Estado de Arte Paper

Albert Montenegro

Contents

Chapter 1

XAI en Marcha

1.1 Estado del Arte en Explainable AI (XAI) para el Análisis de la Marcha

El campo de la Inteligencia Artificial Explicable (XAI) ha demostrado ser fundamental para abordar la opacidad de los modelos de aprendizaje automático aplicados al análisis de la marcha. La capacidad de interpretar los resultados generados por estos modelos facilita su aceptación en contextos clínicos y permite identificar patrones biomecánicos relevantes para diagnósticos y tratamientos. A continuación, se presenta una revisión de los estudios más destacados en esta área.

1.1.1 Explaining the Unique Nature of Individual Gait Patterns with Deep Learning

Este trabajo, publicado por Horst et al. (2019) en *Scientific Reports*, se enfoca en la interpretabilidad de los modelos de aprendizaje profundo aplicados a la marcha. Utilizando la técnica de *Layer-wise Relevance Propagation* (LRP), los autores logran identificar regiones específicas del ciclo de marcha que contribuyen a la clasificación de patrones únicos de individuos. Este enfoque no solo mejora la comprensión de las predicciones, sino que también incrementa la confianza de los clínicos en el uso de redes neuronales convolucionales (CNN) para aplicaciones de rehabilitación y diagnóstico.

1.1.2 Interpretability of Input Representations for Gait Classification in Patients after Total Hip Arthroplasty

Dindorf et al. (2020), en *Sensors*, investigan cómo distintas representaciones de entrada afectan la interpretabilidad de modelos de clasificación de marcha en pacientes con artroplastia total de cadera. Utilizando técnicas como *Local Interpretable Model-Agnostic Explanations* (LIME), el estudio demuestra que las estadísticas descriptivas y las características automáticas generadas con *tsfresh* son clave para identificar movimientos relevantes de cadera y rodilla. Los modelos SVM y MLP lograron desempeños sobresalientes, con precisiones superiores al 97%.

1.1.3 Explainable Machine Learning in Clinical Gait Analysis

Slijepcevic et al. (2021), en *ACM Transactions on Computing for Healthcare*, implementan *Layer-wise Relevance Propagation* (LRP) para analizar patrones de marcha en pacientes con trastornos clínicos. Los modelos evaluados incluyen CNNs, MLPs y SVMs, mostrando que las regiones temporales clave en las fuerzas de reacción al suelo son fundamentales para las predicciones. Este estudio subraya la importancia de validar estadísticamente las explicaciones generadas y su relevancia clínica mediante *Statistical Parametric Mapping* (SPM).

1.1.4 Explainable Gait Recognition with Prototyping Encoder-Decoder

Moon et al. (2022), en *PLOS ONE*, introducen un modelo basado en un *encoder-decoder* para el reconocimiento de marcha en entornos abiertos. Utilizando mapas de atribución generados por Grad-CAM, los autores identifican las señales más relevantes en modalidades como presión y aceleración. Este enfoque destaca por su robustez frente a variaciones en los datos y su capacidad para generalizar en escenarios biométricos y clínicos.

1.1.5 Leveraging Explainable Machine Learning to Identify Gait Biomechanical Parameters Associated with ACL Injury

Kokkotis et al. (2022), en *Scientific Reports*, combinan técnicas de XAI con aprendizaje automático para identificar parámetros biomecánicos asociados a lesiones del ligamento cruzado anterior (ACL). Utilizando SHAP, el estudio demuestra que características como el ángulo de flexión de la rodilla y los momentos articulares de la cadera son esenciales para clasificar pacientes con y sin reconstrucción de ACL. Los modelos SVM y NN alcanzaron precisiones superiores al 92%.

1.1.6 Trustworthy Visual Analytics in Clinical Gait Analysis

Rind et al. (2022), en un taller de la IEEE, desarrollan una herramienta interactiva (gaitXplorer) que combina Grad-CAM con visualizaciones interactivas para analizar patrones de marcha en pacientes con parálisis cerebral. Este enfoque permite a los clínicos interpretar resultados automáticos y comparar datos de pacientes con promedios grupales, incrementando la confianza en los modelos automáticos.

1.1.7 XAI and Wearable Sensor-Based Gait Analysis to Identify Patients with Osteopenia and Sarcopenia in Daily Life

Kim et al. (2022), en *Biosensors*, utilizan XAI para analizar parámetros de marcha en mujeres mayores con osteopenia y sarcopenia. Los modelos ML (e.g., XGBoost, RF) superaron a los modelos DL (e.g., CNN, BiLSTM) en la clasificación, destacando la utilidad de parámetros estadísticos descriptivos. Técnicas como SHAP y Grad-CAM permiten interpretar los resultados y mejorar la confianza en los modelos.

1.1.8 Explainable Machine Learning in Human Gait Analysis: A Study on Children With Cerebral Palsy

Slijepcevic et al. (2023), en *IEEE Access*, investigan patrones de marcha en niños con parálisis cerebral utilizando datos de análisis de marcha tridimensionales (3DGA). Los modelos evaluados incluyen CNNs, SNNs, RF y DT, destacando la importancia de las señales sagitales de tobillo y rodilla para la clasificación. Técnicas como Grad-CAM y la importancia de características permiten explicar las predicciones y orientar la evaluación clínica.

1.1.9 Identification and Interpretation of Gait Analysis Features and Foot Conditions by Explainable AI

Özateş et al. (2024), en *Scientific Reports*, presentan un pipeline de ML para identificar condiciones específicas del pie a partir del análisis de marcha. Los modelos evaluados incluyen SVM, RF y KNN, combinados con técnicas XAI como LIME para identificar las características más relevantes. El estudio demuestra que la combinación de ML y XAI puede automatizar y mejorar la interpretación del análisis de marcha.

1.1.10 A New Method Applied for Explaining the Landing Patterns

Xu et al. (2024), en *Heliyon*, proponen un modelo basado en LRP para explicar patrones de aterrizaje en movimientos biomecánicos. A través de *Statistical Parametric Mapping* (SPM) y análisis del tamaño del efecto, se validan las explicaciones generadas por el modelo. Este enfoque destaca por identificar contribuciones relevantes en las señales de las articulaciones del tobillo y la rodilla, particularmente en el plano sagital.

1.1.11 Conclusión

Los estudios revisados destacan la aplicación de XAI en el análisis de la marcha, permitiendo interpretaciones detalladas y validadas estadísticamente de los modelos de aprendizaje automático. Técnicas como LRP, Grad-CAM y SHAP han demostrado ser fundamentales para mejorar la aceptación clínica y la confianza en los modelos automáticos. Este avance abre nuevas posibilidades para el diagnóstico, rehabilitación y planificación de tratamientos en biomecánica clínica.

Autores	Modelo	XAI	Tarea	Dispositivo	Entrada	Multiclase	Exactitud (%)
Horst et al. (2019)	CNN	LRP	Patrones únicos de marcha individual	Plataforma de fuerza	GRF normalizadas	No	99.4
Dindorf et al. (2020)	SVM, MLP	LIME	Clasificación tras artroplastia total de cadera	IMUs	Características automáticas y estadísticas	No	97.3
Dindorf et al. (2021)	RF, SVM	SHAP	Interpretación de datos de postura espinal	IMUs	Ángulos funcionales	No	95.0
Slijepcevic et al. (2021)	CNN, SVM	LRP	Análisis de marcha clínica en trastornos	GRF	Señales temporales clave	Sí	92.0 / 85.0
Moon et al. (2022)	Autoencoder	Grad-CAM	Reconocimiento de marcha en open set	Plantillas con sensores	Presión, aceleración y rotación	No	90.0
Kim et al. (2022)	XGBoost	SHAP	Identificación de osteopenia y sarcopenia	Sensores portables	Parámetros biomecánicos	No	88.7
Rind et al. (2022)	CNN	Grad-CAM	Patrones de marcha en parálisis cerebral	3DGA	Señales angulares y GRF	Sí	87.0 / 75.0
Kokkotis et al. (2022)	SVM, NN	SHAP	Diagnóstico de lesión de ACL	Plataforma de fuerza	Ángulos y momentos articulares	Sí	94.9 / 89.0
Slijepcevic et al. (2023)	RF, CNN	Grad-CAM	Patrones de marcha en niños con parálisis cerebral	3DGA	Ángulos y GRF	Sí	93.4
Xu et al. (2024)	ANN	LRP	Reconocimiento de patrones de aterrizaje	IMUs	Señales cinéticas y cinemáticas	No	99.5
Özateş et al. (2024)	SVM, RF	LIME	Diagnóstico de condiciones del pie	IMUs	Ángulos funcionales	Sí	87.0

Table 1.1: Resumen del estado del arte en Explainable AI (XAI) aplicado al análisis de la marcha.

1.1.12 Clasificación Multiclase

En el trabajo de **Rind et al. (2022)**, se realiza una tarea de clasificación multiclase. Este artículo aborda el análisis de patrones de marcha en pacientes con parálisis cerebral. Los patrones clasificados incluyen:

- *True Equinus*
- *Jump Gait*
- *Apparent Equinus*
- *Crouch Gait*

El desempeño reportado para la clasificación incluye:

- Exactitud general (*accuracy*) en clasificación binaria: **87%**.
- Precisión en clasificación multiclase: **75%**.

El trabajo de **Slijepcevic et al. (2022)** aborda tanto tareas de clasificación binaria como multiclase en el análisis de patrones de marcha clínica.
El desempeño reportado incluye:

- Clasificación binaria: **92%**.
- Clasificación multiclase: **85%**.

El trabajo de **Kokkotis et al. (2022)** aborda tanto la clasificación binaria como multiclase.
El desempeño reportado incluye:

- **Clasificación binaria (Control vs. ACL):**
  - SVM: Precisión **94.95%**, con métricas de Precision (**96.72%**) y Recall (**97.62%**).
- **Clasificación multiclase (Control, ACLD, ACLR):**
  - SVM: Precisión **89%**.

El trabajo de **Slijepcevic et al. (2023)** explora el análisis de patrones de marcha en niños con parálisis cerebral (CP).
El desempeño reportado incluye:

- **Clasificación multiclase (True Equinus, Crouch Gait, Jump Gait, Apparent Equinus):**
  - Mejor desempeño con Random Forest (RF): Precisión **93.4%**.
  - CNNs y SNNs mostraron menor precisión, siendo más sensibles a la configuración de entrada.

El trabajo de **Erkan et al. (2024)** aborda la identificación automática de condiciones específicas del pie.
El desempeño reportado incluye:

- **Clasificación multiclase (7 clases):**
  - Mejor desempeño con Majority Voting (MV): Precisión balanceada (**Balanced Accuracy**) **87%**.
  - Otros modelos, como KNN, lograron métricas comparables (**Balanced Accuracy** superior a **82%**).

1.2 Explaining the Unique Nature of Individual Gait Patterns with Deep Learning

1.2.1 Detalles del Artículo

- **Título:** Explaining the Unique Nature of Individual Gait Patterns with Deep Learning
- **Autores:** Fabian Horst, Djordje Slijepcevic, Sebastian Lapuschkin, Wiebke Lambrecht, Brian Horsak
- **Revista:** Scientific Reports
- **Nivel:** Publicación revisada por pares, cuartil Q1, índice de impacto 4.379 (2020).
- **Fecha de publicación:** septiembre de 2019.

1.2.2 Problema Principal

El artículo aborda la falta de interpretabilidad en modelos de aprendizaje profundo aplicados al análisis de la marcha. Aunque las redes neuronales profundas pueden capturar patrones únicos de marcha individuales, su naturaleza de "caja negra" dificulta la comprensión y aceptación en contextos clínicos. Este trabajo propone el uso de técnicas de inteligencia artificial explicable (XAI) para desentrañar cómo estos modelos identifican la unicidad de los patrones de marcha.

1.2.3 Arquitectura del Modelo

- **Modelo:**
  - Red Neuronal Convolucional (CNN) de una dimensión.
  - Estructura: Capas convolucionales seguidas de capas densas.
  - Activación: ReLU.
  - Salida: Capa *softmax* para clasificación.
- **Datos:**
  - Conjunto de datos *GaitRec* con mediciones tridimensionales de fuerza de reacción al suelo (GRF).
  - Total: 68 sujetos (38 hombres, 30 mujeres) con varias mediciones individuales.
  - Cada medición: Normalización temporal (ciclo completo de marcha) y amplitud (peso corporal).

1.2.4 Métodos de Explicabilidad

- **Layer-wise Relevance Propagation (LRP):**
  - Identifica las contribuciones de cada entrada a las predicciones.
  - Asigna valores de relevancia a cada punto temporal de las GRF.
- **Aplicación en GRF:**
  - Destaca regiones temporales clave en las predicciones (e.g., fases de contacto inicial y propulsión).
  - Relaciona estas regiones con características biomecánicas relevantes.

1.2.5 Creación y Preprocesamiento de Datos

- **Recolección:**
  - Medición de GRF en 3 direcciones (vertical, anteroposterior y mediolateral).
  - Captura durante un ciclo de marcha completo.
- **Preprocesamiento:**
  - Normalización temporal (100 puntos de muestra por ciclo).
  - Escalado por peso corporal para uniformidad entre sujetos.

1.2.6 Resultados

- **Precisión del Modelo:**
  - Precisión promedio de clasificación: **99.4%** en datos individuales.
  - Precisión general en clasificación de sujetos: **97.6%**.
- **Explicaciones con LRP:**
  - Identificación de patrones únicos en fases específicas del ciclo de marcha.
  - Confirmación por expertos clínicos de la coherencia biomecánica.
- **Impacto:**
  - Reducción de la "caja negra" en modelos de aprendizaje profundo.
  - Incremento de confianza en el uso clínico de modelos automáticos.

1.2.7 Conclusiones Clave

- **Relevancia:** Las técnicas de XAI permiten explicar la unicidad de patrones individuales de marcha.
- **Contribuciones:** Este trabajo demuestra que las redes profundas pueden combinar precisión con interpretabilidad.
- **Implicaciones:** Facilita la aceptación de modelos en aplicaciones médicas, particularmente en rehabilitación y diagnóstico de trastornos de marcha.

1.3 Interpretability of Input Representations for Gait Classification in Patients after Total Hip Arthroplasty

1.3.1 Detalles del Artículo

- **Título:** Interpretability of Input Representations for Gait Classification in Patients after Total Hip Arthroplasty
- **Autores:** Carlo Dindorf, Wolfgang Teufl, Bertram Taetz, Gabriele Bleser, Michael Fröhlich
- **Revista:** Sensors
- **Nivel:** Revista revisada por pares, cuartil Q1, índice de impacto 3.576 (2020).
- **Fecha de publicación:** agosto de 2020.

1.3.2 Problema Principal

El artículo aborda la necesidad de interpretar modelos de clasificación en el análisis de la marcha clínica, especialmente en pacientes después de una artroplastia total de cadera (THA). Se investiga cómo las distintas representaciones de entrada afectan la precisión del modelo y su interpretabilidad.

1.3.3 Arquitectura del Modelo

Se evaluaron múltiples enfoques:

- **SVM Lineal y SVM RBF:** Clasificación basada en vectores de soporte.
- **Random Forest (RF):** Clasificador robusto con múltiples árboles de decisión.
- **Multilayer Perceptron (MLP):** Red neuronal con dos capas ocultas.

1.3.4 Creación y Preprocesamiento de Datos

- **Datos:** 27 sujetos sanos y 20 pacientes con THA, registrados mediante un sistema de unidades de medición inercial (IMU).
- **Preprocesamiento:**
  - División en ciclos de marcha.
  - Normalización temporal (100 pasos) y amplitud.
  - Representaciones de entrada: (1) Olas (V_waves), (2) Estadísticas descriptivas simples (V_simple), y (3) Características automáticas extraídas con *tsfresh* (V_tsfresh).

1.3.5 Resultados

- **Mejor desempeño:**
  - **V_tsfresh**: Precisión promedio (**Acc**) de **100%** con SVM Lineal y MLP.
  - **V_simple**: **Acc** de **97.38%**.
  - **V_waves**: **Acc** de **95.88%** con SVM Lineal sin normalización.
- **Interpretación con LIME:**
  - Identificación de características relevantes como rotación del tobillo y movimientos sagitales de cadera y rodilla.
  - Se confirmaron diferencias significativas entre pacientes y sujetos sanos mediante Statistical Parametric Mapping (SPM).

1.3.6 Conclusiones

El artículo demuestra que las representaciones de entrada afectan tanto la precisión del modelo como su interpretabilidad. La combinación de estadísticas simples y datos de forma de onda proporciona una buena interpretabilidad para aplicaciones clínicas. Además, los métodos XAI como LIME ayudan a identificar patrones de marcha individuales relevantes para la clasificación.

1.4 Classification and Automated Interpretation of Spinal Posture Data Using XAI

1.4.1 Detalles del Artículo

- **Título:** Classification and Automated Interpretation of Spinal Posture Data Using a Pathology-Independent Classifier and Explainable Artificial Intelligence (XAI)
- **Autores:** Carlo Dindorf, Jürgen Konradi, Claudia Wolf, Bertram Taetz, Gabriele Bleser, Janine Huthwelker, Philipp Drees, Michael Fröhlich, et al.
- **Revista:** Sensors
- **Nivel:** Publicación revisada por pares, cuartil Q2 en ciencias aplicadas.
- **Fecha de publicación:** septiembre de 2021.

1.4.2 Problema Principal

El artículo aborda la falta de modelos patológicamente independientes para clasificar datos de postura espinal. Los modelos existentes suelen depender de datos específicos de patologías, lo que limita su generalización. Además, la opacidad de los modelos de aprendizaje automático plantea desafíos en aplicaciones clínicas debido a la necesidad de interpretabilidad y cumplimiento con normativas como GDPR.

1.4.3 Arquitectura del Modelo

- **Clasificador Patológicamente Independiente:** Basado en una Máquina de Vectores de Soporte de Clase Única (OCSVM), que aprende características de sujetos sanos y detecta desviaciones patológicas como anomalías.
- **Transformación de Salidas:** La salida del modelo se ajusta a una distribución probabilística utilizando el método de Platt.
- **Comparación:** Se comparan los resultados con un clasificador binario de Bosque Aleatorio (RF) para evaluar el desempeño relativo.

1.4.4 Métodos de XAI Utilizados

- **Local Interpretable Model-Agnostic Explanations (LIME):**
  - Genera explicaciones locales aproximando predicciones de modelos complejos con modelos simples.
  - Identifica características relevantes específicas de cada sujeto.
- **Aplicación Práctica:**
  - Las explicaciones permiten identificar diferencias patológicas clave, como rotaciones o inclinaciones vertebrales, asociadas con condiciones específicas como fusiones espinales.
  - Los valores de LIME se correlacionan con características anatómicas y biomecánicas relevantes, facilitando la adaptación de terapias personalizadas.

1.4.5 Evaluación de Desempeño

- **Fusión Espinal:** Mejor desempeño, con $F1 = 0.80 \pm 0.12$ y $MCC = 0.57 \pm 0.23$ utilizando OCSVM.
- **Osteoartritis:** Desempeño moderado con $F1 = 0.69 \pm 0.04$.
- **Dolor de Espalda:** Peor desempeño, con $F1 = 0.54 \pm 0.13$, debido a las pequeñas diferencias estáticas en pacientes con esta condición.
- **Datos Sintéticos:** La separación de clases afecta significativamente el desempeño, mostrando mejores resultados con mayores separaciones.

1.4.6 Conclusiones Clave

- **Relevancia Clínica:** La combinación de OCSVM y XAI permite identificar características específicas de sujetos individuales, mejorando la personalización de tratamientos.
- **Limitaciones:** Los datos estáticos pueden no capturar adecuadamente las diferencias dinámicas; se recomienda explorar datos dinámicos en futuros estudios.
- **Implicaciones para la Medicina Personalizada:** Las explicaciones generadas ofrecen una herramienta objetiva para monitorear y ajustar terapias pre y postoperatorias.

1.5 Explaining Machine Learning Models for Clinical Gait Analysis

1.5.1 Detalles del Artículo

- **Título:** Explaining Machine Learning Models for Clinical Gait Analysis
- **Autores:** Djordje Slijepcevic, Fabian Horst, Sebastian Lapuschkin, et al.
- **Revista:** ACM Transactions on Computing for Healthcare
- **Nivel:** Publicación en una revista revisada por pares, cuartil Q2 en computación aplicada a la salud.
- **Fecha de publicación:** diciembre de 2021.

1.5.2 Problema Principal

El artículo aborda la opacidad (carácter de caja negra) de los modelos de aprendizaje automático en el análisis clínico de la marcha. Propone el uso de métodos de inteligencia artificial explicable (XAI) para analizar la relevancia de las características aprendidas en tareas de clasificación de patrones de marcha, incrementando la transparencia y aceptabilidad de los modelos en aplicaciones médicas.

1.5.3 Arquitectura del Modelo

Se emplearon tres enfoques de aprendizaje automático:

- **Red Neuronal Convolucional (CNN):** Procesa señales de reacción al suelo (GRF) con capas convolucionales y una capa densa final para predicción multiclase.
- **Perceptrón Multicapa (MLP):** Red completamente conectada con dos capas ocultas y activación ReLU.
- **Máquina de Soporte Vectorial (SVM):** Clasificador lineal optimizado con regularización L^2.

1.5.4 Función de Explicabilidad

- Se utilizó la técnica **Layer-wise Relevance Propagation (LRP)** para determinar la relevancia de regiones específicas de las señales GRF en la predicción del modelo.
- Las explicaciones destacan qué regiones temporales y características de la señal influyen en la clasificación, vinculándolas a características clínicas relevantes.

1.5.5 Creación y Preprocesamiento de Datos

- Conjunto de datos **GaitRec:** Incluye mediciones tridimensionales de GRF de 132 pacientes con trastornos de marcha y 62 controles sanos.
- Las señales se normalizaron temporalmente (fase de soporte al 100%) y por amplitud (peso corporal al 100%).

1.5.6 Casos de Estudio y Métricas

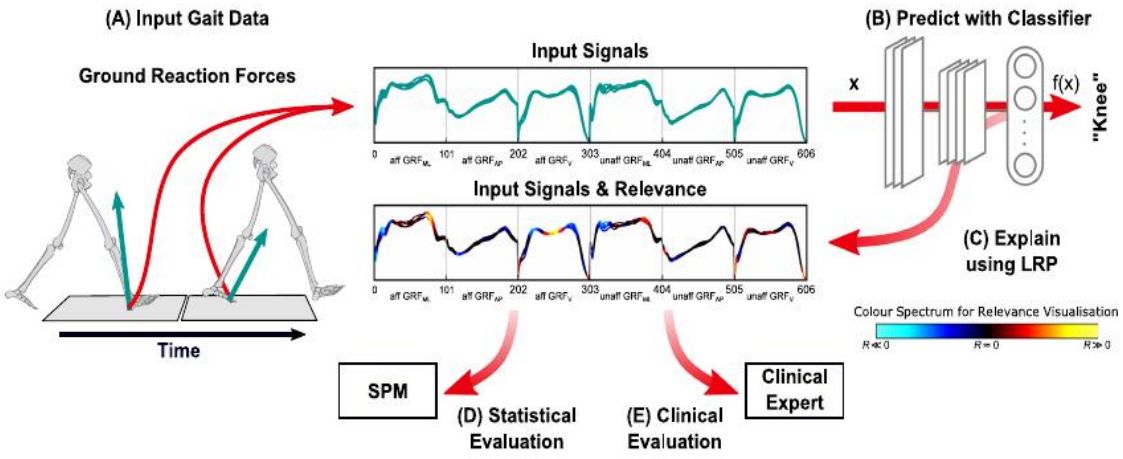
- **Clasificación Binaria (Controles vs. Trastornos):**
  - Mejor desempeño: **SVM** con datos no normalizados.
  - **Accuracy:** 89.1% ± 5.9.
  - **AUC-ROC:** Reportado como alto en configuraciones similares.
  - Regiones relevantes identificadas en señales GRF verticales y horizontales.
- **Clasificación Multiclase:**
  - Mejor desempeño: **SVM** con datos min-max normalizados.
  - **Accuracy:** 59.5% ± 8.5.
  - **F1-Score:** No reportado explícitamente para esta tarea.
  - Las explicaciones mostraron solapamientos entre características relevantes de diferentes trastornos.

1.5.7 Evaluación Estadística y Clínica

- **Estadística:** Se utilizó Statistical Parametric Mapping (SPM) para validar la relevancia estadística de las regiones señaladas por LRP.
- **Clínica:** Expertos clínicos confirmaron la coherencia de las regiones relevantes con anomalías biomecánicas conocidas.

1.5.8 Méritos Adicionales

- **Interpretabilidad clínica:** LRP permitió vincular predicciones del modelo a características biomecánicas específicas.
- **Generalización:** Los resultados sugieren que los métodos XAI pueden facilitar la integración de modelos en prácticas clínicas al aumentar la confianza y transparencia.



1.6 Explainable gait recognition with prototyping encoder-decoder

1.6.1 Detalles del Artículo

- **Título:** Explainable gait recognition with prototyping encoder-decoder
- **Autores:** Jucheol Moon, Yong-Min Shin, Jin-Duk Park, Nelson Hebert Minaya, Won-Yong Shin, Sang-Il Choi
- **Revista:** PLOS ONE
- **Nivel:** Multidisciplinaria, Cuartil Q1, índice de impacto 3.752 (2022-2023).
- **Fecha de publicación:** marzo de 2022.

1.6.2 Problema Principal

El artículo aborda el reconocimiento de marcha en un entorno de *open set*, un desafío donde se busca identificar a individuos conocidos y rechazar a desconocidos. Este enfoque es útil en aplicaciones biométricas y médicas. Además, el trabajo resuelve problemas de sensibilidad a hiperparámetros y falta de interpretabilidad de modelos tradicionales mediante una arquitectura innovadora y herramientas de inteligencia artificial explicable (XAI).

1.6.3 Creación de Datos para el Modelo

Los datos se recolectan usando plantillas de calzado equipadas con sensores (presión, aceleración y rotación). Cada muestra es segmentada en pasos unitarios representativos, con suavizado Gaussiano aplicado para eliminar errores.

1.6.4 Cálculo de Prototipos

Los prototipos para cada sujeto se calculan promediando los vectores embebidos correspondientes a los pasos unitarios en cada modalidad (presión, aceleración y rotación). La fórmula general es:

$$c_m^a = \frac{1}{q} \sum_{i=1}^{q} s_i^m$$

donde c_m^a es el prototipo de la modalidad m para el sujeto a, y q es el número total de pasos.

1.6.5 Arquitectura del Modelo y Flujo de Datos

El sistema propuesto es un prototipo de red *encoder-decoder* con las siguientes características:

- **Encoder:** Transforma los pasos unitarios en vectores embebidos de 128 dimensiones, empleando tres sub-*encoders* con capas convolucionales 1D.
- **Decoder:** Reconstruye los datos originales desde los vectores embebidos. En este modelo, el decoder intenta aproximar el prototipo promedio de las muestras de un sujeto (en cada modalidad: presión, aceleración o rotación) desde el vector embebido generado por el encoder.
- **Funciones de pérdida:**
  - **Pérdida de triplete:** Aumenta la separación entre clases distintas y reduce la dispersión dentro de una misma clase. Su expresión es:

    $$L_{\text{triplet}} = \|\mathbf{v}_{i,a} - \mathbf{v}_{j,a}\|_2^2 - \|\mathbf{v}_{i,a} - \mathbf{v}_{k,b}\|_2^2 + \alpha$$

    donde:
    - $\mathbf{v}_{i,a}$ y $\mathbf{v}_{j,a}$: vectores embebidos de pasos unitarios del mismo sujeto a,
    - $\mathbf{v}_{k,b}$: vector embebido de un paso unitario de un sujeto distinto b,
    - α: margen que define la separación mínima deseada.
  - **Pérdida de prototipo:** Reduce la distancia entre los datos originales y sus prototipos promedio. Su expresión es:

    $$L_{\text{proto}} = \frac{1}{|M|} \sum_{m \in M} \|g(f(s_{i,a}^m)) - c_a^m\|_2^2$$

    donde:
    - $g(f(s_{i,a}^m))$: reconstrucción del prototipo por el *decoder*,
    - c_a^m: prototipo promedio de la modalidad m para el sujeto a,
    - M: conjunto de modalidades (presión, aceleración, rotación).

1.6.6 Interpretabilidad del Modelo (XAI)

El modelo utiliza herramientas de XAI para producir explicaciones visuales y cuantitativas:

- **Mapas de atribución (heatmaps):**
  - Representan la relevancia de cada característica de entrada (presión, aceleración, rotación) en el tiempo.
  - Se visualizan como mapas de calor donde:
    - El eje horizontal representa las características.
    - El eje vertical representa el tiempo.
    - Los colores más intensos indican mayor relevancia.
- **Pruebas de degradación del desempeño:**
  - Ocultan áreas identificadas como relevantes por los mapas de atribución.
  - Evalúan cómo disminuye el desempeño del modelo (medido por precisión, TPR y TNR) al eliminar esas áreas.

Métodos empleados:

- **Análisis de Sensibilidad (SA):** Mide cómo pequeños cambios en el input afectan la salida.
- **Propagación de Relevancia por Capas (LRP):** Redistribuye la relevancia de la salida hacia las entradas para identificar características críticas.

Tipo de explicaciones:

- **Gráficas:** Mapas de calor y degradación del desempeño.
- **Cuantitativas:** Valores numéricos que indican sensibilidad y relevancia.

1.6.7 Uso Práctico del Modelo

El flujo operativo para reconocimiento incluye:

1. **Codificación:** Un paso unitario es mapeado al espacio latente mediante el *encoder*.
2. **Comparación:** El vector embebido es comparado con prototipos conocidos usando una máquina de soporte vectorial ($OSVM$).
3. **Decisión:** Si el vector cae dentro del límite de decisión, se clasifica como una clase conocida; de lo contrario, se rechaza.

1.6.8 Resultados y Evaluación

- **Desempeño:** El sistema logra métricas de precisión (ACC), tasa de verdaderos positivos (TPR) y tasa de verdaderos negativos (TNR) superiores al 90%.
- **Robustez:** La combinación de pérdidas reduce la dependencia del modelo a los hiperparámetros.
- **Explicabilidad:** Los mapas de atribución revelan qué partes de los datos (como fases de contacto con el suelo) son más relevantes para la clasificación.

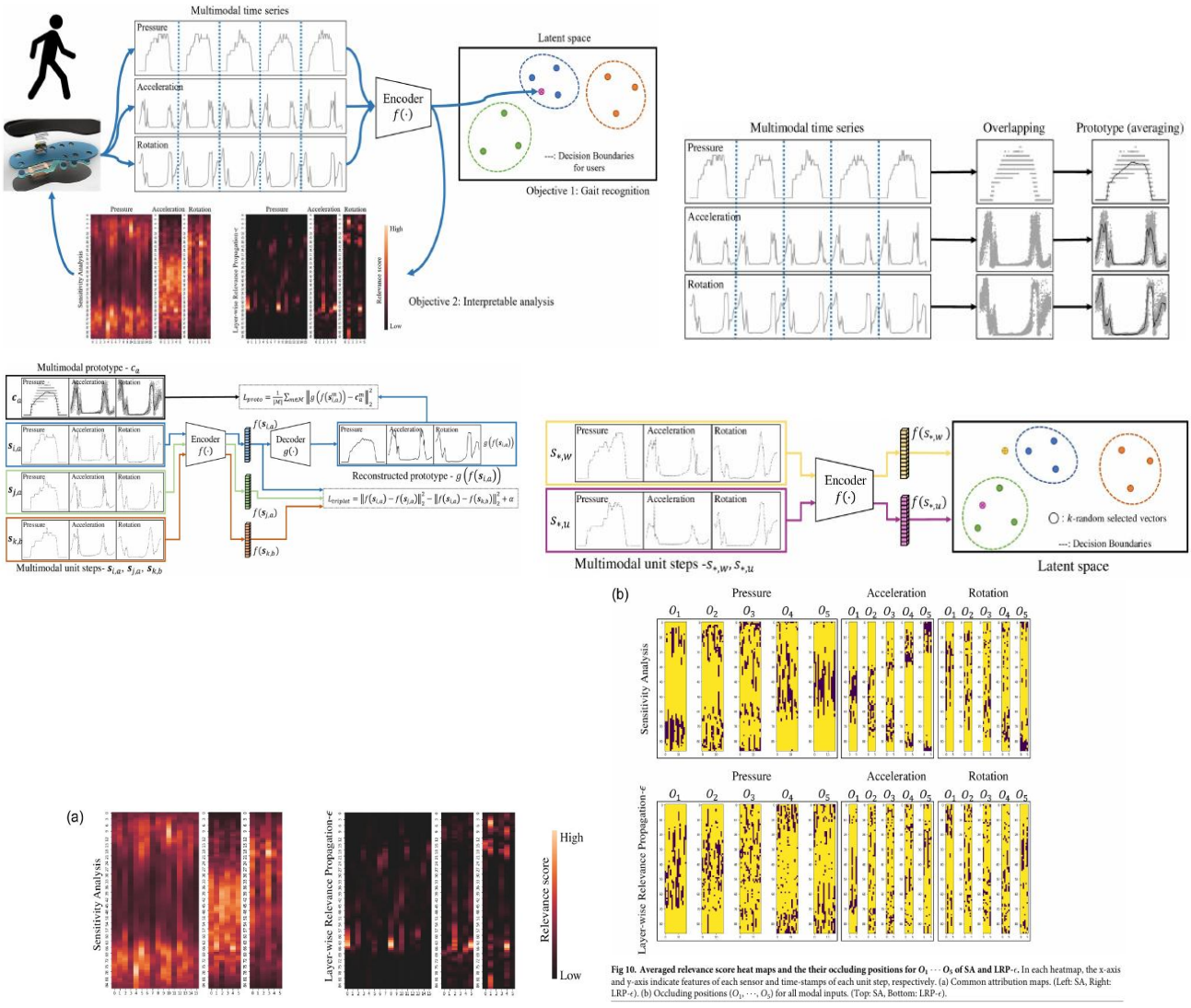


Fig 10. Averaged relevance score heat maps and the their occluding positions for $O_1 \cdots O_5$ of SA and LRP-ϵ. In each heatmap, the x-axis and y-axis indicate features of each sensor and time-stamps of each unit step, respectively. (a) Common attribution maps. (Left: SA, Right: LRP-ϵ). (b) Occluding positions ($O_1, \cdots, O_5$) for all modal inputs. (Top: SA, Bottom: LRP-ϵ).

1.7 XAI and Wearable Sensor-Based Gait Analysis to Identify Patients with Osteopenia and Sarcopenia in Daily Life

1.7.1 Detalles del Artículo

- **Título:** XAI and Wearable Sensor-Based Gait Analysis to Identify Patients with Osteopenia and Sarcopenia in Daily Life
- **Autores:** Jeong-Kyun Kim, Myung-Nam Bae, Kangbok Lee, Jae-Chul Kim, Sang Gi Hong
- **Revista:** Biosensors
- **Nivel:** Revista revisada por pares, cuartil Q2 en bioingeniería y tecnologías de sensores.
- **Fecha de publicación:** marzo de 2022.

1.7.2 Problema Principal

El artículo busca identificar pacientes con osteopenia y sarcopenia utilizando análisis de marcha basado en sensores inerciales. Este enfoque permite evaluar el riesgo de estas condiciones en la vida diaria sin necesidad de equipamiento médico especializado. Además, el estudio utiliza inteligencia artificial explicable (XAI) para analizar la relevancia de los parámetros de marcha en las decisiones del modelo.

1.7.3 Arquitectura del Modelo

Se emplearon múltiples modelos de aprendizaje automático y aprendizaje profundo:

- **Modelos ML:**
  - Random Forest (RF)
  - Extreme Gradient Boosting (XGBoost)
  - Máquina de Soporte Vectorial (SVM)
- **Modelos DL:**
  - Convolutional Neural Network (CNN)
  - Bi-Directional Long Short-Term Memory (BiLSTM)
  - ResNet50 con transferencia de aprendizaje
- **Explicabilidad:** Uso de SHAP, Grad-CAM y Relevance-CAM para interpretar la contribución de los parámetros al modelo.

1.7.4 Creación y Preprocesamiento de Datos

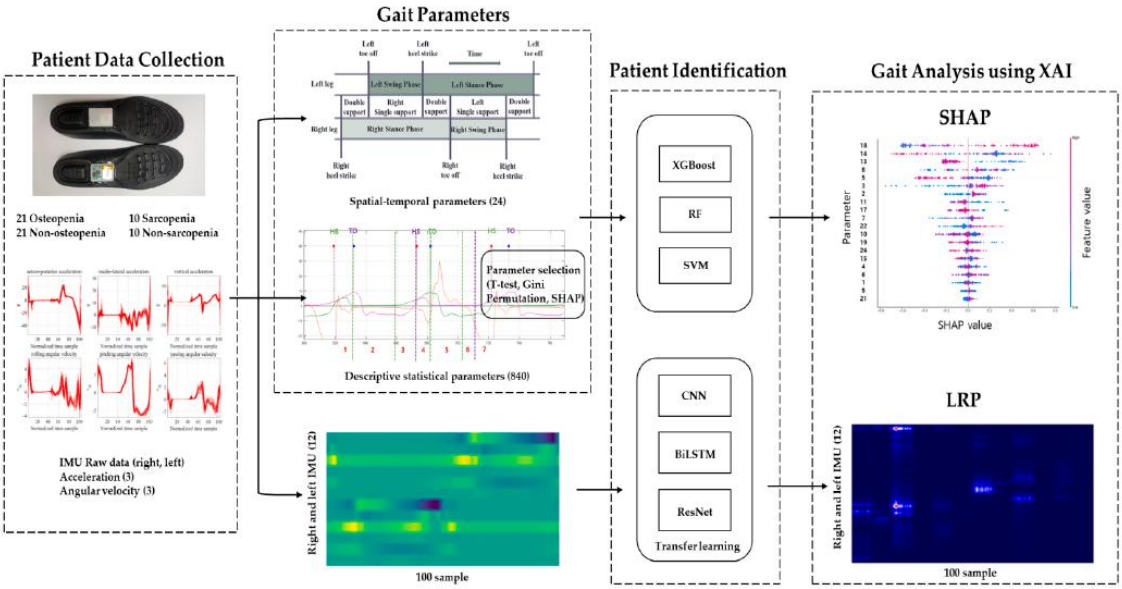
- **Participantes:** 42 mujeres mayores de 65 años:
  - 21 pacientes con osteopenia y 21 sin osteopenia.
  - 10 pacientes con sarcopenia y 10 sin sarcopenia.
- **Datos:** Señales inerciales recolectadas mediante plantillas con sensores IMU.
- **Parámetros:**
  - 24 parámetros espacio-temporales (e.g., fase de soporte, fase de balanceo, cadencia).
  - 100 parámetros estadísticos descriptivos (e.g., promedio, desviación estándar, curtosis).

1.7.5 Resultados y Métricas

- **Osteopenia:**
  - **XGBoost:** Precisión de 88.69% con los 4 parámetros más relevantes.
  - **ResNet (transfer learning):** Precisión de 78.6%, F1-score de 78.7%.
- **Sarcopenia:**
  - **RF:** Precisión de 93.75% con los 18 parámetros más relevantes.
  - **ResNet (transfer learning):** Precisión de 70%, F1-score de 60.6%.

1.7.6 Conclusiones Clave

- Los modelos ML superaron a los modelos DL en la clasificación de osteopenia y sarcopenia.
- Los parámetros estadísticos descriptivos fueron más informativos que los parámetros espacio-temporales.
- XAI proporcionó explicaciones detalladas de cómo los parámetros contribuyen al diagnóstico, mejorando la confiabilidad del modelo.



1.8 Leveraging Explainable Machine Learning to Identify Gait Biomechanical Parameters Associated with ACL Injury

1.8.1 Detalles del Artículo

- **Título:** Leveraging Explainable Machine Learning to Identify Gait Biomechanical Parameters Associated with Anterior Cruciate Ligament Injury
- **Autores:** Christos Kokkotis, Serafeim Moustakidis, Themistoklis Tsatalas, et al.
- **Revista:** Scientific Reports
- **Nivel:** Publicación revisada por pares, cuartil Q1 en multidisciplinario.
- **Fecha de publicación:** abril de 2022.

1.8.2 Problema Principal

El artículo aborda la identificación de parámetros biomecánicos relevantes para diagnosticar lesiones del ligamento cruzado anterior (ACL) y evaluar la efectividad de la reconstrucción (ACLR). Se utiliza aprendizaje automático explicable (XAI) para superar la opacidad de los modelos tradicionales y vincular parámetros biomecánicos a resultados clínicos.

1.8.3 Arquitectura del Modelo

Se diseñó un pipeline de aprendizaje automático con los siguientes componentes:

- **Selección de características:** ReliefF para reducir la dimensionalidad del espacio de características.
- **Modelos ML evaluados:**
  - **SVM:** Mejor desempeño con **Accuracy** de 94.95%.
  - **Red Neuronal (NN):** Segundo mejor desempeño con **Accuracy** de 92.89%.
  - Otros modelos evaluados: XGBoost, Random Forest, Decision Trees, Logistic Regression, Naive Bayes, KNN.
- **Explicabilidad:** SHAP para analizar el impacto de cada característica en la salida del modelo.

1.8.4 Creación y Preprocesamiento de Datos

- **Participantes:** 151 sujetos divididos en tres grupos:
  - CON: Controles sanos.
  - ACLD: Pacientes con ACL lesionado.
  - ACLR: Pacientes después de reconstrucción de ACL.
- **Datos recolectados:** Fuerzas de reacción al suelo (GRF), cinemática y cinética del plano sagital.
- **Normalización:** Escalado a [0,1] y uso de técnicas de filtrado de datos.

1.8.5 Resultados y Métricas

- **Clasificación Multiclase (CON, ACLD, ACLR):**
  - SVM: Mejor desempeño global con **Accuracy** de 94.95%, Precision de hasta 96.72% y Recall de hasta 97.62%.
  - NN: Accuracy de 92.89%.
- **Principales parámetros identificados (SHAP):** K2, H4, A3, GRF4, GRF7, K1, A4.
- **Estadística:** Diferencias significativas en parámetros como K2, H4 y GRF4, confirmadas mediante ANOVA.

1.8.6 Conclusiones Clave

- Los modelos XAI permiten identificar parámetros biomecánicos que los análisis estadísticos tradicionales podrían ignorar.
- Parámetros como el ángulo mínimo de flexión de rodilla (K2) y el momento máximo de flexión de cadera (H4) están vinculados a resultados clínicos.
- Se destaca la necesidad de herramientas más robustas para interpretar interacciones complejas entre parámetros.

1.9 Trustworthy Visual Analytics in Clinical Gait Analysis

1.9.1 Detalles del Artículo

- **Título:** Trustworthy Visual Analytics in Clinical Gait Analysis: A Case Study for Patients with Cerebral Palsy
- **Autores:** Alexander Rind, Djordje Slijepčević, Matthias Zeppelzauer, Fabian Unglaube, Andreas Kranzl, Brian Horsak
- **Revista:** IEEE Workshop on Trust and Expertise in Visual Analytics (TREX)
- **Nivel:** Publicación revisada por pares en un taller de la IEEE.
- **Fecha de publicación:** 19 de diciembre de 2022.

1.9.2 Problema Principal

El análisis clínico de la marcha en pacientes con parálisis cerebral (PC) produce grandes cantidades de datos complejos. Los modelos de aprendizaje automático han demostrado ser útiles para clasificar patrones de marcha, pero su naturaleza de caja negra genera desconfianza entre los clínicos. El artículo propone una solución de analítica visual enriquecida con explicabilidad (gaitXplorer) para aumentar la transparencia y confianza en los modelos.

1.9.3 Arquitectura del Modelo

- **Datos:**
  - Datos tridimensionales de análisis de marcha de 257 niños (355 extremidades afectadas) con PC.
  - Variables como ángulos articulares, momentos articulares y fuerzas de reacción del suelo (GRF).
- **Modelo:**
  - Red Neuronal Convolucional (CNN) de una dimensión.
  - Estructura: 4 capas convolucionales seguidas de capas totalmente conectadas y una capa *softmax* con cuatro clases de salida (true equinus, jump gait, apparent equinus y crouch gait).
  - Función de activación: SELU, con regularización por *alpha dropout*.
- **Explicabilidad:**
  - Algoritmo Grad-CAM adaptado para datos unidimensionales.
  - Identifica regiones relevantes de las series temporales que contribuyen a las predicciones.

1.9.4 Interfaz Visual Interactiva

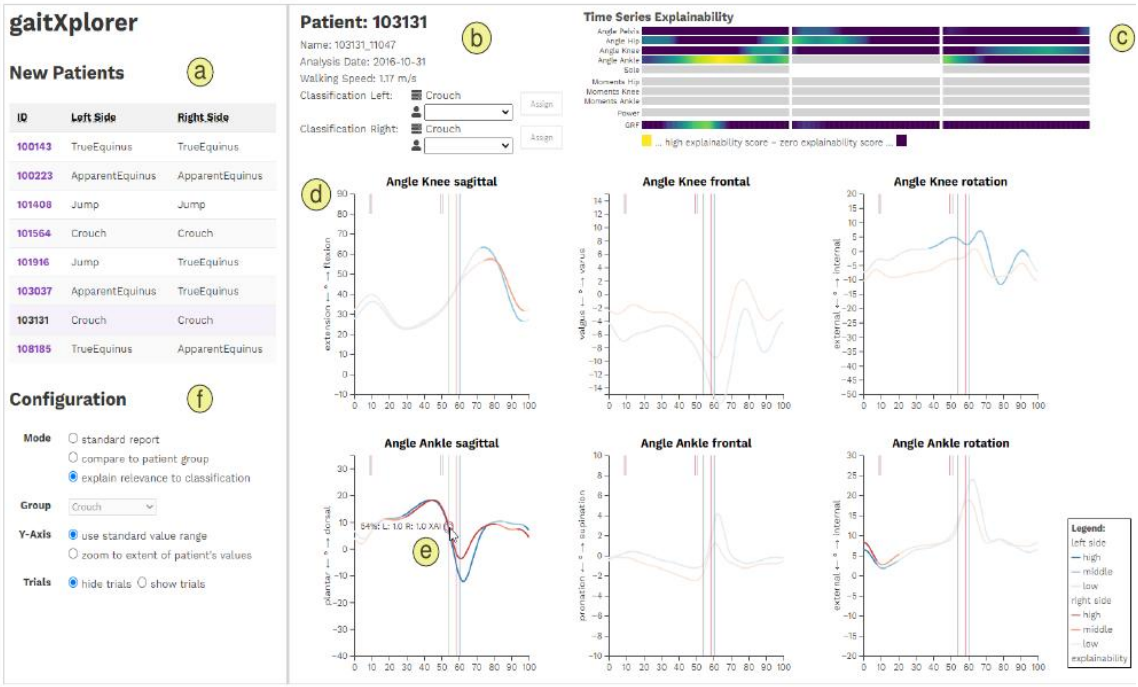
- Interfaz basada en web que permite:
  - Visualizar las series temporales de un paciente y su relevancia en las predicciones.
  - Comparar datos de un paciente con grupos promedio.
  - Modificar clasificaciones automáticas.
- Los datos se presentan en gráficos de líneas organizados por variables biomecánicas y partes del cuerpo.
- Tres modos de visualización:
  - Modo estándar: Muestra datos promedio.
  - Modo de explicabilidad: Resalta regiones relevantes usando Grad-CAM.
  - Modo de comparación grupal: Compara al paciente con promedios de grupos clasificados.

1.9.5 Evaluación de las Explicaciones

- Se realizó un estudio de caso con dos expertos clínicos que evaluaron ocho pacientes.
- Feedback:
  - Se confiaron en explicaciones para cuatro extremidades, mientras que otras fueron consideradas sospechosas por falta de relevancia en áreas esperadas.
  - Ejemplo: Para true equinus, los expertos esperaban alta relevancia en el ángulo sagital del tobillo, pero el modelo destacó el ángulo sagital de la rodilla.
- Las explicaciones fueron calificadas como esenciales para ganar confianza en las clasificaciones automáticas.

1.9.6 Conclusiones Clave

- La integración de Grad-CAM y visualizaciones interactivas mejora la transparencia de los modelos de aprendizaje automático.
- Se identificaron discrepancias entre las expectativas clínicas y las explicaciones generadas, subrayando la necesidad de mejorar los algoritmos explicables.
- El enfoque propuesto tiene potencial para generalizarse a otros contextos clínicos y sistemas de análisis visual.



1.10 Explainable Machine Learning in Human Gait Analysis: A Study on Children With Cerebral Palsy

1.10.1 Detalles del Artículo

- **Título:** Explainable Machine Learning in Human Gait Analysis: A Study on Children With Cerebral Palsy
- **Autores:** Djordje Slijepcevic, Matthias Zeppelzauer, Fabian Unglaube, Andreas Kranzl, Christian Breiteneder, Brian Horsak
- **Revista:** IEEE Access
- **Nivel:** Revista revisada por pares, cuartil Q1 en ingeniería y ciencias de la computación.
- **Fecha de publicación:** junio de 2023.

1.10.2 Problema Principal

El estudio aborda la clasificación de patrones de marcha asociados con la parálisis cerebral (CP) mediante datos de análisis de marcha 3D clínicos (3DGA). Además, utiliza métodos de explicabilidad (XAI) para evaluar la relevancia clínica de las características aprendidas por los modelos, enfrentando los desafíos de opacidad y aceptación clínica de los modelos de aprendizaje automático.

1.10.3 Arquitectura del Modelo

Se evaluaron cuatro tipos de modelos:

- **Modelos de aprendizaje profundo:**
  - Redes Neuronales Convolucionales (CNNs)
  - Redes Neuronales Auto-normalizantes (SNNs)
- **Modelos tradicionales:**
  - Random Forest (RF)
  - Decision Trees (DT)

1.10.4 Datos y Preprocesamiento

- **Conjunto de datos:** 302 pacientes (375 extremidades afectadas) categorizados en cuatro patrones de marcha (true equinus, crouch gait, jump gait, apparent equinus).
- **Captura de datos:** Señales cinemáticas y de fuerzas de reacción al suelo (GRF), normalizadas a un ciclo de marcha completo.
- **Configuraciones de entrada:**
  - Todas las señales de 3DGA.
  - Solo señales cinemáticas.
  - Ángulos sagitales de rodilla y tobillo.
  - Solo señales de GRF.

1.10.5 Métodos XAI y Aplicaciones Prácticas

- **Métodos utilizados:**
  - **Grad-CAM:**
    - Utilizado con CNNs para generar mapas de saliencia que destacan las regiones de las señales más relevantes para la clasificación.
    - Los mapas ayudan a identificar cuáles partes del ciclo de marcha (e.g., fase de soporte, oscilación) son clave para las predicciones del modelo.
  - **Importancia de características (Gini Impurity):**
    - Aplicado en modelos RF y DT para medir la relevancia de cada característica (e.g., ángulos sagitales, GRF).
    - Los valores más altos indican las características más críticas para el modelo.
- **Uso práctico:**
  - Grad-CAM permitió a los clínicos visualizar directamente qué fases y señales específicas de la marcha influyeron en las predicciones del modelo.
  - La importancia de características permitió identificar parámetros biomecánicos clave como los ángulos sagitales de tobillo y rodilla, orientando la evaluación clínica hacia esas señales.
  - Estas explicaciones respaldan decisiones clínicas, facilitando la comprensión del diagnóstico y la planificación del tratamiento.

1.10.6 Resultados y Métricas

- **Mejor desempeño:** Random Forest con precisión de **93.4%** utilizando ángulos sagitales de rodilla y tobillo.
- **Comparación de modelos:**
  - CNNs y SNNs muestran menor precisión y mayor sensibilidad a configuraciones de entrada.
  - Los modelos tradicionales (RF y DT) se enfocan en características clínicamente relevantes con explicaciones más concisas.
- **Explicabilidad:** Grad-CAM destaca regiones temporales en las señales, mientras que los modelos basados en árboles muestran la jerarquía de importancia de características.

1.10.7 Conclusiones Clave

- Los modelos tradicionales superan a las redes neuronales en precisión y robustez.
- Los métodos XAI destacan la relevancia de las señales sagitales (rodilla y tobillo) como características clave para la clasificación.
- La integración de explicabilidad promueve la confianza en los modelos y su potencial uso en la práctica clínica.

1.11 A New Method Applied for Explaining the Landing Patterns

1.11.1 Detalles del Artículo

- **Título:** A New Method Applied for Explaining the Landing Patterns: Interpretability Analysis of Machine Learning
- **Autores:** Datao Xu, Huiyu Zhou, Wenjing Quan, Ukadike Chris Ugbolue, Fekete Gusztav, Yaodong Gu
- **Revista:** Heliyon
- **Nivel:** Revista de acceso abierto revisada por pares, cuartil Q1 en ciencias multidisciplinarias.
- **Fecha de publicación:** 9 de febrero de 2024.

1.11.2 Problema Principal

El artículo aborda la falta de interpretabilidad en los modelos de aprendizaje automático utilizados para reconocer patrones de aterrizaje en biomecánica clínica. Proponen un modelo basado en *Layer-wise Relevance Propagation* (LRP) que no solo mejora la transparencia, sino que también identifica las características más relevantes de las señales biomecánicas.

1.11.3 Arquitectura del Modelo

- **Modelos evaluados:** SVM, ANN, CNN y ZeroR.
- **Modelo final:** Red Neuronal Artificial (ANN) con el mejor rendimiento.
- **Datos de entrada:**
  - Señales cinemáticas y cinéticas tridimensionales de las extremidades inferiores durante la fase de aterrizaje.
  - Datos segmentados por articulación (tobillo, rodilla, cadera) y por plano (sagital, frontal, transversal).

1.11.4 Evaluación de las Explicaciones

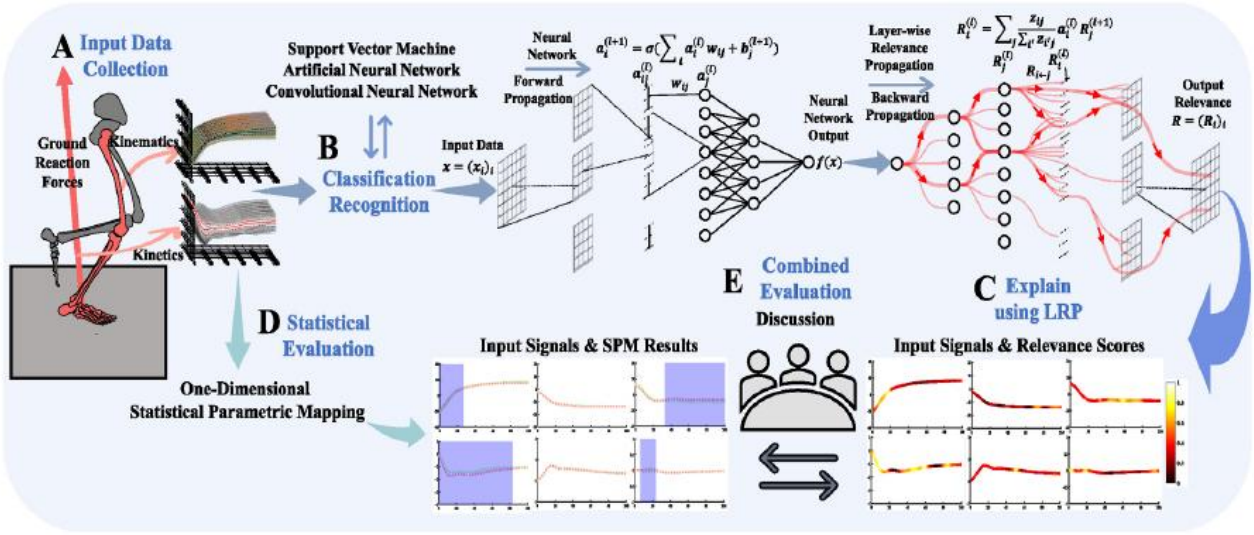
- **Método de explicabilidad:**
  - LRP: Calcula puntajes de relevancia (*Relevance Scores*, RS) para identificar la contribución de cada característica de entrada.
- **Evaluación estadística:**
  - *Statistical Parametric Mapping (SPM):* Verifica diferencias significativas entre patrones utilizando pruebas t pareadas.
  - Tamaño del efecto (*Effect Size*): Mide la magnitud de las diferencias en las señales relevantes.
- **Resultados clave:**
  - Consistencia entre los RS derivados del LRP y los resultados estadísticos en las articulaciones del tobillo y la rodilla.
  - Mayor relevancia en las señales del plano sagital y durante las primeras fases del aterrizaje.

1.11.5 Resultados de Clasificación

- **Desempeño de modelos:**
  - ANN obtuvo una precisión promedio de 99.46% con señales combinadas (cinemática y cinética).
  - Mejor clasificación observada para datos de la rodilla y el plano sagital.

1.11.6 Conclusiones Clave

- El modelo basado en LRP permite una comprensión más profunda de las contribuciones específicas de cada señal al reconocimiento de patrones de aterrizaje.
- El enfoque propuesto puede mejorar significativamente la confianza y transparencia en aplicaciones clínicas.



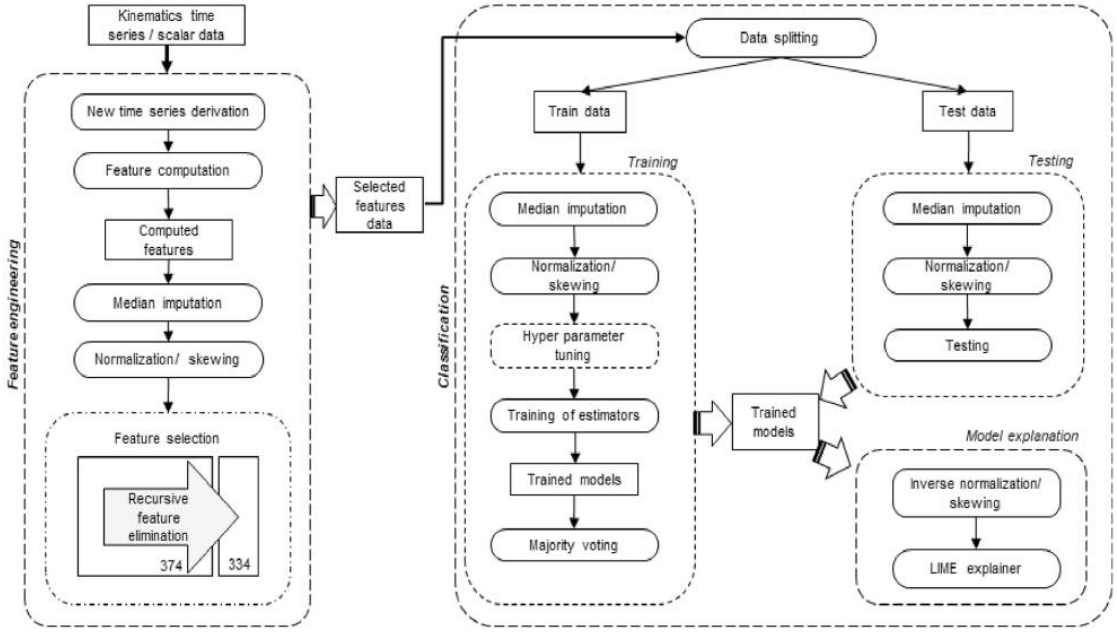
1.12 Identification and Interpretation of Gait Analysis Features and Foot Conditions by Explainable AI

1.12.1 Detalles del Artículo

- **Título:** Identification and Interpretation of Gait Analysis Features and Foot Conditions by Explainable AI
- **Autores:** Mustafa Erkam Özateş, Alper Yaman, Firooz Salami, Sarah Campos, Sebastian I. Wolf, Urs Schneider
- **Revista:** Scientific Reports
- **Nivel:** Revista revisada por pares, cuartil Q1, índice de impacto 4.996 (2023).
- **Fecha de publicación:** marzo de 2024.

1.12.2 Problema Principal

El artículo aborda la complejidad del análisis clínico de la marcha, un proceso crítico para diagnosticar condiciones del pie y planificar cirugías. Propone un pipeline de aprendizaje automático (ML) e inteligencia artificial explicable (XAI) para la selección de características y la identificación automática de condiciones específicas del pie.



1.12.3 Arquitectura del Modelo

Se evaluaron cinco modelos de ML:

- **Support Vector Machines (SVM):** Modelo base para selección de características.
- **Random Forest (RF):** Modelo robusto contra desequilibrio de clases.
- **Logistic Regression (LREG):** Modelo lineal supervisado.
- **K-nearest Neighbors (KNN):** Clasificador basado en distancias.
- **Majority Voting (MV):** Combina los modelos anteriores con votación ponderada.

1.12.4 Creación y Preprocesamiento de Datos

- **Dataset:** 348 sujetos (248 pacientes con 6 condiciones específicas del pie y 100 sujetos control).
- **Ángulos funcionales analizados:** 12 ángulos derivados de un ciclo de marcha (ej. flexión tibiotalar, ángulo del arco medial).
- **Normalización:** Los datos se normalizaron entre 0 y 1, y se imputaron valores faltantes con la mediana.
- **Extracción de características:** Se generaron 374 características, incluyendo derivadas temporales y desviaciones respecto a patrones normales.
- **Selección de características:** Eliminación recursiva de características basada en SVM y RF, reduciendo el conjunto a 334 características.

1.12.5 Resultados

- **Mejores desempeños:**
  - **KNN y MV:** Precisión balanceada (**Balanced Accuracy**), **Recall**, **Precision** y **F1 Score** promedio de 0.87.
  - Otros modelos alcanzaron métricas superiores a 0.82 en promedio.
- **Interpretación con LIME:**
  - Identificación de las cinco características más relevantes por condición del pie.
  - Ejemplo: En *tibiotalar osteoarthritis*, el ángulo mínimo del arco medial fue crítico en las fases de soporte y oscilación.

1.12.6 Conclusiones

El pipeline propuesto demuestra que la combinación de ML y XAI puede automatizar y mejorar la interpretación del análisis de marcha, ofreciendo una herramienta útil para médicos al identificar características relevantes y condiciones específicas del pie.

Chapter 2

Otros

2.1 Assessing Fidelity in XAI Post-hoc Techniques: A Comparative Study with Ground Truth Explanations Datasets

2.1.1 Detalles del Artículo

- **Título:** Assessing Fidelity in XAI Post-hoc Techniques: A Comparative Study with Ground Truth Explanations Datasets
- **Autores:** Miquel Miró-Nicolau, Antoni Jaume-i-Capó, Gabriel Moyà-Alcover
- **Revista:** Preprint, arXiv
- **Nivel:** Investigación preliminar sin revisión por pares.
- **Fecha de publicación:** Noviembre de 2023.

2.1.2 Problema Principal

El artículo aborda la necesidad de evaluar la fidelidad de los métodos de explicabilidad de IA (XAI) post-hoc en comparación con los modelos originales. Dado que no existe una ground truth en explicaciones generadas, se diseñan datasets sintéticos con explicaciones controladas para validar la precisión y consistencia de los métodos XAI.

2.1.3 Definición de Fidelidad

En este artículo, la fidelidad se define como la capacidad de un método XAI para representar de manera precisa las características relevantes que el modelo utiliza para tomar decisiones. En otras palabras, un método es fiel si las explicaciones generadas reflejan de manera confiable los patrones o regiones importantes identificadas por el modelo original durante su proceso de inferencia.

2.1.4 Metodología

- **Conjuntos de datos:**
  - Creación de los datasets TXUXIv1, TXUXIv2, y TXUXIv3 con diferentes niveles de complejidad visual.
  - Las imágenes incluyen ground truth generadas a partir de máscaras controladas, diseñadas para evaluar la correspondencia entre explicaciones y patrones reales.
- **Modelos:** Red Neuronal Convolucional (CNN) preentrenada para tareas de clasificación y regresión.
- **Métricas de evaluación:**
  - **Earth Mover's Distance (EMD):** Mide qué tan diferentes son las distribuciones de saliencia generadas por el método XAI y las "verdades terreno".
  - **Minimum Intersection over Union (MIN):** Evalúa la superposición entre los mapas de saliencia generados y las regiones reales relevantes.

2.1.5 Comparación de Métodos XAI

Se evaluaron 13 métodos de explicabilidad post-hoc, divididos en tres categorías:

- **Métodos de sensibilidad:** LIME, SHAP, RISE.
- **Métodos basados en Class Activation Maps (CAM):** Grad-CAM, Grad-CAM++, Score-CAM, SIDU.
- **Métodos de retropropagación:** Gradient, Guided Backpropagation (GBP), Layer-wise Relevance Propagation (LRP), DeepLIFT, Integrated Gradients, SmoothGrad.

2.1.6 Evaluación de las Explicaciones

- **Resultados en EMD:**
  - Los métodos GBP y LRP obtuvieron los valores más bajos (mejor desempeño), con $EMD \leq 0.035$, indicando alta similitud entre las explicaciones generadas y las "verdades terreno".
  - Métodos como LIME y SHAP presentaron mayores diferencias, lo que sugiere limitaciones en contextos con características visuales complejas.
- **Resultados en MIN:**
  - GBP y LRP alcanzaron los mejores valores, con $MIN \geq 0.202$, demostrando una alta superposición con las regiones relevantes del dataset.
  - Los métodos CAM tuvieron menores puntuaciones debido a mapas de saliencia menos precisos y difusos.
- **Observaciones cualitativas:**
  - Métodos basados en retropropagación generaron mapas más ruidosos pero mejor alineados con las regiones relevantes.
  - Métodos basados en perturbación (LIME, SHAP) mostraron inconsistencias en imágenes con patrones complejos.

2.1.7 Resultados y Métricas

- Métodos basados en retropropagación (GBP, LRP) destacan por su fidelidad, especialmente en escenarios donde la complejidad de las imágenes aumenta.
- Métodos como SHAP y Grad-CAM se desempeñaron peor en EMD y MIN debido a la falta de consistencia en sus explicaciones.

2.1.8 Conclusiones Clave

- Los métodos de retropropagación demostraron ser los más fieles, capturando con precisión las características relevantes para la toma de decisiones del modelo.
- Los datasets TXUXI proporcionaron una plataforma confiable para evaluar métodos XAI en escenarios controlados y realistas.
- Se destaca la necesidad de reducir el ruido en mapas de saliencia para mejorar la interpretabilidad y confianza de los usuarios finales.

2.2 Deep Learning for Case-Based Reasoning through Prototypes

2.2.1 Detalles del Artículo

- **Título:** Deep Learning for Case-Based Reasoning through Prototypes
- **Autores:** Oscar Li, Hao Liu, Chaofan Chen, Cynthia Rudin
- **Revista:** Presentado en la Trigésima Segunda Conferencia de la AAAI sobre Inteligencia Artificial (AAAI-18).
- **Nivel:** Investigación de alto impacto en interpretabilidad de redes neuronales.
- **Fecha de publicación:** 2018 (presentado como preprint en arXiv el 21 de noviembre de 2017).

2.2.2 Problema Principal

El artículo aborda la falta de interpretabilidad en redes neuronales profundas, presentando una arquitectura que combina el aprendizaje profundo con razonamiento basado en casos. La red explica sus predicciones usando prototipos aprendidos, que representan ejemplos específicos en el conjunto de datos.

2.2.3 Arquitectura del Modelo

El modelo combina:

- **Autoencoder:** Reduce la dimensionalidad y aprende representaciones latentes útiles.
- **Capa de prototipos:** Contiene vectores que representan ejemplos cercanos a los datos de entrenamiento en el espacio latente.
- **Clasificación:** Calcula distancias en el espacio latente entre inputs codificados y prototipos, pasando luego por una capa softmax para predecir probabilidades de clase.

2.2.4 Función de Costo

- **Cross-entropy loss (E):** Penaliza errores de clasificación.
- **Reconstruction loss (R):** Evalúa la fidelidad de la reconstrucción en el autoencoder.
- **R1:** Obliga a que cada prototipo sea similar a al menos un ejemplo del conjunto de entrenamiento.
- **R2:** Obliga a que cada ejemplo del conjunto de entrenamiento esté cerca de al menos un prototipo.

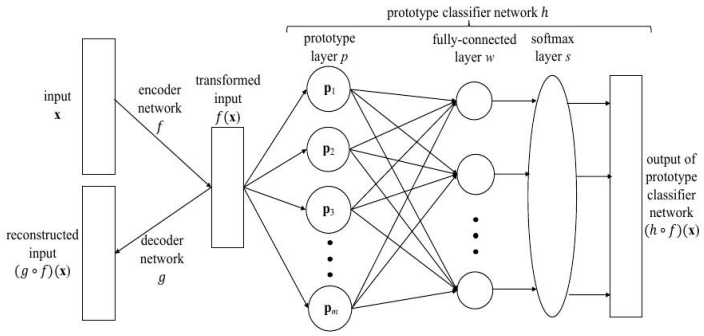


Figure 1: Network Architecture

2.2.5 Explicaciones Generadas

- **Prototipos visuales:** Los prototipos se decodifican a imágenes similares a ejemplos reales del conjunto de datos.
- **Interpretación de predicciones:** La clasificación se basa en la cercanía del input codificado a los prototipos más relevantes.

2.2.6 Casos de Estudio

- **MNIST:** Precisión del 99.22%. Los prototipos muestran diferentes estilos de escritura dentro de la misma clase.
- **Cars dataset:** Clasificación de ángulos de visión de autos. Los prototipos eliminan información irrelevante (como el color) y se enfocan en la forma y orientación.
- **Fashion MNIST:** Precisión del 89.95%. Los prototipos destacan contornos, ignorando detalles como texturas o patrones.

2.2.7 Méritos Adicionales

- **Interpretabilidad integrada:** No se necesita análisis posentrenamiento.
- **Regularización:** Los términos de interpretabilidad actúan como regularizadores, reduciendo el sobreajuste.
- **Flexibilidad:** Permite múltiples prototipos por clase, capturando variaciones dentro de las clases.

2.3 Machine Learning Models to Help Predict Treatment Outcomes in Clinical Gait Analysis

2.3.1 Detalles del Artículo

- **Título:** Machine Learning Models to Help Predict Treatment Outcomes in Clinical Gait Analysis
- **Autores:** Michael Schwartz, Andrew Ries, Andrew Georgiadis
- **Revista:** Gait & Posture
- **Nivel:** Publicación en una revista revisada por pares, cuartil Q1 en biomecánica clínica.
- **Fecha de publicación:** Julio de 2024.

2.3.2 Problema Principal

El artículo aborda la falta de predictibilidad y la estancada precisión de los resultados de tratamiento guiados por análisis clínico de la marcha. Propone el uso de modelos de aprendizaje automático para generar estimaciones honestas y sin sesgo de los resultados de tratamiento, mejorando así el proceso de toma de decisiones clínicas.

2.3.3 Métodos y Diseño del Estudio

- **Diseño del Estudio:** Análisis retrospectivo utilizando datos de pacientes pediátricos ambulatorios sometidos a diferentes tratamientos de extremidades inferiores.
- **Datos:** Incluye 7450 extremidades tratadas y no tratadas, obtenidas de un centro de análisis de la marcha.
- **Objetivos de Predicción:** Cambios estructurales, funcionales y de actividades relacionados con 12 cirugías comunes (representan más del 95% de los procedimientos en la base de datos).
- **Método de Evaluación:** Análisis de cobertura honesta y errores absolutos medios (MAE) para evaluar la precisión de los modelos.

2.3.4 Modelos de Aprendizaje Automático

- **Algoritmos Utilizados:**
  - Direct matching (correspondencia directa): Simple y comprensible para los clínicos.
  - Modelos estadísticos adicionales para ajuste de sesgos y evaluación de cobertura.
- **Evaluación:** Los modelos fueron evaluados en términos de sesgo, amplitud de intervalos de confianza (CIs) y cobertura promedio:
  - **Cobertura:** 86% para extremidades tratadas y 87% para controles.
  - **Errores Absolutos Medios (MAE):** Varían según el tipo de predicción y tratamiento.

2.3.5 Resultados Destacados

- **Ejemplo 1:** Cambios en el eje bimaleolar después de derotación tibial:
  - **Sesgo:** Inexistente.
  - **MAE:** $8^{\underline{o}}$ en extremidades tratadas y $6^{\underline{o}}$ en controles.
  - **Amplitud de CI:** $38^{\underline{o}}$ para extremidades tratadas y $27^{\underline{o}}$ para controles.
- **Ejemplo 2:** Mejora funcional después de osteotomía femoral:
  - Precisión menor en predicciones funcionales debido a datos de entrada menos precisos.

2.3.6 Discusión y Limitaciones

- **Fortalezas:**
  - Modelos honestos y sin sesgo.
  - Cobertura consistente y resultados interpretables para los clínicos.
- **Limitaciones:**
  - Amplitud subjetivamente amplia de intervalos de confianza.
  - Necesidad de datos de entrada más precisos (p. ej., mediciones tridimensionales, detalles quirúrgicos cuantitativos).

2.3.7 Méritos Adicionales

- **Facilidad de Implementación:** Los modelos son simples y comprensibles para profesionales médicos.
- **Impacto Clínico Potencial:** Integración de predicciones de resultados para mejorar asignaciones de tratamiento basadas en evidencia.
- **Recomendaciones Futuras:** Incorporación de datos avanzados como imágenes 3D y mediciones de fuerza para reducir los intervalos de confianza y mejorar la precisión predictiva.

2.4 On the Coherency of Quantitative Evaluation of Visual Explanations

2.4.1 Detalles del Artículo

- **Título:** On the Coherency of Quantitative Evaluation of Visual Explanations
- **Autores:** Benjamin Vandersmissen, José Oramas
- **Revista:** Computer Vision and Image Understanding (CVIU)
- **Nivel:** Revista revisada por pares, cuartil Q1 en visión por computadora.
- **Fecha de publicación:** Octubre de 2023.
- **DOI:** 10.1016/j.cviu.2024.103934

2.4.2 Problema Principal

El artículo aborda la falta de coherencia en los métodos de evaluación cuantitativa de explicaciones visuales generadas por redes neuronales. Propone un análisis exhaustivo de las métricas más utilizadas para evaluar mapas de calor de saliencia en modelos aplicados a tareas de visión por computadora.

2.4.3 Arquitectura del Modelo y Métodos

- Se analizaron los modelos **ResNet-50** y **VGG16**, entrenados con pesos preentrenados en ImageNet.
- Métodos de explicación visual evaluados:
  - Métodos basados en gradiente: SmoothGrad, Grad-CAM, Integrated Gradients (IG).
  - Métodos de retropropagación: Layer-wise Relevance Propagation (LRP).
  - Métodos basados en perturbación: Occlusion, RISE.
  - Otros: AdaSISE, TAME.

2.4.4 Métricas de Evaluación

- **Insertion y Deletion Curves:**
  - AUC para insertar píxeles relevantes: AdaSISE y RISE lideran con 0.642.
  - AUC para eliminar píxeles relevantes: TAME logra 0.176.
- **Average Drop (%):** Métrica para evaluar cómo el puntaje del modelo cambia al usar solo las regiones destacadas:
  - Mejores resultados: RISE (14.02%), Grad-CAM (14.96%).
- **Pointing Game:** Evalúa la localización de los objetos relevantes:
  - Mayor precisión: AdaSISE (92.77%).

2.4.5 Conclusiones Clave

- **Inconsistencias entre métricas:** Métricas como Insertion y Deletion reflejan nociones opuestas de "bondad" en las explicaciones.
- **Impacto de la Sparsidad:** Mapas de calor más densos favorecen métricas como Insertion, mientras que mapas dispersos benefician métricas como Deletion.
- **Recomendaciones:** Se desaconseja el uso de métricas basadas en tareas proxy, como el Pointing Game, debido a su baja correlación con otras métricas.