UNIVERSIDAD
DE GRANADA

ETS Ingeniería Informática y de Telecomunicación

Máster en Ciencia de Datos e Ingeniería de Computadores

trabajo de fin de máster

Mejora de la Interpretabilidad en Modelos de Clasificación de Lesiones Cancerosas en Biopsias de Próstata mediante Técnicas de XAI.

Presentado por:
Carlos Lara Casanova

Tutor:
Francisco Herrera Triguero
Departamento de Ciencias de la Computación e Inteligencia Artificial
Mentor:
Iván Sevillano-García
Departamento de Ciencias de la Computación e Inteligencia Artificial

Curso académico 2024-2025

Mejora de la Interpretabilidad en Modelos de Clasificación de Lesiones Cancerosas en Biopsias de Próstata mediante Técnicas de XAI.

Carlos Lara Casanova

Carlos Lara Casanova *Mejora de la Interpretabilidad en Modelos de Clasificación de Lesiones Cancerosas en Biopsias de Próstata mediante Técnicas de XAI.*.

Trabajo de fin de Máster. Curso académico 2024-2025.

Responsables de tutorización

Francisco Herrera Triguero
Departamento de Ciencias de la Computación e Inteligencia Artificial

Iván Sevillano-García
Departamento de Ciencias de la Computación e Inteligencia Artificial

Máster en Ciencia de Datos e Ingeniería de Computadores

ETS Ingeniería Informática y de Telecomunicación

Universidad de Granada

Declaración de originalidad

D./Dña. Carlos Lara Casanova

Declaro explícitamente que el trabajo presentado como Trabajo de Fin de Máster (TFM), correspondiente al curso académico 2024-2025, es original, entendido esto en el sentido de que no he utilizado para la elaboración del trabajo fuentes sin citarlas debidamente.

En Granada a 26 de agosto de 2025

Fdo: Carlos Lara Casanova

Índice general

Agradecimientos

Gracias a Francisco e Iván por brindarme la oportunidad de trabajar con ellos y guiarme durante la realización de este trabajo. Gracias a mi familia por apoyarme durante el tiempo que me llevó desarrollar este trabajo.

Summary

KEYWORDS: neural networks, xai, explainable ai, prediction, explainability, artifical inteligence, machine learning, computer vision, deep learning, prostate cancer, gleason score, gleason groups.

Aquí va el resumen en inglés

Resumen

PALABRAS CLAVE: redes neuronales, xai, ia explicable, clasificación, explicabilidad, inteligencia artificial, aprendizaje automático, visión por computador, aprendizaje profundo, cáncer de próstata, gleason score, gleason groups.

Resumen en español.

1 Introducción

En 2020, el **cáncer de próstata, (Prostate Cancer, PCa)** fue el segundo tipo de cáncer más frecuente, y el quinto más mortal, en varones [HJR+21]. Una **biopsia de próstata** es una prueba que consiste en la extracción de pequeños tejidos de la próstata para examinar posibles signos de cáncer (Ver Figura 1.1). El **Sistema de Puntuación de Gleason (Gleason Score)** es una calificación, en el rango de 2 a 10, que se da a biopsias de la próstata tras ser examinadas bajo un microscopio [Sha24]. Valores más altos indican cánceres más agresivos y de crecimiento más rápido. En 2014 la **Sociedad Internacional de Patología Urológica (International Society of Urological Pathology, ISUP)** propuso un nuevo sistema basado en la Puntuación de Gleason, que propone cinco grupos ordenados (**Gleason Groups, GGs**) [EARH17]:

- **GG1:** Cáncer de grado bajo. Puntuación de gleason 6 o inferior.
- **GG2:** Cáncer de grado medio. Puntuación de gleason 7.
- **GG3:** Cáncer de grado medio pero más agresivo que GG2. Puntuación de gleason 7 pero percibido más agresivo.
- **GG4:** Cáncer de grado alto. Puntuación de gleason 8.
- **GG5:** Cáncer de grado alto. Puntuación de gleason 9 o 10.

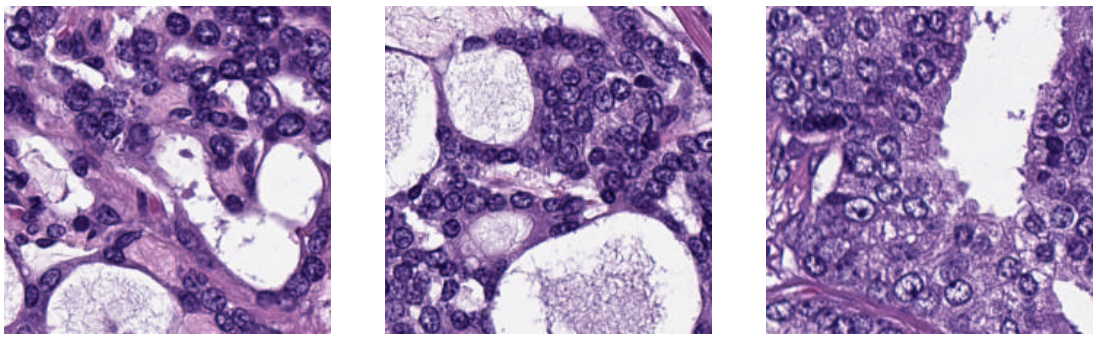

Figura 1.1: Tres imágenes correspondientes a tres biopsias distintas.

En el campo de la **Inteligencia Artificial (IA)**, la necesidad de asegurar transparencia y confiabilidad a dado pie a acuñar el término de **Inteligencia Artificial eXplicable (eXplainable**

Artificial Inteligence, XAI) [LBC+24]. El uso de XAI es especiamente importantes en ámbitos donde la toma de decisiones repercute directamente en la vida de las personas, como es el caso de la medicina. La Unión Europea ha propuesto una regulación para el uso de IA asistida en estos aspectos, entre ellos, que estos sistemas sean capaces de explicar sus decisiones de forma clara y comprensible.

Dentro del campo de la XAI, hay diferentes propuestas para la evaluación y la mejora de las explicaciones generadas. Por un lado, herramientas como **REVEL** [SGLH23a] permiten analizar la calidad de las mismas de manera robusta. Por otro, enfoques como **X-SHIELD** [SGLH23b] buscan mejorar las explicaciones mediante técnicas de regularización que aseguran el buen comportamiento de la generación de explicaciones. Aplicar estas técnicas en imagen médica puede ayudar a mejorar la interpretación de las personas profesionales de la salud de las decisiones de una IA que asista.

En este contexto este Trabajo de Fin de Máster trata de resolver el problema de **construir un modelo de IA de clasificación para lesiones cancerosas en biposias de próstata** y de la **mejora en interpretabilidad del modelo mediante técnicas de XAI**. Para esto se vamos a utilizar las herramientas REVEL y XSHIELD, y propondremos posibles mejoras de XSHIELD.

El TFM se estructura del siguiente modo. En primer lugar vamos a introducir los fundamentos teóricos necesarios para el correcto entendimiento del trabajo realizado (Capítulo 2). Después se va a hacer un estudio del estado del arte (Capítulo 3). A continuación presentamos las propuestas de este TFM (Capítulo 4). En el Capítulo 5 se describe el conjunto de datos con el que trabajamos, los distintos experimentos que realizamos y discutimos los resultados . Finalmente en el Capítulo 6 se incluyen conclusiones y posibles trabajos futuros.

1.1. Objetivos

El objetivo principal del TFM es la utilización (y propuesta) de técnicas de evaluación y mejora de explicaciones para modelos de clasificación de imágenes para detección de tejidos cancerosos en biopsias de próstata. Para el desarrollo del TFM, se divide el objetivo en distintos subobjetivos:

- Revisión del estado del arte Cuál?
- Estudio del código que implementa [SGLH23a, SGLH23b] para su correcto entendimiento.
- Proposición de nuevos métodos XAI para mejora de explicaciones para modelos de explicación.
- Modificación del código para implementar dichas propuestas.
- Diseño de los experimentos a realizar e implementación de estos.

- Evaluación de los resultados y subsecuente discusión.

Aquí introduciría el objetivo principal del TFM y lo dividiría en algunos sub-objetivos (revisión del estado del arte, estudio de las implementaciones de las métricas ReVEL/regularización X-Shield, proposición de mis regularizaciones, implementación de estas, experimentación y finalmente evaluación/discusión de estos).

1.2. Planificación

En esta sección hablo de cómo me he planificado el proyecto según estudio del problema/-diseño de mis modelos/implementación/experimentación (haría una tabla). Luego haría una comparación de mi planificación vs cómo se ha repartido finalmente y haría una estimación final de cuántas horas me ha llevado el TFM y en cuánto se estima el coste del proyecto final.
POR HACER

2 Fundamentos teóricos

En esta sección haría una introducción a la parte teórica de deep learning.

La sección de aprendizaje automático sería los fundamentos.

La sección de deep learning introduce conceptos de deep learning y en particular de visión por computador.

Importante, justo las dos secciones anteriores las tengo ya escritas en mi TFG, que justo trata de una aplicación de visión por computador, me preguntaba si se puede reutilizar lo que tengo en mi TFG o quizás eso no es buena praxis y debería reescribirlo, estoy interesado en saberlo).

Finalmente tendría la sección más relevante para este TFM que sería la de . Primero haría una introducción más general basándome en este paper (https://www.sciencedirect.com/science/article/pii/S156
y luego explicaría LIME para explicar bien lo que son las métricas REVEL más adelante.

2.1. Aprendizaje automático

2.1.1. Problema de clasificación

2.1.2. Problema de regresión

2.1.3. Optimización

Sobreentrenamiento

2.2. Deep Learning

Redes neuronales

Funciones de activación

Batch normalization

Dropout

2.2.1. Redes convolucionales

Capas convolucionales

Capas de pooling

Capas totalmente conectadas

2.3. XAI

2.3.1. LIME

2.3.2. Métricas ReVEL

2.3.3. Regularización X-Shield

3 Estado del arte

En esta sección hago búsquedas en Scopus para ver el estado del arte y analizarlo. Para seros sinceros no tengo muy claro que debería buscar ya que dudo que encuentre artículos que apliquen explicabilidad a detección de cánceres de próstata (aunque está bien hacer la búsqueda para tenerlo claro).

Por otro lado quizás debería buscar trabajos que traten la detección automática de cánceres de próstata o que apliquen explicabilidad a otras tareas médicas a ver que encuentro. Estoy interesado en ver cómo debería enfocar el estado del arte.

4 Métodos

En este capítulo introduciría qué son las métricas revel en profundidad (cálculo/interpretación) + La regularización X-SHIELD. Luego introduciría mis propuestas. Antes de introducir mis propuestas haría un introducción sobre EfficientNet (en particular efficientnet b2) e introduciría su arquitectura y por qué la he elegido.

4.1. Métricas ReVEL

4.2. X-Shield

4.3. Métodos propuestos

4.3.1. EfficientNet

4.3.2. FXShield

4.3.3. FRShield

4.3.4. HShield

4.4. Implementación

Aquí explico de qué código he partido (+ GitHub), breve introducción de cómo está organizado el proyecto (en carpetas/archivos), los datos, cómo me he organizado el proyecto en conda (versión de python + modulos) finalmente pondría un link al proyecto en mi GitHub). Por cierto me interesa saber si prefieres que cree un fork de alguno de tus proyectos en github (Iván) o si me creo uno a parte en el mío (obviamente referenciando tu github como partida), a mi me da igual.

5 Experimentos

5.1. Datos empleados

En esta sección explico el conjunto de datos del que parto. Explico como está organizado, cómo están balanceadas las clases, qué clases hay, qué significan, puedo mostrar algunos datos para mostrar ejemplos (si tengo permiso, claro). Me interesa saber de dónde provienen los datos también para ponerlo. En cuanto a estudio cualitativo de estos no creo que yo pueda aportar mucho así que no creo que haga (además de que hay demasiados datos).

5.2. Experimentos realizados

5.2.1. Separación de datos

Cómo están divididos los datos. He utilizado un conjunto de validación ya que validación cruzada es demasiado costoso en tiempo.

Validation

Optimizador y elección de hiperparámetros

5.3. Métricas

Explico que he utilizado Accurcy, cross entropy error y recuerdo el uso de la métricas revel. También hablo de los test estadísticos que he utilizado y los parámetros escogidos.

Test estadísticos bayesianos

5.4. Resultados

Aquí muestro mis resultados + curvas de aprendizaje y comento brevemente estos.

5.5. Discusión

Aquí es donde discuto mis resultados obtenidos.

6 Conclusiones

Conclusiones del TFM y si he conseguido los objetivos que me propuse al inicio.

6.1. Trabajos futuros

Aquí comento trabajos futuros. Cómo probar con otros datsets o lo que comentamos de sacar las métricas REVEL sobre features (en nuestra primera charla).

Bibliografía

[EARH17] Jonathan I Epstein, Mahul B. MD Amin, Victor E. Reuter, y Peter A. Humphrey. Contemporary gleason grading of prostatic carcinoma: An update with discussion on practical issues to implement the 2014 international society of urological pathology (isup) consensus conference on gleason grading of prostatic carcinoma. *The American Journal of Surgical Pathology*, 2017.

[HJR+21] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, y Bray F. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 2021.

[LBC+24] Luca Longo, Mario Brcic, Federico Cabitza, Jaesik Choi, Roberto Confalonieri, Javier Del Ser, Riccardo Guidotti, Yoichi Hayashi, Francisco Herrera, Andreas Holzinger, Richard Jiang, Hassan Khosravi, Freddy Lecue, Gianclaudio Malgieri, Andrés Páez, Wojciech Samek, Johannes Schneider, Timo Speith, y Simone Stumpf. Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, 106:102301, 2024.

[SGLH23a] Iván Sevillano-García, Julián Luengo, y Francisco Herrera. Revel framework to measure local linear explanations for black-box models: Deep learning image classification case study. *International Journal of Intelligent Systems*, 2023(1):8068569, 2023.

[SGLH23b] Iván Sevillano-García, Julián Luengo, y Francisco Herrera. X-shield: Regularization for explainable artificial intelligence. *ArXiv*, 2023.

[Sha24] Sovrin M. Shah. Gleason grading system [internet]. 2024. Revisado el 17 de Mayo, 2024; accedido el 25 de Agosto, 2025. Revisado por: VeriMed Healthcare Network, David C. Dugdale, Brenda Conaway y A.D.A.M. Inc.