ISSN 1870-4069

Evaluación causal de características en base a explicaciones de clasificadores profundos de imágenes médicas: Un estudio de caso sobre imágenes de cálculos renales ex-vivo

Armando Villegas-Jimenez[1], Daniel Flores-Araiza[2], Francisco Lopez-Tiro[2, 3], Miguel Gonzalez-Mendoza[2], Gilberto Ochoa-Ruiz[2], Christian Daul[3]

[1] Instituto Politécnico Nacional,
México

[2] Tecnologico de Monterrey,
México

[3] Université de Lorraine,
Centre National de la Recherche Scientifique,
Francia

Resumen. Entender las razones detrás de las salidas de los modelos de aprendizaje profundo es crucial en el diagnóstico médico. Aunque existen métodos de inteligencia artificial explicable (XAI) para identificar las causas detrás de las predicciones, las evaluaciones cuantitativas de estas relaciones causales son limitadas. Por ello, proponemos una técnica para medir la relación causal entre las características del área de interés en imágenes de una clase específica y la salida de un clasificador, enfocándonos en imágenes de piedras renales. Nuestro método, llamado Puntuación de Explicación Causal (PEC), se evaluó en un conjunto de datos de imágenes ex-vivo de cálculos renales. Los experimentos demostraron que las relaciones causales medidas son más precisas cuando el área de interés se identifica utilizando un método explicable en lugar de anotaciones humanas en cuadros delimitadores, lo que ayuda a identificar qué explicaciones de los resultados de los modelos de aprendizaje profundo son más confiables en el contexto médico. El método PEC adapta técnicas existentes para trabajar con máscaras de segmentación en lugar de cajas delimitadoras, permitiendo una medición más precisa de las relaciones causales. Además, hemos modificado el método GradCAM para automatizar la extracción de máscaras de segmentación binarias, facilitando la obtención de medidas causales más consistentes que con segmentaciones manuales y facilitando el uso de nuetsro metodo al reducir la dependencia de anotaciones humanas. Los resultados indican que el método PEC permite una evaluación más informada de si las predicciones de un modelo y sus explicaciones se derivan de relaciones causales discernibles o no, lo que indica una dirección prometedora para mejorar el nivel de comprensión y confianza que podemos obtener al usar modelos DL como herramientas parael Diagnóstico Asistido por Computadora (CADx).

Palabras clave: Diagnóstico asistido por computadora (CADx), IA explicable (XAI), aprendizaje profundo (DL), análisis de imágenes médicas, análisis morfoconstitucional (MCA), piedras renales.

Armando Villegas-Jiménez, Daniel Flores-Araiza, Francisco López-Tiro, et al.

Causal Evaluation of Features from Explanations of Deep Classifiers of Medical Images: A Case Study on Ex-vivo Kidney Stone Imaging

Abstract. Understanding the reasons behind deep learning models' outputs is crucial in medical diagnosis. Although explainable artificial intelligence (XAI) methods exist to identify the causes behind predictions, quantitative evaluations of these causal relationships are limited. Therefore, we propose a technique to measure the causal relationship between the characteristics of the area of interest in images of a specific class and the output of a classifier, focusing on images of kidney stones. Our method, called Causal Explanation Scoring (PEC), was evaluated on a data set of ex-vivo images of kidney stones. Experiments demonstrated that measured causal relationships are more accurate when the area of interest is identified using an explainable method rather than human annotations of bounding boxes, helping to identify which explanations of the outputs of deep learning models are more reliable in the medical context. The PEC method adapts existing techniques to work with segmentation masks instead of bounding boxes, allowing for more precise measurement of causal relationships. Additionally, we have modified the GradCAM method to automate the extraction of binary segmentation masks, making it easier to obtain more consistent causal measurements than with manual segmentations and facilitating the use of our method by reducing the dependence on human annotations. The results indicate that the PEC method allows for a more informed assessment of whether a model's predictions and explanations are derived from discernible causal relationships or not. This indicates a promising direction for improving the level of understanding and confidence we can gain by using DL models as tools for Computer-Aided Diagnostics (CADx).

Keywords: Computer-aided diagnosis (CADx), explainable AI (XAI), deep learning (DL), medical image analysis, morpho-constitutional analysis (AMC), kidney stones.

1. Introducción

La identificación temprana del tipo de piedra renal depende de la necesidad de dicha clasificación por parte de un urólogo para determinar e iniciar el tratamiento. Además, varios países desarrollados, como señalan [8] y [6], informan de una incidencia significativa de litiasis urinaria (formación o presencia de cálculos renales), con alrededor de un 10 % de su población experimentando un episodio de cálculos renales al menos una vez en su vida. Además, hay una tasa de recurrencia notablemente alta del 40 % en estos países. Comúnmente, el procedimiento de análisis y clasificación de los cálculos renales, conocido como Análisis Morfo-Constitucional (AMC) [2] es largo, caro y requiere una gran experiencia.

Además, se ha demostrado que el análisis de imágenes médicas, así como el AMC, dependen en gran medida del operador [3, 18, 20]. Además de estas dificultades, debido al creciente número de pacientes cada año y a la gran diversidad natural de casos

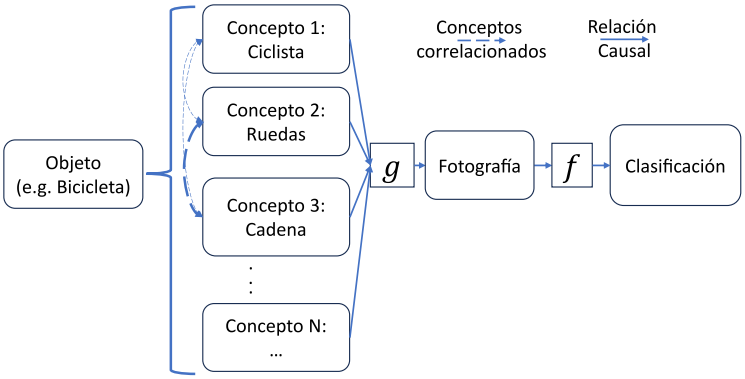


Fig. 1. Gráfico causal que relaciona instancias (e.g. bicicleta), sus conceptos de alto nivel (e.g. ciclista, ruedas, cadena etc.), con la imagen resultante (fotografía) y un clasificador de Deep Learning f, que da una clasificación. El borde discontinuo indica conceptos correlacionados, independientemente de su tipo de relación. Las aristas que conectan los conceptos con la imagen, a través de g, corresponden al proceso natural de generación de imágenes (e.g una cámara al tomar una fotografía).

médicos, el ámbito de la medicina necesita constantemente métodos más precisos y rápidos [17]. Los recientes avances en el campo de la IA, específicamente los avances en Deep Learning (DL) han impulsado la adopción temprana de modelos DL en imágenes médicas [9, 7]. En el contexto de la AMC, se han propuesto métodos desarrollados para el reconocimiento in-vivo o ex-vivo de cálculos renales [12, 10, 5, 19]. Aunque la mayoría de los métodos basados en DL superan a los no basados en DL en términos de precisión, carecen de capacidad de explicación, ya que los modelos DL simplemente emiten una clasificación para una entrada, independientemente de si es por las razones adecuadas o no.

Sin embargo, dado que el campo del análisis de imágenes médicas conlleva decisiones de alto riesgo, principalmente el diagnóstico, que tiene un impacto directo y profundo en la vida de los pacientes, no se puede exagerar la necesidad de un análisis automatizado robusto de imágenes médicas para el Computer-Aided Diagnosis (CADx) [15]. Por lo tanto, los especialistas médicos necesitan entender cómo las características de la imagen de entrada causaron la salida del modelo DL [4]. Precisamente, el campo de la eXplainable AI (XAI) busca proporcionar una comprensión del comportamiento de un modelo DL.

Bajo este objetivo principal, la mayoría de los métodos XAI propuestos relacionados a clasificación de imágenes destacan las causas de la salida de un modelo a partir de su entrada [1], es decir, las explicaciones al señalar la parte de la imagen de entrada que se cálculo causa la salida del modelo DL f se espera refleje el proceso inverso al natural de generación g de la imagen, representado en la Fig.1.

Sin embargo, esta relación causal presuntamente presente en las explicaciones se ha dejado sin una medida cuantitativa. Por ello, para abordar dichas carencias, en este trabajo hemos I) adaptado un método [11], en un conjunto de datos ex-vivo de piedras renales [2], para la puntuación causal, de la relación entre las características latentes en un clasificador de imágenes y su clasificación de salida.

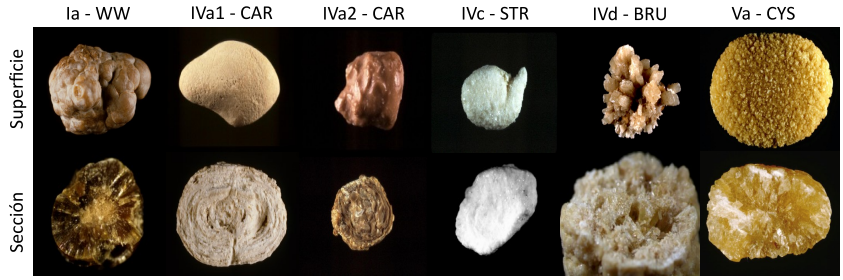


Fig. 2. Ejemplos de los seis subtipos diferentes de cálculos renales del conjunto de datos [2]. El nombre completo de las piedras renales son Whewellite (Ia - WW), Carbapatite (IVa1 & IVa2, CAR), Struvite (IVc - STR), Brushite (IVd -BRU), y Cystine (Va - CYS). Aquí mostramos ejemplos de ambos tipos de vistas en el conjunto de datos por clase, vista de la "Superficie" de las piedras, y vista de la"Sección" transversal.

La adaptación propuesta permite una medición más precisa puesto que trabaja con máscaras de segmentación, en lugar de cajas delimitadoras como lo hacen trabajos previos [11]. Adicionalmente, nuestra adaptación produce una medida causal entre 0 y 1, en lugar de cualquier valor positivo. También, II) modificamos un método de IA explicable (Grad-CAM [16]), para automatizar la extracción de máscaras de segmentación binarias de la región de mayor interés para el modelo, permitiendo obtener medidas causales más consistentes que con máscaras de segmentación anotadas por humanos. Con nuestro método: Puntuación de Explicación Causal (PEC), proporcionamos una forma de validar causalmente las salidas de un modelo DL basado en las áreas de la imagen de entrada indicadas por las explicaciones generadas para la misma entrada y salida del modelo.

Y lo que es más importante, mostramos resultados que indican que nuestro método (PEC) obtiene mejores resultados que cuando se utilizan máscaras de segmentación de los objetos de interés anotadas por humanos. Así pues, nuestro trabajo se enfoca en permitir que los especialistas de la salud aprovechen los hallazgos de los modelos DL para el diagnóstico, comprendiendo la lógica que subyace a dichos resultados, al tiempo que se facilita la aplicación responsable de estas potentes tecnologías de IA en el diagnóstico médico, logrando un equilibrio crucial entre la eficiencia de las máquinas y la responsabilidad humana.

2. Conjunto de datos y métodos

2.1. Conjunto de datos de piedras renales

Nuestro conjunto de datos ex-vivo, Fig.2, se divide en 209 imágenes de superficie y 157 de sección, que en total suman 366 imágenes. Estas imágenes se adquirieron con una cámara digital (CCD) en condiciones de iluminación controladas y con un fondo uniforme. El conjunto de datos está clasificado por los subtipos de cálculos renales, seis en nuestro caso. Estos subtipos, como se muestran en la Fig.2, son la Whewellite, subtipo Ia (Ia - WW), Carbapatite subtipo IVa1 (IVa1 - CAR), Carbapatite subtipo IVa2 (IVa2 - CAR), Struvite subtipo IVc (IVc - STR), Brushite subtipo IVd (IVd - BRU) y Cystine subtipo Va (Va - CYS).

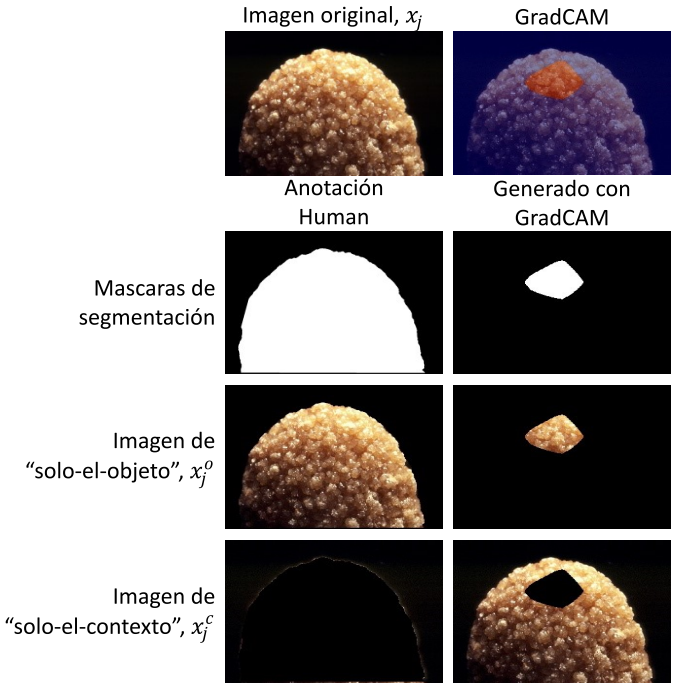
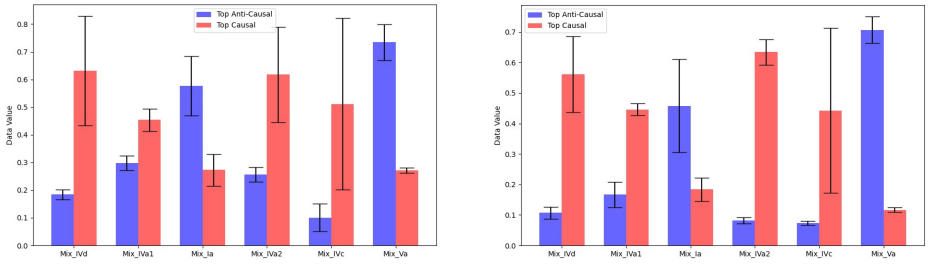


Fig. 3. Arriba a la izquierda: una imagen de ejemplo x_j del conjunto de datos. Arriba a la derecha: Grad-CAM para la clase predicha correspondiente, con umbral para mantener el 30 % de los valores más altos, en rojo. A partir de las segmentaciones del conjunto de datos anotadas por humanos y de las segmentaciones Grad-CAM con umbral, se obtuvieron las imágenes de "solo-el-objeto" x_j^o y "solo-el-contexto" x_j^c.

2.2. Método: Puntuación de explicación causal (PEC)

Inspirándonos en [11], modificamos ligeramente su propuesta "Feature Ratio (FR)" [11] para comprobar la relevancia de las características causales/anticausales señaladas por una segunda red denominada "Neural Causation Coefficient (NCC)". Nuestra modificación es la transformación de las puntuaciones FR originales para que estén acotadas entre 0 y 1, como se ve en la Ec.2, para facilitar su comparación entre diferentes FRs. El modelo NCC es un clasificador binario que indica si la activación de una característica de la última capa convolucional de una CNN se considera causal o anticausalmente relacionada con la salida del modelo.

Para establecer resultados de referencia fácilmente comparables en este trabajo y para futuras implementaciones, empleamos ResNet18, el cual es el modelo utilizado como función f en la Fig. 1. El modelo de ResNet18 f es una red neuronal convolucional que se ha utilizado ampliamente en este campo como clasificador sobre el cual el modelo NCC evaluará las puntuaciones causales. Las puntuaciones causales se obtienen de cada una de las 512 características de activación, $f_l \in \mathbb{R}^{512}$, ya que este es el tamaño de salida de la ultima capa convolucional de la ResNet18 f, considerada como el resultado del extractor de características del modelo.



(a) Puntajes de "Objeto" PEC en base a mascaras anotadas por humanos.

(b) Puntajes de "Objeto" PEC en base a mascaras automáticas con Grad-CAM.

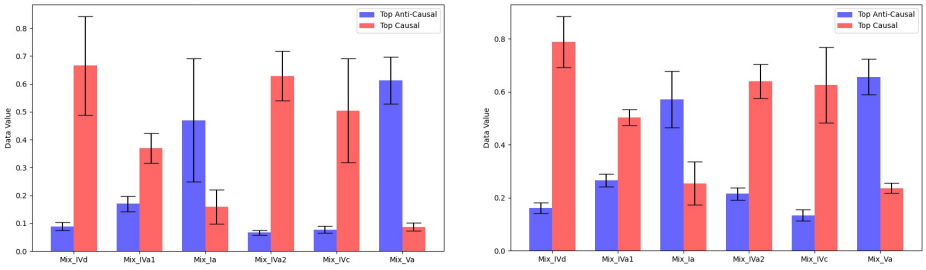
Fig. 4. Puntuación de Explicación Causal (PEC) para las imágenes de "Object-only". En 4a se utilizaron máscaras anotadas manualmente y en 4b se obtuvieron máscaras a partir del método Grad-CAM adaptado. En ambos casos se obtuvo la puntuación PEC para el "objeto".

La arquitectura NCC se entrena siguiendo el trabajo previo [11], utilizando los hiperparámetros y configuración de la implementación en [14]. El conjunto de datos utilizado para la prueba fue el conjunto de datos Tübingen, versión 1.0 [13], que es una colección de ciento ocho muestras observacionales causa-efecto del mundo real. Ya que el conjunto de datos Tübingen se utiliza habitualmente como referencia estándar en el campo de la inferencia causal. El modelo NCC entrenado obtuvo una precisión del 72 %, en lugar del 79 % comunicado anteriormente, a la hora de clasificar las relaciones causales y anticausales.

Aunque esta diferencia pone de manifiesto un área de mejora, es útil para nuestro objetivo actual de demostrar si una puntuación causal es más precisa cuando el "objeto de interés" se identifica mediante un método explicable, en lugar de ser anotado por humanos. Los mapas de características obtenidos de las m imágenes de entrada x_j y salida y_j, por clase $k \in \{1, \ldots, 6\}$, se procesan como un conjunto de pares de entrada (X, Y) por la red NCC. De esta manera, obtenemos las puntuaciones de NCC para cada mapa de características de cada imagen de entrada y las promediamos por mapa de características.

Para cada categoría k y el top 1 % de las característica f_l de acuerdo a sus puntuaciones causales y anticausales, determinaremos su relevancia como una característica de objeto f_l^o o característica de contexto f_l^c. Para ello, preparamos dos versiones alternativas de cada imagen de entrada x_j, las imágenes "solo-el-objeto" x_j^o y "solo-el-contexto" x_j^c. La imagen "solo-el-objeto" x_j^o, indicado por el área blanca de la máscara de segmentación, denota la sección de la imagen original que contiene el "objeto" correspondiente a la clase de la imagen x_j.

De forma complementaria, el recorte "solo-el-contexto" x_j^c, es el área negra de la máscara de segmentación e indica el "contexto", que podemos considerar el fondo de la imagen de entrada x_j. Un ejemplo de ambas imágenes, "solo-el-objeto" x_j^o y "solo-el-contexto" x_j^c, para ambos tipos de segmentación (anotada manualmente y Grad-CAM) pueden verse en la Fig.3. Las características f_l con puntuaciones causales y anticausales en el top 1 % más alto, para los mapas de características promediados por clase k, se seleccionan para informar de sus Feature Ratios (FRs) de acuerdo con la Ec. 1.



(a) Puntajes de "Contexto" PEC en base a mascaras anotadas por humanos

(b) Puntajes de "Contexto" PEC en base a mascaras automáticas con Grad-CAM

Fig. 5. Puntuación de Explicación Causal (PEC) para las imágenes de "solo-el-contexto". En (a) se utilizaron máscaras anotadas manualmente y en (b) se obtuvieron máscaras a partir del método Grad-CAM adaptado. En ambos casos se obtuvo la puntuación PEC para el "contexto".

Estos FRs nos permiten determinar en qué medida cada característica f_l es imputable a la máscara de segmentación del objeto (ratio de la característica del objeto, s_j^o) de la categoría k, o del contexto (ratio de la característica del contexto, s_j^c):

$$s_j^o = \frac{\sum_{j=1}^{m} |f_{jl}^c - f_{jl}|}{\sum_{j=1}^{m} |f_{jl}|}, \qquad s_j^c = \frac{\sum_{j=1}^{m} |f_{jl}^o - f_{jl}|}{\sum_{j=1}^{m} |f_{jl}|}. \tag{1}$$

Adaptación de Feature Ratio (FR): En esta propuesta, los ratios anteriores de "Objeto" s_j^o y "Contexto" s_j^c se adaptan para ser transformadas en valores positivos, acotados entre 0 y 1, en la siguiente Ec.2:

$$\sigma_o(s_j^o) = \frac{2}{1 + e^{-s_j^o}} - 1, \qquad \sigma_c(s_j^c) = \frac{2}{1 + e^{-s_j^c}} - 1. \tag{2}$$

La ResNet18 utilizada como clasificador se entrenó con el conjunto de datos de piedras renales descrito en la Sec.2.1. El entrenamiento consistió en 30 épocas, utilizando el optimizador Adam, con una tasa de aprendizaje de 0,0001. Esta ResNet18 tiene dos capas totalmente conectadas con 512 neuronas y la capa de salida final tiene 6 neuronas para la clasificación de las 6 clases de cálculos renales. De este modelo clasificador ResNet18 f, se utilizaron sus capas convolucionales como extractor de características.

Máscaras de segmentación anotadas manualmente: Las imágenes de "solo-el-objeto" y "solo-el-contexto" de los cálculos renales se obtuvieron a partir de máscaras de segmentación anotadas manualmente con valores 0 o 1 para el contexto y el objeto en la imagen respectivamente. La imagen de "solo-el-objeto" se obtiene multiplicando la imagen de entrada original y su correspondiente máscara de segmentación. A continuación, restamos de la imagen original la imagen "solo-el-objeto", lo que da como resultado la imagen "solo-el-contexto".

Máscaras de segmentación con Grad-CAM: Grad-CAM se caracteriza por calcular un mapa de calor, a partir de las activaciones de un modelo DL y sus gradientes. Sin embargo, las explicaciones producidas por Grad-CAM pueden llegar a tener todos sus valores iguales a cero para algunas entradas, para remediar esta situación aplicamos la modificación de elevar al cuadrado los elementos de la matriz de saliencia de Grad-CAM en lugar del paso de activación con la función ReLU [16]. Con ello se pretende mantener en el mapa de calor de Grad-CAM los valores más salientes, independientemente de su signo original.

Los mapas de Grad-CAM indican el área más relevante de la imagen de entrada para su correspondiente clasificación [16]. En estos mapas de calor se utilizó un umbral, para retener el 30 % de las activaciones más altas en el mapa de calor (el área más importante) como la porción "solo-el-objeto", con valores de 1, y el área restante, considerada menos importante, como la porción "solo-el-contexto", con valores de 0. De esta manera, obtuvimos máscaras de segmentación a partir de Grad-CAM. Este proceso se repite para todas las imágenes del conjunto de datos, con lo que obtenemos un conjunto de segmentaciones generadas por Grad-CAM.

Finalmente, se aplica el mismo proceso para las "Máscaras de segmentación anotadas manualmente" utilizando las segmentaciones generadas por Grad-CAM para obtener sus correspondientes imágenes "solo-el-objeto" y "solo-el-contexto", un ejemplo de los resultados obtenidos se puede ver en la Fig.3. La modificación de los FR "s" en la Ec.2, y el uso de máscaras de segmentación obtenidas automáticamente a partir del Grad-CAM adaptado, es nuestro método de Puntuación de Explicación Causal (PEC) propuesto.

3. Resultados y discusiones

Como se observa en la Fig.4 y la Fig.5, las mediciones causales/anticausales basadas en máscaras anotadas manualmente y las explicaciones de mapas de calor son posibles, incluso con menos varianza que con las máscaras de segmentación anotadas manualmente, que requieren mucho tiempo. Además, los resultados obtenidos entre las máscaras de segmentación anotadas manualmente Fig. 4a y los resultados de las máscaras generadas con segmentación Grad-CAM, Fig.4b fueron notablemente similares en los valores de sus medias, así como también para los resultados de "solo-el-contexto" en Fig. 5a y Fig. 5b. En los resultados, la máscara de segmentación Grad-CAM presentan la ventaja de una menor varianza que los resultados de las anotaciones manuales que recortan la piedra completa para las puntuaciones "solo-el-objeto".

Limitaciones: Ni el trabajo previo utilizado como inspiración [11], ni nuestra propuesta PEC hasta el momento consideran el caso para la identificación de correlaciones entre pares de características de activación f_l del modelo clasificador f. Es necesario realizar mediciones adicionales con diferentes Redes Neuronales Convolucionales (CNNs) para analizar que tan consistentes son los resultados. El uso de un valor umbral arbitrario podría estar limitando los resultados obtenidos con las máscaras de segmentación Grad-CAM. Un hallazgo clave es que para la mayoría de las clases, 4 de 6, la puntuación causal FR de "objecto" es predominante, como se muestra

en Fig.4, contradictoriamente al hallazgo en [11]. Esta diferencia clave puede deberse a la menor puntuación de rendimiento de NCC para la clasificación de señales causales, y a la limitada cantidad de muestras de datos en el conjunto de datos, de solo 366 para las 6 clases.

4. Conclusiones y trabajo futuro

Las mediciones causales basadas en las características más relevantes de la imagen de entrada son favorables. Nuestro método, PEC, demuestra que es posible automatizar las mediciones causales teniendo acceso a los pesos del modelo DL. Además, con nuestra propuesta, PEC, es posible dar a los especialistas una indicación de qué características de un modelo son las más relevantes y si éstas guardan una relación causal o anticausal con la salida. No obstante, son necesarios más experimentos.

Como trabajo futuro, deberían explorarse diferentes niveles de umbrales de segmentación para identificar el valor óptimo para evaluar tanto las puntuaciones causales como las anticausales. Para nuestro conjunto de datos, en particular, esto es relevante, debido a que las grandes áreas originales de fondo negro y las máscaras de segmentación obtenidas de Grad-CAM son empíricamente pequeñas, como se observa en Fig. 3.

Una dirección interesante a explorar para mejoras es modificar las puntuaciones FR de “Objeto” y “Contexto” para que se basen en un enfoque de aprendizaje métrico (metric learning) sobre el espacio latente extraído por las capas convolucionales del clasificador de imágenes, en lugar del cambio de activación de la imagen original frente a los recortes de “Objeto” o “Contexto” únicamente. Por último, igualmente la aplicación de distintos métodos XAI para obtener las máscaras de segmentación se deja para futuros trabajos.

Acknowledgments. The authors wish to acknowledge the Mexican Council for Science and Technology (CONAHCYT) for the support in terms of postgraduate scholarships in this project, and the Data Science Hub at Tecnologico de Monterrey for their support on this project. This work has been supported by Azure Sponsorship credits granted by Microsoft’s AI for Good Research Lab through the AI for Health program. The project was also supported by the French-Mexican ANUIES CONAHCYT Ecos Nord grant 322537.

Compliance with ethical approval. The images were captured in medical procedures following the ethical principles outlined in the Helsinki Declaration of 1975, as revised in 2000, with the consent of the patients.

Referencias

1. Borys, K., Schmitt, Y. A., Nauta, M., Seifert, C., Krämer, N., Friedrich, C. M., Nensa, F.: Explainable AI in medical imaging: An overview for clinical practitioners – Beyond saliency-based XAI approaches. European Journal of Radiology, vol. 162, pp. 110786 (2023) doi: 10.1016/j.ejrad.2023.110786

2. Corrales, M., Doizi, S., Barghouthy, Y., Traxer, O., Daudon, M.: Classification of stones according to Michel Daudon: A narrative review. European Urology Focus, vol. 7, no. 1, pp. 13–21 (2021) doi: 10.1016/j.euf.2020.11.004
3. De-Coninck, V., Keller, E. X., Traxer, O.: Metabolic evaluation: Who, when and how often. Current Opinion in Urology, vol. 29, no. 1, pp. 52–64 (2019) doi: 10.1097/mou.0000000000000562
4. Flores-Araiza, D., Lopez-Tiro, F., El-Beze, J., Hubert, J., Gonzalez-Mendoza, M., Ochoa-Ruiz, G., Daul, C.: Deep prototypical-parts ease morphological kidney stone identification and are competitively robust to photometric perturbations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 295–304 (2023) doi: 10.48550/arXiv.2304.04077
5. Gonzalez-Zapata, J., Lopez-Tiro, F., Villalvazo-Avila, E., Flores-Araiza, D., Hubert, J., Ochoa-Ruiz, G., Daul, C., Mendez-Vazquez, A.: A metric learning approach for endoscopic kidney stone identification. Expert Systems with Applications, vol. 255, pp. 124711 (2024) doi: 10.1016/j.eswa.2024.124711
6. Hall, P. M.: Nephrolithiasis: Treatment, causes, and prevention. Cleveland Clinic Journal of Medicine, vol. 76, no. 10, pp. 583–591 (2009) doi: 10.3949/ccjm.76a.09043
7. Jiang, H., Diao, Z., Shi, T., Zhou, Y., Wang, F., Hu, W., Zhu, X., Luo, S., Tong, G., Yao, Y. D.: A review of deep learning-based multiple-lesion recognition from medical images: classification, detection and segmentation. Computers in Biology and Medicine, vol. 157, pp. 106726 (2023) doi: 10.1016/j.compbiomed.2023.106726
8. Kasidas, G. P., Samuell, C. T., Weir, T. B.: Renal stone analysis: Why and how? Annals of Clinical Biochemistry: International Journal of Laboratory Medicine, vol. 41, no. 2, pp. 91–97 (2004) doi: 10.1258/000456304322879962
9. Lee, L. I. T., Kanthasamy, S., Ayyalaraju, R. S., Ganatra, R.: The current state of artificial intelligence in medical imaging and nuclear medicine. BJR—Open, vol. 1, no. 1, pp. 20190037 (2019) doi: 10.1259/bjro.20190037
10. Lopez, F., Varelo, A., Hinojosa, O., Mendez, M., Trinh, D. H., ElBeze, Y., Hubert, J., Estrade, V., Gonzalez, M., Ochoa, G., Daul, C.: Assessing deep learning methods for the identification of kidney stones in endoscopic images. In: Proceedings of the 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 2778–2781 (2021) doi: 10.1109/EMBC46164.2021.9630211
11. Lopez-Paz, D., Nishihara, R., Chintala, S., Scholkopf, B., Bottou, L.: Discovering causal signals in images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017) doi: 10.48550/arXiv.1605.08179
12. Lopez-Tiro, F., Estrade, V., Hubert, J., Flores-Araiza, D., Gonzalez-Mendoza, M., Ochoa, G., Daul, C.: On the in vivo recognition of kidney stones using machine learning. IEEE Access, vol. 12, pp. 10736–10759 (2024) doi: 10.1109/access.2024.3351178
13. Mooij, J. M., Peters, J., Janzing, D., Zscheischler, J., Schölkopf, B.: Distinguishing cause from effect using observational data: Methods and benchmarks. Journal of Machine Learning Research, vol. 17, no. 32, pp. 1103–1204 (2016)
14. Park, S.: Neural-causation-coefficient (2023) github.com/euphoria0-0/Neural-Causation-Coefficient
15. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, vol. 1, no. 5, pp. 206–215 (2019) doi: 10.1038/s42256-019-0048-x
16. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: IEEE International Conference on Computer Vision (2017) doi: 10.1109/iccv.2017.74
17. Topol, E.: Deep medicine: How artificial intelligence can make healthcare human again. Basic Books (2019)

18. Zhu, C., Doyle, T. E., Noseworthy, M. D.: Ultrasound operator variance classification for agency in artificial intelligence support of cyber-physical systems. In: IEEE Canadian Conference on Electrical and Computer Engineering, IEEE, pp. 446–451 (2022) doi: 10.1109/ccece49351.2022.9918266
19. Zhu, W., Zhou, R., Yuan, Y., Timothy, C., Jain, R., Luo, J.: Segprompt: Using segmentation map as a better prompt to finetune deep models for kidney stone classification (2023) doi: 10.48550/arXiv.2303.08303
20. Åkesson, L., Svensson, A., Edenbrandt, L.: Operator dependent variability in quantitative analysis of myocardial perfusion images. Clinical Physiology and Functional Imaging, vol. 24, no. 6, pp. 374–379 (2004) doi: 10.1111/j.1475-097x.2004.00574.x