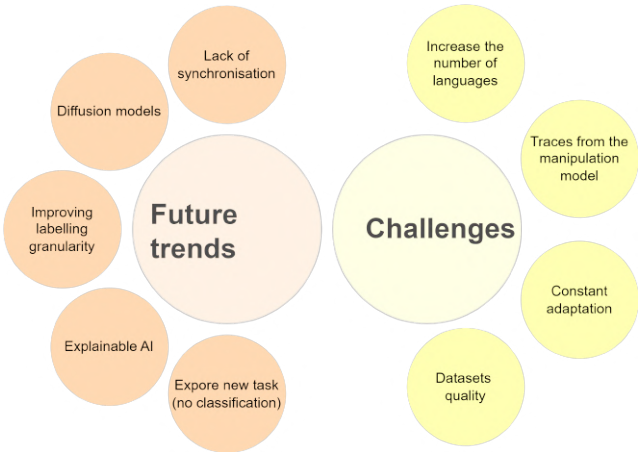


Figure 25: Future trends and challenges of forensics techniques on multimedia data.

Throughout this review of the state of the art, several trends have been observed, and it has also been possible to see which ways will establish the **future trends** in this field. We will highlight the most promising research lines for the near future.

1. ***Diffusion models*** have achieved excellent performance in image and video generation in recent years (see, for example, Blattmann et al. [277] or Ho et al. [278]) and is expected to become a widely used technique for multimedia data manipulation in the near future. These techniques have already begun to be used for image manipulation with promising results [279]. In the task of generating manipulated multimedia data, diffusion models can be used to manipulate existing videos and generate modified versions of them. For example, they can be used to perform tasks such as changing styles, removing or adding objects, altering facial attributes, and manipulating the appearance of objects, among others. The diffusion process in these models involves iterating over the input data and gradually refining the generation to fit a target distribution or to achieve the desired effect. At each iteration step, small perturbations are added to the initial data to generate a new version and refined as the stages progress. These diffusion models can be trained using techniques such as supervised learning or reinforcement learning, depending on the specific task to be addressed. In addition, they can be combined with other techniques, such as the use of GANs, to further improve the quality and diversity of manipulated multimedia generations. If you want to manipulate the audio for the sample as well, you will have to apply audio signal manipulation techniques and finally synchronise them carefully, so that no inconsistencies arise between the two modalities, video and audio.

2. Regarding the ***development of systems for the detection of manipulated data***, it is expected that the trend observed throughout this survey will continue and that ***multimodal classification systems*** will begin to increase. This approach provides greater richness and complexity of information. When audio and video are combined, a more complete representation of the information is obtained. Audio provides audio details such as speech, music, and sound effects, while video captures visual information such as facial expressions, movement, and interaction between objects. By working with both modalities, a wider range of features and contexts can be captured, resulting in a richer and more detailed representation of the data. By focussing on both modalities, any user will be able to pick up subtle signs of manipulation, such as inconsistencies and lack of temporal continuity, among others. Furthermore, unlike work that focusses only on manipulating visual information, multimodal systems provide a new variety of situations where only one modality is manipulated while holding the other modality constant or manipulating both. Training detection systems that include this variety will be closer to reality and will have greater generalisability.
3. The labelling used in most datasets, public and private, is at the level of the entire sample, video, audio, or multimodal; however, there is another possibility, which has only been observed in some articles, which is the temporal localisation, ***improving labelling granularity*** of the samples, which will improve the quality of the datasets, increase the available information, and help improve the explainability of the systems. Although this approach may seem an arduous task, labelling tools have been developed that would allow us to label the different segments of the video, both visual and acoustic information, as real or manipulated. An example of an interesting tool with many opportunities is Label Studio [280].
4. Finally, the lack of ***explainable AI*** (XAI) techniques in most articles is striking. Considering who the end-users are, non-expert internet users, information teachers, among others, the application of explainable AI techniques is essential for users to trust the systems, due to the opacity they present as "black box" algorithms. It is a fact that the application of these techniques could greatly improve the quality of classification or localisation systems when dealing with dynamic samples. Regarding the explainability techniques that can be applied, there is a wide variety, but we have to be careful with the selected techniques, as inappropriate visualisations may cause the reverse effect and confuse the user more, thus compromising the reliability of the system [281].
5. Manipulation techniques still make mistakes, such as the lack of synchronisation between the different modalities, video and audio, and between manipulated and real segments. These inconsistencies are used in multimodal forensics techniques. Future datasets will have to ***pay special attention*** to video ***fluency***, trying to ***minimise inconsistencies*** between real and manipulated fragments and between different modalities. Synchronisation should be not only between the different modalities, video and audio, but also between the manipulated and real fragments, as these are weaknesses that make it easier for forensics systems to work. New generative and manipulation models will have to pay special attention to this and try to improve the quality of the datasets.

Secondly, we will look at the different challenges that the field of multimedia content forensics will face. Throughout this review, we have observed a number of weaknesses that need to be overcome and present a **challenge** for future work in the area.

1. Most audio and multimodal work tends to use a single language, usually English and occasionally Japanese and Chinese. This limitation of the languages of the datasets may create bias, making it difficult for the system to detect signals of manipulation if the language is different. Due to the lack of analysis on this aspect, it would be necessary to conduct in-depth studies on whether bias is being created and, if so, to consider ***multilinguality*** in future datasets. Furthermore, the acoustic information provided can help to understand situations and contexts, depending on culture. It is not a simple task to include different languages in future datasets, but it would be very interesting as it would avoid biases and provide a wider variety of different contexts and situations from diverse cultures, as they belong to different population groups.
2. The ***quality*** of current ***datasets*** is limited, most of them focus on the subjects of the dataset, and many are recorded in controlled environments, show traces of manipulation techniques and the number of techniques used is limited. This makes it difficult for forensics systems not to be able to generalise correctly in real situations, so it will be very important to generate new, more complex datasets, with a wider variety of manipulation techniques that will allow the development of new quality detection techniques applicable to new problems.
3. As we have seen throughout this survey, there are a wide variety of manipulation techniques, focussing on the transformation of the face, changing the colour or background, adding or removing elements from the sample, etc. These techniques can be generated with different models and architectures, each of them generating ***representative traces of the model*** in the samples. These features can generate a bias in forensics systems. If a system has been trained on samples generated with a single model, it will not be able to generalise correctly to samples generated with other models. Therefore, datasets have to be generated with a wide variety of methodologies or models, even for

the same manipulation technique, allowing the model to improve the generalisability for methodologies that it has not seen previously and to be able to better detect samples close to reality.

4. It is clear that this line of research consists of two tracks, manipulation techniques and forensics techniques. Malicious actors are increasingly creating manipulated samples that are more precise and difficult to detect. On the other hand, researchers are trying to develop new detection systems that can detect new manipulation techniques. This domain is based on a race between the creation of new manipulation techniques and the development of new detection systems. Therefore, it is very important to be in ***constant adaptation*** and to develop new data sets that are increasingly closer to reality and of better quality that allow us to develop new and more effective detection systems that allow us to fight disinformation, which is becoming an increasing problem in social networks.

8.3. Conslusion

In recent years, the manipulation of multimedia content has become a serious problem in social networks. Within this field we can find both manipulation of image (or frames from a video), video, audio, or multimodal manipulation of the visual and acoustic information of the sample. In order to fight against this type of disinformation, we need quality datasets that are as close as possible to the real world and effective forensic systems that are capable of detecting multiple types of manipulation and can be applied in real situations. Although there is still a long way to go, we strongly believe that this is the right way forward to fight the manipulation of multimedia content and thus disinformation on the internet.

Advances in Deep Learning in recent years have provided us with the necessary techniques to cope with the manipulation of multimedia content. This review aims to achieve this goal by describing the following items for the reader:

1. The most common manipulation techniques and available datasets.
2. The most widely used forensic techniques in the state of the art.
3. The tools already implemented and available to users.
4. An analysis of the evolution of the domain in the last six years.

Multimedia content manipulation and detection techniques have been critically analysed with the aim of defining the lines of research to be followed in the coming years and the current challenges in the domain. From the information analysed in this survey, we were able to extract the main ideas of the domain to facilitate and stimulate research on the detection of multimedia content manipulation.

Table 21
Download links to all the datasets included in this survey.

	Dataset	Link
Video	UADFV	www.cs.albany.edu/âĹlsw/downloads.html
	HOHA-based dataset	www.irisa.fr/vista/actions
	DeepFake TIMIT	www.idiap.ch/en/dataset/deepfaketimit
	FaceForensics++	github.com/ondyari/FaceForensics
	Celeb-DF	github.com/yuezunli/celeb-deepfakeforensics
	DFDC-Preview	cvlab.cse.msu.edu/project-ffd.html
	DFDC	cvlab.cse.msu.edu/project-ffd.html
	DeeperForensics	github.com/EndlessSora/DeeperForensics-1.0
	WildDeepFake	github.com/deepfakeinthewild/deepfake-in-the-wild
	ForgeryNet	yinanhe.github.io/projects/forgerynet.html
	KoDF	deepbrainai-research.github.io/kodf/
	FFIW 10K	github.com/tfzhou/FFIW
Audio	ASVspoof 2019	www.asvspoof.org
	ASVspoof 2021	www.asvspoof.org
	FoR	bil.eecs.yorku.ca/datasets
	WaveFake	zenodo.org/record/5642694#.ZBJ1THbMKUk
	FAD	zenodo.org/record/6635521#.ZBJkgnbMKUm
	IWA	deepfake-demo.aisec.fraunhofer.de/in_the_wild
Multimodal	LAV-DF	https://github.com/ControlNet/LAV-DF
	FakeAVCeleb	sites.google.com/view/fakeavcelebdash-lab/
	VideoSham	github.com/adobe-research/VideoSham-dataset

Table 22
Download links to other datasets not included in section 5.

	Article	Link
Video	Guo et al. [196]	github.com/EricGzq/Hybrid-Fake-Face-Dataset
	Dang et al. [205]	cvlab.cse.msu.edu/dffd-dataset.html
	Kolagati et al. [230]	github.com/dessa-oss/DeepFake-Detection
Audio	Zhang et al. [236]	github.com/Amforever/FMFCC-A
	Zhang et al. [246]	github.com/nii-yamagishilab/PartialSpoof
Multimodal	Yang et al. [258]	pan.baidu.com/s/1MckHs-H57jTma5v0o6XYMA

Table 23
Download links to the available codes and models of the analysed works.

	Article	Link
Video	Guo et al. [196]	github.com/EricGzq/AMTENnet
	Chen et al. [202]	github.com/LightningChan/DLFMNet
	Dang et al. [205]	github.com/JStehouwer/FFD_CVPR2020
	Zhao et al. [208]	github.com/yoctta/multiple-attention
	Lu et al. [224]	github.com/Yujiang-Lu/CWSA-tensorflow
	Cozzolino et al. [225]	github.com/grip-unina/id-reveal
	Pu et al. [227]	github.com/PWB97/Deepfake-detection
Audio	Dongre et al. [232]	github-dev.cs.illinois.edu/athimma2/deepfake-audio-classifier
	Wang and Yamagishi [235]	github.com/nii-yamagishilab/project-NN-Pytorch-scripts
	Zhang et al. [236]	github.com/Amforever/FMFCC-A
	Wang and Yamagishi [238]	github.com/nii-yamagishilab/project-NN-Pytorch-scripts
	Tak et al. [239]	github.com/eurecom-asp/rawnet2-antispoofing
	Hua et al. [241]	github.com/ghua-ac/end-to-end-synthetic-speech-detection
	Kawa et al. [242]	github.com/piotrkawa/specrnet
	Zhang et al. [246]	github.com/nii-yamagishilab/PartialSpoof
	Jung et al. [248]	github.com/clovaai/aasist
Multimodal	Lewis et al. [253]	github.com/jklewis99/MultimodalDeepfakeDetection
	Cozzolino et al. [264]	github.com/grip-unina/poi-forensics
	Cai et al. [145]	github.com/ControlNet/LAV-DF
	Yang et al. [258]	github.com/SYSU-DISG/AVoiD-DF

A. Datasets download links

In this section we have included all the download links for the datasets described in the survey (Table 21) and for those that have been used to create forensic systems (Table 22 but not described in Section 5.

B. Codes availables from forensics systems

The available codes for the different works described in the survey, Section 6, are shown in Table 23.

References

[1] A. Zareie, R. Sakellariou, Minimizing the spread of misinformation in online social networks: A survey, Journal of Network and Computer Applications 186 (2021) 103094.

[2] C. Ireton, J. Posetti, Journalism, fake news & disinformation: handbook for journalism education and training, Unesco Publishing, 2018.

[3] S. Zannettou, T. Caulfield, J. Blackburn, E. De Cristofaro, M. Sirivianos, G. Stringhini, G. Suarez-Tangil, On the origins of memes by means of fringe web communities, in: Proceedings of the Internet Measurement Conference 2018, 2018, pp. 188–202.

[4] X. Zhu, Y. Kim, H. Park, Do messages spread widely also diffuse fast? examining the effects of message characteristics on information diffusion, Computers in human behavior 103 (2020) 37–47.

[5] M. Hameleers, T. E. Powell, T. G. Van Der Meer, L. Bos, A picture paints a thousand lies? the effects and mechanisms of multimodal disinformation and rebuttals disseminated via social media, Political Communication 37 (2020) 281–301.

[6] S. Tyagi, D. Yadav, A detailed analysis of image and video forgery detection techniques, The Visual Computer (2022) 1–21.

[7] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, H. Malik, Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward, Applied Intelligence (2022) 1–53.

[8] A. S. Abdulreda, A. J. Obaid, A landscape view of deepfake techniques and detection methods, International Journal of Nonlinear Analysis and Applications 13 (2022) 745–755.

[9] A. Mitra, S. P. Mohanty, P. Corcoran, E. Kougianos, A machine learning based approach for deepfake detection in social media through key video frame extraction, SN Computer Science 2 (2021) 1–18.

[10] J. Hu, X. Liao, W. Wang, Z. Qin, Detecting compressed deepfake videos in social networks using frame-temporality two-stream convolutional network, IEEE Transactions on Circuits and Systems for Video Technology 32 (2021) 1089–1102.

[11] H. Stiff, F. Johansson, Detecting computer-generated disinformation, International Journal of Data Science and Analytics 13 (2022) 363–383.

[12] J. Mallet, R. Dave, N. Seliya, M. Vanamala, Using deep learning to detecting deepfakes, arXiv preprint arXiv:2207.13644 (2022).

[13] C. Papastergiopoulos, A. Vafeiadis, I. Papadimitriou, K. Votis, D. Tzovaras, On the generalizability of two-dimensional convolutional neural networks for fake speech detection, in: Proceedings of the 1st International Workshop on Multimedia AI against Disinformation, 2022, pp. 3–9.

[14] N. Ljubešić, I. Mozetič, P. K. Novak, Quantifying the impact of context on the quality of manual hate speech annotation, Natural Language Engineering (2022) 1–14.

[15] M. Popa-Wyatt, J. L. Wyatt, Slurs, roles and power, Philosophical Studies 175 (2018) 2879–2906.

[16] S. Ullmann, M. Tomalin, Quarantining online hate speech: technical and ethical perspectives, Ethics and Information Technology 22 (2020) 69–80.

[17] Q.-T. Tran, T.-P. Tran, M.-S. Dao, T.-V. La, A.-D. Tran, D. T. Dang Nguyen, A textual-visual-entailment-based unsupervised algorithm for cheapfake detection, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 7145–7149.

[18] S. Papadopoulos, G. Kordopatis-Zilos, M. Zampoglou, O. Papadopoulou, Dataset column: Datasets for online multimedia verification, ACM SIGMultimedia Records 11 (2022) 1–1.

[19] D. Dagar, D. K. Vishwakarma, A literature review and perspectives in deepfakes: generation, detection, and applications, International Journal of Multimedia Information Retrieval (2022) 1–71.

[20] K. E. Ak, Y. Sun, J. H. Lim, Learning by imagination: A joint framework for text-based image manipulation and change captioning, IEEE Transactions on Multimedia (2022).

[21] Y. Mirsky, W. Lee, The creation and detection of deepfakes: A survey, ACM Computing Surveys (CSUR) 54 (2021) 1–41.

[22] S. Chen, L. Xiao, A. Kumar, Spread of misinformation on social media: What contributes to it and how to combat it, Computers in Human Behavior (2022) 107643.

[23] M. Albahar, J. Almalki, Deepfakes: Threats and countermeasures systematic review, Journal of Theoretical and Applied Information Technology 97 (2019) 3242–3250.

[24] B. Chesney, D. Citron, Deep fakes: A looming challenge for privacy, democracy, and national security, Calif. L. Rev. 107 (2019) 1753.

[25] F. Juefei-Xu, R. Wang, Y. Huang, Q. Guo, L. Ma, Y. Liu, Countering malicious deepfakes: Survey, battleground, and horizon, International Journal of Computer Vision (2022) 1–57.

[26] J. Jing, H. Wu, J. Sun, X. Fang, H. Zhang, Multimodal fake news detection via progressive fusion networks, Information processing & management 60 (2023) 103120.

[27] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Communications of the ACM 63 (2020) 139–144.

[28] T. Zhang, Deepfake generation and detection, a survey, Multimedia Tools and Applications 81 (2022) 6259–6276.

[29] F.-A. Croitoru, V. Hondru, R. T. Ionescu, M. Shah, Diffusion models in vision: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence (2023).

[30] J. Cao, P. Qi, Q. Sheng, T. Yang, J. Guo, J. Li, Exploring the role of visual content in fake news detection, Disinformation, Misinformation, and Fake News in Social Media (2020) 141–161.

[31] K. Shu, A. Bhattacharjee, F. Alatawi, T. H. Nazer, K. Ding, M. Karami, H. Liu, Combating disinformation in a social media age, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 10 (2020) e1385.

[32] F. Alam, S. Cresci, T. Chakraborty, F. Silvestri, D. Dimitrov, G. D. S. Martino, S. Shaar, H. Firooz, P. Nakov, A survey on multimodal disinformation detection, arXiv preprint arXiv:2103.12541 (2021).

[33] S. B. Parikh, P. K. Atrey, Media-rich fake news detection: A survey, in: 2018 IEEE conference on multimedia information processing and retrieval (MIPR), IEEE, 2018, pp. 436–441.

[34] M. Choraś, K. Demestichas, A. Giełczyk, Á. Herrero, P. Ksieniewicz, K. Remoundou, D. Urda, M. Woźniak, Advanced machine learning techniques for fake news (online disinformation) detection: A systematic mapping study, Applied Soft Computing 101 (2021) 107050.

[35] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, J. Ortega-Garcia, Deepfakes and beyond: A survey of face manipulation and fake detection, Information Fusion 64 (2020) 131–148.

[36] X. Ju, An overview of face manipulation detection, Journal of Cybersecurity 2 (2020) 197.

[37] S. Pashine, S. Mandiya, P. Gupta, R. Sheikh, Deep fake detection: Survey of facial manipulation detection solutions, arXiv preprint arXiv:2106.12605 (2021).

[38] P. Yu, Z. Xia, J. Fei, Y. Lu, A survey on deepfake video detection, Iet Biometrics 10 (2021) 607–624.

[39] M. Weerawardana, T. Fernando, Deepfakes detection methods: A literature survey, in: 2021 10th International Conference on Information and Automation for Sustainability (ICIAfS), IEEE, 2021, pp. 76–81.

[40] A. Malik, M. Kuribayashi, S. M. Abdullahi, A. N. Khan, Deepfake detection for human face images and videos: A survey, Ieee Access 10 (2022) 18757–18775.

[41] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, H. Malik, Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward, Applied intelligence 53 (2023) 3974–4026.

[42] C. Comito, L. Caroprese, E. Zumpano, Multimodal fake news detection on social media: a survey of deep learning techniques, Social Network Analysis and Mining 13 (2023) 1–22.

[43] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, et al., The prisma 2020 statement: an updated guideline for reporting systematic reviews, International journal of

surgery 88 (2021) 105906.
[44] M. Ferrara, A. Franco, D. Maltoni, Decoupling texture blending and shape warping in face morphing, in: 2019 international conference of the biometrics special interest group (BIOSIG), IEEE, 2019, pp. 1–5.
[45] H. Zhang, S. Venkatesh, R. Ramachandra, K. Raja, N. Damer, C. Busch, Mipgan—generating strong and high quality morphing attacks using identity prior driven gan, IEEE Transactions on Biometrics, Behavior, and Identity Science 3 (2021) 365–383.
[46] L. Moser, J. Selfe, D. Hendler, D. Roble, Dynamic neural face morphing for visual effects, in: SIGGRAPH Asia 2021 Technical Communications, 2021, pp. 1–4.
[47] N. Damer, A. M. Saladie, A. Braun, A. Kuijper, Morgan: Recognition vulnerability and attack detectability of face morphing attacks created by generative adversarial network, in: 2018 IEEE 9th international conference on biometrics theory, applications and systems (BTAS), IEEE, 2018, pp. 1–10.
[48] Z. Blasingame, C. Liu, Diffusion models for stronger face morphing attacks, arXiv preprint arXiv:2301.04218 (2023).
[49] S. Venkatesh, H. Zhang, R. Ramachandra, K. Raja, N. Damer, C. Busch, Can gan generated morphs threaten face recognition systems equally as landmark based morphs?-vulnerability and detection, in: 2020 8th International Workshop on Biometrics and Forensics (IWBF), IEEE, 2020, pp. 1–6.
[50] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 4401–4410.
[51] N. Zhang, X. Liu, X. Li, G.-J. Qi, Morphganformer: Transformer-based face morphing and de-morphing, arXiv preprint arXiv:2302.09404 (2023).
[52] D. A. Hudson, L. Zitnick, Generative adversarial transformers, in: International conference on machine learning, PMLR, 2021, pp. 4487–4499.
[53] J. Song, C. Meng, S. Ermon, Denoising diffusion implicit models, arXiv preprint arXiv:2010.02502 (2020).
[54] K. Sun, S. Wu, Z. Huang, N. Zhang, Q. Wang, H. Li, Controllable 3d face synthesis with conditional generative occupancy fields, arXiv preprint arXiv:2206.08361 (2022).
[55] P. Zhuang, L. Ma, S. Koyejo, A. Schwing, Controllable radiance fields for dynamic face synthesis, in: 2022 International Conference on 3D Vision (3DV), IEEE, 2022, pp. 1–11.
[56] J. Sun, H. Yu, J. J. Zhang, J. Dong, H. Yu, G. Zhong, Face image-sketch synthesis via generative adversarial fusion, Neural Networks 154 (2022) 179–189.
[57] N. K. Yadav, S. K. Singh, S. R. Dubey, Csa-gan: Cyclic synthesized attention guided generative adversarial network for face synthesis, Applied Intelligence 52 (2022) 12704–12723.
[58] T. Yoshikawa, Y. Endo, Y. Kanamori, Diversifying detail and appearance in sketch-based face image synthesis, The Visual Computer 38 (2022) 3121–3133.
[59] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, T. Aila, Analyzing and improving the image quality of stylegan, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 8110–8119.
[60] B. Zhang, S. Gu, B. Zhang, J. Bao, D. Chen, F. Wen, Y. Wang, B. Guo, Styleswin: Transformer-based gan for high-resolution image generation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 11304–11314.
[61] H. Zhou, Y. Liu, Z. Liu, P. Luo, X. Wang, Talking face generation by adversarially disentangled audio-visual representation, in: Proceedings of the AAAI conference on artificial intelligence, volume 33, 2019, pp. 9299–9306.
[62] C. Du, Q. Chen, T. He, X. Tan, X. Chen, K. Yu, S. Zhao, J. Bian, Dae-talker: High fidelity speech-driven talking face generation with diffusion autoencoder, arXiv preprint arXiv:2303.17550 (2023).
[63] M. Stypulkowski, K. Vougioukas, S. He, M. Zikeba, S. Petridis, M. Pantic, Diffused heads: Diffusion models beat gans on talking-face generation, arXiv preprint arXiv:2301.03396 (2023).
[64] W. Cao, T. Wang, A. Dong, M. Shu, Transfs: Face swapping using transformer, in: 2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG), IEEE, 2023, pp. 1–8.
[65] I. Perov, D. Gao, N. Chervoniy, K. Liu, S. Marangonda, C. Umé, M. Dpfks, C. S. Facenheim, L. RP, J. Jiang, et al., Deepfacelab: Integrated, flexible and extensible face-swapping framework, arXiv preprint arXiv:2005.05535 (2020).
[66] Y. Wang, X. Chen, J. Zhu, W. Chu, Y. Tai, C. Wang, J. Li, Y. Wu, F. Huang, R. Ji, Hififace: 3d shape and semantic prior guided high fidelity face swapping, arXiv preprint arXiv:2106.09965 (2021).
[67] Y. Nirkin, I. Masi, A. T. Tuan, T. Hassner, G. Medioni, On face segmentation, face swapping, and face perception, in: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE, 2018, pp. 98–105.
[68] D. Bitouk, N. Kumar, S. Dhillon, P. Belhumeur, S. K. Nayar, Face swapping: automatically replacing faces in photographs, in: ACM SIGGRAPH 2008 papers, 2008, pp. 1–8.
[69] G. Gao, H. Huang, C. Fu, Z. Li, R. He, Information bottleneck disentanglement for identity swapping, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 3404–3413.
[70] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, J. Choo, Stargan: Unified generative adversarial networks for multi-domain image-to-image translation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8789–8797.
[71] Y. Xu, B. Deng, J. Wang, Y. Jing, J. Pan, S. He, High-resolution face swapping via latent semantics disentanglement, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 7642–7651.
[72] Y. Li, P. Sun, H. Qi, S. Lyu, Toward the creation and obstruction of deepfakes, in: Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks, Springer International Publishing Cham, 2022, pp. 71–96.
[73] L. Li, J. Bao, H. Yang, D. Chen, F. Wen, Faceshifter: Towards high fidelity and occlusion aware face swapping, arXiv preprint arXiv:1912.13457 (2019).
[74] Q. Li, W. Wang, C. Xu, Z. Sun, Learning disentangled representation for one-shot progressive face swapping, arXiv preprint arXiv:2203.12985 (2022).
[75] Y. Nirkin, Y. Keller, T. Hassner, Fsganv2: Improved subject agnostic face swapping and reenactment, IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (2022) 560–575.
[76] C. Shu, H. Wu, H. Zhou, J. Liu, Z. Hong, C. Ding, J. Han, J. Liu, E. Ding, J. Wang, Few-shot head swapping in the wild, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10789–10798.
[77] R. Chen, X. Chen, B. Ni, Y. Ge, Simswap: An efficient framework for high fidelity face swapping, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 2003–2011.
[78] G.-S. J. Hsu, H.-Y. Wu, Pose-guided and style-transferred face reenactment, in: 2021 IEEE International Conference on Image Processing (ICIP), IEEE, 2021, pp. 2458–2462.
[79] X. Fu, X. Wang, J. Liu, W. Liu, J. Dai, J. Han, Makeitsmile: Detail-enhanced smiling face reenactment, in: 2022 International Joint Conference on Neural Networks (IJCNN), IEEE, 2022, pp. 1–8.
[80] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Niessner, P. Pérez, C. Richardt, M. Zollhöfer, C. Theobalt, Deep video portraits, ACM Transactions on Graphics (TOG) 37 (2018) 1–14.
[81] C. Hu, X. Xie, L. Wu, Face reenactment via generative landmark guidance, Image and Vision Computing 130 (2023) 104611.
[82] S. Bounareli, C. Tzelepis, V. Argyriou, I. Patras, G. Tzimiropoulos, Stylemask: Disentangling the style space of stylegan2 for neural face reenactment, in: 2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG), IEEE, 2023, pp. 1–8.

[83] C. Chan, S. Ginosar, T. Zhou, A. A. Efros, Everybody dance now, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 5933–5942.

[84] J. Ren, M. Chai, S. Tulyakov, C. Fang, X. Shen, J. Yang, Human motion transfer from poses in the wild, in: Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16, Springer, 2020, pp. 262–279.

[85] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, Z. Ling, The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods, arXiv preprint arXiv:1804.04262 (2018).

[86] L. Liu, W. Xu, M. Zollhoefer, H. Kim, F. Bernard, M. Habermann, W. Wang, C. Theobalt, Neural animation and reenactment of human actor videos, arXiv preprint arXiv:1809.03658 (2018).

[87] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, B. Catanzaro, High-resolution image synthesis and semantic manipulation with conditional gans, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8798–8807.

[88] T. Ki, D. Min, Stylelipsync: Style-based personalized lip-sync video generation, arXiv preprint arXiv:2305.00521 (2023).

[89] K. Vougioukas, S. Petridis, M. Pantic, End-to-end speech-driven realistic facial animation with temporal gans., in: CVPR Workshops, 2019, pp. 37–40.

[90] K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, C. Jawahar, A lip sync expert is all you need for speech to lip generation in the wild, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 484–492.

[91] P. KR, R. Mukhopadhyay, J. Philip, A. Jha, V. Namboodiri, C. Jawahar, Towards automatic face-to-face translation, in: Proceedings of the 27th ACM international conference on multimedia, 2019, pp. 1428–1436.

[92] S. Suwajanakorn, S. M. Seitz, I. Kemelmacher-Shlizerman, Synthesizing obama: learning lip sync from audio, ACM Transactions on Graphics (ToG) 36 (2017) 1–13.

[93] Y. Shalev, L. Wolf, End to end lip synchronization with a temporal autoencoder, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 341–350.

[94] L. Song, W. Wu, C. Qian, R. He, C. C. Loy, Everybody's talkin': Let me talk as you want, IEEE Transactions on Information Forensics and Security 17 (2022) 585–598.

[95] A. Lahiri, V. Kwatra, C. Frueh, J. Lewis, C. Bregler, Lipsync3d: Data-efficient learning of personalized 3d talking faces from video using pose and lighting normalization, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 2755–2764.

[96] G. Wang, P. Zhang, L. Xie, W. Huang, Y. Zha, Attention-based lip audio-visual synthesis for talking face generation in the wild, arXiv preprint arXiv:2203.03984 (2022).

[97] Y. Shen, C. Yang, X. Tang, B. Zhou, Interfacegan: Interpreting the disentangled face representation learned by gans, IEEE transactions on pattern analysis and machine intelligence 44 (2020) 2004–2018.

[98] H. Kim, Y. Choi, J. Kim, S. Yoo, Y. Uh, Exploiting spatial dimensions of latent in gan for real-time image editing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 852–861.

[99] A. P. Fard, M. H. Mahoor, S. A. Lamer, T. Sweeny, Ganalyzer: Analysis and manipulation of gans latent space for controllable face synthesis, arXiv preprint arXiv:2302.00908 (2023).

[100] S. H. Mohammadi, A. Kain, An overview of voice conversion systems, Speech Communication 88 (2017) 65–82.

[101] R. Aihara, R. Takashima, T. Takiguchi, Y. Ariki, Gmm-based emotional voice conversion using spectrum and prosody features, American Journal of Signal Processing 2 (2012) 134–138.

[102] Z. Yue, X. Zou, Y. Jia, H. Wang, Voice conversion using hmm combined with gmm, in: 2008 Congress on Image and Signal Processing, volume 5, IEEE, 2008, pp. 366–370.

[103] V. Popa, H. Silen, J. Nurminen, M. Gabbouj, Local linear transformation for voice conversion, in: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2012, pp. 4517–4520.

[104] P. Song, Y. Bao, L. Zhao, C. Zou, Voice conversion using support vector regression, Electronics letters 47 (2011) 1045–1046.

[105] S. Kannan, P. R. Raju, R. S. S. Madhav, S. Tripathi, Voice conversion using spectral mapping and td-psola, in: Advances in Computing and Network Communications: Proceedings of CoCoNet 2020, Volume 2, Springer, 2021, pp. 193–205.

[106] L.-H. Chen, Z.-H. Ling, L.-J. Liu, L.-R. Dai, Voice conversion using deep neural networks with layer-wise generative training, IEEE/ACM Transactions on Audio, Speech, and Language Processing 22 (2014) 1859–1872.

[107] E. Azarov, M. Vashkevich, D. Likhachov, A. A. Petrovsky, Real-time voice conversion using artificial neural networks with rectified linear units., in: INTERSPEECH, 2013, pp. 1032–1036.

[108] K.-S. Lee, Restricted boltzmann machine-based voice conversion for nonparallel corpus, IEEE signal processing letters 24 (2017) 1103–1107.

[109] Y. Y. Lin, C.-M. Chien, J.-H. Lin, H.-y. Lee, L.-s. Lee, Fragmentvc: Any-to-any voice conversion by end-to-end extracting and fusing fine-grained voice fragments with attention, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2021, pp. 5939–5943.

[110] T. Kaneko, H. Kameoka, K. Tanaka, N. Hojo, Stargan-vc2: Rethinking conditional methods for stargan-based voice conversion, arXiv preprint arXiv:1907.12279 (2019).

[111] J. Lian, C. Zhang, D. Yu, Robust disentangled variational speech representation learning for zero-shot voice conversion, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2022, pp. 6572–6576.

[112] Y.-H. Chen, D.-Y. Wu, T.-H. Wu, H.-y. Lee, Again-vc: A one-shot voice conversion using activation guidance and adaptive instance normalization, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2021, pp. 5954–5958.

[113] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, M. A. Ponti, Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone, in: International Conference on Machine Learning, PMLR, 2022, pp. 2709–2720.

[114] B. Nguyen, F. Cardinaux, Nvc-net: End-to-end adversarial voice conversion, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2022, pp. 7012–7016.

[115] J. Kim, J. Kong, J. Son, Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech, in: International Conference on Machine Learning, PMLR, 2021, pp. 5530–5540.

[116] H. Kameoka, T. Kaneko, K. Tanaka, N. Hojo, Acvae-vc: Non-parallel voice conversion with auxiliary classifier variational autoencoder, IEEE/ACM Transactions on Audio, Speech, and Language Processing 27 (2019) 1432–1443.

[117] H. Tachibana, K. Uenoyama, S. Aihara, Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention, in: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, 2018, pp. 4784–4788.

[118] N. Li, S. Liu, Y. Liu, S. Zhao, M. Liu, Neural speech synthesis with transformer network, in: Proceedings of the AAAI conference on artificial intelligence, volume 33, 2019, pp. 6706–6713.

[119] Y. Liu, Z. Xu, G. Wang, K. Chen, B. Li, X. Tan, J. Li, L. He, S. Zhao, Delightfultts: The microsoft speech synthesis system for blizzard challenge 2021, arXiv preprint arXiv:2110.12612 (2021).

[120] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, et al., Natural tts synthesis by conditioning wavenet on mel spectrogram predictions, in: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, 2018, pp. 4779–4783.

[121] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu, Wavenet:

A generative model for raw audio, arXiv preprint arXiv:1609.03499 (2016).

[122] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, M. Kudinov, Grad-tts: A diffusion probabilistic model for text-to-speech, in: International Conference on Machine Learning, PMLR, 2021, pp. 8599–8608.

[123] R. Prenger, R. Valle, B. Catanzaro, Waveglow: A flow-based generative network for speech synthesis, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 3617–3621.

[124] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. Oord, S. Dieleman, K. Kavukcuoglu, Efficient neural audio synthesis, in: International Conference on Machine Learning, PMLR, 2018, pp. 2410–2419.

[125] J. Kong, J. Kim, J. Bae, Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis, Advances in Neural Information Processing Systems 33 (2020) 17022–17033.

[126] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, T.-Y. Liu, Fastspeech 2: Fast and high-quality end-to-end text to speech, arXiv preprint arXiv:2006.04558 (2020).

[127] J. Donahue, S. Dieleman, M. Bińkowski, E. Elsen, K. Simonyan, End-to-end adversarial text-to-speech, arXiv preprint arXiv:2006.03575 (2020).

[128] A. Łańcucki, Fastpitch: Parallel text-to-speech with pitch prediction, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2021, pp. 6588–6592.

[129] Y. Lei, S. Yang, J. Cong, L. Xie, D. Su, Glow-wavegan 2: High-quality zero-shot text-to-speech synthesis and any-to-any voice conversion, arXiv preprint arXiv:2207.01832 (2022).

[130] R. Huang, Z. Zhao, H. Liu, J. Liu, C. Cui, Y. Ren, Prodiff: Progressive fast diffusion model for high-quality text-to-speech, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 2595–2605.

[131] M. Jeong, H. Kim, S. J. Cheon, B. J. Choi, N. S. Kim, Diff-tts: A denoising diffusion model for text-to-speech, arXiv preprint arXiv:2104.01409 (2021).

[132] T. Sadekova, V. Gogoryan, I. Vovk, V. Popov, M. Kudinov, J. Wei, A unified system for voice cloning and voice conversion through diffusion probabilistic modeling, Proc. Interspeech 2022 (2022) 3003–3007.

[133] C. Jemine, Real-time-voice-cloning, University of Liege, Liege, Belgium (2019).

[134] S. Arik, J. Chen, K. Peng, W. Ping, Y. Zhou, Neural voice cloning with a few samples, Advances in neural information processing systems 31 (2018).

[135] Q. Chen, M. Tan, Y. Qi, J. Zhou, Y. Li, Q. Wu, V2c: Visual voice cloning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 21242–21251.

[136] P. Cano, A. Loscos, J. Bonada, M. de Boer, X. Serra, Voice morphing system for impersonating in karaoke applications., in: ICMC, 2000.

[137] C. Orphanidou, I. Moroz, S. Roberts, Wavelet-based voice morphing (2004).

[138] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, H. Li, Spoofing and countermeasures for speaker verification: A survey, speech communication 66 (2015) 130–153.

[139] S. Abdelnabi, R. Hasan, M. Fritz, Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 14940–14949.

[140] T.-Y. Wang, I. Kawaguchi, H. Kuzuoka, M. Otsuki, Effect of manipulated amplitude and frequency of human voice on dominance and persuasiveness in audio conferences, Proceedings of the ACM on human-computer interaction 2 (2018) 1–18.

[141] A. Dixit, N. Kaur, S. Kingra, Review of audio deepfake detection techniques: Issues and prospects, Expert Systems (2023) e13322.

[142] H. Khalid, S. Tariq, M. Kim, S. S. Woo, Fakeavceleb: A novel audio-video multimodal deepfake dataset, arXiv preprint arXiv:2108.05080 (2021).

[143] J. Wang, Z. Li, C. Zhang, J. Chen, Z. Wu, L. S. Davis, Y.-G. Jiang, Fighting malicious media data: A survey on tampering detection and deepfake detection, arXiv preprint arXiv:2212.05667 (2022).

[144] T. Mittal, R. Sinha, V. Swaminathan, J. Collomosse, D. Manocha, Video manipulations beyond faces: A dataset with human-machine analysis, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 643–652.

[145] Z. Cai, S. Ghosh, T. Gedeon, A. Dhall, K. Stefanov, M. Hayat, " glitch in the matrix!": A large scale benchmark for content driven audio-visual forgery detection and localization, arXiv preprint arXiv:2305.01979 (2023).

[146] S. Oh, M. Kang, H. Moon, K. Choi, B. S. Chon, A demand-driven perspective on generative audio ai, arXiv preprint arXiv:2307.04292 (2023).

[147] D. Bigioi, S. Basak, H. Jordan, R. McDonnell, P. Corcoran, Speech driven video editing via an audio-conditioned diffusion model, arXiv preprint arXiv:2301.04474 (2023).

[148] Z. Tang, Z. Yang, C. Zhu, M. Zeng, M. Bansal, Any-to-any generation via composable diffusion, arXiv preprint arXiv:2305.11846 (2023).

[149] Y. Li, M.-C. Chang, S. Lyu, In ictu oculi: Exposing ai created fake videos by detecting eye blinking, in: 2018 IEEE International workshop on information forensics and security (WIFS), IEEE, 2018, pp. 1–7.

[150] D. Güera, E. J. Delp, Deepfake video detection using recurrent neural networks, in: 2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS), IEEE, 2018, pp. 1–6.

[151] P. Korshunov, S. Marcel, Deepfakes: a new threat to face recognition? assessment and detection, arXiv preprint arXiv:1812.08685 (2018).

[152] C. Sanderson, B. C. Lovell, Multi-region probabilistic histograms for robust and scalable identity inference, in: Advances in Biometrics: Third International Conference, ICB 2009, Alghero, Italy, June 2-5, 2009. Proceedings 3, Springer, 2009, pp. 199–208.

[153] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, Faceforensics++: Learning to detect manipulated facial images, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 1–11.

[154] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, M. Nießner, Face2face: Real-time face capture and reenactment of rgb videos, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2387–2395.

[155] J. Thies, M. Zollhöfer, M. Nießner, Deferred neural rendering: Image synthesis using neural textures, Acm Transactions on Graphics (TOG) 38 (2019) 1–12.

[156] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, Faceforensics: A large-scale video dataset for forgery detection in human faces, arXiv preprint arXiv:1803.09179 (2018).

[157] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, C. C. Ferrer, The deepfake detection challenge (dfdc) preview dataset, arXiv preprint arXiv:1910.08854 (2019).

[158] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, C. C. Ferrer, The deepfake detection challenge (dfdc) dataset, arXiv preprint arXiv:2006.07397 (2020).

[159] Y. Li, X. Yang, P. Sun, H. Qi, S. Lyu, Celeb-df: A large-scale challenging dataset for deepfake forensics, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 3207–3216.

[160] A. Aravkin, J. V. Burke, L. Ljung, A. Lozano, G. Pillonetto, Generalized kalman smoothing: Modeling and algorithms, Automatica 86 (2017) 63–86.

[161] E. Reinhard, M. Adhikhmin, B. Gooch, P. Shirley, Color transfer between images, IEEE Computer graphics and applications 21 (2001) 34–41.

[162] L. Jiang, R. Li, W. Wu, C. Qian, C. C. Loy, Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection, in: Proceedings of the IEEE/CVF conference on computer vision and

pattern recognition, 2020, pp. 2889–2898.
[163] B. Zi, M. Chang, J. Chen, X. Ma, Y.-G. Jiang, Wilddeepfake: A challenging real-world dataset for deepfake detection, in: Proceedings of the 28th ACM international conference on multimedia, 2020, pp. 2382–2390.
[164] Y. He, B. Gan, S. Chen, Y. Zhou, G. Yin, L. Song, L. Sheng, J. Shao, Z. Liu, Forgerynet: A versatile benchmark for comprehensive forgery analysis, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 4360–4369.
[165] P. Kwon, J. You, G. Nam, S. Park, G. Chae, Kodf: A large-scale korean deepfake detection dataset, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10744–10753.
[166] H. Zhou, Y. Sun, W. Wu, C. C. Loy, X. Wang, Z. Liu, Pose-controllable talking face generation by implicitly modularized audio-visual representation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 4176–4186.
[167] Y. Nirkin, Y. Keller, T. Hassner, Fsgan: Subject agnostic face swapping and reenactment, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 7184–7193.
[168] Y. Lu, J. Chai, X. Cao, Live speech portraits: real-time photorealistic talking-head animation, ACM Transactions on Graphics (TOG) 40 (2021) 1–17.
[169] T. Zhou, W. Wang, Z. Liang, J. Shen, Face forensics in the wild, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 5778–5788.
[170] Y. Li, X. Yang, P. Sun, H. Qi, S. Lyu, Celeb-df (v2): a new dataset for deepfake forensics, arXiv preprint arXiv:1909.12962 (2019).
[171] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, K. A. Lee, Asvspoof 2019: Future horizons in spoofed and fake audio detection, arXiv preprint arXiv:1904.05441 (2019).
[172] J. Yamagishi, C. Veaux, K. MacDonald, et al., Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92), University of Edinburgh. The Centre for Speech Technology Research (CSTR) (2019).
[173] R. Reimao, V. Tzerpos, For: A dataset for synthetic speech detection, in: 2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), IEEE, 2019, pp. 1–10.
[174] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, J. Miller, Deep voice 3: 2000-speaker neural text-to-speech, proc. ICLR (2018) 214–217.
[175] J. Kominek, A. W. Black, The cmu arctic speech databases, in: Fifth ISCA workshop on speech synthesis, 2004.
[176] K. Ito, L. Johnson, The lj speech dataset, https://keithito.com/LJ-Speech-Dataset/, 2017.
[177] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch, et al., Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild, IEEE/ACM Transactions on Audio, Speech, and Language Processing (2023).
[178] J. Frank, L. Schönherr, Wavefake: A data set to facilitate audio deepfake detection, arXiv preprint arXiv:2111.02813 (2021).
[179] R. Sonobe, S. Takamichi, H. Saruwatari, Jsut corpus: free large-scale japanese speech corpus for end-to-end speech synthesis, arXiv preprint arXiv:1711.00354 (2017).
[180] K. Kumar, R. Kumar, T. De Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, A. C. Courville, Melgan: Generative adversarial networks for conditional waveform synthesis, Advances in neural information processing systems 32 (2019).
[181] G. Yang, S. Yang, K. Liu, P. Fang, W. Chen, L. Xie, Multi-band melgan: Faster waveform generation for high-quality text-to-speech, in: 2021 IEEE Spoken Language Technology Workshop (SLT), IEEE, 2021, pp. 492–498.
[182] D. P. Kingma, P. Dhariwal, Glow: Generative flow with invertible 1x1 convolutions, Advances in neural information processing systems 31 (2018).
[183] N. M. Müller, P. Czempin, F. Dieckmann, A. Froghyar, K. Böttinger, Does audio deepfake detection generalize?, arXiv preprint arXiv:2203.16263 (2022).
[184] H. Ma, J. Yi, C. Wang, X. Yan, J. Tao, T. Wang, S. Wang, L. Xu, R. Fu, Fad: A chinese dataset for fake audio detection, arXiv preprint arXiv:2207.12308 (2022).
[185] N. Perraudin, P. Balazs, P. L. Søndergaard, A fast griffin-lim algorithm, in: 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, IEEE, 2013, pp. 1–4.
[186] H. Kawahara, Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds, Acoustical science and technology 27 (2006) 349–353.
[187] I. Korshunova, W. Shi, J. Dambre, L. Theis, Fast face-swap using convolutional neural networks, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 3677–3685.
[188] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. Lopez Moreno, Y. Wu, et al., Transfer learning from speaker verification to multispeaker text-to-speech synthesis, Advances in neural information processing systems 31 (2018).
[189] K. Uddin, Y. Yang, B. T. Oh, Double compression detection in hevc-coded video with the same coding parameters using picture partitioning information, Signal Processing: Image Communication 103 (2022) 116638.
[190] J. H. Hong, Y. Yang, B. T. Oh, Detection of frame deletion in hevc-coded video in the compressed domain, Digital Investigation 30 (2019) 23–31.
[191] J. Zhang, J. Ni, H. Xie, Deepfake videos detection using self-supervised decoupling network, in: 2021 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2021, pp. 1–6.
[192] C. Q. Huamán, A. L. S. Orozco, L. J. G. Villalba, Authentication and integrity of smartphone videos through multimedia container structure analysis, Future Generation Computer Systems 108 (2020) 15–33.
[193] A. A. Pokroy, A. D. Egorov, Efficientnets for deepfake detection: Comparison of pretrained models, in: 2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus), IEEE, 2021, pp. 598–600.
[194] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1251–1258.
[195] S. Kingra, N. Aggarwal, N. Kaur, Lbpnet: Exploiting texture descriptor for deepfake detection, Forensic Science International: Digital Investigation 42 (2022) 301452.
[196] Z. Guo, G. Yang, J. Chen, X. Sun, Fake face detection via adaptive manipulation traces extraction network, Computer Vision and Image Understanding 204 (2021) 103170.
[197] E. Kim, S. Cho, Exposing fake faces through deep neural networks combining content and trace feature extractors, IEEE Access 9 (2021) 123493–123503.
[198] Z. Xu, J. Liu, W. Lu, B. Xu, X. Zhao, B. Li, J. Huang, Detecting facial manipulated videos based on set convolutional neural networks, Journal of Visual Communication and Image Representation 77 (2021) 103119.
[199] D. Afchar, V. Nozick, J. Yamagishi, I. Echizen, Mesonet: a compact facial video forgery detection network, in: 2018 IEEE international workshop on information forensics and security (WIFS), IEEE, 2018, pp. 1–7.
[200] M. Yu, S. Ju, J. Zhang, S. Li, J. Lei, X. Li, Patch-dfd: Patch-based end-to-end deepfake discriminator, Neurocomputing 501 (2022) 583–595.
[201] G. Mazaheri, A. K. Roy-Chowdhury, Detection and localization of facial expression manipulations, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 1035–1045.
[202] P. Chen, J. Liu, T. Liang, C. Yu, S. Zou, J. Dai, J. Han, Dlfmnet: End-to-end detection and localization of face manipulation using multi-domain features, in: 2021 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2021, pp. 1–6.

[203] J. Deng, J. Guo, E. Ververas, I. Kotsia, S. Zafeiriou, Retinaface: Single-shot multi-level face localisation in the wild, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 5203–5212.

[204] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.

[205] H. Dang, F. Liu, J. Stehouwer, X. Liu, A. K. Jain, On the detection of digital face manipulation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition, 2020, pp. 5781–5790.

[206] Y. Qian, G. Yin, L. Sheng, Z. Chen, J. Shao, Thinking in frequency: Face forgery detection by mining frequency-aware clues, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII, Springer, 2020, pp. 86–103.

[207] F. Li, B. Zhang, B. Liu, Ternary weight networks, arXiv preprint arXiv:1605.04711 (2016).

[208] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, N. Yu, Multi-attentional deepfake detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 2185–2194.

[209] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, M. Paluri, A closer look at spatiotemporal convolutions for action recognition, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2018, pp. 6450–6459.

[210] X. H. Nguyen, T. S. Tran, K. D. Nguyen, D.-T. Truong, et al., Learning spatio-temporal features to detect manipulated facial videos created by the deepfake techniques, Forensic Science International: Digital Investigation 36 (2021) 301108.

[211] A. Das, S. Das, A. Dantcheva, Demystifying attention mechanisms for deepfake detection, in: 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), IEEE, 2021, pp. 1–7.

[212] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al., The kinetics human action video dataset, arXiv preprint arXiv:1705.06950 (2017).

[213] A. Bulat, G. Tzimiropoulos, How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks), in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 1021–1030.

[214] F. Chamot, Z. Geradts, E. Haasdijk, Deepfake forensics: Cross-manipulation robustness of feedforward-and recurrent convolutional forgery detection methods, Forensic Science International: Digital Investigation 40 (2022) 301374.

[215] A. Chintha, A. Rao, S. Sohrawardi, K. Bhatt, M. Wright, R. Ptucha, Leveraging edges and optical flow on faces for deepfake detection, in: 2020 IEEE international joint conference on biometrics (IJCB), IEEE, 2020, pp. 1–10.

[216] A. Chintha, B. Thai, S. J. Sohrawardi, K. Bhatt, A. Hickerson, M. Wright, R. Ptucha, Recurrent convolutional structures for audio spoof and video deepfake detection, IEEE Journal of Selected Topics in Signal Processing 14 (2020) 1024–1037.

[217] S. J. Sohrawardi, A. Chintha, B. Thai, S. Seng, A. Hickerson, R. Ptucha, M. Wright, Poster: Towards robust open-world detection of deepfakes, in: Proceedings of the 2019 ACM SIGSAC conference on computer and communications security, 2019, pp. 2613–2615.

[218] D. M. Montserrat, H. Hao, S. K. Yarlagadda, S. Baireddy, R. Shao, J. Horváth, E. Bartusiak, J. Yang, D. Guera, F. Zhu, et al., Deepfakes detection with automatic face weighting, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 2020, pp. 668–669.

[219] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International conference on machine learning, PMLR, 2019, pp. 6105–6114.

[220] J. Deng, J. Guo, N. Xue, S. Zafeiriou, Arcface: Additive angular margin loss for deep face recognition, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 4690–4699.

[221] T. Fernando, C. Fookes, S. Denman, S. Sridharan, Detection of fake and fraudulent faces via neural memory networks, IEEE Transactions on Information Forensics and Security 16 (2020) 1973–1988.

[222] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[223] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using rnn encoder-decoder for statistical machine translation, arXiv preprint arXiv:1406.1078 (2014).

[224] Y. Lu, Y. Liu, J. Fei, Z. Xia, Channel-wise spatiotemporal aggregation technology for face video forensics, Security and Communication Networks 2021 (2021) 1–13.

[225] D. Cozzolino, A. Rössler, J. Thies, M. Nießner, L. Verdoliva, Id-reveal: Identity-aware deepfake video detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 15108–15117.

[226] V. Blanz, T. Vetter, A morphable model for the synthesis of 3d faces, in: Proceedings of the 26th annual conference on Computer graphics and interactive techniques, 1999, pp. 187–194.

[227] W. Pu, J. Hu, X. Wang, Y. Li, S. Hu, B. Zhu, R. Song, Q. Song, X. Wu, S. Lyu, Learning a deep dual-level network for robust deepfake detection, Pattern Recognition 130 (2022) 108832.

[228] A. Agarwal, A. Agarwal, S. Sinha, M. Vatsa, R. Singh, Md-csdnetwork: Multi-domain cross stitched network for deepfake detection, in: 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), IEEE, 2021, pp. 1–8.

[229] G. Wang, Q. Jiang, X. Jin, W. Li, X. Cui, Mc-lcr: Multimodal contrastive classification by locally correlated representations for effective face forgery detection, Knowledge-Based Systems 250 (2022) 109114.

[230] S. Kolagati, T. Priyadharshini, V. M. A. Rajam, Exposing deepfakes using a deep multilayer perceptron–convolutional neural network model, International Journal of Information Management Data Insights 2 (2022) 100054.

[231] F. Iqbal, A. Abbasi, A. R. Javed, Z. Jalil, J. Al-Karaki, Deepfake audio detection via feature engineering and machine learning (2022).

[232] V. Dongre, A. T. Reddy, N. Reddeddy, Adaptive re-calibration of channel-wise features for adversarial audio classification, arXiv preprint arXiv:2210.11722 (2022).

[233] M. H. Rahman, M. Graciarena, D. Castan, C. Cobo-Kroenke, M. McLaren, A. Lawson, Detecting synthetic speech manipulation in real audio recordings, in: 2022 IEEE International Workshop on Information Forensics and Security (WIFS), IEEE, 2022, pp. 1–6.

[234] A. Pianese, D. Cozzolino, G. Poggi, L. Verdoliva, Deepfake audio detection by speaker verification, in: 2022 IEEE International Workshop on Information Forensics and Security (WIFS), IEEE, 2022, pp. 1–6.

[235] X. Wang, J. Yamagishi, Investigating active-learning-based training data selection for speech spoofing countermeasure, in: 2022 IEEE Spoken Language Technology Workshop (SLT), IEEE, 2023, pp. 585–592.

[236] Z. Zhang, Y. Gu, X. Yi, X. Zhao, Fmfcc-a: a challenging mandarin dataset for synthetic speech detection, in: Digital Forensics and Watermarking: 20th International Workshop, IWDW 2021, Beijing, China, November 20–22, 2021, Revised Selected Papers, Springer, 2022, pp. 117–131.

[237] T. Chen, A. Kumar, P. Nagarsheth, G. Sivaraman, E. Khoury, Generalization of audio deepfake detection., in: Odyssey, 2020, pp. 132–137.

[238] X. Wang, J. Yamagishi, A comparative study on recent neural spoofing countermeasures for synthetic speech detection, arXiv preprint arXiv:2103.11326 (2021).

[239] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, A. Larcher, End-to-end anti-spoofing with rawnet2, in: ICASSP 2021-2021

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2021, pp. 6369–6373.

[240] J.-w. Jung, S.-b. Kim, H.-j. Shim, J.-h. Kim, H.-J. Yu, Improved rawnet with feature map scaling for text-independent speaker verification using raw waveforms, arXiv preprint arXiv:2004.00526 (2020).

[241] G. Hua, A. B. J. Teoh, H. Zhang, Towards end-to-end synthetic speech detection, IEEE Signal Processing Letters 28 (2021) 1265–1269.

[242] P. Kawa, M. Plata, P. Syga, Specrnet: Towards faster and more accessible audio deepfake detection, arXiv preprint arXiv:2210.06105 (2022).

[243] R. Wang, F. Juefei-Xu, Y. Huang, Q. Guo, X. Xie, L. Ma, Y. Liu, Deepsonar: Towards effective and robust detection of ai-synthesized fake voices, in: Proceedings of the 28th ACM international conference on multimedia, 2020, pp. 1207–1216.

[244] A. Gomez-Alanis, J. A. Gonzalez-Lopez, S. P. Dubagunta, A. M. Peinado, M. M. Doss, On joint optimization of automatic speaker verification and anti-spoofing in the embedding space, IEEE Transactions on Information Forensics and Security 16 (2020) 1579–1593.

[245] J. Khochare, C. Joshi, B. Yenarkar, S. Suratkar, F. Kazi, A deep learning framework for audio deepfake detection, Arabian Journal for Science and Engineering (2021) 1–12.

[246] L. Zhang, X. Wang, E. Cooper, N. Evans, J. Yamagishi, The partialspoof database and countermeasures for the detection of short fake speech segments embedded in an utterance, IEEE/ACM Transactions on Audio, Speech, and Language Processing (2022).

[247] Z. Zhang, X. Yi, X. Zhao, Fake speech detection using residual network with transformer encoder, in: Proceedings of the 2021 ACM workshop on information hiding and multimedia security, 2021, pp. 13–22.

[248] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, N. Evans, Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2022, pp. 6367–6371.

[249] W. Ge, H. Tak, M. Todisco, N. Evans, Can spoofing countermeasure and speaker verification systems be jointly optimised?, arXiv preprint arXiv:2303.07073 (2023).

[250] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.

[251] Y. Kwon, H.-S. Heo, B.-J. Lee, J. S. Chung, The ins and outs of speaker recognition: lessons from voxsrc 2020, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2021, pp. 5809–5813.

[252] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, et al., Wavlm: Large-scale self-supervised pre-training for full stack speech processing, IEEE Journal of Selected Topics in Signal Processing 16 (2022) 1505–1518.

[253] J. K. Lewis, I. E. Toubal, H. Chen, V. Sandesera, M. Lomnitz, Z. Hampel-Arias, C. Prasad, K. Palaniappan, Deepfake video detection based on spatial, spectral, and temporal inconsistencies using multimodal deep learning, in: 2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), IEEE, 2020, pp. 1–9.

[254] Y. M. Assael, B. Shillingford, S. Whiteson, N. De Freitas, Lipnet: End-to-end sentence-level lipreading, arXiv preprint arXiv:1611.01599 (2016).

[255] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen, et al., Deep speech 2: End-to-end speech recognition in english and mandarin, in: International conference on machine learning, PMLR, 2016, pp. 173–182.

[256] L. Shang, Z. Kou, Y. Zhang, D. Wang, A multimodal misinformation detector for covid-19 short videos on tiktok, in: 2021 IEEE International Conference on Big Data (Big Data), IEEE, 2021, pp. 899–908.

[257] H. Cheng, Y. Guo, T. Wang, Q. Li, X. Chang, L. Nie, Voice-face homogeneity tells deepfake, arXiv preprint arXiv:2203.02195 (2022).

[258] W. Yang, X. Zhou, Z. Chen, B. Guo, Z. Ba, Z. Xia, X. Cao, K. Ren, Avoid-df: Audio-visual joint learning for detecting deepfake, IEEE Transactions on Information Forensics and Security 18 (2023) 2015–2029.

[259] G. Wang, P. Zhang, L. Xie, W. Huang, Y. Zha, Y. Zhang, An audio-visual attention based multimodal network for fake talking face videos detection, arXiv preprint arXiv:2203.05178 (2022).

[260] H. Ilyas, A. Javed, K. M. Malik, Avfakenet: A unified end-to-end dense swin transformer deep learning model for audio–visual deepfakes detection, Applied Soft Computing 136 (2023) 110124.

[261] C. Feng, Z. Chen, A. Owens, Self-supervised video forensics by audio-visual anomaly detection, arXiv preprint arXiv:2301.01767 (2023).

[262] K. Hara, H. Kataoka, Y. Satoh, Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2018, pp. 6546–6555.

[263] K. Chatfield, K. Simonyan, A. Vedaldi, A. Zisserman, Return of the devil in the details: Delving deep into convolutional nets, arXiv preprint arXiv:1405.3531 (2014).

[264] D. Cozzolino, M. Nießner, L. Verdoliva, Audio-visual person-of-interest deepfake detection, arXiv preprint arXiv:2204.03083 (2022).

[265] M. Lomnitz, Z. Hampel-Arias, V. Sandesara, S. Hu, Multimodal approach for deepfake detection, in: 2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), IEEE, 2020, pp. 1–9.

[266] V. Panayotov, G. Chen, D. Povey, S. Khudanpur, Librispeech: an asr corpus based on public domain audio books, in: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, 2015, pp. 5206–5210.

[267] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, D. Manocha, Emotions don't lie: An audio-visual deepfake detection method using affective cues, in: Proceedings of the 28th ACM international conference on multimedia, 2020, pp. 2823–2832.

[268] B. Hosler, D. Salvi, A. Murray, F. Antonacci, P. Bestagini, S. Tubaro, M. C. Stamm, Do deepfakes feel emotions? a semantic approach to detecting deepfakes via emotional inconsistencies, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 1013–1022.

[269] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, M. Pantic, Avec'19: Audio/visual emotion challenge and workshop, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 2718–2719.

[270] InVID., Invid verification application, Retrieved from https://www.invid-project.eu/invid-verification-application/, Accessed on, 2023.

[271] Intel., Intel real-time deepfake detector, Retrieved from https://www.intel.com/content/www/us/en/newsroom/news/intel-introduces-real-time-deepfake-detector.html, Accessed on, 2023.

[272] Y. Li, C. Zhang, P. Sun, L. Ke, Y. Ju, H. Qi, S. Lyu, Deepfake-o-meter: An open platform for deepfake detection, in: 2021 IEEE Security and Privacy Workshops (SPW), IEEE, 2021, pp. 277–281.

[273] R. Defender, Detect deepfakes stop misinformation., Retrieved from https://realitydefender.com/stop-misinformation, Accessed on, 2023.

[274] Microsoft, Microsoft video authenticator, Retrieved from https://blogs.microsoft.com/on-the-issues/2020/09/01/disinformation-deepfakes-newsguard-video-authenticator/, Accessed on, 2023.

[275] DuckDuckGoose, Deepdetector software, Retrieved from https://www.duckduckgoose.ai/detector, Accessed on, 2023.

[276] DuckDuckGoose, Deepfakeproof: Real-time online deepfake detection, Retrieved from https://www.duckduckgoose.ai/, Accessed on, 2023.

[277] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, K. Kreis, Align your latents: High-resolution video synthesis with latent diffusion models, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

[278] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, D. J. Fleet, Video diffusion models, arXiv preprint arXiv:2204.03458 (2022).

[279] C. Kong, D. Jeon, O. Kwon, N. Kwak, Leveraging off-the-shelf diffusion model for multi-attribute fashion image manipulation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 848–857.

[280] M. Tkachenko, M. Malyuk, A. Holmanyuk, N. Liubimov, Label Studio: Data labeling software, 2020-2022. URL: https://github.com/heartexlabs/label-studio, open source software available from https://github.com/heartexlabs/label-studio.

[281] P. Gohel, P. Singh, M. Mohanty, Explainable ai: current status and future directions, arXiv preprint arXiv:2107.07045 (2021).

Appendices

PUBLICATION 4

Publication 4: Testing the Performance, Adequacy, and Applicability of an Artificial Intelligent Model for Pediatric Pneumonia Diagnosis

J4: Domínguez-Rodríguez, S., Liz-López, H., Panizo, A., Ballesteros, Á., Dagan, R., Greenberg, D., ... & Camacho, D. (2023). Testing the performance, adequacy, and applicability of an Artificial intelligent model for pediatric pneumonia diagnosis. Computer Methods and Programs in Biomedicine, 107765.
DOI: 10.1016/j.cmpb.2023.107765
Impact factor: 6.1 (JCR, 2022) [Q1, 15/110 CS, Theory & Methods; Q1, 22/96 Engineering, Biomedical; Q1, 7/31, Medical Informatics]

- Overall contribution: A study is conducted on the use of artificial intelligence in the diagnosis of paediatric pneumonia. The main contributions of the study are the demonstration that convolutional neural networks (CNNs) can be effective in detecting pneumonia in chest X-rays of children, and the identification of the most important features for pneumonia detection in these images. Furthermore, the study provides a comparison of the results of pneumonia diagnosis with artificial intelligence with traditional methods and discusses the clinical implications of these findings for the diagnosis and treatment of pneumonia in children.

- Contribution of the PhD candidate:

  - Second author of the paper.

  - Implementation of the presented system.

  - Co-author of the manuscript, figures and tables.

ARTICLE IN PRESS

Computer Methods and Programs in Biomedicine xxx (xxxx) xxx

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Testing the performance, adequacy, and applicability of an artificial intelligence model for pediatric pneumonia diagnosis

Sara Domínguez-Rodríguez [a], Helena Liz-López [b], Angel Panizo-LLedot [b,*], Álvaro Ballesteros [a], Ron Dagan [c], David Greenberg [c,d], Lourdes Gutiérrez [a], Pablo Rojo [a,e], Enrique Otheo [f], Juan Carlos Galán [g], Sara Villanueva [e], Sonsoles García [e], Pablo Mosquera [e], Alfredo Tagarro [a,h,i], Cinta Moraleda [a,e], David Camacho [b]

[a] *Pediatric Research and Clinical Trials Unit (UPIC). Instituto de Investigación Sanitaria Hospital 12 de Octubre (imas12), Fundación para la Investigación Biomédica del Hospital 12 de Octubre, Madrid, Spain*
[b] *Computer Systems Engineering Department, Universidad Politécnica de Madrid, Spain*
[c] *Faculty of Health Sciences, Ben-Gurion University of the Negev, Beer-Sheva, Israel*
[d] *Soroka University Medical Center, Beer-Sheva, Israel*
[e] *Pediatric Infectious Diseases Unit. Department of Pediatrics, Hospital Universitario 12 de Octubre, Madrid, Spain*
[f] *Hospital Universitario Ramón y Cajal. Pediatrics Department, Madrid, Spain*
[g] *Hospital Universitario Ramón y Cajal, Microbiology Department, Madrid, Spain*
[h] *Fundación para la Investigación e Innovación Biomédica del Hospital Universitario Infanta Sofía y Hospital Universitario del Henares. Madrid, Spain*
[i] *Pediatrics Research Group. Universidad Europea de Madrid. Pediatrics, Madrid, Spain*

ARTICLE INFO

Keywords:
Pneumonia
CNNs
Deep-learning
Chest X-ray

ABSTRACT

Background: Community-acquired Pneumonia (CAP) is a common childhood infectious disease. Deep learning models show promise in X-ray interpretation and diagnosis, but their validation should be extended due to limitations in the current validation workflow. To extend the standard validation workflow we propose doing a pilot test with the next characteristics. First, the assumption of perfect ground truth (100% sensitive and specific) is unrealistic, as high intra and inter-observer variability have been reported. To address this, we propose using Bayesian latent class models (BLCA) to estimate accuracy during the pilot. Additionally, assessing only the performance of a model without considering its applicability and acceptance by physicians is insufficient if we hope to integrate AI systems into day-to-day clinical practice. Therefore, we propose employing explainable artificial intelligence (XAI) methods during the pilot test to involve physicians and evaluate how well a Deep Learning model is accepted and how helpful it is for routine decisions as well as analyze its limitations by assessing the etiology. This study aims to apply the proposed pilot to test a deep Convolutional Neural Network (CNN)-based model for identifying consolidation in pediatric chest-X-ray (CXR) images already validated using the standard workflow.

Methods: For the standard validation workflow, a total of 5856 public CXRs and 950 private CXRs were used to train and validate the performance of the CNN model. The performance of the model was estimated assuming a perfect ground truth. For the pilot test proposed in this article, a total of 190 pediatric chest-X-ray (CXRs) images were used to test the CNN model support decision tool (SDT). The performance of the model on the pilot test was estimated using extensions of the two-test Bayesian Latent-Class model (BLCA). The sensitivity, specificity, and accuracy of the model were also assessed. The clinical characteristics of the patients were compared according to the model performance. The adequacy and applicability of the SDT was tested using XAI techniques. The adequacy of the SDT was assessed by asking two senior physicians the agreement rate with the SDT. The applicability was tested by asking three medical residents before and after using the SDT and the agreement between experts was calculated using the kappa index.

Results: The CRXs of the pilot test were labeled by the panel of experts into consolidation (124/176, 70.4%) and no-consolidation/other infiltrates (52/176, 29.5%). A total of 31/176 (17.6%) discrepancies were found between the model and the panel of experts with a kappa index of 0.6. The sensitivity and specificity reached a median of

* Corresponding author.
E-mail address: angel.panizo@upm.es (A. Panizo-LLedot).

https://doi.org/10.1016/j.cmpb.2023.107765
Received 24 March 2023; Received in revised form 7 July 2023; Accepted 13 August 2023
Available online 9 September 2023

Please cite this article as: Sara Domínguez-Rodríguez et al., *Computer Methods and Programs in Biomedicine*, https://doi.org/10.1016/j.cmpb.2023.107765

90.9 (95% Credible Interval (CrI), 81.2–99.9) and 77.7 (95% CrI, 63.3–98.1), respectively. The senior physicians reported a high agreement rate (70%) with the system in identifying logical consolidation patterns. The three medical residents reached a higher agreement using SDT than alone with experts (0.66±0.1 vs. 0.75±0.2).
Conclusions: Through the pilot test, we have successfully verified that the deep learning model was underestimated when a perfect ground truth was considered. Furthermore, by conducting adequacy and applicability tests, we can ensure that the model is able to identify logical patterns within the CXRs and that augmenting clinicians with automated preliminary read assistants could accelerate their workflows and enhance accuracy in identifying consolidation in pediatric CXR images.

1. Introduction

Community-acquired Pneumonia (CAP) is the most common infectious disease in childhood. Establishing CAP etiology can often be challenging due to the time-consuming process of the microbiology diagnosis. Given the heterogeneity of the symptoms and uncertain sensitivity or specificity of physical examination, physicians often rely on radiographic findings [1]. However, the process of reading Chest X-ray (CXRs) images can be subjective or inaccurate since the interpretation depends on the expertise. The CXRs are classified into: normal; consolidation, which corresponds to alveolar pneumonia; and non-consolidation/other infiltrates, corresponding to non-alveolar pneumonia [2]. The consolidation pattern in the CXRs is associated with bacterial pneumonia and the other infiltrates are likely associated with viral pneumonia. This pattern in the CXRs is followed by an empirical decision of antibiotics prescription [2] which often results in an over-treatment with antibiotics reinforcing the need for better tools to find an appropriate diagnosis [3].

Researchers have proposed several computer algorithms to analyze CXRs images [4,5]. Deep Learning has demonstrated high performance in image interpretation and diagnosis [6,7]. Specifically, Convolutional Neural Networks (CNNs) [8,9], which are inspired by the overall learning process of the visual cortex and are commonly used in the field of computer vision. Several decision-support tools (SDT) that include this technology have been developed to assist physicians in making decisions [10,11]. However, these SDTs were validated using the standard workflow which is insufficient if we want to include these models into the day-to-day clinical practice. In the field of Deep Learning, the dataset is usually divided into two subsets: the first is used for training the model; and the second, composed of samples that the system has never processed, is used to validate the performance of the system. This process, although common in the field of Deep Learning, presents some limitations. Therefore, we propose extending the standard validation workflow with a pilot test that reinforces the evaluation of the SDT's performance, adequacy, and applicability. Specifically, the pilot test will address two points: the assumption of imperfect ground truth, i.e., in which every label is not known beyond doubt; and employing explainable artificial intelligence (XAI) to involve physicians and evaluate how well a Deep Learning model is accepted and how helpful it is for routine decisions.

Regarding the ground truth, in the field of computer vision is common to validate the models considering a perfect ground truth (100% sensitive and specific) [10–13]. Needless to say, this consideration may hold for tasks like Optical Character Recognition but is unrealistic when dealing with CXRs as high intra and inter-observer variability has been reported [14,15]. If the error rates of the ground truth, are ignored during the evaluation of new diagnostic tests or support-decision systems, the accuracy of new tests or systems can be underestimated [16–18]. To overcome this pitfall, we propose using Bayesian latent class models (BLCA) as a method to further estimate the accuracy [19]. Nevertheless, we are aware that having several readings from different experts for each CXR in order to estimate the inter-observer variability is not feasible given the large amount of data required to train a deep-learning model. Hence, including this approach in the classic evaluation workflow is unrealistic. However, it is not less true that this requirement can be fulfilled for a test pilot that extends the classical evaluation workflow since the sample size can be smaller for such an evaluation. There is no consensus on the sample size needed for test pilots and feasibility studies, [26] reported that the median sample size for these studies should be around 30–36 per group, which will indicate that 190 CXR will be enough for a pilot study to test the feasibility of an already trained/validated model.

Regarding involving physicians, most of the studies found on the state of the art usually evaluate their models without analyzing their applicability and adequacy, i.e., without ensuring that the model would be usable by end-users and can identify logical medical patterns [10,20]. Doing this is of vital importance in fields like medicine and health care. Especially when using Deep Learning because it is difficult for humans to understand how the system has reached a particular conclusion (this phenomenon is called the "Blackbox syndrome"). In response to this problem, the field of Explainable Artificial Intelligence (XAI) was born [21–23]. The use of XAI allows evaluating a model beyond common quantitative analysis increasing the trust between humans and AI. Assessing whether the patterns found by an ML model are in line with the logic followed by senior physicians or if these same patterns are of use for less experienced ones, is mandatory for any model that aims to be included in the clinical routines. As with ground truth, we are aware that it is not possible to have several experts analyzing the patterns of the hundreds or thousands of samples usually used to validate a deep learning model. However, it is no less true that such a study can be carried out with a smaller but very representative subset of samples, as is the case in the pilot test we propose in this paper.

Therefore, due to the two aforementioned limitations, this work proposes enhancing the standard validation workflow with a pilot test that does not assume a perfect ground truth and uses XAI techniques to test its applicability and acceptance by physicians. The pilot test proposed in this work uses an SDT tool previously trained and validated by some members of our group following the standard workflow [24]. The SDT tool has shown promising results. Several experiments, including transfer learning, were performed to find the best architecture using public (5856 CXRs) and private (950 CXRs) datasets. Due to the good results obtained by the aforementioned model, this study aims to further test the performance, applicability, and adequacy of the CXRs CNN ensemble system tool (SDT) by doing a pilot test following the above considerations and using an independent pediatric CXRs set (Fig. 1). We believe that the proposed pilot test enhances the standard validation workflows and applying them will be able to accelerate the integration of these tools into medical routines, which is of particular importance in the case of pediatrics since a recent systematic review enhances the problem that despite recent improvements in the quality of AI products and AI applications in radiology, only a few thoracic AI systems are specifically designed for the pediatric population [25,42–45]. However, they do not conduct a pilot test with expert radiologists like the one proposed in this work.

The main contribution of this work can be briefly summarized as follows:

- Extending the classic validation workflow to deal with imperfect ground truth by using Bayesian latent class models (BLCA) to estimate accuracy.

- Extending the classic validation workflow to assess the applicability and acceptance of a deep learning model by employing explainable AI (XAI) methods to involve physicians and evaluate model acceptance and usefulness in routine clinical decisions.
- In depth study of an SDT specifically tailored to pediatric CAP.

2. Material and methods

2.1. Study population

Four different datasets have been used throughout the creation and validation of the SDT. Firstly, two datasets were used for the training and standard validation workflow, one of them public and one private provided by the Ben-Gunion University. Finally, two other datasets, PCAPE and VALSDANCE, were used for the pilot test. In total, more than 7000 radiographs were used for the training, standard validation flow, and pilot test, see Table 3. To test the performance of the SDT, through the pilot test, a total of 190 pediatric CXRs images were collected from the observational multi-center prospective cohort studies VALSDANCE and PCAPE [27]. The radiographs in this study were not in DICOM format, as they are obtained, but were in other formats more commonly used by physicians, which allows them to open them with different devices such as mobiles. Eligible participants were children and adolescents from 1 month to 17 years hospitalized with CAP. The interpretation of the CXRs was performed following the standards of the "WHO Vaccine Trial Investigators Radiology Working Group" [28] classifying the CXRs images in consolidation or other infiltrates. A panel of experts composed of two pediatricians (AT and EO) interpreted the CXRs.

pattern identified by the SDT was correct. In addition, to evaluate the applicability of the system, an independent subset of 40 pediatric CXRs images provided by Ben-Gurion University (Israel) were randomly selected. These CRXs images were labeled by two senior pediatricians (RD y DG) and a radiologist from Ben-Gurion University (Israel). The experts also had access to lateral radiographs of the patients to increase the precision of the labeling of each case. The CXR reading project was approved by the Soroka University Ethics Committee (Number 3075).

This study was approved by the Ethics Board of coordinating hospitals (codes Hospital Infanta Sofía (Madrid, Spain) 320/11 and Hospital 12 Octubre (Madrid, Spain) 17/311).

Finally, an extensive microbiological workup was performed. All the laboratory methods were described in Supplementary Material Appendix 1.

2.2. Algorithm description

The SDT used in this study has already been designed and validated by our team following the standard validation workflow. This section presents a summary of the architecture for a detailed description of the training procedure, the architecture of the ensemble model, and its performance we refer the reader to the original paper [24]. For reproducibility purposes, all the code needed to reproduce the experimentation is available at this Github repository.[1] Firstly, in preprocessing, we reduced the three channels of the images to the first one and normalize the shape to 150 × 150. Then, images were normalized by diving them by the average pixel value of the images. To overcome the overfitting, we apply data augmentation techniques with different transformations: shearing (0.2), zoom (0.05), rotation (0.2), and horizontal shift (0.1).

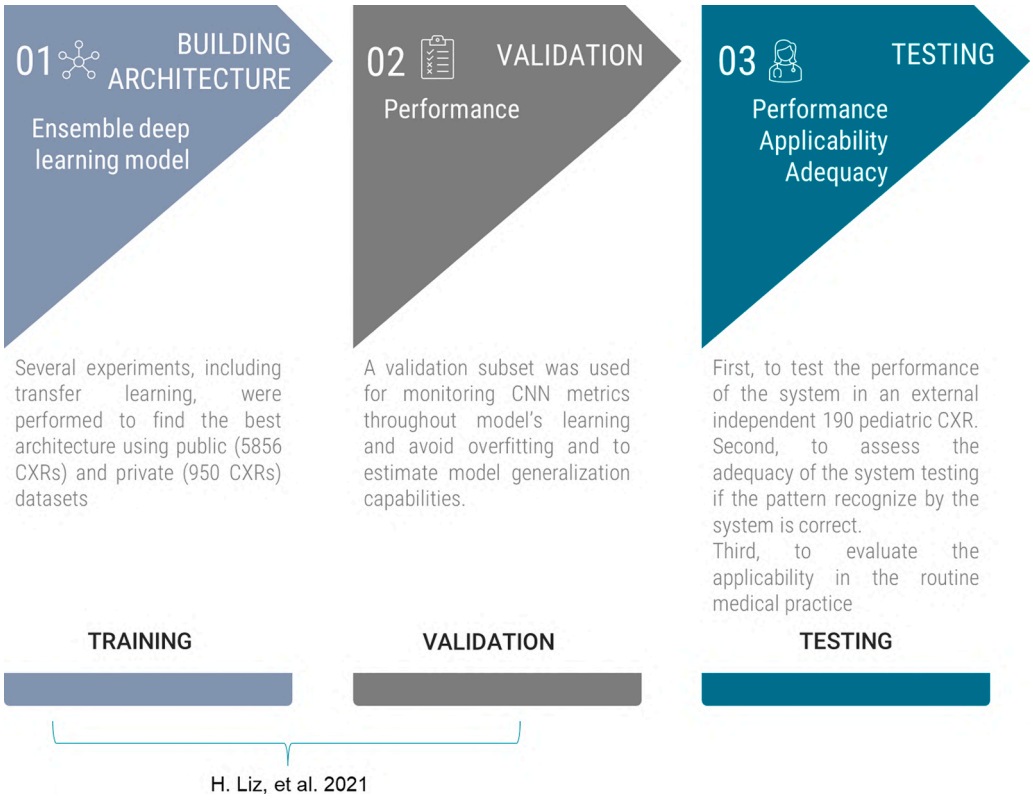


Fig. 1. Workload presented in this study.

For the second objective which aims to evaluate the adequacy of the system, a subset of 40/190 of the above pediatric CXRs images labeled as consolidation from both SDT and the experts were shown to two independent senior pediatricians (CM and PR) to test if the consolidation

[1] https://github.com/helenalizlopez/Evaluation-of-Convolutional-Neural-Networks-models-for-pneumonia-diagnosis-master

The SDT is composed of an ensemble of 5 CNNs. The architecture of each CNN is composed of 4 Convolutional Blocks, each one with a convolutional layer with 32 neurons, a kernel of size NxN, ReLu activation, 2×2 Max Pooling, and Batch Normalization. In addition, the kernel size of each Convolutional Block progressively decreases from 11×11 in the first one, 7×7 in the second, 5×5 in the third, and 3×3 in the final one. The output of the last convolutional layer is flattened and a dropout of 70% is applied. Then, this information is processed by a Fully Connected Layer composed of 128 neurons with an L2 regularization at a strength of 0.01. Finally, the classification layer consists of a Fully Connected Layer of two neurons with Softmax activation. Please note that to reach a final decision by the ensemble the average probability prediction across the five CNNs is used.

To train the five neural networks that form the ensemble, a dataset composed of 950 CRXs from the Ben-Gurion University was used. This dataset was split in two following a stratified random split procedure, 70% for train and validation and 30% for testing. Then, the 70% reserved for train and validation was further split into 5 train/validation sets using stratified 5-fold cross-validation. Finally, each of the 5 ensembles was trained and validated using these 5 folds and the Adam optimizer with a learning rate of 1e-4. In addition, the complete ensemble was tested using the 30% reserved at the first random split of the dataset.

2.3. Evaluation

In the SDT generation, we evaluate the performance of the classification system, considering the labels as perfect ground truth, using two metrics: AUC and TPR. We use TPR instead of TNR because it is more dangerous not to treat a patient who needs antibiotics with antibiotics than to give antibiotics to a person who does not need them. For the pilot test, the performance of the model was estimated considering an imperfect ground truth and using extensions of the two-test, two BLCA [29] implemented in a simplified interface application [18]. We divide the overall sample into two subpopulations according to age. This is because BLCA needs to estimate true disease prevalence, and a 2×2 summary table of two diagnostic tests applied to one population does not provide enough data for this calculation [19,30]. In the event that two diagnostic tests were applied together to one population, it is possible to divide a single population data set into multiple population data sets with different prevalences based on specific variables [31]. Non-informative prior distribution was used for all parameter estimations. The model was set with 2 Markov chain Monte Carlo (MCMC) chains, 5000 burn-in iterations, 25,000 total iterations, and 10 thinning intervals. The convergence was checked using tracing plots and the fitness of the model was checked through observed and predicted frequencies comparison. To assess the probability of observed frequencies, the dataset was replicated 20,000 times and selected only 2000 times (thin=10). To evaluate the performance of the model we used confusion matrices to assess sensitivity, specificity, false-positive rate, true positive rate, and the area under the operating curve (AUC). We also provided the results for conventional methods where the gold standard was considered 100% sensible and specific.

The discrepancies were described according to their sociodemographic, clinical, and microbiology features. The discrepancies were described as false positive (FP) if the expert labeled as consolidation the CXR and the system did not and false-negative (FN) if the expert labeled as other infiltrates/non-consolidation the CXR and the system as consolidation. In summary tables of continuous variables, interquartile ranges, and medians were assessed. Shapiro-Wilk test was performed to test normality. In summary tables of categorical variables, counts and percentages were used. Chi-squared and Fisher Tests (expected count <5) were performed when categorical variables. For continuous variables, Kruskal Wallis was performed. All hypothesis testing was carried out at the 5% significance level. All the analyses were performed using R software.

To test the adequacy and applicability of the SDT, the patterns found by the CNN ensemble were analyzed. To extract these patterns in a human-friendly way, XAI techniques were applied. Specifically, heatmaps were used to indicate the areas of the CRX where the system has found evidence that supports a particular diagnosis (consolidation | other infiltrates/non-consolidation) using the visualize cam function from the Keras vis package [32]. We used two output neurons to obtain a separate heatmap and generate the desired visualization for each of the two classes.

The adequacy of the model was tested by asking two pediatric infectious disease experts (PR and CM) if the pattern recognized as consolidation by the model in a subset of 40 CXR and pointed in the heatmap was a plausible consolidation pattern. The agreement was rated using a 0–5 Likert scale meaning 1: strongly disagree, 2: disagree, 3: Neutral, 4: Agree, and 5: Strongly agree.

The applicability of the model was tested using a subset of the training set ($n = 40$) including 20 of CXRs classified as consolidation and 20 as non-consolidation. These 40 CRXs were randomly selected from the dataset of 190 chest x-rays. A total of three medicine interns in pediatrics classified the 40 CXRs. After that, they again classified these de-identified 40 CXRs (without knowing that they were the same) but using the decision-support tool, including the heatmaps. Any change of opinion after using the tool was analyzed. If this modification means an agreement with the panel experts we reported it as a success, if this modification means a disagreement with the panel of experts we reported it as a failure, and if no change of opinion we reported it as an agreement.

3. Results

The results of the pilot test are presented in this section.

3.1. Performance

A total of 190 CRXs images were used as a testing set to evaluate the SDT performance. Of those, 14/190 (7.4%) were excluded because the CXR was not crisp enough and the lung was not identifiable.

The CRXs were labeled by the panel of experts into consolidation (124/176, 70.4%) and no-consolidation/other infiltrates (52/176, 29.5%). A total of 31/176 (17.6%) discrepancies were found between the model and the panel of experts with a kappa index of 0.6. From those discrepancies, 18/176 (10.2%) were misclassified as other infiltrates/no-consolidation being FN. Of those, 3/18 (16.7%) raised doubts to some of the experts. A total of 13/176 (7.4%) were misclassified as consolidation being FP. The FN patients' features were like those classified as having consolidation in the CXR. Both FP and FN were mostly with viral etiology, and with no clinical complications. The discrepancies were summarized in Table 1.

The probability of consolidation predicted by the algorithm was slightly higher in the CXRs images that were classified as consolidation by both system and experts than the FP, however, these differences were not statistically significant (Fig. 2). Likewise, the CXRs images that were classified as other infiltrates/no-consolidation presented no statistically significant differences with the FN.

The etiology according to the performance of the model was summarized in Fig. 3. All the patients with typical bacteria such as S. *pneumoniae*, S. *pyogenes*, and S. *aureus* were equally classified by both system and expert physicians as having consolidated CXR. However, 2 patients with the atypical bacteria M.*pneumoniae* were FN. A total of 7/46 (15%) patients with an atypical bacterial etiology were discrepancies between the SDT and the expert, 5/47 (10.9%) were misclassified as other infiltrates, and only 2/47 (4.3%) were misclassified as consolidation. A total of 23/102 (22.6%) patients with a viral etiology were discrepancies between the SDT and the expert. 12/102 (11.8%) were misclassified as other infiltrates and 11/102 (10.8%) were misclassified as consolidation.

Table 1
Summary descriptive table by performance.

System Panel of experts	Consolidation Consolidation *N = 106*	Other infiltrates/no consolidation Consolidation *N = 18*	p-value	Other infiltrates/no consolidation Other infiltrates/no consolidation *N = 39*	Consolidation Other infiltrates/no consolidation *N = 13*	*p*-value
Gender:			0.539			0.309
Male	47 (44.3%)	6 (33.3%)		28 (71.8%)	7 (53.8%)	
Female	59 (55.7%)	12 (66.7%)		11 (28.2%)	6 (46.2%)	
Age (months)	40.6 [18.2;72.7]	28.7 [17.0;54.2]	0.338	20.0 [12.8;32.8]	31.5 [13.3;38.6]	0.665
Bacterial aetiology			0.539			1.000
Yes	56 (52.8%)	6 (33.3%)		8 (20.5%)	2 (15.4%)	
No	50 (47.2%)	12 (66.7%)		31 (79.5%)	11 (84.6%)	
Aetiology:			0.449			0.854
Atypical bacteria	34 (32.1%)	5 (27.8%)		5 (12.8%)	2 (15.4%)	
Typical bacteria	24 (22.6%)	2 (11.1%)		3 (7.69%)	0 (0.00%)	
Virus	48 (45.3%)	11 (61.1%)		31 (79.5%)	11 (84.6%)	
Complications:			0.153			0.561
No	76 (71.7%)	16 (88.9%)		35 (89.7%)	13 (100%)	
Yes	30 (28.3%)	2 (11.1%)		4 (10.3%)	0 (0.00%)	
PICU admission:			0.212			1.000
No	93 (87.7%)	18 (100%)		35 (89.7%)	13 (100%)	
Yes	13 (12.3%)	0 (0.00%)		4 (10.3%)	0 (0.00%)	

System: CNNs deep learning algorithm. Panel of experts: two physicians and one radiologist.

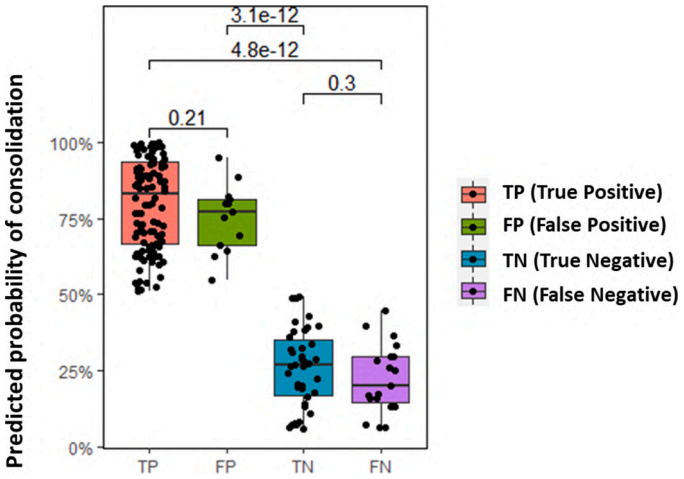


Fig. 2. Predicted probability of consolidation given by the system according to the agreement system-expert
TP: True positive (both classifications of CXRs as "consolidation"); FP: False Positive (classified by expert as "Other infiltrates/no pneumonia" and as consolidated by the System); TN: True Negative (both classifications of CXRs as "Other infiltrates/no pneumonia"); FN: False Negative (classified by expert as "consolidation" and as "Other infiltrates/no pneumonia" by the System).

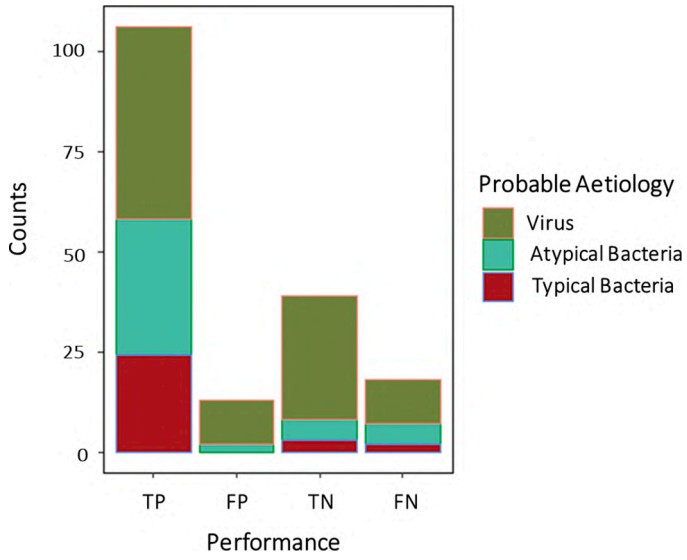


Fig. 3. Performance according to probable aetiology
TP: True positive (both classifications of CXRs as "consolidation"); FP: False Positive (classified by expert as "Other infiltrates/no pneumonia" and as consolidated by the System); TN: True Negative (both classifications of CXRs as "Other infiltrates/no pneumonia"); FN: False Negative (classified by expert as "consolidation" and as "Other infiltrates/no pneumonia" by the System).

Overall, the SDT presented an 82.3% (95% CI, 75.9–87.7%) accuracy, 0.89 of precision, 0.85 recall, and an F1 score of 0.87. The AUC of the model was 0.79 and the precision-recall curve 0.51. To perform the BLCA, the study population was divided according to age. A total of 134/176 (76%) were children below 5 years with a 64.9 (95% CI, 56.2–72.8) estimated prevalence of having consolidation in the CXR. The children older than 5 years old 42/176 (24%) presented a higher prevalence of having consolidation in the CXR (88.1 (95% CI, 73.6–95.5)). By the conventional method, which assumes that the interpretation by the panel of experts is 100% sensible and 100% specific, the sensitivity of the system reached 85.5 (95% CI, 77.8 - 90.9) and 75.0 (95% CI, 60.8–85.5) specificity. However, with BLCA, the sensitivity and specificity reached a median of 90.9 (95% Credible Interval (CrI), 81.2–99.9), and 77.7 (95% CrI, 63.3–98.1) respectively (Table 2). The BLCA model converges (Supplementary Figure 1, Supplementary Figure 2) and the fitness was acceptable with Bayesian posterior probability close to 0.5 (describing the observed data well [18]) (Supplementary Table 1).

3.2. Adequacy

The adequacy was tested by asking two senior physicians if they agree with the pattern observed and colored in the heatmap as consolidation by the System in 40 randomly selected CXRs. A total of 28/40 (70%) CXRs were rated as agree or strongly agree. Only 9/40 (22%) were classified as "strongly disagree". An example of the heatmaps shown to the physicians is available in Fig. 4

3.3. Applicability

A total of 40 CXRs images were interpreted by three medical residents. These CXRs were interpreted with and without using the support decision system and the kappa index for the agreement was calculated. For the three residents, the interpretation using the SDT increased their agreement with the WHO experts panel in the validation set. The overall

Table 2
Diagnostic test performance subpopulations by age.

Parameters	Physician assessment as a perfect gold standard (%)*	Bayesian latent class model (%) **
Prevalence of consolidation		
Patients ≤5 years	64.9 (56.2 - 72.8)	60.2 (45.7 - 74.1)
Patients >5 years	88.1 (73.6 - 95.5)	85.0 (67.8 - 94.7)
Image read by physicians		
Sensitivity	100 (100 - 100)	98.6 (88.2 - 100)
Specificity	100 (100 - 100)	83.2 (60.7 - 99.9)
PPV	100 (100 - 100)	92.1 (76.0 - 100)
NPV	100 (100 - 100)	97.0 (71.2 - 100)
Deep learning		
Sensitivity	85.5 (77.8 - 90.9)	90.9 (81.2 - 99.9)
Specificity	75.0 (60.8 - 85.5)	77.7 (63.3 - 98.1)
PPV	89.1 (81.7 - 93.8)	88.7 (77.9 - 99.2)
NPV	68.4 (54.6 - 79.7)	81.5 (60.1 - 99.9)

* Conventional method assumed that the interpretation by the physicians is perfect (100% sensitivity and 100% specificity; all patients with gold standard test positive presented x-ray with consolidation and all patients with gold standard test negative presented x-ray with other infiltrates/no-pneumonia). Values shown are estimated means with 95% confidence interval.

** Bayesian latent class model does not assume that any test is perfect. Values shown are estimated median with 95% credible interval. To cope with the unknown accuracy of the gold standard, variable "true" classification is included into the model. The variable "true" classification contains two mutually exclusive categories, "consolidation" or "other infiltrates", and the real value of this variable for each subject is considered unobserved (i.e. latent). When the unobservable variable is categorical, the term "latent class analysis" applies. The chance that the test will be consolidation, if a subject has "consolidation" is the true sensitivity (Se), and the chance that the test will be other infiltrates if a subject is "other infiltrates" is true specificity (Sp). Therefore, this kind of model did not assume that any test is perfect, but consider true accuracy of each test for diagnosing the true classification ("consolidation" or "other infiltrates").

Table 3
Datasets used in training, standard validation workflow and pilot test.

Dataset	SDT Generation Training	SDT Generation Validation	Standard validation flow	Pilot test
Ben-Gunion	532	133	285	40
Public dataset	3279	820	1757	0
VALSDANCE & PCAPE	0	0	0	190

inter-observer agreement was good (kappa=0.7) and the median agreement between residents and WHO panel experts was (kappa=0.7) before using the SDT and median kappa=0.8 after using the SDT (Fig. 5). All the residents found the tool useful and answered "yes" to the questions "Will you use it if finally implemented?" and "Do you feel confident to decide to use the tool?".

4. Discussion

In this study, we conducted a pilot test on an automated support decision tool for the classification of chest X-rays (CXRs) into consolidation and other infiltrates/no consolidation, according to the WHO classification standardization. Upon piloting this tool, we observed that the deep learning model underestimated its performance when considering a perfect ground truth. Additionally, the deep learning model demonstrated the ability to identify logical patterns in the CXRs. Furthermore, we found that the utilization of this tool by medical residents led to increased knowledge and decreased decisional conflicts.

Images account for increasing amounts of medical data but require extensive manual interpretation. Providers need technologies to aid them in extracting, visualizing, and interpreting them. The radiological

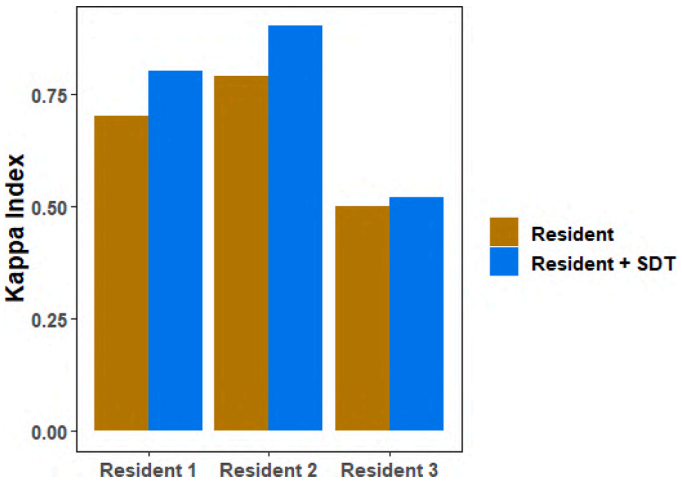


Fig. 5. Agreement between medical residents and experts before (yellow) and after (blue) using SDT.

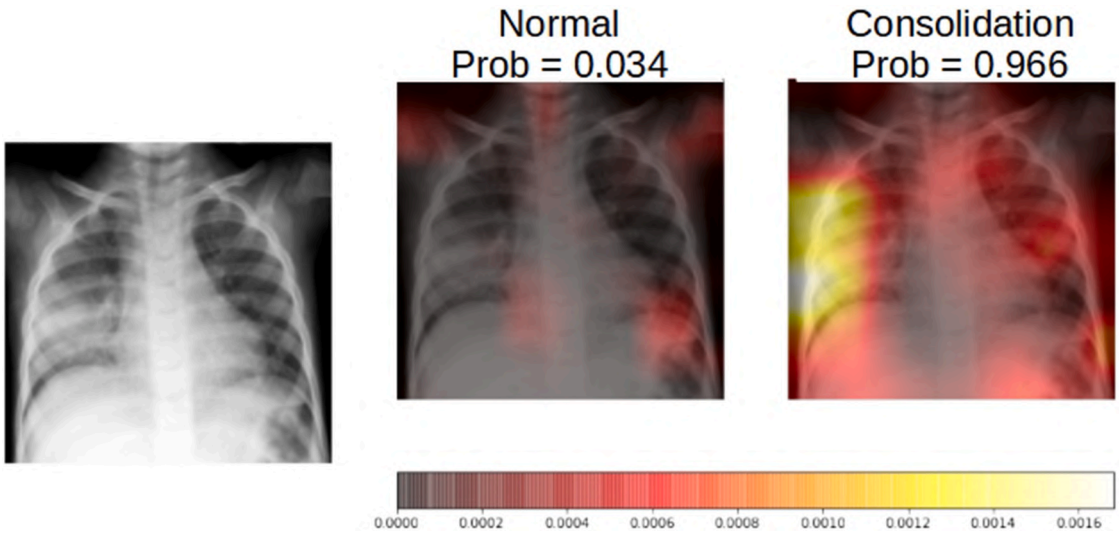


Fig. 4. Heatmaps of a consolidation sample 1. Left side, Neuron 0, shows the heatmaps for the non-consolidation class, whereas the right side, Neuron 1, shows the heatmaps for consolidation class for five individual CNNs.

examination plays an important role in the diagnosis of CAP and treatment decisions [33]. However, there is a higher subjectivity and higher inter and intraobserver variability in the interpretation of CXRs [34–36]. Recently, several artificial intelligence (AI) algorithms have emerged to offer assistance in interpretation [37–39]. However, the evaluation methodology used for assessing these algorithms only used popular metrics such as the area under the curve (AUC), or F-scores for label-based precision and recall evaluations [43]. In this study, we further evaluated the tool for clinical purposes based on sensitivity, specificity, and clinical predictive value rather than label-based evaluations. This tool presented high sensitivity and high specificity compared with the routine CXRs interpretation made by the attending physician. Importantly, there were only 2 patients with a typical bacterial etiology where the SDT yield discrepant results according to the expert, and none of them were driven by *S.pneumoniae*. In other words, none of the FN was critical (with clinical complications of PICU admitted).

The evaluation of new diagnostic tools usually relied on a perfect ground truth/gold standard, which is obtained by a loose consensus process such as a majority vote from a few board-certified radiologists/expert physicians. However, if the error rates of a gold standard are ignored during the evaluation of new diagnostic tests or tools, the accuracy of these new systems can be underestimated, and the prevalence can be either under-or over-estimated. In this study, we have evaluated the performance using a Bayesian latent analysis allowing us to provide a better estimate accounting for this interpretation variability [18,31,40].

The generation and evaluation of a model follow the step of training, validation, and testing. The training procedure is defined as the design and fit of the ML model using a training set. On the other side, the validation and testing steps are usually interchanged in the literature, defined as a process of evaluation of the model [41]. However, the validation step is often used for the model tune and hypothesis testing, interfering with the final model selection, and possibly adding some bias to the results. The testing process, on the contrary, is an unbiased evaluation of the performance of the final model in real conditions. We have performed a testing step based on XAI techniques to evaluate the adequacy of the model and the applicability for the target users. Overall, senior physicians evaluated the ability of the system to identify logical patterns rating correctly the majority of the CXRs labeled as consolidation. On the other hand, the results of the questionnaire study show that most of the resident respondents liked using the tool and feel confident using it in their clinical duties. This tool drove them to a better decision according to the interpretation by WHO panel experts

This study has several limitations. First, the assessment of the tool's impact on knowledge and decisional conflict is limited by the fact that there was no control group comparing usual care to use of the tool, but the pre-and post-results were encouraging. Second, this is a pilot testing study but further testing in several scenarios as low-income countries should be tested. Third, more senior experts and medical experts are needed for inferring better quality controls in applicability and adequacy and refining the system image segmentation. Fourth, the images were collected from the clinical practice and not specifically for research purposes. These images are sometimes blurred or with low contrast because the median age of the patients was 2 years, and it is difficult to have good quality images in this population. For this reason, a total of 14 images were discarded due to the bad quality of the image.

Although the pediatric population is understudied compared to the adult population [25], other studies have proposed the use of SDT for pediatric pneumonia, such as those by [43,44]. However, these studies have certain limitations that hinder their direct applicability to day-to-day clinical practice. The former study developed a CNN model with good performance in diagnosing the etiology (viral, bacterial, or mycoplasma) in children using clinical data (C Reactive protein and age) and CXR. Nevertheless, C Reactive protein is not always available in middle-to-low-income countries. Furthermore, the study did not provide a specific classification of the etiology, such as typical or atypical bacterial pneumonia, which plays a crucial role in determining the appropriate antibiotic treatment, as we have done. The latter study presented an accurate CNN model for pediatric pneumonia diagnosis. However, the images used were not obtained from a real-world scenario; instead, they were pre-selected images of good quality from a single hospital. In contrast, our dataset consists of CXR images of children hospitalized with pneumonia from two hospitals in Spain, collected from routine in-hospital procedures that involve physicians. Additionally, both papers did not involve the medical community in testing the model and did not analyze its applicability in routine decision support services.

The development of these tools requires a coordinated effort. This work shows that it takes a multidisciplinary team effort where physicians, data managers, machine learning researchers, medical imaging specialists, software developers, and statisticians all need to work together to do a systematic end-to-end study, design, and execution. Further research is required to assess the clinical acceptance of artificial intelligence in the real world and quantify the benefit of such automation to physicians and their patients.

Summary Table

what was already known on the topic	*what this study added to our knowledge*
• Deep Learning, has demonstrated high performance in image interpretation and diagnosis • Convolutional Neural Network algorithms have been used and evaluated with high accuracy for pneumonia diagnosis using Chest-X rays images • The pattern in the CXRs is followed by an empirical decision of antibiotics prescription which often results in an over-treatment with antibiotics reinforcing the need for better tools to find an appropriate diagnosis	• AI system that is specifically designed and tested for the pediatric population, which is under • This study assesses not only the performance in terms of AUC or F-score but also the adequacy and applicability of the model • This study has an extensive evaluation of the performance taking into account not only the number of false-positive and negative results but also characterizing them in terms of clinical relevance and pathogenicity

Funding

This work has been partially supported by the following grants and funding agencies: a European Society for Paediatric Diseases (ESPID) Small Grant Award 2019; The Spanish Ministry of Science and Innovation under FightDIS (PID2020–117263GB-100), funded by MCIN/AEI/10.13039/501100011033 and, as appropriate, by "ERDF A way of making Europe", by the "European Union" or by the "European Union NextGenerationEU/PRTR"; Comunidad Autónoma de Madridunder S2018/TCS-4566 grant (CYNAMON); Finally, David Camacho has been supported by the Comunidad Autónoma de Madrid under: "Convenio Plurianual with the Universidad Politécnica de Madrid in the actuation line of Programa de Excelencia para el Profesorado Universitario".

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used Grammarly and DeepL in order to check the spelling and grammar of the document. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

CRediT authorship contribution statement

Sara Domínguez-Rodríguez: Conceptualization, Methodology, Formal analysis, Writing – original draft. **Helena Liz-López:** Methodology, Software, Formal analysis, Writing – review & editing. **Angel Panizo-LLedot:** Methodology, Software, Formal analysis, Writing – review & editing. **Álvaro Ballesteros:** Data curation. **Ron Dagan:**

Resources, Investigation. **David Greenberg:** Resources, Investigation. **Lourdes Gutiérrez:** Resources, Investigation. **Pablo Rojo:** Resources, Investigation. **Enrique Otheo:** Resources, Investigation. **Juan Carlos Galán:** Resources, Investigation. **Sara Villanueva:** Resources, Investigation. **Sonsoles García:** Resources, Investigation. **Pablo Mosquera:** Resources, Investigation. **Alfredo Tagarro:** Resources, Investigation. **Cinta Moraleda:** Resources, Investigation. **David Camacho:** Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank all the patients and families for their participation in this cohort, and the staff members who cared for them.

VALS-DANCE Study Group: Ana Barrios, MD, PhD (Pediatrics Department, Hospital Universitario Infanta Sofía, Pediatrics Research Group, Universidad Europea de Madrid, Madrid, Spain), Mónica Pacheco, MD (Pediatrics Department, Hospital Universitario Infanta Sofía, Pediatrics Research Group, Universidad Europea de Madrid, Madrid, Spain), Carmen Arquero, MD (Pediatrics Department, Hospital Universitario Infanta Sofía, Pediatrics Research Group, Universidad Europea de Madrid, Madrid, Spain), María Bernardino (Pediatrics Department, Hospital Universitario Infanta Sofía, Pediatrics Research Group, Universidad Europea de Madrid, Madrid, Spain), Alfonso Cañete, MD, PhD (Pediatrics Department, Hospital Universitario Infanta Sofía, Pediatrics Research Group, Universidad Europea de Madrid, Madrid, Spain), Luis Prieto, MD, PhD (Pediatric Infectious Diseases Unit, Department of Pediatrics, Hospital Universitario 12 de Octubre), Lourdes Gutiérrez [Pediatric Research and Clinical Trials Unit (UPIC), Instituto de Investigación Sanitaria Hospital 12 de Octubre (IMAS12), Madrid, Spain; Fundación para la Investigación, Biomédica del Hospital 12 de Octubre, Madrid, Spain; RITIP (Traslational Research Network in Pediatric Infectious Diseases)], Cristina Epalza, MD (Pediatric Infectious Diseases Unit, Department of Pediatrics, Hospital Universitario 12 de Octubre), Pablo Rojo, MD, PhD (Pediatric Infectious Diseases Unit, Department of Pediatrics, Hospital Universitario 12 de Octubre; Pediatric Research and Clinical Trials Unit (UPIC), Instituto de Investigación Sanitaria Hospital 12 de Octubre (IMAS12), Madrid, Spain; Fundación para la Investigación, Biomédica del Hospital 12 de Octubre, Madrid, Spain; RITIP (Traslational Research Network in Pediatric Infectious Diseases; Pediatrics Department, Universidad Complutense de Madrid, Madrid, Spain, Lidia Oviedo, MD (Pediatric Infectious Diseases Unit, Department of Pediatrics, Hospital Universitario 12 de Octubre), Miquel Serna-Pascual [Pediatric Research and Clinical Trials Unit (UPIC), Instituto de Investigación Sanitaria Hospital 12 de Octubre (IMAS12), Madrid, Spain; Fundación para la Investigación, Biomédica del Hospital 12 de Octubre, Madrid, Spain; RITIP (Traslational Research Network in Pediatric Infectious Diseases)], Raquel Ramos Corral (Microbiology Department, Laboratorio BR Salud, Hospital Universitario Infanta Sofía, San Sebastián de los Reyes, Madrid, Spain), María Dolores Folgueira, MD, PhD (Microbiology Department, Hospital Universitario 12 de Octubre, Instituto de Investigación Sanitaria Hospital 12 de Octubre (IMAS12), Madrid, Spain), Alfredo Pérez-Rivilla, MD, PhD (Microbiology Department, Hospital Universitario 12 de Octubre, Instituto de Investigación Sanitaria Hospital 12 de Octubre (IMAS12), Madrid, Spain), Carmen Vázquez, MD (Pediatrics Department, Hospital Universitario Ramón y Cajal), Nathalia Gerig, MD (Pediatrics Department, Hospital Universitario Ramón y Cajal), María Pilar Romero, MD, PhD (Microbiology Department, Hospital Universitario La Paz, Madrid, Spain), Arantxa Berzosa, MD (Pediatrics Department, Hospital Universitario de Getafe, Getafe, Madrid, Spain), Beatriz Soto, MD, PhD (Pediatrics Department, Hospital Universitario de Getafe, Getafe, Madrid, Spain), Sara Guillén, MD, PhD (Pediatrics Department, Hospital Universitario de Getafe, Getafe, Madrid, Spain), Francisco José Sanz de Santaeufemia, MD, PhD (Pediatrics Department, Hospital Universitario Niño Jesús, Madrid, Spain), Ana Belén Jiménez, MD, PhD (Pediatrics Department, Hospital Universitario Fundación Jiménez Díaz, Madrid, Spain), Elvira Martín, MD, PhD (Pediatrics Department, Hospital Universitario Fundación Jiménez Díaz, Madrid, Spain), Talía Sainz, MD, PhD [Pediatrics, Infectious and Tropical Diseases, Hospital Universitario La Paz. Instituto Investigación Hospital La Paz (IDIPAZ), Madrid, Spain; RITIP (Traslational Research Network in Pediatric Infectious Diseases)], Cristina Calvo, MD, PhD [Pediatrics, Infectious and Tropical Diseases, Hospital Universitario La Paz. Instituto Investigación Hospital La Paz (IDIPAZ), Madrid, Spain; RITIP (Traslational Research Network in Pediatric Infectious Diseases)], Marta Llorente, MD, PhD (Pediatrics Department, Hospital Universitario del Sureste, Arganda del Rey, Madrid, Spain), Elisa Garrote, MD, PhD (Pediatrics Department, Hospital Universitario Basurto, Bilbao, Vizcaya, Spain), Cristina Muñoz, MD, PhD (Pediatrics Department, Hospital General de Villalba, Villalba, Madrid, Spain), Paula Sánchez, MD, PhD (Pediatric Infectious Diseases, Immunology and Rheumatology Unit, University Hospital Virgen del Rocío, Instituto de Biomedicina de Sevilla (IBIS), Sevilla, Spain), Mar Santos, MD, PhD (Pediatric Infectious Diseases Unit, Hospital Universitario Gregorio Marañón, Madrid, Spain), José-Tomás Ramos, MD, PhD (Pediatrics Department, Hospital Clínico San Carlos, Madrid, Spain; Pediatrics Department, Universidad Complutense de Madrid, Madrid, Spain), Marta Illán, MD (Pediatrics Department, Hospital Clínico San Carlos, Madrid, Spain), David Molina, MD, PhD (Microbiology Department, Hospital Universitario de Getafe, Getafe, Madrid, Spain), Manuel Imaz, MD, PhD (Microbiology Department, Hospital Universitario Basurto, Bilbao, Vizcaya, Spain), on behalf of VALS-DANCE Working Group.

Other members of VALS-DANCE Working Group: Rut del Valle (Hospital Universitario Infanta Sofía), Julia Yebra (Hospital Universitario Infanta Sofía), Rosa Batista (Hospital Universitario Infanta Sofía), Teresa Raga (Hospital Universitario Infanta Sofía), Maria García-Baró (Hospital Universitario Infanta Sofía), Magdalena Hawkins (Hospital Universitario Infanta Sofía, Universidad Europea), Daniel Blázquez (Hospital Universitario 12 de Octubre), Manuel Gijón (Hospital Universitario 12 de Octubre), Lucía Figueroa (Fundación para la Investigación Biomédica del Hospital 12 de Octubre), Nazaret del Amo (BR Salud, Hospital Universitario Infanta Sofía), Ana Méndez-Echeverría (Hospital Universitario La Paz), Mercedes Alonso-Sanz (Hospital Universitario Niño Jesús), Esther Casado Verrier (Hospital de Villalba), María José Cilleruelo (Hospital Universitario Puerta de Hierro), María Luz Golmayo (Hospital Universitario Puerta de Hierro), María Isabel Sánchez (Hospital Universitario Puerta de Hierro), Teresa del Rosal (Hospital Universitario La Paz), Alfonso Rodríguez-Albarrán (Hospital de Arganda), Ana Coca (Hospital Universitario Ramón y Cajal), Raquel Buenache (Hospital Universitario Ramón y Cajal), Inmaculada Mota (Hospital Universitario Ramón y Cajal).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cmpb.2023.107765.

References

[1] W.G. Boersma, J.M.A. Daniels, A. Löwenberg, W.J. Boeve, E.J. van de Jagt, «Reliability of radiographic findings and the relation to etiologic agents in community-acquired pneumonia», Respir. Med. 100 (5) (2006) 926–932, https://doi.org/10.1016/J.RMED.2005.06.018.

[2] T. Cherian, et al., «Standardized interpretation of paediatric chest radiographs for the diagnosis of pneumonia in epidemiological studies», Bull. World Health Organ. 83 (5) (2005) 353–359. S0042-96862005000500011.

[3] G. Tomsony, I. Vlad, «The need to look at antibiotic resistance from a health systems perspective», Upsala J. Med. Sci. 119 (2) (2014) 117–124, https://doi.org/10.3109/03009734.2014.902879, nInforma Healthcare.
[4] F. Piccialli, V. Di Somma, F. Giampaolo, S. Cuomo, G. Fortino, «A survey on deep learning in medicine: why, how and when?», Inf. Fusion 66 (2021) 111–137, https://doi.org/10.1016/j.inffus.2020.09.006.
[5] A.K. Jaiswal, P. Tiwari, S. Kumar, D. Gupta, A. Khanna, J.J.P.C. Rodrigues, «Identifying pneumonia in chest X-rays: a deep learning approach», Measurement 145 (2019) 511–518, https://doi.org/10.1016/j.measurement.2019.05.076.
[6] H. Jiang, et al., A review of deep learning-based multiple-lesion recognition from medical images: classification, detection and segmentation, Comput. Biol. Med. 106726 (2023), https://doi.org/10.1016/j.compbiomed.2023.106726.
[7] S. Usama, K. Bukhari, S. Safwan, K. Bukhari, A. Syed, S. Sajid, H. Shah, «The diagnostic evaluation of Convolutional Neural Network (CNN) for the assessment of chest X-ray of patients infected with COVID-19», medRxiv 03 (26) (2020), 20044610 https://doi.org/10.1101/2020.03.26.20044610.
[8] D. Varshni, K. Thakral, L. Agarwal, R. Nijhawan, A. Mittal, «Pneumonia detection using CNN based feature extraction», in: Proceedings of the IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), IEEE, 2019, pp. 1–7, https://doi.org/10.1109/ICECCT.2019.8869364.
[9] T. Rahman, et al., «Transfer learning with deep Convolutional Neural Network (CNN) for pneumonia detection using chest X-ray», Appl. Sci. 10 (9) (2020) 3233, https://doi.org/10.3390/app10093233.
[10] N. Mahomed, et al., «Computer-aided diagnosis for world health organization-defined chest radiograph primary-endpoint pneumonia in children», Pediatr. Radiol. 50 (4) (2020) 482–491, https://doi.org/10.1007/s00247-019-04593-0.
[11] D.S. Kermany, et al., «Identifying medical diagnoses and treatable diseases by image-based deep learning», Cell 172 (5) (2018) 1122–1131, https://doi.org/10.1016/j.cell.2018.02.010.
[12] W.H.K. Chiu, et al., «Detection of COVID-19 using deep learning algorithms on chest radiographs», J. Thorac. Imaging 35 (6) (2020) 369–376, https://doi.org/10.1097/RTI.0000000000000559.
[13] B. Li, G. Kang, K. Cheng, N. Zhang, «Attention-guided convolutional neural network for detecting pneumonia on chest X-rays», in: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, Institute of Electrical and Electronics Engineers Inc, 2019, pp. 4851–4854, https://doi.org/10.1109/EMBC.2019.8857277.
[14] M.N. Albaum, et al., «Interobserver reliability of the chest radiograph in community-acquired pneumonia», Chest 110 (2) (1996) 343–350, https://doi.org/10.1378/chest.110.2.343.
[15] E. Sarria, G.B. Fischer, J.A.B. Lima, S.S. Menna Barreto, J.A.M. Flôres, R. Sukiennik, [Interobserver agreement in the radiological diagnosis of lower respiratory tract infections in children], J. Pediatr. (Rio J.) 79 (6) (2003) 497–503.
[16] A.A. Buck, J.J. Gart, «Comparison of a screening test and a reference test in epidemiologic studies: I. Indices of agreement and their relation to prevalence», Am. J. Epidemiol. 83 (3) (1966) 586–592, https://doi.org/10.1093/oxfordjournals.aje.a120609.
[17] J.J. Gart, A.A. Buck, «Comparison of a screening test and a reference test in epidemiologic studies: II. A probabilistic model for the comparison of diagnostic tests», Am. J. Epidemiol. 83 (3) (1966) 593–602, https://doi.org/10.1093/oxfordjournals.aje.a120610.
[18] C. Lim, et al., «Using a web-based application to define the accuracy of diagnostic tests when the gold standard is imperfect», PLoS ONE 8 (11) (2013) e79489, https://doi.org/10.1371/journal.pone.0079489.
[19] L. Joseph, T.W. Gyorkos, L. Coupal, «Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard», Am. J. Epidemiol. 141 (3) (1995) 263–272, https://doi.org/10.1093/oxfordjournals.aje.a117428.
[20] L. Qiang, R.M. Nishikawa, Computer-Aided Detection and Diagnosis in Medical Imaging, 1 ed., CRC Press, 2015.
[21] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R.M. Summers, «Chest X-ray 8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases», in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, https://doi.org/10.1109/CVPR.2017.369.
[22] Avola D., Bacciu A., Cinque L., Fagioli A., Marini M.R., Taiello R., «Study on transfer learning capabilities for pneumonia classification in chest-X-rays image», oct. 2021, doi: 10.48550/arxiv.2110.02780.
[23] L.O. Teixeira, et al., «Impact of lung segmentation on the diagnosis and explanation of COVID-19 in chest X-ray images», Sensors 21 (21) (2021) 7116, https://doi.org/10.3390/s21217116.
[24] H. Liz, M. Sánchez-Montañés, A. Tagarro, S. Domínguez-Rodríguez, R. Dagan, D. Camacho, «Ensembles of Convolutional Neural Networks for pediatric pneumonia diagnosis», Future Gener. Comput. Syst. 122 (2020) 220–233, https://doi.org/10.1016/j.future.2021.04.007.
[25] S. Schalekamp, W.M. Klein, K.G. van Leeuwen, «Current and emerging artificial intelligence applications in chest imaging: a pediatric perspective», Pediatr. Radiol. (2021) 1–11, https://doi.org/10.1007/S00247-021-05146-0/FIGURES/3.
[26] S.A. Billingham, A.L. Whitehead, S.A. Julious, An audit of sample sizes for pilot and feasibility trials being undertaken in the United Kingdom registered in the United Kingdom Clinical Research Network database, BMC Med. Res. Methodol. 13 (2013) 104, https://doi.org/10.1186/1471-2288-13-104.
[27] A. Tagarro, et al., «A tool to distinguish viral from bacterial pneumonia», Pediatr. Infect. Dis. J. (2021) https://doi.org/10.1097/INF.0000000000000334.
[28] T. Cherian, et al., «Standardized interpretation of paediatric chest radiographs for the diagnosis of pneumonia in epidemiological studies», Bull. World Health Organ 83 (5) (2005) 353–359. S0042-96862005000500011.
[29] S.L. Hui, S.D. Walter, «Estimating the error rates of diagnostic tests», Biometrics 36 (1) (1980) 167, https://doi.org/10.2307/2530508.
[30] N. Dendukuri, L. Joseph, «Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests», Biometrics 57 (1) (2001) 158–167, https://doi.org/10.1111/j.0006-341X.2001.00158.x.
[31] N. Toft, E. Jørgensen, S. Højsgaard, «Diagnosing diagnostic tests: evaluating the assumptions underlying the estimation of sensitivity and specificity in the absence of a gold standard», Prev. Vet. Med. (2005) 19–33, https://doi.org/10.1016/j.prevetmed.2005.01.006. Elsevier.
[32] Neural Network Visualization Toolkit For Keras», «GitHub - raghakot/keras-vis:, GitHub, 2022 https://github.com/raghakot/keras-vis. accedido 23 de marzo de.
[33] A.J. Alario, P.L. McCarthy, R. Markowitz, P. Kornguth, N. Rosenfield, J. M. Leventhal, «Usefulness of chest radiographs in children with acute lower respiratory tract disease», J. Pediatr. 111 (2) (1987) 187–193, https://doi.org/10.1016/S0022-3476(87)80065-3.
[34] Y. Balabanova, et al., «Variability in interpretation of chest radiographs among Russian clinicians and implications for screening programmes: observational study», Br. Med. J. 331 (7513) (2005) 379–382, https://doi.org/10.1136/bmj.331.7513.379.
[35] M.I. Neuman, et al., «Variability in the interpretation of chest radiographs for the diagnosis of pneumonia in children», J. Hosp. Med. 7 (4) (2012) 294–298, https://doi.org/10.1002/jhm.955.
[36] V. Novack, L.S. Avnon, A. Smolyakov, R. Barnea, A. Jotkowitz, F. Schlaeffer, «Disagreement in the interpretation of chest radiographs among specialists and clinical outcomes of patients hospitalized with suspected pneumonia», Eur. J. Intern. Med. 17 (1) (2006) 43–47, https://doi.org/10.1016/j.ejim.2005.07.008.
[37] M. Annarumma, S.J. Withey, R.J. Bakewell, E. Pesce, V. Goh, G. Montana, «Automated triaging of adult chest radiographs with deep artificial neural networks», Radiology 291 (1) (2019) 196–202, https://doi.org/10.1148/radiol.2018180921.
[38] K.C.L. Wong, M. Moradi, J. Wu, T. Syeda-Mahmood, «Identifying disease-free chest X-ray images with deep transfer learning», arXiv (2019) https://doi.org/10.1117/12.2513164.
[39] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, et al., CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison, in: 33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, AAAI Press, 2019, pp. 590–597.
[40] W.O. Johnson, G. Jones, I.A. Gardner, «Gold standards are out and Bayes is in: implementing the cure for imperfect reference tests in diagnostic accuracy studies», Prev. Vet. Med. 167 (2019) 113–127, https://doi.org/10.1016/j.prevetmed.2019.01.010.
[41] M. Kuhn, Applied predictive modeling, 26, Springer, New York, 2013.
[42] T.T. Tran, et al., Learning to automatically diagnose multiple diseases in pediatric chest radiographs using deep convolutional neural networks, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.
[43] S. Hu, Y. Zhu, D. Dong, et al., Chest radiographs using a context-fusion Convolution Neural Network (CNN): can it distinguish the etiology of community-acquired pneumonia (CAP) in children?[J], J. Digit. Imaging 35 (5) (2022) 1079–1090.
[44] Mohammed, I., Singh, N., & Venkatasubramanian, M. (2019). Computer-assisted detection and diagnosis of pediatric pneumonia in chest X-ray images.
[45] H.H Pham, N.H Nguyen, T.T Tran, T.N.M Nguyen, H.Q Nguyen, "PediCXR: an open, large-scale chest radiograph dataset for interpretation of common thoracic diseases in children, Scient. Data (2023) 240, 10.1.

Bibliography

[1] S. Domínguez-Rodríguez, H. Liz-López, A. Panizo, Á. Ballesteros, R. Dagan, D. Greenberg, L. Gutiérrez, P. Rojo, E. Otheo, J. C. Galán, et al., Testing the performance, adequacy, and applicability of an artificial intelligent model for pediatric pneumonia diagnosis, Computer Methods and Programs in Biomedicine (2023) 107765.

[2] S. Atito, M. Awais, J. Kittler, Sit: Self-supervised vision transformer, arXiv preprint arXiv:2104.03602 (2021).

[3] W. Yu, P. Zhou, S. Yan, X. Wang, Inceptionnext: When inception meets convnext, arXiv preprint arXiv:2303.16900 (2023).

[4] M. Gu, Y. Zhang, Y. Wen, G. Ai, H. Zhang, P. Wang, G. Wang, A lightweight convolutional neural network hardware implementation for wearable heart rate anomaly detection, Computers in Biology and Medicine (2023) 106623.

[5] V. Leiva, J. Mazucheli, B. Alves, A novel regression model for fractiles: Formulation, computational aspects, and applications to medical data, Fractal and Fractional 7 (2023) 169.

[6] A. Hannan, P. Pal, Detection and classification of kidney disease using convolutional neural networks, J Neurol Neurorehab Res. 2023; 8 (2) 136 (2023).

[7] Y. Li, H. Liu, Q. Wu, F. Mu, J. Yang, J. Gao, C. Li, Y. J. Lee, Gligen: Open-set grounded text-to-image generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 22511–22521.

[8] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, ImageNet Large Scale Visual Recognition Challenge, International Journal of Computer Vision (IJCV) 115 (2015) 211–252.

[9] A. Krizhevsky, Learning multiple layers of features from tiny images, Technical Report, 2009.

[10] Y. LeCun, C. Cortes, C. Burges, Mnist handwritten digit database, ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist 2 (2010).

[11] H. Xiao, K. Rasul, R. Vollgraf, Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017. arXiv:cs.LG/1708.07747.

[12] J. Han, J. Pei, H. Tong, Data mining: concepts and techniques, Morgan kaufmann, 2022.

[13] A. Mohammed, R. Kora, A comprehensive review on ensemble deep learning: Opportunities and challenges, Journal of King Saud University-Computer and Information Sciences (2023).

[14] S. Salcedo-Sanz, P. Ghamisi, M. Piles, M. Werner, L. Cuadra, A. Moreno-Martínez, E. Izquierdo-Verdiguier, J. Muñoz-Marí, A. Mosavi, G. Camps-Valls, Machine learning information fusion in earth observation: A comprehensive review of methods, applications and data sources, Information Fusion 63 (2020) 256–272.

[15] H. Zhang, H. Xu, X. Tian, J. Jiang, J. Ma, Image fusion meets deep learning: A survey and perspective, Information Fusion 76 (2021) 323–336.

[16] K. Bayoudh, R. Knani, F. Hamdaoui, A. Mtibaa, A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets, The Visual Computer (2021) 1–32.

[17] S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, M. P. Lungren, Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines, NPJ digital medicine 3 (2020) 136.

[18] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, L.-P. Morency, Efficient low-rank multimodal fusion with modality-specific factors, arXiv preprint arXiv:1806.00064 (2018).

[19] H. R. V. Joze, A. Shaban, M. L. Iuzzolino, K. Koishida, Mmtm: Multimodal transfer module for cnn fusion, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 13289–13299.

[20] M. A. Ganaie, M. Hu, A. Malik, M. Tanveer, P. Suganthan, Ensemble deep learning: A review, Engineering Applications of Artificial Intelligence 115 (2022) 105151.

[21] B. Krawczyk, L. L. Minku, J. Gama, J. Stefanowski, M. Woźniak, Ensemble learning for data stream analysis: A survey, Information Fusion 37 (2017) 132–156.

[22] S. González, S. García, J. Del Ser, L. Rokach, F. Herrera, A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities, Information Fusion 64 (2020) 205–237.

[23] V.-L. Nguyen, E. Hüllermeier, M. Rapp, E. Loza Mencía, J. Fürnkranz, On aggregation in ensembles of multilabel classifiers, in: International Conference on Discovery Science, Springer, 2020, pp. 533–547.

[24] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, F. Herrera, Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence, Information Fusion 99 (2023) 101805.

[25] M. Xu, S. Yoon, A. Fuentes, D. S. Park, A comprehensive survey of image augmentation techniques for deep learning, Pattern Recognition (2023) 109347.

[26] J. Huertas-Tato, A. Martin, D. Camacho, Bertuit: Understanding spanish language in twitter through a native transformer, arXiv preprint arXiv:2204.03465 (2022).

[27] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, et al., Wavlm: Large-scale self-supervised pre-training for full stack speech processing, IEEE Journal of Selected Topics in Signal Processing 16 (2022) 1505–1518.

[28] B. Zi, M. Chang, J. Chen, X. Ma, Y.-G. Jiang, Wilddeepfake: A challenging real-world dataset for deepfake detection, in: Proceedings of the 28th ACM international conference on multimedia, 2020, pp. 2382–2390.

[29] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, Information fusion 58 (2020) 82–115.

[30] L. Zou, H. L. Goh, C. J. Y. Liew, J. L. Quah, G. T. Gu, J. J. Chew, M. P. Kumar, C. G. L. Ang, A. W. A. Ta, Ensemble image explainable ai (xai) algorithm for severe community-acquired pneumonia and covid-19 respiratory infections, IEEE Transactions on Artificial Intelligence 4 (2022) 242–254.

[31] U. Schlegel, H. Arnout, M. El-Assady, D. Oelke, D. A. Keim, Towards a rigorous evaluation of xai methods on time series, in: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), IEEE, 2019, pp. 4197–4201.

[32] A. Das, P. Rad, Opportunities and challenges in explainable artificial intelligence (xai): A survey, arXiv preprint arXiv:2006.11371 (2020).

[33] T. Gomez, T. Fréour, H. Mouchère, Metrics for saliency map evaluation of deep learning explanation methods, in: International Conference on Pattern Recognition and Artificial Intelligence, Springer, 2022, pp. 84–95.

[34] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.

[35] H. Liz, M. Sánchez-Montañés, A. Tagarro, S. Domínguez-Rodríguez, R. Dagan, D. Camacho, Ensembles of convolutional neural network models for pediatric pneumonia diagnosis, Future Generation Computer Systems 122 (2021) 220–233.

[36] H. Liz, J. Huertas-Tato, M. Sánchez-Montañés, J. Del Ser, D. Camacho, Deep learning for understanding multilabel imbalanced chest x-ray datasets, Future Generation Computer Systems 144 (2023) 291–306.

[37] H. Liz-López, J. Huertas-Tato, D. Camacho, Transparency in medicine: How explainable ai is revolutionizing patient care, in: Proceedings of 2023 International Conference on Network, Multimedia and Information Technology (NMITCON), 2023.

[38] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, et al., Identifying medical diagnoses and treatable diseases by image-based deep learning, cell 172 (2018) 1122–1131.

[39] K. Alomar, H. I. Aysel, X. Cai, Data augmentation in classification and segmentation: A survey and new strategies, Journal of Imaging 9 (2023) 46.

[40] F. Chollet, et al., Keras, 2015. URL: https://github.com/fchollet/keras.

[41] E. Ghafourian, F. Samadifam, H. Fadavian, P. Jerfi Canatalay, A. Tajally, S. Channumsin, An ensemble model for the diagnosis of brain tumors through mris, Diagnostics 13 (2023) 561.

[42] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning, PMLR, 2019, pp. 6105–6114.

[43] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.

[44] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818–2826.

[45] C. Szegedy, S. Ioffe, V. Vanhoucke, A. A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Thirty-first AAAI conference on artificial intelligence, 2017.

[46] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1251–1258.

[47] A. Bustos, A. Pertusa, J.-M. Salinas, M. de la Iglesia-Vayá, Padchest: A large chest x-ray image dataset with multi-label annotated reports, Medical image analysis 66 (2020) 101797.

[48] J. Islam, Y. Zhang, Towards robust lung segmentation in chest radiographs with deep learning, arXiv preprint arXiv:1811.12638 (2018).

[49] S. Reza, O. B. Amin, M. Hashem, Transresunet: Improving u-net architecture for robust lungs segmentation in chest x-rays, in: 2020 IEEE Region 10 Symposium (TENSYMP), IEEE, 2020, pp. 1592–1595.

[50] F. Charte, A. J. Rivera, M. J. del Jesus, F. Herrera, Remedial-hwr: Tackling multilabel imbalance through label decoupling and data resampling hybridization, Neurocomputing 326 (2019) 110–122.

[51] H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, et al., The era5 global reanalysis, Quarterly Journal of the Royal Meteorological Society 146 (2020) 1999–2049.

[52] H. Baars, M. Radenz, A. A. Floutsi, R. Engelmann, D. Althausen, B. Heese, A. Ansmann, T. Flament, A. Dabas, D. Trapon, et al., Californian wildfire smoke over europe: A first example of the aerosol observing capabilities of aeolus compared to ground-based lidar, Geophysical Research Letters 48 (2021) e2020GL092194.

[53] T. Haiden, I. Sandu, G. Balsamo, G. Arduini, A. Beljaars, Addressing biases in near-surface forecasts, ECMWF Newsletter 157 (2018) 20–25.

[54] I. Alados, I. Foyo-Moreno, F. Olmo, L. Alados-Arboledas, Relationship between net radiation and solar radiation for semi-arid shrub-land, Agricultural and Forest Meteorology 116 (2003) 221–227.

[55] R. Hogan, Radiation quantities in the ecmwf model and mars, ECMWF, 2016 (2015).

[56] M. Yuan, T. Leirvik, M. Wild, Global trends in downward surface solar radiation from spatial interpolated ground observations during 1961–2019, Journal of Climate 34 (2021) 9501–9521.

[57] S. Solomon, K. Dube, K. Stone, P. Yu, D. Kinnison, O. B. Toon, S. E. Strahan, K. H. Rosenlof, R. Portmann, S. Davis, et al., On the stratospheric chemistry of midlatitude wildfire smoke, Proceedings of the National Academy of Sciences 119 (2022) e2117325119.

[58] N. Gobron, B. Pinty, M. M. Verstraete, J.-L. Widlowski, Advanced vegetation indices optimized for up-coming sensors: Design, performance, and applications, IEEE Transactions on Geoscience and Remote Sensing 38 (2000) 2489–2505.

[59] C. Van Wagner, et al., Development and structure of the Canadian forest fire weather index system, volume 35, 1987.

[60] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16000–16009.

[61] C. Feichtenhofer, H. Fan, Y. Li, K. He, Masked autoencoders as spatiotemporal learners, arXiv preprint arXiv:2205.09113 (2022).

[62] Y. Pang, W. Wang, F. E. Tay, W. Liu, Y. Tian, L. Yuan, Masked autoencoders for point cloud self-supervised learning, arXiv preprint arXiv:2203.06604 (2022).

[63] Y. Kim, Y. Choi, D. Widemann, T. Zohdi, A fast and accurate physics-informed neural network reduced order model with shallow masked autoencoder, Journal of Computational Physics 451 (2022) 110841.

[64] P. Jain, S. C. Coogan, S. G. Subramanian, M. Crowley, S. Taylor, M. D. Flannigan, A review of machine learning applications in wildfire science and management, Environmental Reviews 28 (2020) 478–505.

[65] F. Abid, A survey of machine learning algorithms based forest fires prediction and detection systems, Fire Technology 57 (2021) 559–590.

[66] K. Bot, J. G. Borges, A systematic review of applications of machine learning techniques for wildfire management decision support, Inventions 7 (2022) 15.

[67] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, et al., The prisma 2020 statement: an updated guideline for reporting systematic reviews, International journal of surgery 88 (2021) 105906.

[68] Y. Li, M.-C. Chang, S. Lyu, In ictu oculi: Exposing ai created fake videos by detecting eye blinking, in: 2018 IEEE International workshop on information forensics and security (WIFS), IEEE, 2018, pp. 1–7.

[69] D. Güera, E. J. Delp, Deepfake video detection using recurrent neural networks, in: 2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS), IEEE, 2018, pp. 1–6.

[70] P. Korshunov, S. Marcel, Deepfakes: a new threat to face recognition? assessment and detection, arXiv preprint arXiv:1812.08685 (2018).

[71] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, Faceforensics++: Learning to detect manipulated facial images, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 1–11.

[72] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, C. C. Ferrer, The deepfake detection challenge (dfdc) preview dataset, arXiv preprint arXiv:1910.08854 (2019).

[73] Y. Li, X. Yang, P. Sun, H. Qi, S. Lyu, Celeb-df: A large-scale challenging dataset for deepfake forensics, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 3207–3216.

[74] L. Jiang, R. Li, W. Wu, C. Qian, C. C. Loy, Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 2889–2898.

[75] Y. He, B. Gan, S. Chen, Y. Zhou, G. Yin, L. Song, L. Sheng, J. Shao, Z. Liu, Forgerynet: A versatile benchmark for comprehensive forgery analysis, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 4360–4369.

[76] P. Kwon, J. You, G. Nam, S. Park, G. Chae, Kodf: A large-scale korean deepfake detection dataset, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10744–10753.

[77] T. Zhou, W. Wang, Z. Liang, J. Shen, Face forensics in the wild, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 5778–5788.

[78] Y. Li, X. Yang, P. Sun, H. Qi, S. Lyu, Celeb-df (v2): a new dataset for deepfake forensics, arXiv preprint arXiv:1909.12962 (2019).

[79] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, C. C. Ferrer, The deepfake detection challenge (dfdc) dataset, arXiv preprint arXiv:2006.07397 (2020).

[80] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, K. A. Lee, Asvspoof 2019: Future horizons in spoofed and fake audio detection, arXiv preprint arXiv:1904.05441 (2019).

[81] R. Reimao, V. Tzerpos, For: A dataset for synthetic speech detection, in: 2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), IEEE, 2019, pp. 1–10.

[82] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch, et al., Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild, IEEE/ACM Transactions on Audio, Speech, and Language Processing (2023).

[83] J. Frank, L. Schönherr, Wavefake: A data set to facilitate audio deepfake detection, arXiv preprint arXiv:2111.02813 (2021).

[84] N. M. Müller, P. Czempin, F. Dieckmann, A. Froghyar, K. Böttinger, Does audio deepfake detection generalize?, arXiv preprint arXiv:2203.16263 (2022).

[85] H. Ma, J. Yi, C. Wang, X. Yan, J. Tao, T. Wang, S. Wang, L. Xu, R. Fu, Fad: A chinese dataset for fake audio detection, arXiv preprint arXiv:2207.12308 (2022).

[86] H. Khalid, S. Tariq, M. Kim, S. S. Woo, Fakeavceleb: A novel audio-video multimodal deepfake dataset, arXiv preprint arXiv:2108.05080 (2021).

[87] T. Mittal, R. Sinha, V. Swaminathan, J. Collomosse, D. Manocha, Video manipulations beyond faces: A dataset with human-machine analysis, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 643–652.

[88] Z. Cai, S. Ghosh, T. Gedeon, A. Dhall, K. Stefanov, M. Hayat, " glitch in the matrix!": A large scale benchmark for content driven audio-visual forgery detection and localization, arXiv preprint arXiv:2305.01979 (2023).

[89] J. H. Hong, Y. Yang, B. T. Oh, Detection of frame deletion in hevc-coded video in the compressed domain, Digital Investigation 30 (2019) 23–31.

[90] C. Q. Huamán, A. L. S. Orozco, L. J. G. Villalba, Authentication and integrity of smartphone videos through multimedia container structure analysis, Future Generation Computer Systems 108 (2020) 15–33.

[91] J. Zhang, J. Ni, H. Xie, Deepfake videos detection using self-supervised decoupling network, in: 2021 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2021, pp. 1–6.

[92] K. Uddin, Y. Yang, B. T. Oh, Double compression detection in hevc-coded video with the same coding parameters using picture partitioning information, Signal Processing: Image Communication 103 (2022) 116638.

[93] H. Dang, F. Liu, J. Stehouwer, X. Liu, A. K. Jain, On the detection of digital face manipulation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition, 2020, pp. 5781–5790.

[94] Y. Qian, G. Yin, L. Sheng, Z. Chen, J. Shao, Thinking in frequency: Face forgery detection by mining frequency-aware clues, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII, Springer, 2020, pp. 86–103.

[95] A. A. Pokroy, A. D. Egorov, Efficientnets for deepfake detection: Comparison of pretrained models, in: 2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus), IEEE, 2021, pp. 598–600.

[96] P. Chen, J. Liu, T. Liang, C. Yu, S. Zou, J. Dai, J. Han, Dlfmnet: End-to-end detection and localization of face manipulation using multi-domain features, in: 2021 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2021, pp. 1–6.

[97] Z. Guo, G. Yang, J. Chen, X. Sun, Fake face detection via adaptive manipulation traces extraction network, Computer Vision and Image Understanding 204 (2021) 103170.

[98] E. Kim, S. Cho, Exposing fake faces through deep neural networks combining content and trace feature extractors, IEEE Access 9 (2021) 123493–123503.

[99] A. Mitra, S. P. Mohanty, P. Corcoran, E. Kougianos, A machine learning based approach for deepfake detection in social media through key video frame extraction, SN Computer Science 2 (2021) 1–18.

[100] Z. Xu, J. Liu, W. Lu, B. Xu, X. Zhao, B. Li, J. Huang, Detecting facial manipulated videos based on set convolutional neural networks, Journal of Visual Communication and Image Representation 77 (2021) 103119.

[101] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, N. Yu, Multi-attentional deepfake detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 2185–2194.

[102] M. Yu, S. Ju, J. Zhang, S. Li, J. Lei, X. Li, Patch-dfd: Patch-based end-to-end deepfake discriminator, Neurocomputing 501 (2022) 583–595.

[103] G. Mazaheri, A. K. Roy-Chowdhury, Detection and localization of facial expression manipulations, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 1035–1045.

[104] S. Kingra, N. Aggarwal, N. Kaur, Lbpnet: Exploiting texture descriptor for deepfake detection, Forensic Science International: Digital Investigation 42 (2022) 301452.

[105] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, M. Paluri, A closer look at spatiotemporal convolutions for action recognition, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2018, pp. 6450–6459.

[106] D. M. Montserrat, H. Hao, S. K. Yarlagadda, S. Baireddy, R. Shao, J. Horváth, E. Bartusiak, J. Yang, D. Guera, F. Zhu, et al., Deepfakes detection with automatic face weighting, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 2020, pp. 668–669.

[107] T. Fernando, C. Fookes, S. Denman, S. Sridharan, Detection of fake and fraudulent faces via neural memory networks, IEEE Transactions on Information Forensics and Security 16 (2020) 1973–1988.

[108] A. Chintha, A. Rao, S. Sohrawardi, K. Bhatt, M. Wright, R. Ptucha, Leveraging edges and optical flow on faces for deepfake detection, in: 2020 IEEE international joint conference on biometrics (IJCB), IEEE, 2020, pp. 1–10.

[109] A. Das, S. Das, A. Dantcheva, Demystifying attention mechanisms for deepfake detection, in: 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), IEEE, 2021, pp. 1–7.

[110] D. Cozzolino, A. Rössler, J. Thies, M. Nießner, L. Verdoliva, Id-reveal: Identity-aware deepfake video detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 15108–15117.

[111] X. H. Nguyen, T. S. Tran, K. D. Nguyen, D.-T. Truong, et al., Learning spatio-temporal features to detect manipulated facial videos created by the deepfake techniques, Forensic Science International: Digital Investigation 36 (2021) 301108.

[112] J. Hu, X. Liao, W. Wang, Z. Qin, Detecting compressed deepfake videos in social networks using frame-temporality two-stream convolutional network, IEEE Transactions on Circuits and Systems for Video Technology 32 (2021) 1089–1102.

[113] Y. Lu, Y. Liu, J. Fei, Z. Xia, Channel-wise spatiotemporal aggregation technology for face video forensics, Security and Communication Networks 2021 (2021) 1–13.

[114] A. Agarwal, A. Agarwal, S. Sinha, M. Vatsa, R. Singh, Md-csdnetwork: Multi-domain cross stitched network for deepfake detection, in: 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), IEEE, 2021, pp. 1–8.

[115] S. Kolagati, T. Priyadharshini, V. M. A. Rajam, Exposing deepfakes using a deep multilayer perceptron–convolutional neural network model, International Journal of Information Management Data Insights 2 (2022) 100054.

[116] F. Chamot, Z. Geradts, E. Haasdijk, Deepfake forensics: Cross-manipulation robustness of feedforward-and recurrent convolutional forgery detection methods, Forensic Science International: Digital Investigation 40 (2022) 301374.

[117] W. Pu, J. Hu, X. Wang, Y. Li, S. Hu, B. Zhu, R. Song, Q. Song, X. Wu, S. Lyu, Learning a deep dual-level network for robust deepfake detection, Pattern Recognition 130 (2022) 108832.

[118] G. Wang, Q. Jiang, X. Jin, W. Li, X. Cui, Mc-lcr: Multimodal contrastive classification by locally correlated representations for effective face forgery detection, Knowledge-Based Systems 250 (2022) 109114.

[119] F. Iqbal, A. Abbasi, A. R. Javed, Z. Jalil, J. Al-Karaki, Deepfake audio detection via feature engineering and machine learning (2022).

[120] V. Dongre, A. T. Reddy, N. Reddeddy, Adaptive re-calibration of channel-wise features for adversarial audio classification, arXiv preprint arXiv:2210.11722 (2022).

[121] M. H. Rahman, M. Graciarena, D. Castan, C. Cobo-Kroenke, M. McLaren, A. Lawson, Detecting synthetic speech manipulation in real audio recordings, in: 2022 IEEE International Workshop on Information Forensics and Security (WIFS), IEEE, 2022, pp. 1–6.

[122] A. Pianese, D. Cozzolino, G. Poggi, L. Verdoliva, Deepfake audio detection by speaker verification, in: 2022 IEEE International Workshop on Information Forensics and Security (WIFS), IEEE, 2022, pp. 1–6.

[123] X. Wang, J. Yamagishi, Investigating active-learning-based training data selection for speech spoofing countermeasure, in: 2022 IEEE Spoken Language Technology Workshop (SLT), IEEE, 2023, pp. 585–592.

[124] T. Chen, A. Kumar, P. Nagarsheth, G. Sivaraman, E. Khoury, Generalization of audio deepfake detection., in: Odyssey, 2020, pp. 132–137.

[125] X. Wang, J. Yamagishi, A comparative study on recent neural spoofing countermeasures for synthetic speech detection, arXiv preprint arXiv:2103.11326 (2021).

[126] A. Gomez-Alanis, J. A. Gonzalez-Lopez, S. P. Dubagunta, A. M. Peinado, M. M. Doss, On joint optimization of automatic speaker verification and anti-spoofing in the embedding space, IEEE Transactions on Information Forensics and Security 16 (2020) 1579–1593.

[127] R. Wang, F. Juefei-Xu, Y. Huang, Q. Guo, X. Xie, L. Ma, Y. Liu, Deepsonar: Towards effective and robust detection of ai-synthesized fake voices, in: Proceedings of the 28th ACM international conference on multimedia, 2020, pp. 1207–1216.

[128] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, A. Larcher, End-to-end anti-spoofing with rawnet2, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2021, pp. 6369–6373.

[129] G. Hua, A. B. J. Teoh, H. Zhang, Towards end-to-end synthetic speech detection, IEEE Signal Processing Letters 28 (2021) 1265–1269.

[130] J. Khochare, C. Joshi, B. Yenarkar, S. Suratkar, F. Kazi, A deep learning framework for audio deepfake detection, Arabian Journal for Science and Engineering (2021) 1–12.

[131] L. Zhang, X. Wang, E. Cooper, N. Evans, J. Yamagishi, The partialspoof database and countermeasures for the detection of short fake speech segments embedded in an utterance, IEEE/ACM Transactions on Audio, Speech, and Language Processing (2022).

[132] Z. Zhang, Y. Gu, X. Yi, X. Zhao, Fmfcc-a: a challenging mandarin dataset for synthetic speech detection, in: Digital Forensics and Watermarking: 20th International Workshop, IWDW 2021, Beijing, China, November 20–22, 2021, Revised Selected Papers, Springer, 2022, pp. 117–131.

[133] P. Kawa, M. Plata, P. Syga, Specrnet: Towards faster and more accessible audio deepfake detection, arXiv preprint arXiv:2210.06105 (2022).

[134] Z. Zhang, X. Yi, X. Zhao, Fake speech detection using residual network with transformer encoder, in: Proceedings of the 2021 ACM workshop on information hiding and multimedia security, 2021, pp. 13–22.

[135] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, N. Evans, Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2022, pp. 6367–6371.

[136] W. Ge, H. Tak, M. Todisco, N. Evans, Can spoofing countermeasure and speaker verification systems be jointly optimised?, arXiv preprint arXiv:2303.07073 (2023).

[137] J. K. Lewis, I. E. Toubal, H. Chen, V. Sandesera, M. Lomnitz, Z. Hampel-Arias, C. Prasad, K. Palaniappan, Deepfake video detection based on spatial, spectral, and temporal inconsistencies using multimodal deep learning, in: 2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), IEEE, 2020, pp. 1–9.

[138] L. Shang, Z. Kou, Y. Zhang, D. Wang, A multimodal misinformation detector for covid-19 short videos on tiktok, in: 2021 IEEE International Conference on Big Data (Big Data), IEEE, 2021, pp. 899–908.

[139] H. Cheng, Y. Guo, T. Wang, Q. Li, X. Chang, L. Nie, Voice-face homogeneity tells deepfake, arXiv preprint arXiv:2203.02195 (2022).

[140] G. Wang, P. Zhang, L. Xie, W. Huang, Y. Zha, Y. Zhang, An audio-visual attention based multimodal network for fake talking face videos detection, arXiv preprint arXiv:2203.05178 (2022).

[141] D. Cozzolino, M. Nießner, L. Verdoliva, Audio-visual person-of-interest deepfake detection, arXiv preprint arXiv:2204.03083 (2022).

[142] H. Ilyas, A. Javed, K. M. Malik, Avfakenet: A unified end-to-end dense swin transformer deep learning model for audio–visual deepfakes detection, Applied Soft Computing 136 (2023) 110124.

[143] W. Yang, X. Zhou, Z. Chen, B. Guo, Z. Ba, Z. Xia, X. Cao, K. Ren, Avoid-df: Audio-visual joint learning for detecting deepfake, IEEE Transactions on Information Forensics and Security 18 (2023) 2015–2029.

[144] M. Lomnitz, Z. Hampel-Arias, V. Sandesara, S. Hu, Multimodal approach for deepfake detection, in: 2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), IEEE, 2020, pp. 1–9.

[145] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, D. Manocha, Emotions don't lie: An audio-visual deepfake detection method using affective cues, in: Proceedings of the 28th ACM international conference on multimedia, 2020, pp. 2823–2832.

[146] B. Hosler, D. Salvi, A. Murray, F. Antonacci, P. Bestagini, S. Tubaro, M. C. Stamm, Do deepfakes feel emotions? a semantic approach to detecting deepfakes via emotional inconsistencies, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 1013–1022.

[147] InVID., Invid verification application, Retrieved from `https://www.invid-project.eu/invid-verification-application/`, Accessed on, 2023.

[148] Intel., Intel real-time deepfake detector, Retrieved from `https://www.intel.com/content/www/us/en/newsroom/news/intel-introduces-real-time-deepfake-detector.html`, Accessed on, 2023.

[149] Y. Li, C. Zhang, P. Sun, L. Ke, Y. Ju, H. Qi, S. Lyu, Deepfake-o-meter: An open platform for deepfake detection, in: 2021 IEEE Security and Privacy Workshops (SPW), IEEE, 2021, pp. 277–281.

[150] R. Defender, Detect deepfakes stop misinformation., Retrieved from `https://realitydefender.com/stop-misinformation`, Accessed on, 2023.

[151] Microsoft, Microsoft video authenticator, Retrieved from `https://blogs.microsoft.com/on-the-issues/2020/09/01/disinformation-deepfakes-newsguard-video-authenticator/`, Accessed on, 2023.

[152] DuckDuckGoose, Deepdetector software, Retrieved from `https://www.duckduckgoose.ai/detector`, Accessed on, 2023.

[153] DuckDuckGoose, Deepfakeproof: Real-time online deepfake detection, Retrieved from `https://www.duckduckgoose.ai/`, Accessed on, 2023.

[154] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, K. Kreis, Align your latents: High-resolution video synthesis with latent diffusion models, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

[155] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, D. J. Fleet, Video diffusion models, arXiv preprint arXiv:2204.03458 (2022).

[156] C. Kong, D. Jeon, O. Kwon, N. Kwak, Leveraging off-the-shelf diffusion model for multi-attribute fashion image manipulation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 848–857.

[157] M. Tkachenko, M. Malyuk, A. Holmanyuk, N. Liubimov, Label Studio: Data labeling software, 2020-2022. URL: `https://github.com/heartexlabs/label-studio`, open source software available from https://github.com/heartexlabs/label-studio.