Diagnosti-CAT, Herramienta de predicción de enfermedades de salud mental mediante machine learning

Miguel Ángel Calderón, Rubén Martínez, Alba Martínez, Runtian Wang,
Álvaro Álvarez, Marcos Gallardo
Universidad Autónoma de Madrid

Resumen

Según los datos recogidos por la Organización Mundial de la Salud (OMS) uno de cada ocho personas padece algún problema de salud mental [24]. Entre ellos destacan la ansiedad, la depresión, la esquizofrenia o el trastorno de bipolaridad. No obstante, la principal preocupación de la salud mental deriva principalmente en los años perdidos por discapacidad. Por estas razones en este trabajo se ha propuesto el desarrollo de Diagnosti-CAT, una herramienta que identifica y cuantifica la gravedad de la depresión, la ansiedad la esquizofrenia y el trastorno de bipolaridad mediante la realización de un cuestionario en un centro médico. A lo largo de este procedimiento el paciente fue grabado tanto de manera audiovisual con el objetivo de obtener las transcripciones clínicas y los gestos y las expresiones faciales del paciente. Gracias a la utilización de Swin transformers y BERT (Representación de Codificación Bidireccional de Transformadores) se obtienen las probabilidades de padecer una de las cuatro patologías bajo estudio. Los resultados conseguidos son prometedores, con un elevado rendimiento en las métricas empleadas tanto sensibilidad, especificidad como precisión, del mismo modo que avalan la efectividad de la técnica creada.

Palabras clave: Depresión, Ansiedad, Trastorno de bipolaridad, Esquizofrenia, Vídeo, Transcripciones clínicas, Swin Transformer, BERT.

1. Introducción

El campo de la salud mental ha obtenido un gran interés estos últimos años debido a la creciente conciencia sobre la importancia del mismo en el bienestar de la población. Un porcentaje, cada vez mayor, de la población tiene algún tipo de problema de salud mental o lo padecerá a lo largo de su vida. Por lo que, la detección temprana y precisa de enfermedades como la depresión, la ansiedad, la esquizofrenia y el trastorno bipolar es crucial para asegurar intervenciones efectivas y poder brindar el tratamiento adecuado a cada paciente lo antes posible.

Asimismo, se ha observado una creciente tendencia en el aumento de técnicas de aprendizaje automático en el campo de la salud mental. Con el uso de estas técnicas, tareas como el análisis de grandes conjuntos de datos clínicos, así como la extracción de patrones complejos que pueden llegar a ser difíciles de identificar con los métodos tradicionales, toman otra dimensión en términos de dificultad. Desde el análisis de texto de notas médicas, hasta el procesamiento de imágenes cerebrales, el aprendizaje automático ha demostrado ser de gran utilidad en diversos aspectos de la detección y evaluación de enfermedades mentales.

Sin embargo, aún persisten grandes desafíos y de vital importancia dentro de este campo. Una de las limitaciones clave es la falta de modelos integrales que no sólo identifiquen la existencia de una enfermedad mental en el paciente, sino que también evalúen su gravedad. La mayoría de los estudios se centran

en la detección binaria de enfermedades, lo que no proporciona una imagen completa del estado de salud mental de un paciente. Además, la interpretación clínica de los resultados de los modelos de aprendizaje automático sigue siendo un área de investigación y desarrollo activa. Por otra parte, la escasez en términos de datos de carácter multimodal es una traba en la elaboración de modelos más completos. Esto se debe a varios factores como la dificultad de acceso a este tipo de datos debido a la tan delicada situación en la que se encuentran los pacientes, la regularización de la información por la ley de protección de datos o a la falta de confianza en este tipo de proyectos.

Por lo que este trabajo plantea abordar estas limitaciones y capitalizar las posibles oportunidades en ambos campos, salud mental y aprendizaje automático. En primer lugar, el desarrollo de un modelo multimodal que mediante datos en formato de video y texto sea capaz de clasificar al paciente como enfermo a la vez que cuantifica la gravedad de esta afección, es toda una novedad en el campo de investigación y permitirá optimizar los tratamientos además de reducir tiempos de espera en los casos más graves. Por otra parte, la investigación en un área tan estigmatizada como la salud mental ayudará a normalizar este tipo de trastornos. Asimismo, al ofrecer una herramienta de evaluación precisa y accesible a través de una aplicación basada en preguntas, motiva su uso en personas que nunca han tenido contacto cercano con expertos de la salud mental, aumentando la probabilidad de que éstas busquen tratamiento en caso de que les sea necesario.

Esta investigación hace frente a diversas limitaciones que deben ser abordadas con cuidado. Una de las principales preocupaciones del estudio es el cumplimiento estricto de las regulaciones de protección de datos, especialmente al manejar información sensible relacionada con la salud mental de los individuos de la muestra. Asimismo, el desafío a la hora de tratar el error del modelo en la predicción de enfermedades mentales es una preocupación constante, ya que cualquier inexactitud podría tener alguna consecuencia en términos de diagnóstico y tratamiento, aunque esto sólo sea una herramienta la cual debe ir siempre acompañada de la opinión de un experto. Otro desafío a abordar es la representatividad de la muestra de datos utilizada para el entrenamiento del modelo, lo que puede introducir sesgos y afectar la generalización de los resultados a poblaciones más amplias.

El principal objetivo de este trabajo de investigación es desarrollar y validar un un modelo de aprendizaje automático que sea capaz de predecir la probabilidad y evaluar la gravedad de diversas enfermedades de salud mental, como la ansiedad, la depresión, la esquizofrenia y el trastorno bipolar, a partir de respuestas estructuradas proporcionadas por los pacientes en un cuestionario específico y las imágenes de los mismos obtenidas mediante grabaciones. Con esto se busca ofrecer una herramienta no invasiva, accesible y precisa que pueda ser utilizada por profesionales de la salud mental con el fin de detectar estas dolencias y su gravedad antes de tener un primer acercamiento, consiguiendo así, optimizar el proceso de detección y permitiendo tratar a tiempo los problemas más graves.

En relación a las novedosas aportaciones de esta investigación, el desarrollo de una herramienta capaz de no sólo calcular la probabilidad de padecer una enfermedad mental, sino también poder medir la gravedad de la misma, es un paso innovador en este campo. Asimismo, el mecanismo de recolección de datos de distintas tipologías como video o audio transcrito a texto aporta más novedad al trabajo.

Por otro lado, consiguiendo mejorar la detección temprana y la evaluación precisa de enfermedades de salud mental, este estudio tiene el potencial de impactar de forma muy positiva en la salud pública al optimizar el proceso de diagnóstico consiguiendo así intervenciones más tempranas y efectivas para los pacientes de mayor riesgo.

El estudio se estructura conforme a una introducción que contextualiza la importancia de la detección precoz de enfermedades mentales y las limitaciones actuales en este campo. Luego, se realiza una revisión exhaustiva de la literatura para identificar los problemas y las oportunidades de mejora. El objetivo principal del paper será desarrollar un modelo predictivo multidimensional que pueda predecir tanto la probabilidad de padecer enfermedades mentales como la gravedad de las mismas. La metodología detalla el diseño del cuestionario, las técnicas de preprocesamiento de datos y la arquitectura del modelo

de aprendizaje automático propuesto. En las conclusiones se analizará el impacto, las limitaciones y las implicaciones éticas. Por último, se tratará el trabajo a futuro y las posibles mejoras de este estudio. Las referencias y apéndices se incluirán al final del trabajo según sea necesario para completar la información dada.

2. Estado actual del arte

Para realizar el estado del arte, se ha seguido la metodología propuesta por las pautas PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses). Esta metodología proporciona un marco estructurado y con gran detalle para la realización de revisiones sistemáticas y la presentación de esta información de una manera trasparente. Este enfoque es capaz de asegurar rigor y reproducibilidad en la identificación, selección y análisis de la literatura relevante relacionada con la detección de enfermedades mentales mediante machine learning.

El proceso de revisión sistemática se dividió en varias etapas clave, estas incluyen la definición de la cuestión que nos ha llevado a esta investigación, la identificación de fuentes de información fiables, la selección de los estudios según los criterios predefinidos en las búsquedas, la extracción de datos e información más relevante y la evaluación de la calidad metodológica de los estudios incluidos. Asimismo, se hace uso de bases de datos especializadas y recursos bibliográficos reconocidos para garantizar la rigurosidad de la búsqueda bibliográfica.

2.1. Identificación y descripción del dominio

En el dominio de la salud mental, el uso de técnicas avanzadas de aprendizaje automático para el análisis y diagnóstico de trastornos mentales se ha convertido en un campo de investigación intensivo y prometedor. Este enfoque tecnológico responde a la necesidad creciente de herramientas de diagnóstico precisas y no invasivas que puedan identificar y evaluar trastornos como la depresión, la ansiedad, la esquizofrenia y el trastorno bipolar. Los estudios recientes en este campo han explorado diversas metodologías, como la evaluación multimodal que integra audio, vídeo y datos textuales para captar una gama más amplia de indicadores de trastornos mentales. Por ejemplo, estudios como los que utilizan los conjuntos de datos DAIC-WOZ y AVEC [45, 9, 33, 20, 34], han demostrado la viabilidad de aplicar modelos de aprendizaje profundo para el análisis comportamental y emocional, mientras que otros estudios han destacado el valor de las transcripciones de interacciones clínicas y el análisis de patrones de comunicación en la detección de síntomas de trastornos como la esquizofrenia y la depresión leve. Estos esfuerzos reflejan un campo en evolución que busca superar las barreras de los métodos tradicionales mediante la adopción de soluciones tecnológicas que mejoran la detección temprana y la personalización del tratamiento en salud mental, enfocándose en la creación de modelos que no solo identifiquen la presencia de trastornos, sino que también cuantifiquen su severidad y progresión.

2.2. Identificación de problemas similares al planteado en el dominio, explicación de diferencias

Al revisar meticulosamente los resúmenes de artículos relevantes para el proyecto Diagnosti-CAT, hemos encontrado que, si bien muchos estudios presentan enfoques innovadores en el ámbito de la salud mental, no todos se alinean con los objetivos específicos de nuestro proyecto multimodal. Por ejemplo, el estudio *Unsupervised Deep Learning to Detect Agitation From Videos in People With Dementia* [22] introduce técnicas de aprendizaje profundo no supervisado para la detección de agitación en personas con demencia. Aunque es una contribución valiosa, su enfoque en un trastorno específico contrasta con nuestra necesidad de abordar un espectro más amplio de trastornos, como la ansiedad, la depresión, la bipolaridad y la esquizofrenia.

En una línea similar, otros estudios, como *The Reproducibility of Bio-Acoustic Features is Associated With Sample Duration, Speech Task, and Gender* [5] y *Analyzing Acoustic and Prosodic Fluctuations in Free Speech to Predict Psychosis Onset in High-risk*

Youths [2], exploran meticulosamente las características bioacústicas y prosódicas del habla. Estos análisis ofrecen detalles significativos sobre las condiciones específicas de los sujetos estudiados. Sin embargo, a pesar de su profundidad analítica, ambos artículos centran sus hallazgos en aspectos que no integran plenamente las dimensiones lingüística, visual y de transcripción de audio que nuestro proyecto persigue combinar en un modelo de aprendizaje automático unificado.

Además, algunos estudios se centran en metodologías que, aunque pertinentes para ciertos aspectos de la salud mental, divergen de nuestras metas de diagnóstico integral. Por ejemplo, *ECoNet: Estimating Everyday Conversational Network From Free-Living Audio for Mental Health Applications* [27] propone un innovador enfoque para estimar redes conversacionales a partir de grabaciones de audio, correlacionando estas con la salud mental. Si bien el enfoque es prometedor para la estimación de redes sociales, se desvía del objetivo de Diagnosti-CAT de realizar diagnósticos clínicos basados en características multimodales.

Por otro lado, estudios como *A Hybrid Deep Feature Selection Framework for Emotion Recognition from Human Speeches* [31] y *Federated Learning for Violence Incident Prediction in a Simulated Cross-Institutional Psychiatric Setting* [8] introducen marcos de trabajo interesantes para la selección de características emocionales en el habla y para la predicción de incidentes de violencia, respectivamente. Estos enfoques son útiles para aplicaciones específicas dentro de la salud mental, pero su aplicación se centra en ámbitos que no cumplen con la amplitud y diversidad de diagnósticos que nuestro proyecto pretende ofrecer.

Dual-Stream Multiple Instance Learning for Depression Detection With Facial Expression Videos[42] presenta un método basado en aprendizaje múltiple instancia supervisado para detectar depresión a través de videos de expresiones faciales, destacando en precisión pero limitado al uso de un solo tipo de dato y en un entorno supervisado, lo que contrasta con la integración de datos multimodales y el aprendizaje menos supervisado que busca Diagnosti-CAT. Asimismo, *Temporal Facial Features for Depression Screening*[14] y *Additive Cross-Modal Attention Network (ACMA) for Depression Detection Based on Audio and Textual Features*[20] se centran únicamente en la depresión y utilizan tecnologías que divergen de las arquitecturas de transformers y BERT planeadas para nuestro proyecto, además de no abordar los múltiples trastornos que Diagnosti-CAT intenta diagnosticar.

Finalmente, *An Ambient Intelligence-Based Approach for Longitudinal Monitoring of Verbal and Vocal Depression Symptoms*[34] y *Exploring Language Markers of Mental Health in Psychiatric Stories*[46] ofrecen enfoques innovadores en el monitoreo y análisis lingüístico de la depresión, pero su aplicabilidad es limitada por el enfoque en recurrencia de síntomas y el análisis de datos en un solo idioma, respectivamente, lo cual no cumple con las necesidades de análisis en tiempo real y en múltiples idiomas de Diagnosti-CAT. Además, *EANDC: An explainable attention network based deep adaptive clustering model for mental health treatment*[3] y *An Improved Bagging Ensemble in Predicting Mental Disorder Using Hybridized Random Forest - Artificial Neural Network Model*[1] proponen métodos basados en emociones y modelos híbridos que, aunque efectivos para sus propósitos específicos, no alinean con las tecnologías avanzadas y el enfoque multimodal y multitrastorno requerido por Diagnosti-CAT.

Tras todo este análisis, se ha prestado una especial atención a los artículos [10, 45, 35, 54, 24, 44], debido a su relevancia directa y a la profundidad con que abordan problemas similares a los de nuestro proyecto, como estudios que integran evaluaciones multimodales y utilizan aprendizaje profundo para analizar señales lingüísticas, acústicas, visuales e incluso de transcripción de audio en el diagnóstico de las patologías objetivo. De hecho, no solo cubren estas enfermedades, sino que también demuestran metodologías avanzadas y resultados prometedores que pueden potenciar el diagnóstico y seguimiento en el contexto de Diagnosti-CAT; ofreciendo un marco sólido para la detección temprana y la intervención en salud mental. Por ejemplo, el estudio sobre depresión y predicción de recaídas utiliza técnicas de aprendizaje automático y datos audiovisuales para diagnosticar la depresión, similar a nuestra propuesta integral [35].

Además, la evaluación de síntomas de esquizofrenia también emplea un enfoque multimodal que incluye análisis lingüístico, acústico y visual, alineándose con las técnicas que se pretende utilizar [10]. Del mismo modo, el análisis mediante representaciones espectrales de video en [45].utiliza primitivas de comportamiento facial para detectar la depresión, destacando la utilidad de las técnicas de procesamiento de imágenes.

Sin embargo, todavía se han identificado múltiples problemas en los estudios anteriores. Estos incluyen la variabilidad y subjetividad en las evaluaciones clínicas, que pueden diferir sustancialmente entre los evaluadores debido a interpretaciones humanas subjetivas, como se ha observado en la evaluación de síntomas de la esquizofrenia [10]. Otro desafío significativo es la limitada integración de las capacidades del aprendizaje profundo, que restringe la eficacia y precisión de los diagnósticos automáticos, especialmente en la detección y análisis de la bipolaridad [44]. Los estudios también han señalado la escasez de datos multimodales, exacerbada por dificultades en la recopilación de datos sensibles debido a restricciones éticas y de privacidad, lo que limita la capacidad de entrenar modelos robustos y comprensivos [54]. Además, la falta de consideración del contexto del comportamiento en las mediciones puede llevar a análisis superficiales que no reflejan adecuadamente la complejidad de los estados mentales [45]. Existen diferencias fundamentales en cómo se planea expandir y mejorar estos enfoques. A diferencia de los estudios mencionados que se centran en un único trastorno, Diagnosti-CAT estará diseñado para identificar múltiples trastornos: ansiedad, depresión, bipolaridad, esquizofrenia y diferenciarlos de pacientes sanos, además de medir la gravedad de las mismas. Esto representa una ampliación significativa del alcance diagnóstico, proporcionando una herramienta más versátil y comprensiva que puede adaptarse a diversos entornos clínicos y demográficos. Con respecto a los problemas, se plantea mitigarlos de forma metodológica, además, mientras que cada uno de los estudios previos se enfoca en metodologías específicas o conjuntos de datos limitados, Diagnosti-CAT integrará múltiples tipos de datos y modalidades para ofrecer un análisis más robusto y detallado, superando las limitaciones de estudios que se enfocan en un único aspecto o trastorno [45, 10].

2.3. Comparativa de datasets utilizados

Los estudios revisados en el ámbito de la salud mental adoptan una variedad de propuestas multimodales y automáticas para abordar el análisis y la detección de diferentes trastornos, empezando con el estudio [35]. utiliza el conjunto de datos DAIC-WOZ, que incluye entrevistas clínicas biomédicas audiovisuales, adoptando una estrategia de entrenamiento *Leave-One-Subject-Out* específicamente enfocada en la depresión, que refleja un interés por comprender y predecir la recaída en pacientes previamente diagnosticados, lo cual es crucial para el tratamiento a largo plazo y la gestión de la depresión. En contraste, el estudio [10] examina múltiples conjuntos de datos que incorporan modalidades como audio, video, para detectar una gama más amplia de trastornos mentales. Este enfoque no solo amplía el espectro de trastornos estudiados sino que también diversifica las fuentes de datos utilizadas, lo que puede enriquecer la precisión y la robustez de los modelos de diagnóstico desarrollados. Además, el estudio [54]se centra en transcripciones de entrevistas reales entre consejeros y pacientes, utilizando datos que incluyen detalles semánticos y paralingüísticos para detectar ansiedad y depresión leve, que destaca la importancia de los matices comunicativos en la evaluación de la salud mental. Por otro lado, la [10] involucra grabaciones de entrevistas psiquiátricas con pacientes esquizofrénicos, utilizando modalidades lingüísticas, acústicas y visuales en un modelo de fusión profunda para evaluar la severidad de los síntomas. Este método busca capturar la complejidad de la esquizofrenia, que a menudo requiere una evaluación multidimensional para un diagnóstico preciso mientras que en [45] se emplea los conjuntos de datos AVEC 2013 y AVEC 2014, que contienen videos etiquetados con escalas de depresión BDI-II, usando primitivas de comportamiento facial para el análisis de la depresión. Este estudio ilustra cómo las tecnologías de procesamiento visual y el análisis de gestos pueden ser cruciales para identificar signos de depresión. Se puede observar de manera más concreta

las diferencias y similitudes entre los diversos estudios analizados en cuadro 1.

2.4. Comparativa de modelos utilizados y mejores resultados hasta el momento

Nuestro trabajo de investigación busca desarrollar un modelo de aprendizaje automático multimodal para identificar cuatro trastornos mentales. Para ello, se ha comprobado que el uso de Swin transformer para analizar información de vídeo y de BERT para procesar texto transcrito de audio ofrece muy buenos resultados. En investigaciones existentes, hay modelos multimodales de aprendizaje profundo basados en transformers (vídeo, audio, registros textuales) para evaluar la severidad de la esquizofrenia [10]. Estos modelos se fundamentan en cuatro redes neuronales Transformer preentrenadas, cada una dedicada a un modal específico: lenguaje semántico, lenguaje sintáctico, sonido y vídeo. El modelo principal es textual, con audio y vídeo como apoyos, y ha demostrado una alta precisión en los resultados (MAE = 0.534, MSE = 0.685), validando la efectividad del modelo multimodal. Más específicamente, en el ámbito del reconocimiento por vídeo, investigadores [45] han usado Unidades de Acción Facial (UA), posturas de cabeza y direcciones de la mirada para identificar la depresión. Además, modelos de aprendizaje profundo como CNN requieren entradas de tamaño fijo. Sin embargo, como los vídeos tienen longitudes variables, este estudio también empleó representaciones espectrales y utilizó dos métodos de alineación de frecuencias para crear espectros de tamaño uniforme, solucionando así el problema de la variabilidad en la duración de los vídeos. Los resultados del estudio se verificaron usando MAE y RMSE, con valores entre 5.95 y 6.16 para MAE y entre 7.15 y 8.1 para RMSE, superando significativamente a modelos similares. Por otro lado, otros investigadores han utilizado modelos multimodales que combinan información de vídeo proporcionada por UA y audio de redes VGGish para la identificación de depresión [35]. Estos datos se transformaron en características mediante un modelo de aprendizaje profundo, y luego se introdujeron en el modelo MoN. En el estudio, la codificación de características de los pacientes y la distancia al modelo de normalidad (MoN) se usaron para determinar si los sujetos tenían depresión, alcanzando una precisión del 87.4 % y un F1 de 82.3 %, lo que demuestra la efectividad del modelo. En cuanto a los modelos de reconocimiento de información textual, investigaciones de Zubiaga, Irune [54] y otros proporcionan un marco de referencia. Transformaron las conversaciones de sujetos con diferentes estados psicológicos en información textual, seleccionaron características y luego ingresaron los datos en modelos de aprendizaje automático como bosques aleatorios, redes neuronales y BERT para clasificación. Los resultados mostraron que BERT tenía un rendimiento comparable al de las redes neuronales. Es importante destacar que las características de la paralingüística (como las pausas para respirar profundamente, la risa o las pausas mientras se habla) estaban significativamente relacionadas con los resultados del modelo. Finalmente, estudios de Khoo, Lin Sze [24] y otros han resumido sistemáticamente 184 métodos de detección de salud mental utilizando aprendizaje automático y sensores pasivos (datos multimodales). Los resultados indican que las características relacionadas con ciertos modos (como audio, vídeo, uso de redes sociales, etc.) pueden variar significativamente según las características demográficas y los rasgos de personalidad del individuo, lo que proporciona insights valiosos para nuestra investigación. Véase una descripción de medidas de evaluación del problema en el cuadro 3.

2.5. Oportunidad científica y motivación del proyecto

Tras haber visto los mejores resultados del estado del arte, se identifica una oportunidad de combinar técnicas de BERT vistas en Chuang et al [10], Khandker et al. [23] o Zubiaga et al. [54] y de Swin transformers (Chuang et al. [10], Hwang et al [19], Duan et al [13]) para obtener un modelo multimodal novedoso de diagnóstico de diversas patologías mentales. Véase un resumen comparativo en el cuadro 2. En particular, resulta de especial interés la posibilidad de aplicar dichas técnicas a las siguientes enfermedades: Depresión, ansiedad, esquizofrenia y trastorno

de bipolaridad. De esta forma, se puede desarrollar una herramienta nueva que ayude a los hospitales a diagnosticar pacientes nuevos y medir la posibilidad de padecer estas enfermedades mentales. A su vez, se hará uso de este mismo modelo para calcular la gravedad de dichas patologías, lo que en definitiva resulta una verdadera nueva oportunidad.

3. Metodología y propuesta

El proyecto Diagnosti-CAT consiste en el diseño de un modelo y una herramienta para la identificación y determinación de la gravedad de la ansiedad, la depresión, el trastorno de bipolaridad y la esquizofrenia frente a pacientes sanos en la población adulta. Diagnosti-CAT se compone de un test de 20 preguntas que serán respondidas por los pacientes en una sesión clínica de duración temporal de 15 minutos. Las preguntas serían iguales para todos los pacientes, que abordaron una serie de temas de naturaleza variada con el propósito de reconocer problemas de salud mental. La finalidad de este cuestionario será evaluar las respuestas aportadas por el usuario mediante el análisis textovisual con un novedoso método multimodal basado en herramientas de machine learning como BERT o Swin transformers. Esta plataforma aportará un resultado de la probabilidad de padecer alguna de estas cuatro patologías (o su ausencia), así como de dar una estimación de la gravedad. Con los resultados extraídos, el especialista será responsable de la toma de decisión final, y si es necesario de derivar al paciente a un centro de atención especializada.

De esta forma, una de las principales ventajas de esta idea es la accesibilidad a un diagnóstico médico. Muchas veces la principal limitación que padecen los pacientes es la imposibilidad de acceder a consultas como consecuencia de la dependencia de una tercera persona para poder desplazarse a un centro de atención especializada o la incompatibilidad laboral o el horario de atención sanitaria. De este modo, este software ofrecerá la posibilidad de realizar una estimación previa de la probabilidad de padecer algún problema de salud mental en los centros de atención primaria. En función de los resultados extraídos será el especialista el encargado de derivarlo al psiquiatra. Igualmente, otro aspecto positivo será la disminución de las listas de espera. Este hecho posibilitará la reducción de la carga de trabajo de los médicos, centrándose en aquellos pacientes con mayor probabilidad de presentar alguna patología en función de su gravedad.

Así mismo, se beneficiarían como usuarios todas aquellas personas que requieran un diagnóstico médico o por medio de la consulta de familiares o médicos de atención primaria.

3.1. Objetivos principales

De este modo, se establecen los siguientes objetivos (en orden):

- Creación de una base de datos mulitmodal (texto y vídeo) estructurada de pacientes adultos con depresión, ansiedad, trastorno bipolar, esquizofrenia y sujetos sanos.
- Desarrollo de la plataforma Diagnosti-CAT mediante la utilización de algoritmos de machine learning como BERT y Swin transformers para la identificación y cuantificación de la gravedad·
- Obtención de métricas de rendimiento, como la precisión, sensibilidad, especificidad, entre otras, para la evaluación y comparación de nuestro modelo con métodos actuales.

3.2. Datasets a utilizar

Uno de los apartados más novedosos de este modelo es la fuente, etiquetado y concepción de los datos a utilizar. Para el estudio, se han realizado una serie de consultas psiquiátricas grabadas en hospitales con profesionales de la salud mental. Se han seleccionado 2 perfiles distintos de candidatos: sanos y con diagnóstico en cada una de las patologías mentales objetivo. Para la selección de entrevistados, se establecieron una serie de criterios de inclusión. Para el conjunto de pacientes sanos, se buscaron candidatos adultos entre 18 y 60 años sin diagnóstico previo de patologías mentales. Para el conjunto de pacientes diagnosticados, se buscaron también adultos entre 18 y 60 años, con un historial clínico positivo en una

de las siguientes enfermedades: depresión, ansiedad, trastorno de bipolaridad o esquizofrenia.

Estas consultas, realizadas gracias a una serie de convenios con hospitales de todo el país, se desarrollan con un esquema fijo, en el que el psiquiatra propone 20 preguntas fijas redactadas previamente por un conjunto de expertos. Se graba exclusivamente al entrevistado, enfocando la parte superior del cuerpo. El propósito de estas preguntas de respuesta libre es detectar indicios suficientes de los trastornos psiquiátricos objetivo (depresión, la ansiedad, trastorno de bipolaridad y esquizofrenia) de tal forma que al concluir la entrevista, el profesional sea capaz de etiquetar al entrevistado en una de las 5 clases disponibles (añadiendo al paciente sano). Tras esto, se procede a transcribir las respuestas del entrevistado, y a guardar la grabación. De esta forma, se obtienen dos datasets asociados: uno compuesto de audio con las preguntas y las respuestas, y otro basado en vídeo con las expresiones faciales y gestos. El dataset de audio habrá que procesarlo para pasarlo a texto.

Existen similitudes de este dataset con varios ya existentes. Khan et al. [22] usa grabaciones de vídeo de la Unidad Especializada de Demencia en Toronto, Canadá. Llegando un poco más allá, Chuang et al. [10] hace uso de entrevistas psiquiátricas grabadas que constan de audio, vídeo y transcripciones textuales sobre pacientes con esquizofrenia. Sin embargo, el enfoque propuesto en este artículo difiere tanto en la forma (las entrevistas constan de nuevas preguntas desarrollada por un conjunto de especialistas) como en el fondo; centrándose en el reconocimiento de varias patologías simultáneamente. Sin contar que no existen datasets con estas características realizados sobre pacientes españoles. En el campo gestual y expresivo, la cultura resulta de vital importancia.

3.3. Atributos a inferir

En este trabajo, se han utilizado tanto datos visuales como transcripciones clínicas, obtenidas del audio durante la realización de la prueba. Las grabaciones audiovisuales son registros de duración de 15 minutos, que muestran las expresiones faciales y gestos efectuados por los voluntarios. Con el fin de mejorar el rendimiento y la calidad de nuestros resultados se ha introducido el contenido textual de las respuestas aportadas por los usuarios del cuestionario.

Una vez obtenidas las bases de datos, se realizan una serie de transformaciones sobre las mismas. En particular, se necesita transcribir el audio de las entrevistas a texto. Para llevarlo a cabo, se hace uso de la herramienta Google Cloud Platform. Cabe destacar que en las transcripciones del audio, se conservan elementos vocativos del habla, como interjecciones. En Zubiaga et al. [54] se puede observar que estos elementos contienen mucha información.

En lo que al aspecto visual de los datos corresponde, el objetivo es inferir a partir de gestos faciales y gesticulación manual. Existen trabajos que hacen uso de esto mismo. Se usan expresiones faciales en Zhang et al. [52], Casado et al. [9] y más. También se utilizan atributos gestuales diversos (Francese et al. [15]). Estos atributos son muy extendidos en el estado del arte. La potencia de Diagnosti-CAT será combinar atributos textovisuales para mejorar el rendimiento.

3.4. Métodos a utilizar

Para cumplir los objetivos fijados, se va a hacer uso de un modelo multimodal, basado en dos arquitecturas principales: BERT y Swin transformers. La técnica BERT será utilizada para el procesamiento de las transcripciones clínicas de texto, mientras que para analizar los datos de vídeo de grabaciones de la cara y la parte superior del cuerpo, usaremos Swin transformers. De esta forma se plantea aprovechar al máximo la propiedad multimodal misma de la base de datos.

Los Swin transformers, como se vio en la sección 2, han sido popularmente utilizados para el análisis de expresiones faciales y gestuales como Chuang et al. [10], Hwang et al [19], Duan et al [13], entre otros. Los Swin transformers son una herramienta potente creada específicamente para extraer y procesar características visuales.

Respecto a BERT también se ha comprobado un amplio respaldo científico para el análisis del contenido textual. BERT es capaz de manejar clases desbalanceadas sin necesitar aumentación de datos, lo cual encaja perfectamente en nuestro dataset. Entre varios artículos revisados, Chuang et al [10], Khandker

et al. [23] o Zubiaga et al. [54] usan BERT.

En virtud al interés de cuantificar la gravedad de las patologías objetivo, se ha prestado atención a varios artículos como Milintsevich et al. [33], Jenei et al. [21] o Song et al. [45]. Mediante la utilización de los métodos anteriormente mencionados y el cuestionario creado por expertos de la salud mental se pretenderá obtener dicha estimación de manera objetiva.

Debido al enorme volumen de datos necesario para entrenar estos modelos, se realizará fine-tuning en ambos. Para BERT, se usará el modelo preentrenado [12]. Para preentrenar el Swin transformer, se usará [30]. Este proceso de fine-tuning se encuentra presente en las referencias anteriormente mencionadas.

Con esto, se realizará la combinación de ambos métodos de análisis multimodal. El punto fuerte de esta metolodogía se basará en la fusión de ambas tareas (probabilidad y gravedad) que seguirán caminos diferentes dentro de las redes excepto en la última etapa que se unificarán para alcanzar sendas estimaciones. Después, con el fin de conocer el rendimiento de nuestro modelo se calcularán la precisión, sensibilidad, especificidad. Así se contruye un modelo nuevo, que fusiona características textovisuales, y que le permitirá a Diagnosti-CAT ser más eficaz y preciso para la obtención tanto de la probabilidad como la gravedad de presentar alguna de las cuatro patologías bajo análisis.

4. Propuesta de experimentación

Como ya se ha mencionado en apartados anteriores, la herramienta Diagnosti-CAT conseguirá una serie de resultados sobre 20 preguntas para la obtención de un cuadro clínico. Estas preguntas han sido realizadas por un equipo de psiquiatras colegiados sobre pacientes con las dolencias especificadas más arriba ya diagnosticadas. También se realizará el test en personas que no tengan estos problemas psicológicos para eliminar, en la medida de lo posible, los falsos positivos.

Una vez recopiladas y registradas las respuestas, se realizará una limpieza de los datos, entre los que se incluye la eliminación de datos repetidos, para después usar técnicas de *Machine Learning* sobre los datos.

El volumen esperado de datos para la prueba de los datos será de una cantidad para nada despreciable. Al ser una herramienta que ayude a la comunidad de médicos de naturaleza psicológica, esperamos en este proyecto colaboren una gran cantidad de médicos de la salud mental y que, con suerte, cada uno aporte información de al menos 20 pacientes.

Esto hará que la base de datos tenga una cantidad de 21 columnas, una columna por pregunta y una adicional con la dolencia del paciente.

Gracias a la obtención de una gran cantidad de formularios de personas con las dolencias seleccionadas y personas "sanas", se pueden realizar particiones **train/test** de gran magnitud y además ir cambiando en varias iteraciones los conjuntos de entrenamiento y prueba para mejorar los modelos.

Los resultados que se esperan será superior al de otros modelos parecidos anteriormente referenciados. Esto es por el modo de obtener los datos para entrenar los modelos. Al haber sido preguntas que han sido diseñadas por médicos especialistas permitirá la identificación de la dolencia, o ausencia de ella, así como estimación de la gravedad.

Al ser un formulario extenso, los datos recopilados son muchos haciendo más fácil poder clasificar y con ello poder pasar a la etapa de diagnóstico.

5. Conclusiones

En definitiva, el modelo propuesto puede ser de gran ayuda para el diagnóstico de enfermedades mentales tan difíciles de detectar incluso por médicos especialistas. A su vez ofrecerá una medida de la gravedad de las mismas.

El resultado de la aplicación puede ser de gran ayuda a los médicos mentales de todo el mundo brindándoles una herramienta con la que poder diagnosticar rápidamente alguna de estas enfermedades mentales. Además, se espera que con esta herramienta se ayude a la visibilización de este tipo de patologías a la sociedad.

Este programa resultará, en definitiva, de gran ayuda para los especialistas psiquiátras para el diagnósti-

co de las enfermedades mentales objetivo del programa Diagnosti-CAT. Con ello, se pretende agilizar los tiempos de consulta y diagnóstico por parte de los especialistas.

No obstante, como en todo trabajo sobre salud, este proyecto tiene sus limitaciones. Una de ellas es la ley de protección de datos, que puede dificultar la obtención e incluso obligar a la eliminación de datos de las bases de datos, haciendo que las predicciones sean menos precisas. Los datos médicos resultan siempre de una gran sensibilidad. También es posible que el paciente no conteste de forma esperable a las preguntas, ya sea de manera accidental o intencionada.

Como trabajo futuro se propone el aumento de dolencias a detectar por la herramienta Diagnosti-CAT que se encuentran fuera del abanico inicial de la aplicación. Aunque ya se esté usando un formulario de 20 preguntas para el diagnóstico de algunas enfermedades mentales, se podría aumentar el número de preguntas para aumentar la precisión de la aplicación.

También se puede investigar sobre el diagnóstico prematuro de las enfermedades mentales para que el impacto en los pacientes sea el menor posible.

6. Anexos

Cuadro 1: Comparativa de Datasets Utilizados en la Investigación de Trastornos Mentales

Estudio	**Dataset**	**Modalidades**	**Enfoque de Trastorno**	**Características Específicas del Dataset**
A Model of Normality Inspired Deep Learning Framework for Depression Relapse Prediction Using Audiovisual Data	DAIC-WOZ	Audiovisual	Depresión	Datos de entrevistas clínicas con estrategia de entrenamiento Leave-One-Subject-Out
Machine Learning for Multimodal Mental Health Detection: A Systematic Review of Passive Sensing Approaches	Varios	Multimodal audio, video, datos de dispositivos	Trastornos mentales diversos	Datos que incluyen expresiones faciales, movimientos de cabeza, texto, y más
Multimodal Feature Evaluation and Fusion for Emotional Well-Being Monitorization	Transcripciones de entrevistas reales	Multimodal: Audio, Transcripciones Textuales	Ansiedad y Depresión	Incluye detalles paralingüísticos como música, pasos, inhalaciones, risas
Multimodal Assessment of Schizophrenia Symptom Severity From Linguistic, Acoustic, and Visual Cues	Entrevistas de hospital	Multimodal: Lingüística, Acústica, Visual	Esquizofrenia	Incluye grabaciones de video y audio de entrevistas psiquiátricas, con transcripciones textuales. Fusión de modalidades mediante transformadores.
Spectral Representation of Behaviour Primitives for Depression Analysis	AVEC 2013 y 2014	Audiovisual	Depresión	Videos etiquetados con escalas de depresión BDI-II; uso de primitivas de comportamiento facial
Multimodal Machine Learning Framework to Detect the Bipolar Disorder	Bipolar Disorder Corpus dataset	Audiovisual	Trastorno de bipolaridad	46 entrevistas estructuradas a pacientes turcos

Cuadro 2: Resumen de Parámetros Utilizados en Estudios de Salud Mental

Título del Artículo	**Variables Predictivas**	**Descripción de las Variables**
Multimodal Assessment of Schizophrenia Symptom Severity From Linguistic, Acoustic, and Visual Cues	TLC, PANSS	Escalas psicológicas para evaluar la severidad de síntomas esquizofrénicos, incluyendo aspectos de pensamiento y síntomas positivos/negativos.
Spectral Representation of Behaviour Primitives for Depression Analysis	AUs	Unidades de acción que capturan movimientos faciales relacionados con depresión para análisis temporal.
A Model of Normality Inspired Deep Learning Framework for Depression Relapse Prediction Using Audiovisual Data	Action Units, VGGish, MoN	Información visual y auditiva procesada para predecir recaídas en depresión usando el modelo MoN.
Multimodal Feature Evaluation and Fusion for Emotional Well-Being Monitorization	Random Forest, BERT	Uso de Random Forest y BERT para evaluar características textuales en la clasificación del bienestar emocional.
Machine Learning for Multimodal Mental Health Detection: A Systematic Review of Passive Sensing Approaches	Various Parameters	Revisión de técnicas de aprendizaje automático para detectar problemas de salud mental mediante sensores pasivos.
Multimodal Machine Learning Framework to Detect the Bipolar Disorder	Paragraph Vectors, Denoising autoencoders	Uso multimodal de PV y autoencoders para transcripciones de entrevistas y análisis audiovisual

Cuadro 3: Resumen de Medidas de Evaluación Utilizadas en Estudios de Salud Mental

Título del Estudio	**Medidas de Evaluación**	**Descripción de las Medidas**
Multimodal Assessment of Schizophrenia Symptom Severity From Linguistic, Acoustic, and Visual Cues	MAE, MSE	MAE = 0.534 y MSE = 0.685, indican una precisión superior del modelo en comparación con otros estudios.
Spectral Representation of Behaviour Primitives for Depression Analysis	MAE, RMSE	Uso de MAE y RMSE para validar el modelo. En diferentes bases de datos, MAE varió entre 5.95 y 6.16; RMSE entre 7.15 y 8.1.
A Model of Normality Inspired Deep Learning Framework for Depression Relapse Prediction Using Audiovisual Data	Accuracy, F1 Score	Accuracy = 87.4 % y F1 = 82.3 %, reflejan un rendimiento robusto en predicciones de recaída.
Multimodal Feature Evaluation and Fusion for Emotional Well-Being Monitorization	F-score, Accuracy, Precision, Recall, Loss	Diversas métricas para evaluar el modelo, incluyendo F-score, Accuracy, Precision y Recall.
Machine Learning for Multimodal Mental Health Detection: A Systematic Review of Passive Sensing Approaches	Accuracy, Precision, Recall	Se mencionan métricas como Accuracy, Precision y Recall, pero no se comparan cuantitativamente los modelos.
Multimodal Machine Learning Framework to Detect the Bipolar Disorder	Loss, Accuracy, Precision, Recall, F-score	Se compara el modelo con Naive Bayes y Decision Trees. El modelo mejora resultados en todas las medidas propuestas

Referencias

[1] Oluwashola David Adeniji, Samuel Oladele Adeyemi, and Sunday Adeola Ajagbe. An improved bagging ensemble in predicting mental disorder using hybridized random forest-artificial neural network model. *Informatica*, 46(4), 2022.

[2] Carla Agurto, Mary Pietrowicz, Raquel Norel, Elif K Eyigoz, Emma Stanislawski, Guillermo Cecchi, and Cheryl Corcoran. Analyzing acoustic and prosodic fluctuations in free speech to predict psychosis onset in high-risk youths. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 5575–5579. IEEE, 2020.

[3] Usman Ahmed, Gautam Srivastava, Unil Yun, and Jerry Chun-Wei Lin. Eandc: An explainable attention network based deep adaptive clustering model for mental health treatment. *Future Generation Computer Systems*, 130:106–113, 2022.

[4] Asmaa Alayed, Manar Alrabie, Sarah Aldumaiji, Ghaida Allhyani, Sahar Siyam, and Reem Qaid. An arabic intelligent diagnosis assistant for psychologists using deep learning. *International Journal of Advanced Computer Science and Applications*, 14(6), 2023.

[5] Shaykhah A Almaghrabi, Dominic Thewlis, Simon Thwaites, Nigel C Rogasch, Stephan Lau, Scott R Clark, and Mathias Baumert. The reproducibility of bio-acoustic features is associated with sample duration, speech task, and gender. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:167–175, 2022.

[6] Nadim AA Atiya, Quentin JM Huys, Raymond J Dolan, and Stephen M Fleming. Explaining distortions in metacognition with an attractor network model of decision uncertainty. *PLoS Computational Biology*, 17(7):e1009201, 2021.

[7] Casey C Bennett, Mindy K Ross, EuGene Baek, Dohyeon Kim, and Alex D Leow. Predicting clinically relevant changes in bipolar disorder outside the clinic walls based on pervasive technology interactions via smartphone typing dynamics. *Pervasive and Mobile Computing*, 83:101598, 2022.

[8] Thomas Borger, Pablo Mosteiro, Heysem Kaya, Emil Rijcken, Albert Ali Salah, Floortje Scheepers, and Marco Spruit. Federated learning for violence incident prediction in a simulated cross-institutional psychiatric setting. *Expert Systems with Applications*, 199:116720, 2022.

[9] Constantino Álvarez Casado, Manuel Lage Cañellas, and Miguel Bordallo López. Depression recognition using remote photoplethysmography from facial videos. *IEEE Transactions on Affective Computing*, 2023.

[10] Chih-Yuan Chuang, Yi-Ting Lin, Chen-Chung Liu, Lue-En Lee, Hsin-Yang Chang, An-Sheng Liu, Shu-Hui Hung, and Li-Chen Fu. Multimodal assessment of schizophrenia symptom severity from linguistic, acoustic and visual cues. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2023.

[11] Jetli Chung and Jason Teo. Single classifier vs. ensemble machine learning approaches for mental health prediction. *Brain informatics*, 10(1):1, 2023.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[13] Longteng Duan, Wei Shao, and Linqi Song. Facial expression recognition based on data augmentation and swin-transformer. pages 1–5, 2022.

[14] Ricardo Flores, ML Tlachac, Avantika Shrestha, and Elke Rundensteiner. Temporal facial features for depression screening. In *Adjunct Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2022 ACM International Symposium on Wearable Computers*, pages 488–493, 2022.

[15] Rita Francese and Pasquale Attanasio. Emotion detection for supporting depression screening. *Multimedia Tools and Applications*, 82(9):12771–12795, 2023.

[16] Yanrong Guo, Chenyang Zhu, Shijie Hao, and Richang Hong. Automatic depression detection via learning and fusing features from visual cues. *IEEE transactions on computational social systems*, 2022.

[17] Lang He, Chenguang Guo, Prayag Tiwari, Rui Su, Hari Mohan Pandey, and Wei Dang. Depnet: An automated industrial intelligent system using deep learning for video-based depression analysis. *International Journal of Intelligent Systems*, 37(7):3815–3835, 2022.

[18] Pufang Huang. A mental disorder prediction model with the ability of deep information expression using convolution neural networks technology. *Scientific Programming*, 2022:1–8, 2022.

[19] Hyeonbin Hwang, Soyeon Kim, Parque Wei-Jin, Jiho Seo, Kyungtae Ko, and Hyeon Yeo. Vision transformer equipped with neural resizer on facial expression recognition task. pages 2614–2618, 2023.

[20] Ngumimi Karen Iyortsuun, Soo-Hyung Kim, Hyung-Jeong Yang, Seung-Won Kim, and Min Jhon. Additive cross-modal attention network (acma) for depression detection based on audio and textual features. *IEEE Access*, 2024.

[21] Attila Zoltán Jenei and Gábor Kiss. Severity estimation of depression using convolutional neural network. *Periodica Polytechnica Electrical Engineering and Computer Science*, 65(3):227–234, 2021.

[22] Shehroz S Khan, Pratik K Mishra, Nizwa Javed, Bing Ye, Kristine Newman, Alex Mihailidis, and Andrea Iaboni. Unsupervised deep learning to detect agitation from videos in people with dementia. *IEEE Access*, 10:10349–10358, 2022.

[23] Rezaul K Khandker, Md Rakibul Islam Prince, Farid Chekani, Paul Richard Dexter, Malaz A Boustani, and Zina Ben Miled. Digital-reported outcome from medical notes of schizophrenia and bipolar patients using hierarchical bert. *Information*, 14(9):471, 2023.

[24] Lin Sze Khoo, Mei Kuan Lim, Chun Yong Chong, and Roisin McNaney. Machine learning for multimodal mental health detection: A systematic review of passive sensing approaches. *Sensors*, 24(2):348, 2024.

[25] Rohan kumar Gupta and Rohit Sinha. An investigation on the audio-video data based estimation of emotion regulation difficulties and their association with mental disorders. *IEEE Access*, 2023.

[26] S Lalitha, Deepa Gupta, Mohammed Zakariah, and Yousef Ajami Alotaibi. Mental illness disorder diagnosis using emotion variation detection from continuous english speech. *Computers, Materials & Continua*, 69(3), 2021.

[27] Bishal Lamichhane, Nidal Moukaddam, Ankit B Patel, and Ashutosh Sabharwal. Econet: Estimating everyday conversational network from free-living audio for mental health applications. *IEEE Pervasive Computing*, 21(2):32–40, 2022.

[28] Keke Li and Weifang Yu. A mental health assessment model of college students using intelligent technology. *Wireless Communications and Mobile Computing*, 2021:1–10, 2021.

[29] Shuang Li and Yu Liu. Deep learning-based mental health model on primary and secondary school students' quality cultivation. *Computational Intelligence and Neuroscience*, 2022, 2022.

[30] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[31] Aritra Marik, Soumitri Chattopadhyay, and Pawan Kumar Singh. A hybrid deep feature selection framework for emotion recognition from human speeches. *Multimedia Tools and Applications*, 82(8):11461–11487, 2023.

[32] Ridha Mezzi, Aymen Yahyaoui, Mohamed Wassim Krir, Wadii Boulila, and Anis Koubaa. Mental health intent recognition for arabic-speaking patients using the mini international neuropsychiatric interview (mini) and bert model. *Sensors*, 22(3):846, 2022.

[33] Kirill Milintsevich, Kairit Sirts, and Ga

l Dias. Towards automatic text-based estimation of depression through symptom prediction. *Brain Informatics*, 10(1):4, 2023.

[34] Alice Othmani and Muhammad Muzammel. An ambient intelligence-based approach for longitudinal monitoring of verbal and vocal depression symptoms. In *International Workshop on PRedictive Intelligence In MEdicine*, pages 206–217. Springer, 2023.

[35] Alice Othmani, Assaad-Oussama Zeghina, and Muhammad Muzammel. A model of normality inspired deep learning framework for depression relapse prediction using audiovisual data. *Computer Methods and Programs in Biomedicine*, 226:107132, 2022.

[36] Mois'es R Pacheco-Lorenzo, Sonia M Valladares-Rodr'ıguez, Luis E Anido-Rif'on, and Manuel J Fern'andez-Iglesias. Smart conversational agents for the detection of neuropsychiatric disorders: A systematic review. *Journal of biomedical informatics*, 113:103632, 2021.

[37] Dabin Park, Semin Lim, Yurim Choi, and Hayoung Oh. Depression emotion multi-label classification using everytime platform with dsm-5 diagnostic criteria. *IEEE Access*, 2023.

[38] Diogo Ramalho, Pedro Constantino, Hugo Plácido Da Silva, Miguel Constante, and João Sanches. An augmented teleconsultation platform for depressive disorders. *IEEE Access*, 10:130563–130571, 2022.

[39] Emil Rijcken, Uzay Kaymak, Floortje Scheepers, Pablo Mosteiro, Kalliopi Zervanou, and Marco Spruit. Topic modeling for interpretable text classification from ehrs. *Frontiers in big Data*, 5:846930, 2022.

[40] Katarzyna Rojewska, Stella Ma'ckowska, Michał Ma'ckowski, Agnieszka R'o.za'nska, Klaudia Bara'nska, Mariusz Dzieciatko, and Dominik Spinczyk. Natural language processing and machine learning supporting the work of a psychologist and its evaluation on the example of support for psychological diagnosis of anorexia. *Applied Sciences*, 12(9):4702, 2022.

[41] A Saranya and R Anandan. Figs-deaf: an novel implementation of hybrid deep learning algorithm to predict autism spectrum disorders using facial fused gait features. *Distributed and Parallel Databases*, 40(4):753–778, 2022.

[42] Zixuan Shangguan, Zhenyu Liu, Gang Li, Qiongqiong Chen, Zhijie Ding, and Bin Hu. Dual-stream multiple instance learning for depression detection with facial expression videos. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:554–563, 2022.

[43] Yijun Shao, Yan Cheng, Srikanth Gottipati, and Qing Zeng-Treitler. Outcome prediction for patients with bipolar disorder using prodromal and onset data. *Applied Sciences*, 13(3):1552, 2023.

[44] Lingeswari Sivagnanam and NK Visalakshi. Multimodal machine learning framework to detect the bipolar disorder. *Advances in Parallel Computing Algorithms, Tools and Paradigms*, 41:138, 2022.

[45] Siyang Song, Shashank Jaiswal, Linlin Shen, and Michel Valstar. Spectral representation of behaviour primitives for depression analysis. *IEEE Transactions on Affective Computing*, 13(2):829–844, 2020.

[46] Marco Spruit, Stephanie Verkleij, Kees de Schepper, and Floortje Scheepers. Exploring language markers of mental health in

psychiatric stories. *Applied Sciences*, 12(4):2179, 2022.

[47] ML Tlachac, Avantika Shrestha, Mahum Shah, Benjamin Litterer, and Elke A Rundensteiner. Automated construction of lexicons to improve depression screening with text messages. *IEEE Journal of Biomedical and Health Informatics*, 2022.

[48] PM Durai Raj Vincent, Nivedhitha Mahendran, Jamel Nebhen, Natarajan Deepa, Kathiravan Srinivasan, and Yuh-Chung Hu. Performance assessment of certain machine learning models for predicting the major depressive disorder among it professionals during pandemic times. *Computational intelligence and neuroscience*, 2021, 2021.

[49] Hao Xiong, Shlomo Berkovsky, Mia Romano, Roneel V Sharan, Sidong Liu, Enrico Coiera, and Lauren F McLellan. Prediction of anxiety disorders using a feature ensemble based bayesian neural network. *Journal of Biomedical Informatics*, 123:103921, 2021.

[50] Weizhe Xu, Weichen Wang, Jake Portanova, Ayesha Chander, Andrew Campbell, Serguei Pakhomov, Dror Ben-Zeev, and Trevor Cohen. Fully automated detection of formal thought disorder with time-series augmented representations for detection of incoherent speech (tardis). *Journal of biomedical informatics*, 126:103998, 2022.

[51] Faming Yin, Jing Du, Xinzhou Xu, and Li Zhao. Depression detection in speech using transformer and parallel convolutional neural networks. *Electronics*, 12(2):328, 2023.

[52] Junjie Zhang, Guangmin Sun, Kun Zheng, Sarah Mazhar, Xiaohui Fu, Yu Li, and Hui Yu. Ssgnn: A macro and microfacial expression recognition graph neural network combining spatial and spectral domain features. *IEEE Transactions on Human-Machine Systems*, 52(4):747–760, 2022.

[53] Yan Zhao, Zhenlin Liang, Jing Du, Li Zhang, Chengyu Liu, and Li Zhao. Multi-head attention-based long short-term memory for depression detection from speech. *Frontiers in Neurorobotics*, 15:684037, 2021.

[54] Irune Zubiaga and Raquel Justo. Multimodal feature evaluation and fusion for emotional well-being monitorization. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 242–254. Springer, 2022.