Predicción Conforme aplicada a Modelos de Aprendizaje Profundo de Clasificación y Regresión: Cuantificación Automática de la Incertidumbre en Imágenes Histopatológicas

Diego Andrés Clavijo Granados y Santiago Gelvez Gonzalez

Trabajo de grado presentado como requisito parcial para optar al título de Ingeniero de Sistemas

Director
David Romo Bucheli, Ph.D
Doctor en Ingeniería - Ingeniería Eléctrica

Universidad Industrial de Santander
Facultad de Ingenierías Fisicomecánicas
Escuela de Ingeniería de Sistemas e Informática
Bucaramanga
2025

Dedicatoria

A mi madre Olga Cecilia, por brindarnos a mis hermanos y a mí un hogar lleno de amor, apoyo incondicional y comodidades que nos permitieron progresar libremente. Por la confianza depositada en mí en cada momento y por esa alegría que me demuestras por cada logro que realizo, por más pequeño que sea. Te estaré eternamente agradecido.

A mi padre Carlos Alberto, por ese gran ejemplo de honestidad y por el esfuerzo que realizaste día a día para permitirme estudiar con toda comodidad. También por er quien insistió en que estudiara mi carrera universitaria, ya que probablemente sin ti no la habría comenzado. Por todo eso y por todo lo que vendrá, muchas gracias.

A mi hermano Carlos Daniel, por establecer ese camino profesional admirable que me motiva a continuar mejorando cada día. Por todos esos momentos lindos mientras crecimos juntos que siempre atesoraré y por cada consejo que me continúas dando aun en la distancia. Cada vez que podemos reencontrarnos, me llena de alegría.

A mi hermana Nathalia, por esa cercanía y cariño que me has dado desde siempre. Por tu bondad y amabilidad incomparables. Por ese cariño que siempre me demuestras y por cuidarme y ser una especie de segunda madre para mí, a pesar de ser menor. Sabes que puedes contar conmigo, así como yo contigo.

Diego Andrés

A mi madre Mayra Cristina, por ser la mujer que con amor y cariño me enseñó los valores que me convirtieron en la persona que soy hoy en día; por darme la bendición todos los días y que con eso sé que todo estará bien. Desde pequeño siempre te miré con admiración, y le agradezco a Dios por darme una mamá tan maravillosa. Eres la persona que más amo en el mundo.

A mi padre Juan Carlos, por nunca permitir que me faltara algo; por trabajar hasta la última gota de sudor por mi felicidad; por creer siempre en mí y estar pendiente a mi progreso; por acompañarme en mi pasión y celebrar mis logros.

A mi hermana Cristina, por ser mi compañía incondicional desde pequeño; por ser mi compinche y enseñarme el valor de la lealtad. Eres la persona con la que puedo contar para toda la vida y siempre te tendré en cuenta. No sabes cuánto te quiero y valoro.

A mi nona Estrella, por ser la persona que más me ha amado en este mundo; por haberme consentido hasta el último momento; por nunca soltarme la mano y siempre ser mi consuelo; por enseñarme que tener un corazón bonito es importante. Todos los días te pienso y te extraño cada vez más. Sé que desde el cielo me estás viendo y me estás cuidando.

A mis tíos, Kenny y Cristian, por nunca descuidarme y siempre estar pendiente a mí; por ayudarme y aportar a mi felicidad desde pequeño; por ser esa compañía con la que puedo hablar de todo y transmitirme tranquilidad ante cualquier situación.

A mi nono Mario, por criar a una gran madre, y que a pesar de no compartir con el todo lo que hubiese querido, se que siempre buscó hacerme felíz.

A mis nonos, Octaciano y Maria Eugenia, por darme lo que tuvieron a su disposición para no hacerme sentir solo y colaborarme con lo que les pidiera. Su compañía la valoraré por siempre.

A mis tias, Monica, Claudia y Sonia, por siempre escucharme cuando lo necesite y darme una ayuda ante cualquier situación en las que las necesitara.

A mi mascota Luna, por nunca despegarse de mi en mi tiempo frente al computador y recibirme con una felicidad indescriptible cada vez que llegaba cansado de la universidad.

Cada uno dio lo mejor para verme triunfar, y en cada logro que consiga voy a pensar en ustedes.

Santiago

Agradecimientos

Agradezco a todos mis familiares por su apoyo constante e incondicional. En particular a Jose Luis y Franci por brindarme generosamente su hogar y apoyo durante la realización de mis prácticas.

Agradezco a mi director David Romo por su recurrente ayuda e inmenso apoyo durante todo este trabajo.

Diego Andrés

Quiero agradecerle a Dios por todas las bendiciones que me dio, por mi familia y las experiencias que he vivido. No cambiaría mi vida por nada.

Agradezco a toda mi familia, que son bastantes, porque sé que si los llamo me van a contestar y harían todo lo posible por ayudarme. Gracias por llevar el valor de la colaboración siempre presente.

Le agradezco a Catalina todo su amor y apoyo desde el primer día. Gracias por cada salida juntos y cada muestra de amor que me has dado. Gracias por acompañarme en las buenas y en las malas, me has subido el ánimo en momentos donde más lo necesitaba. Nunca voy a olvidar todo lo que has hecho por mí.

Agradezco a la Universidad Industrial de Santander por recibirme y formarme académicamente; por brindarme experiencias únicas y ayudarme a sacar la mejor versión mía.

Agradezco a mi director David Romo por su apoyo y guía en este trabajo; por brindarme siempre un espacio para discutir mis ideas y plasmarlas de la mejor manera.

Agradezco a todos los amigos que hice en este camino por sacarme una sonrisa y darme muchos momentos lindos vividos. Son pocos los que siguen conmigo, pero cada uno aportó un granito de arena en este proceso.

Santiago

Índice general

Índice de figuras

Índice de tablas

Resumen

Título: Predicción conforme aplicada a modelos de aprendizaje profundo de clasificación y regresión: Cuantificación automática de la incertidumbre en imágenes histopatológicas

Autores: Diego Andrés Clavijo Granados y Santiago Gelvez Gonzalez

Palabras clave: Placas virtuales completas, Aprendizaje profundo, Predicción conforme, Incertidumbre.

Descripción:

La incertidumbre es un aspecto inherente en los modelos de aprendizaje automático debido a que su entrenamiento se realiza con muestras representativas de datos reales asociadas al fenómeno bajo estudio. El muestreo, debido a su carácter estocástico intrínseco, induce variabilidad en las predicciones. La ausencia de una medida de incertidumbre en estos modelos puede ser altamente riesgosa, especialmente en entornos críticos como el sector de la salud, donde la precisión y confiabilidad son fundamentales para la toma de decisiones. El avance de la patología digital ha permitido mejorar el diagnóstico médico mediante el análisis de imágenes histopatológicas digitalizadas. A medida que los modelos de aprendizaje automático se integran en esta área, surgen nuevos desafíos relacionados con la interpretabilidad de los resultados y la necesidad de cuantificar la incertidumbre de predicciones, garantizando así su confiabilidad en escenarios clínicos.

En este trabajo se exploraron distintas técnicas de predicción conforme aplicadas a imágenes histopatológicas, con el objetivo de evaluar y comparar su desempeño en la cuantificación de la incertidumbre. El objetivo fue evaluar su efectividad en la generación de intervalos de confianza y sets de predicciones, así como valorar su potencial de aplicación práctica en entornos médicos. En clasificación, APS alcanzó las mayores coberturas, superando el 90 % en ambas arquitecturas, a cambio de sets más grandes. Además, la variante Mondrian mejoró la equidad en la cobertura al reducir la varianza entre distintos subgrupos, reforzando la robustez del método en escenarios clínicos heterogéneos. Para regresión, la predicción conforme permitió cuantificar la incertidumbre con intervalos. En prueba, la mejor variante fue la predicción conforme inductiva con EfficientNet-B3 (76.76 % cobertura, tamaño promedio del intervalo = 0.2758). En la predicción conforme de Mondrian, EfficientNet-B3 obtuvo 67.03 % (0.2079) . En la predicción conforme basada en regresión por cuantiles se invirtió la tendencia: DenseNet-169 logró mayor cobertura (74.59 %) que EfficientNet-B3 (66.49 %), con intervalos más amplios (0.2822 vs. 0.2634).

---

* Trabajo de Grado de Pregrado

** Facultad de Ingenierías Fisicomecánicas. Escuela de Ingeniería de Sistemas e Informática.

Director: David Romo Bucheli, PhD.

Abstract

Title: Conformal prediction applied to deep learning models of classification and regression: Automatic quantification of uncertainty in histopathological images

Authors: Diego Andrés Clavijo Granados and Santiago Gelvez Gonzalez

Keywords: Whole slide images, Deep learning, Conformal prediction, Uncertainty.

Description:

Uncertainty is an inherent aspect of machine learning models since their training relies on representative samples of real-world data associated with the phenomenon under study. Sampling, due to its intrinsic stochastic nature, induces variability in predictions. The absence of an uncertainty measure in these models can be highly risky, especially in critical domains such as healthcare, where accuracy and reliability are fundamental for decision-making. Advances in digital pathology have enabled improvements in medical diagnosis through the analysis of digitized histopathological images. As machine learning models become integrated into this field, new challenges arise regarding the interpretability of results and the need to quantify predictive uncertainty, thereby ensuring their reliability in clinical scenarios.

In this work, different conformal prediction techniques were explored on histopathological images with the aim of evaluating and comparing their performance in uncertainty quantification. The objective was to assess their effectiveness in generating confidence intervals and prediction sets, as well as to examine their potential for practical application in medical contexts. In classification, APS achieved the highest coverage, surpassing 90 % in both architectures, at the cost of larger prediction sets. Moreover, the Mondrian variant improved fairness in coverage by reducing variance across different subgroups, strengthening the robustness of the method in heterogeneous clinical scenarios. For regression, conformal prediction enabled uncertainty quantification through well-defined intervals. In testing, the best-performing variant was inductive conformal prediction with EfficientNet-B3 (76.76 % coverage, average interval size = 0.2758). In Mondrian conformal prediction, EfficientNet-B3 achieved 67.03 % (0.2079). In quantile regression-based conformal prediction, the trend was reversed: DenseNet-169 obtained higher coverage (74.59 %) than EfficientNet-B3 (66.49 %), with wider intervals (0.2822 vs. 0.2634).

---

* Bachelor Thesis

** Facultad de Ingenierías Fisicomecánicas. Escuela de Ingeniería de Sistemas e Informática.

Director: David Romo Bucheli, PhD.

Introducción

En el campo de la salud, donde las decisiones erróneas pueden tener consecuencias críticas, es esencial que los modelos no solo sean precisos, sino también confiables. Los profesionales médicos requieren herramientas que no solo emitan un diagnóstico, sino que también indiquen el nivel de certeza asociado a cada predicción (Cresswell et al. 2024).

La patología digital ha transformado radicalmente el diagnóstico de enfermedades, permitiendo un análisis más exhaustivo y preciso de tejidos a través de imágenes digitales (Tizhoosh y Pantanowitz 2018). No obstante, la creciente sofisticación de los modelos de aprendizaje automático utilizados en este ámbito presenta nuevos retos, siendo uno de los más relevantes la necesidad de cuantificar la incertidumbre inherente a sus predicciones.

Diferentes metodologías para la predicción conforme han sido propuestas, abarcando enfoques tanto para clasificación como para regresión. Sin embargo, aún no se ha realizado una evaluación exhaustiva de su efectividad en términos de cobertura, lo que deja abierta la necesidad de comparar su desempeño y determinar cuáles ofrecen las mejores garantías en distintos escenarios clínicos.

Para abordar este desafío, implementamos predicción conforme en modelos de clasificación y regresión aplicados a imágenes histopatológicas. Nuestra contribución principal es una comparación sistemática, dentro de un mismo marco experimental, de varias variantes de predicción conforme en clasificación y en regresión. Evaluamos cobertura y tamaño de conjuntos/intervalos, junto con métricas complementarias y un análisis a nivel de paciente. Esta comparación ayuda a identificar qué combinación técnica–arquitectura resulta más confiable según el objetivo clínico (diagnóstico por clases frente a cuantificación continua). En la práctica, la predicción conforme aporta conjuntos o intervalos que comunican de forma explícita la incertidumbre del modelo; esto permite decidir cuándo confiar en la salida automática y cuándo recurrir a la revisión de un experto. En nuestro estudio, el procedimiento fue transparente y replicable, y mostró cómo las predicciones se ajustan al nivel de confianza especificado.

El documento se divide en los siguientes capítulos importantes: en el capítulo 1 tenemos los objetivos que trazamos para la realización de este trabajo. Luego en el capítulo 2 tenemos todo lo relacionado con el marco teórico del tema. En el capítulo 3 tenemos todo el procedimiento completo de la predicción conforme para la tarea de clasificación, y de igual manera en el capítulo 4 tenemos el procedimiento completo de predicción conforme para la tarea de regresión. Y finalmente en el capítulo 5 tenemos las conclusiones generales del trabajo.

Planteamiento y Justificación del problema

Los modelos de aprendizaje profundo han demostrado un desempeño sobresaliente en tareas de clasificación, regresión y segmentación. Sin embargo, su aplicación en entornos críticos, como la salud, enfrenta un desafío fundamental: la incertidumbre en sus predicciones. Dado que estos modelos aprenden a partir de una muestra de datos finita, sus estimaciones pueden ser erróneas sin que exista un mecanismo claro para expresar su grado de confianza. Esta falta de transparencia limita su aplicabilidad clínica y puede comprometer la seguridad en la toma de decisiones médicas.

La patología digital ha impulsado avances significativos en el análisis de imágenes histopatológicas, mejorando la eficiencia y precisión del diagnóstico. No obstante, el uso de modelos de aprendizaje automático en esta área plantea nuevos retos, especialmente en tareas como la graduación histopatológica, donde la confiabilidad de las predicciones es esencial. La adopción de estos modelos en la práctica clínica depende no solo de su exactitud, sino también de su capacidad para expresar incertidumbre de manera interpretable.

La predicción conforme surge como una solución a esta problemática, ya que permite asociar cada predicción con un nivel de confianza cuantificable. Existen diferentes enfoques para implementarla, cada uno con ventajas y limitaciones que pueden afectar su desempeño en imágenes histopatológicas. Sin embargo, no está claro cuál de estos métodos es el más adecuado para evaluar la certidumbre y confiabilidad de los modelos en esta área. Así surge la pregunta de investigación que guía este estudio:

¿Qué ventajas y desventajas ofrecen los resultados de diferentes métodos de predicción conforme para informar la certidumbre y confiabilidad de los modelos de clasificación y regresión en imágenes histopatológicas?

Este trabajo exploró el uso de diferentes métodos de predicción conforme para conocer las ventajas y desventajas de su aplicación en imágenes de patología digital.

1. Objetivos

1.1. Objetivo General

Desarrollar y evaluar algoritmos de predicción conforme aplicados a modelos de aprendizaje de máquina enfocados en tareas de clasificación y regresión en imágenes histopatológicas.

1.2. Objetivos Específicos

- Seleccionar, documentar y procesar bases de datos abiertas de histopatología para las tareas de clasificación y regresión.
- Implementar modelos de clasificación y regresión basados en arquitecturas de redes neuronales disponibles en la literatura.
- Calibrar los modelos automáticos para la realización de prediccion conforme siguiendo diferentes algoritmos disponibles en la literatura.
- Evaluar los métodos de predicción conforme aplicados a los modelos de clasificación y regresión, en términos de la adaptabilidad de los conjuntos conformes, e intervalos de confianza obtenidos.

2. Marco Referencial

Este capítulo contiene descripciones de conceptos utilizados en nuestro trabajo. En la sección 2.1 se presenta la la patología digital y su impacto en la práctica clínica. La sección 2.2 distingue las tareas de clasificación y regresión en aprendizaje automático. En la sección 2.3 se introducen las arquitecturas de redes neuronales convolucionales empleadas en este trabajo. Luego, la sección 2.4 describe los principios de la predicción conforme. Finalmente, la sección 2.5 describe algunos trabajos en los cuales se aplica la predicción conforme en imágenes histopatológicas.

2.1. Patología Digital

La patología es la rama de la medicina dedicada al estudio de las enfermedades, enfocándose en la identificación de sus causas, mecanismos de desarrollo, alteraciones estructurales en células, tejidos y órganos, y consecuencias funcionales en el organismo (Orchard y Nation 2018). Esta disciplina se apoya en el análisis de muestras biológicas como sangre, orina y tejidos, informando el diagnóstico y tratamiento de enfermedades, lo que la convierte en un pilar fundamental para la atención médica.

Dentro de la patología, la histopatología se centra en el estudio microscópico de tejidos biológicos para identificar cambios morfológicos asociados a enfermedades como inflamación, infecciones y cáncer (Orchard y Nation 2018). Esta disciplina requiere una interpretación experta, ya que las alteraciones pueden ser sutiles y subjetivas. Con el avance de la tecnología, las muestras de tejido pueden digitalizarse, generando imágenes histopatológicas que preservan la arquitectura tisular. Estas representaciones digitales facilitan el análisis de la distribución celular y la interacción entre estructuras, ofreciendo ventajas como la accesibilidad, el almacenamiento eficiente y la posibilidad de aplicar herramientas computacionales para su procesamiento (Madabhushi y Lee 2016).

La evolución tecnológica ha permitido la transición hacia la patología digital, que consiste en el uso de tecnologías digitales para la adquisición, gestión y análisis de imágenes de tejidos con fines diagnósticos (Bera et al. 2019). Este enfoque ha transformado la práctica de la patología tradicional al convertir láminas histopatológicas en imágenes digitales de alta resolución, facilitando diagnósticos más rápidos, precisos y reproducibles. Además, permite un almacenamiento eficiente y acceso remoto a datos, lo que resulta crucial para investigaciones longitudinales y colaboraciones internacionales en el ámbito médico.

En la actualidad, la integración de tecnologías avanzadas como el aprendizaje automático (con sus siglas en inglés “ML’) está revolucionando la patología digital (Niazi; Parwani y Gurcan 2019). Estas herramientas automatizan tareas complejas, como la clasifi-

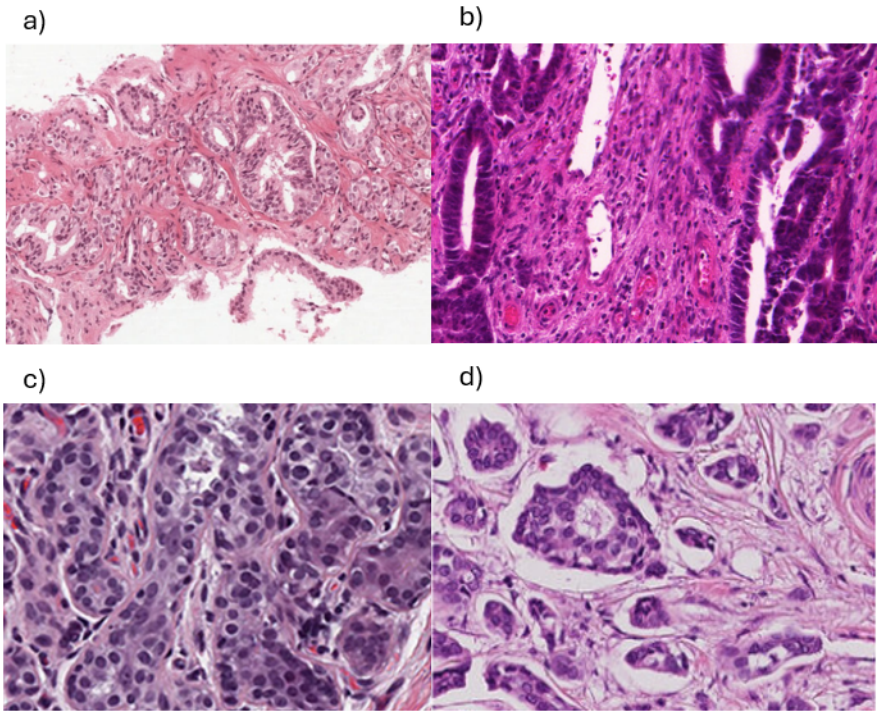
cación de tejidos, la detección de patrones anormales y la predicción de resultados clínicos, reduciendo la carga de trabajo de los patólogos y minimizando errores humanos. La patología computacional puede utilizarse para automatizar tareas que los patólogos ya realizan en la práctica diaria y para descubrir biomarcadores morfológicos para los resultados clínicos de interés (Song et al. 2023). Por ejemplo, el análisis automático de características histopatológicas como la densidad celular o la proliferación tumoral no solo brinda una evaluación cuantitativa y objetiva, sino que también permite identificar correlaciones morfológicas con parámetros clínicos clave, como la agresividad tumoral o la respuesta a tratamientos.

A pesar de su potencial, la implementación de la patología digital enfrenta retos significativos. Entre ellos, destacan la validación de algoritmos, la interoperabilidad entre sistemas y la aceptación de estas herramientas por parte de los profesionales de la salud. Garantizar la privacidad y seguridad de los datos de los pacientes sigue siendo una prioridad fundamental para su adopción en entornos clínicos. Asimismo, la interpretabilidad de los modelos es esencial para generar confianza y facilitar su integración en la práctica médica.

La predicción conforme emerge como una técnica clave para abordar la incertidumbre en los modelos predictivos. Al proporcionar intervalos de confianza o conjuntos conformes que cuantifican la certidumbre en cada predicción, estas técnicas fortalecen la confianza en los sistemas automatizados y facilitan la integración de la patología digital en la práctica médica (Fontana; Zeni y Vantini 2023). Esto resulta especialmente relevante en este proyecto, que busca desarrollar modelos confiables e interpretables para apoyar el diagnóstico en imágenes histopatológicas.

En la Figura 1 se ilustran algunas de las tareas más comunes abordadas mediante patología digital, utilizando imágenes provenientes de bases de datos abiertas. Estas tareas incluyen clasificación, segmentación y regresión, y representan distintos desafíos que han motivado el desarrollo de soluciones automatizadas en la comunidad científica.

Figura 1: *Imágenes de bases de datos abiertas en patología digital enfocadas en diferentes tareas: (a) Graduación de cáncer de próstata (PANDA challenge), (b) Segmentación de glándulas en cáncer colorrectal (GlaS), (c) Cuantificación de celularidad tumoral en tejido mamario (BreastPathQ), (d) Detección de mitosis en tejido mamario (TUPAC16).*



2.2. Aprendizaje automático: Clasificación y regresión

Dos tareas que resaltan a la hora de poner a prueba modelos de aprendizaje automático son las tareas de clasificación y regresión. En el caso de clasificación tendremos que predecir una clase discreta (o categoría) que le pertenezca a la entrada. Los modelos de clasificación constan de una entrada y varias clases asociadas a esta. Es la tarea de los modelos aprender y, basado en lo que aprendió, asignarle una clase a una nueva entrada. Normalmente, cada clase va a tener asignado un valor (como una probabilidad o un puntaje) con el que comparará con las demás clases, y es con esta información que se realiza la predicción (Kotsiantis; Zaharakis y Pintelas 2006). Un ejemplo de esto puede ser: predecir si una imagen de un animal es un “perro” o un “gato”.

Para los modelos de regresión, ahora tenemos que realizar una predicción de un valor continuo. A diferencia de la clasificación, donde la salida es una etiqueta discreta, en la regresión la salida es un número real que puede tomar cualquier valor dentro de un rango determinado (Kumar y Bhatnagar 2022). Normalmente este tipo de tareas se realiza para predecir el valor de algo, por ejemplo, predecir el valor de una casa basándose en sus características (donde el valor puede ser, por ejemplo, $250.000 o $1.000.000), o predecir la

temperatura de una zona. Solo en casos específicos (como predecir una probabilidad o un porcentaje, como el porcentaje de celularidad), el valor predicho estaría limitado al rango entre 0 y 1.

Tabla 1: *Comparación concisa entre las Tareas de Clasificación y Regresión*

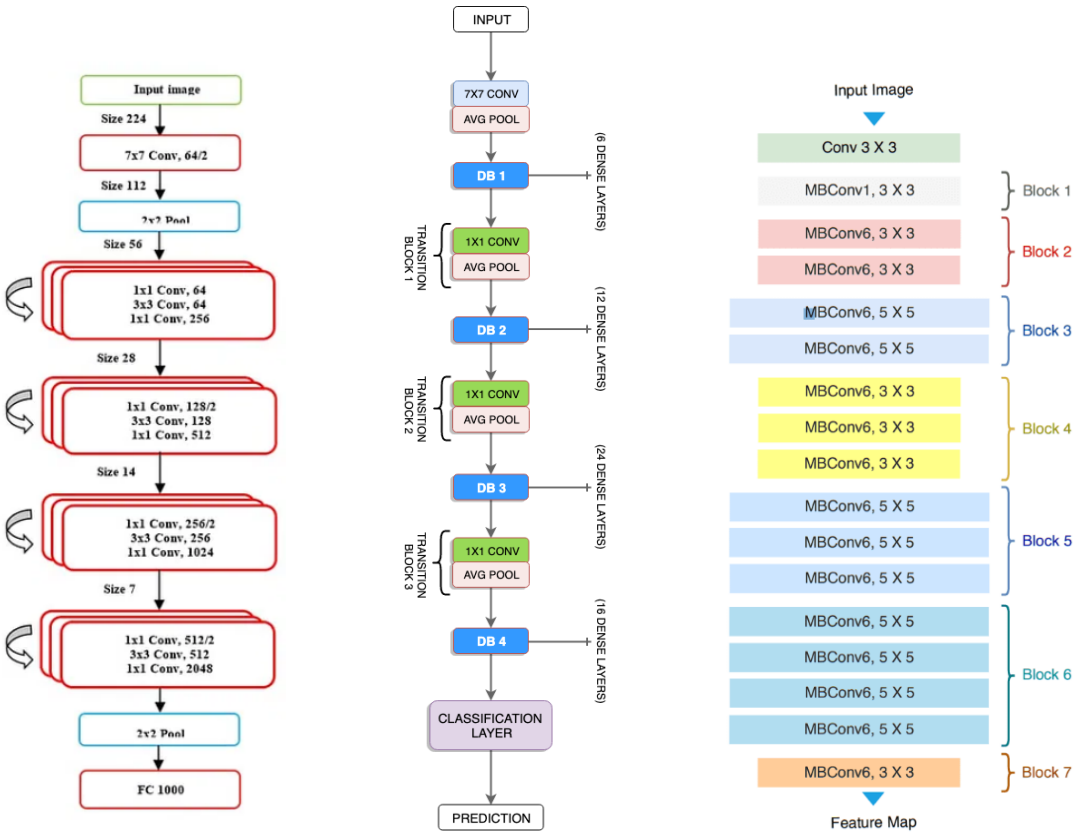
Característica	Tarea de Clasificación	Tarea de Regresión
Objetivo Principal	Predecir una etiqueta discreta o clase.	Predecir un valor continuo (número real).
Dominio de Salida	Conjunto finito y predefinido de categorías.	Rango de valores, típicamente infinito (p. ej., $\mathbb{R}$ o un subconjunto).
Ejemplos de Salida	"Perro", "Gato", "Clase A".	\$250.000, $15.7°C$, 0.85 (probabilidad).
Aplicación Común	Diagnóstico médico, reconocimiento de imágenes.	Predicción de precios, pronóstico del tiempo, predicción de celularidad.

2.3. Redes neuronales convolucionales

Las redes neuronales convolucionales (CNN, por sus siglas en inglés) son un tipo de arquitectura especializada en el procesamiento y análisis de datos que poseen una estructura espacial, como las imágenes. Estas redes están diseñadas para extraer características jerárquicas de los datos, comenzando por patrones simples como bordes y texturas en las primeras capas, hasta características más complejas en las capas profundas. Esta capacidad las convierte en herramientas indispensables en tareas de clasificación y regresión, donde la identificación precisa de patrones es crucial.

Las CNN han transformado significativamente campos como la visión por computadora y la patología digital, facilitando la automatización de procesos complejos. Dentro de este marco, arquitecturas como EfficientNet y DenseNet han emergido como soluciones innovadoras para superar limitaciones comunes en redes profundas y optimizar su desempeño en escenarios clínicos y biomédicos.

Figura 2: *Arquitecturas de Redes Neuronales utilizadas frecuentemente para tareas de clasificación: (a) ResNet: Imagen tomada de Mukherjee (2022), (b) DenseNet: Imagen tomada de Khanna (2020), y (c) EfficientNet: Imagen tomada de Ahmed y Sabab (2021).*



ResNet es una arquitectura de red neuronal profunda diseñada para abordar el problema de la degradación o vanishing gradient en redes muy profundas. Este problema, común en redes tradicionales, limita su rendimiento cuando la profundidad aumenta significativamente. ResNet resuelve este desafío mediante conexiones residuales, que permiten a las capas aprender los "residuos" entre la entrada y la salida esperada, en lugar de intentar aprender la función completa. Estas conexiones no solo mitigan la desaparición del gradiente, sino que también facilitan el entrenamiento de redes mucho más profundas. Esto asegura que las redes sean fáciles de optimizar y eviten altos errores de entrenamiento que suelen ocurrir en redes convencionales a medida que aumenta la profundidad. La arquitectura de ResNet puede observarse en la Figura 2 sección a, donde se ilustran las características clave de este diseño (He et al. 2016).

Las conexiones residuales de ResNet tienen una ventaja significativa en la optimización, ya que no introducen parámetros ni complejidad computacional adicionales (He et al. 2016). Esta característica hace de ResNet una opción sólida tanto para tareas de clasificación como de regresión en imágenes médicas. A través de estas conexiones, ResNet mejora la

estabilidad del modelo, permitiendo que las redes sean entrenadas de manera más eficiente, incluso a profundidades mucho mayores que las de redes tradicionales.

DenseNet, al igual que ResNet, mejora el flujo de información y gradientes durante el entrenamiento mediante conexiones densas y se puede observar en la Figura 2 sección b . A diferencia de las arquitecturas tradicionales, en DenseNet, cada capa está conectada a todas las capas anteriores, lo que permite que cada una reciba como entrada las características acumuladas de las capas precedentes. Esto fomenta la reutilización de características, reduce la redundancia y hace que las redes sean más compactas y eficientes. DenseNet ha demostrado un desempeño sobresaliente en tareas que requieren un análisis detallado de imágenes complejas, como las histopatológicas (Huang; Liu; Pleiss et al. 2022).

Una de las principales ventajas de DenseNet es su capacidad para aliviar el problema del desvanecimiento del gradiente (vanishing gradient), un desafío común en redes neuronales profundas. Además, fortalece la propagación de características, lo que mejora significativamente el flujo de información y gradientes a través de la red, haciéndola más fácil de entrenar. Cada capa tiene acceso directo a los gradientes de la función de pérdida y a la señal de entrada original, lo que lleva a una supervisión profunda implícita, ayudando a optimizar el proceso de aprendizaje (Huang; Liu; Van Der Maaten et al. 2017).

Una característica contraintuitiva de DenseNet es que, a pesar de su diseño profundo y denso, requiere menos parámetros que las redes convolucionales tradicionales, ya que no es necesario volver a aprender mapas de características redundantes. Esta eficiencia en el uso de parámetros, junto con la reutilización de características, contribuye a que DenseNet produzca modelos condensados que son fáciles de entrenar y altamente eficientes (Huang; Liu; Van Der Maaten et al. 2017).

EfficientNet es una arquitectura de red neuronal profunda que introduce un método de escalado compuesto para optimizar el rendimiento de redes convolucionales en tareas de clasificación y regresión y su arquitectura puede observarse en la Figura 2 sección c. A diferencia de las arquitecturas tradicionales, que escalan independientemente las dimensiones de profundidad, ancho y resolución, EfficientNet propone un enfoque unificado para escalar todas estas dimensiones de manera conjunta. Este enfoque permite una mejora significativa en la eficiencia del modelo, garantizando un rendimiento óptimo con un número reducido de parámetros (Tan y Le 2019).

El método de escalado propuesto en EfficientNet es altamente efectivo, ya que utiliza un coeficiente compuesto simple para ajustar las tres dimensiones de forma equilibrada: la profundidad de la red, el ancho de las capas y la resolución de las imágenes de entrada. Esta estrategia de escalado uniforme no solo mejora la precisión del modelo, sino que también reduce los costos computacionales, haciendo que la arquitectura sea más eficiente en compa-

ración con otras redes neuronales profundas (Tan y Le 2019).

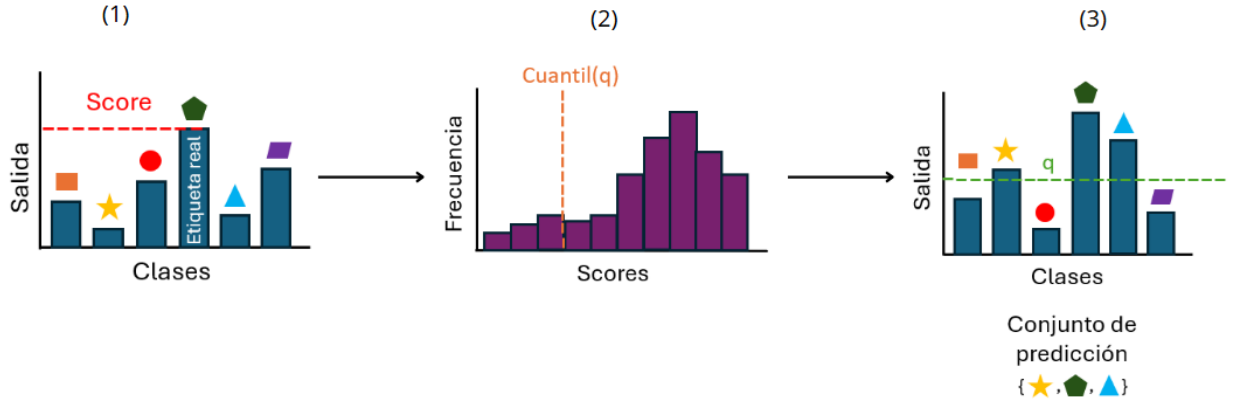
2.4. Nociones de predicción conforme

La Predicción Conforme (PC) es una metodología que genera predicciones asociadas con un nivel de confianza o una medida de certidumbre. Esta técnica se basa en los conceptos de no conformidad y calibración de predicciones, permitiendo que los modelos de aprendizaje automático no solo produzcan una predicción, sino que también indiquen qué tan seguros están de que dicha predicción es correcta. El principio establece que para cualquier nivel de confianza dado (por ejemplo, 90 %), la PC garantiza que la verdadera respuesta estará dentro de un intervalo de predicción con al menos ese porcentaje de confianza.

La incertidumbre representa la medida en que un modelo está seguro o inseguro acerca de sus predicciones. En los modelos de aprendizaje automático, es relevante identificar y comunicar esta incertidumbre, ya que un modelo que estima un alto valor de probabilidad en una predicción no necesariamente tiene un grado alto de confianza en su predicción. La predicción conforme aborda este problema al transformar cualquier noción heurística de incertidumbre de un modelo en una medida rigurosa (Angelopoulos y Bates 2021), mejorando así la interpretabilidad y confiabilidad del modelo.

El planteamiento general de la predicción conforme se basa en el concepto de "conformidad" o "no conformidad", que mide qué tan bien o mal se ajusta una observación nueva a un conjunto de datos previamente observado. La predicción conforme puede aplicarse tanto a tareas de clasificación como de regresión, y utiliza los datos conocidos para determinar qué tan similar es un nuevo ejemplo a los ejemplos previamente etiquetados. La técnica genera intervalos de predicción que contienen la respuesta verdadera con alta probabilidad.

Figura 3: *Esquema general de predicción conforme en tres etapas: (1) entrenamiento del modelo base, (2) cálculo de scores de no conformidad a partir de observaciones etiquetadas, y (3) construcción del conjunto/intervalo de predicción para cada nueva muestra.*



La predicción conforme se compone de tres pasos (Figura 3):

(1) Modelo base. Se entrena un predictor f (red neuronal, SVM, etc.) con los datos disponibles. El modelo devuelve puntajes por clase (clasificación) o valores continuos (regresión).

(2) Scores de no conformidad. Se fija una función $A(x, y; f)$ que mide el "desajuste" entre la salida del modelo y una etiqueta y. Con las observaciones ya etiquetadas (históricas o las que se van incorporando en el tiempo) se calculan los *scores* $\{s_i\}_{i=1}^{n}$, $s_i = A(x_i, y_i; f)$, y se obtiene un umbral $\hat{q}$ como el cuantil empírico al nivel $(1 - \alpha)$:

$$\hat{q} \;=\; s_{(\lceil (n+1)(1-\alpha) \rceil)},$$

usando la corrección finito–muestral estándar.

(3) Conjunto/intervalo por muestra nueva. Para un nuevo x se construye el conjunto conforme con el mismo criterio de no conformidad:

$$\widehat{\mathcal{Y}}(x) \;=\; \{\, y : \; A(x, y; f) \le \hat{q} \,\},$$

que en regresión se traduce en un intervalo $C(x)$ (p. ej., $[\hat{y}(x) - \hat{q},\ \hat{y}(x) + \hat{q}]$ o, si se usan cuantiles, $[\hat{q}_{\text{inf}}(x) - \hat{q},\ \hat{q}_{\text{sup}}(x) + \hat{q}]$). Cada vez que se dispone de nuevas etiquetas, se pueden añadir sus scores al conjunto $\{s_i\}$ y actualizar $\hat{q}$; el procedimiento sigue siendo el mismo.

2.5. Predicción conforme en imágenes histopatológicas

La aplicación de la predicción conforme ha demostrado un alto potencial para fortalecer la fiabilidad de los modelos de aprendizaje profundo en el análisis de imágenes médicas, particularmente en el ámbito de la patología digital. Al proporcionar una estimación cuantitativa de la incertidumbre, la PC permite que los sistemas de inteligencia artificial (IA) identifiquen predicciones poco confiables y las deriven a revisión por parte de especialistas humanos. Este mecanismo no solo se ajusta de manera natural a los flujos de trabajo clínicos, sino que también contribuye a mejorar la confianza en la adopción de estas tecnologías.

Un ejemplo concreto se observa en la histopatología del cáncer de próstata, donde la integración de la PC en modelos de aprendizaje profundo mostró mejoras notables en la fiabilidad diagnóstica y la seguridad del paciente. En un estudio de referencia, un sistema aumentado con PC logró reducir la tasa de errores del 2 % al 0.1 % en pruebas de dominio controlado. La utilidad del marco se evidenció aún más en escenarios fuera de distribución, como el análisis de tejido prostático atípico. En este contexto, el sistema con predicción conforme cometió solo 3 errores (2 %), frente a los 44 errores (25 %) del modelo sin PC, al detectar automáticamente como inciertos el 80 % de los casos más complejos y señalarlos para revisión humana(Olsson et al. 2022). Este resultado subraya el valor de la PC como un componente de apoyo a la toma de decisiones clínicas, reduciendo riesgos asociados con la sobreconfianza de los modelos.

Un beneficio similar se reportó en la clasificación del estatus HER2 en cáncer de mama a partir de imágenes de inmunohistoquímica (IHQ). En este caso, el uso de la predicción conforme permitió obtener predicciones altamente fiables en casos claros, al mismo tiempo que identificaba automáticamente los escenarios ambiguos. Con un nivel de significancia de 0.05, el clasificador basado en predicción conforme logró señalar el 69.8 % de los casos con puntuación equívoca dela etiqueta 2+ como candidatos a pruebas adicionales. Este enfoque de predicción selectiva representa un avance práctico hacia la evaluación automatizada de HER2, pues concentra la automatización en los casos de alta confianza y establece un protocolo claro para los casos inciertos. En conjunto, la predicción conforme aporta un marco que equilibra precisión, interpretabilidad y seguridad, facilitando su integración en entornos médicos reales(Pintawong et al. 2025).

3. Predicción conforme para clasificación: Graduación en cáncer de próstata

En este capítulo nos centramos en la aplicación de la predicción conforme a problemas de clasificación, donde el objetivo del modelo es asignar una etiqueta discreta (categoría) a una imagen. Mediante el uso de la predicción conforme, asignaremos un set de predicciones que garantice un nivel de confianza predefinido, lo que resulta crucial en tareas donde la certeza sobre las decisiones del modelo es fundamental. Esta técnica permite no solo hacer una clasificación, sino también evaluar la incertidumbre asociada a cada predicción, lo que añade un valor significativo en aplicaciones sensibles.

Este capítulo está estructurado de la siguiente manera. En la sección 3.1 se presenta el dataset PANDA utilizado para evaluar los métodos de predicción conforme en la tarea de clasificación. Luego, en la sección 3.2 se detallan las arquitecturas de redes neuronales convolucionales, las métricas de evaluación para tareas de clasificación, y finalmente las variantes de predicción conforme aplicadas a la clasificación y las métricas asociadas a la predicción conforme. A continuación, en la sección 3.3 se describen las configuraciones experimentales utilizadas en la evaluación de la predicción conforme para tareas de clasificación. Finalmente, en la sección 3.4 se exponen los resultados de la predicción conforme en este tipo de tareas, y en la sección 3.5, se presentan una breve discusión de resultados. Para culminar, en la sección 3.6 se relatan las consideraciones finales.

3.1. Materiales

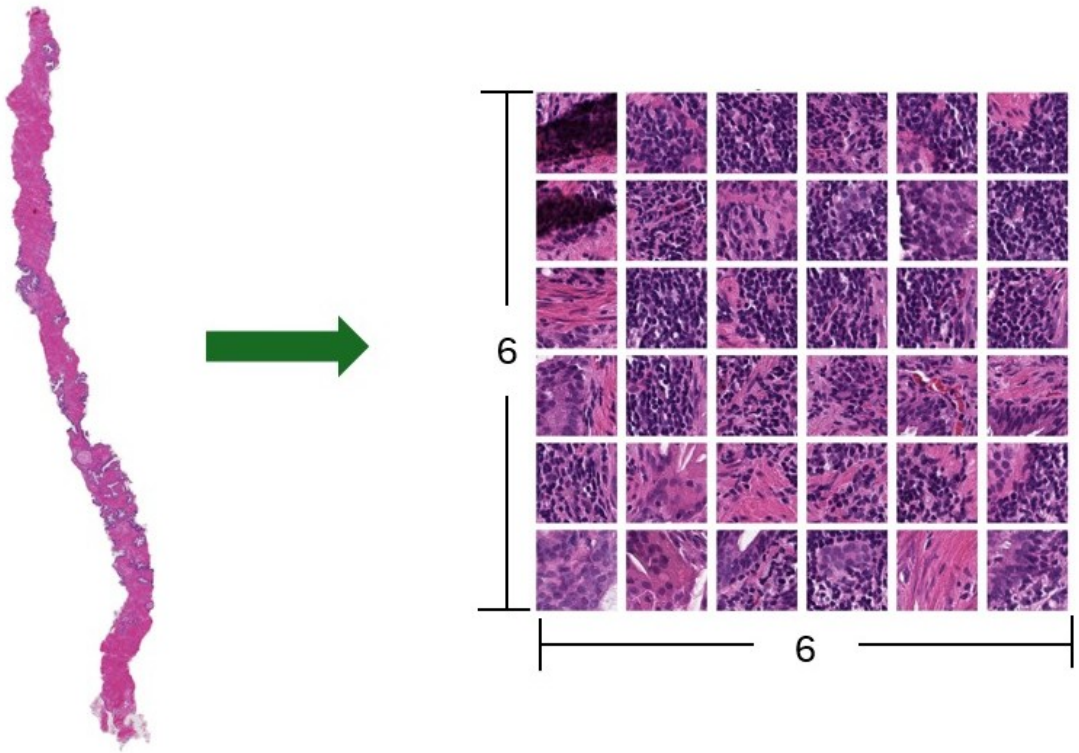
El acceso a bases de datos abiertas en histopatología ha sido clave para el avance de la patología digital, particularmente en el desarrollo de modelos predictivos que usan imágenes y etiquetas necesarias para entrenar y validar modelos de aprendizaje automático. Estas bases de datos abiertas promueven la replicabilidad y la comparación de metodologías.

Entre las bases de datos más relevantes en la actualidad está el proporcionado para el PANDA challenge (Prostate cANcer Grade Assessment), que se ha consolidado como uno de los recursos más influyentes en la comunidad de la patología digital. Este desafío internacional reunió más de 10.000 biopsias de cáncer de próstata provenientes de dos centros médicos líderes(Radbound University Medical Center y Karolinska Institute). Las biopsias fueron anotadas por patólogos expertos bajo los sistemas de graduacioń histológica ISUP (International Society of Urological Pathology) y Gleason (Kaggle 2020). El objetivo principal del PANDA challenge fue desarrollar algoritmos para la graduación automática de la severidad del cáncer de próstata, que es una tarea con impacto en el diagnóstico oportuno y la elección del tratamiento clínico más adecuado.

En la figura 4 se observa un ejemplo de una WSI del PANDA Challenge y su co-

rrespondiente extracción de mosaicos realizada en el preprocesamientode las imagenes del PANDA Challenge.

Figura 4: *Ejemplo WSI del Panda Challenge y la extracción de sus mosaicos*



El Gleason Score es un sistema de clasificación en el cual los patólogos asignan un grado del 1 al 5 a las dos áreas más predominantes del tumor y luego suman estos dos grados para tener así el valor total. Por ejemplo, un puntaje de 7 puede ser producto de gleason score de 4+3 o de 3+4. Esta es una de las limitaciones del sistema, ya que a pesar de ambos sumar 7, la disposición de los patrones tiene un pronóstico diferente ya que el 4 es más agresivo que el 3. Por otro lado, el ISUP grade es un sistema de puntuación que surge de la motivación de mejorar la claridad que faltaba en el Gleason score. Este nuevo sistema clasifica los tumores de próstata en una escala de 1 a 5, siendo el grado 1 el menos agresivo y 5 el más agresivo. Este sistema se basa en la evaluación del patrón arquitectónico de las células cancerosas y es una simplificación y mejora del sistema de Gleason. El ISUP se utiliza para dar un pronóstico y orientar las decisiones terapéuticas y, a diferencia de la puntuación Gleason, que puede ser más compleja, la escala ISUP tiene una mejor correlación con los resultados clínicos. La siguiente tabla muestra la correlación directa entre el sistema de clasificación de Gleason y la nueva escala simplificada del Grupo de la Sociedad Internacional de Uropatología (ISUP).

Tabla 2: *Correlación entre el Sistema de Puntuación de Gleason y los Grupos de Grado ISUP (International Society of Urological Pathology)*

Grupos de Grado ISUP	Puntuación de Gleason	Riesgo
Grupo de Grado 1	Puntuación de Gleason ≤ 6	Bajo Riesgo
Grupo de Grado 2	Puntuación de Gleason $3+4=7$	Riesgo Intermedio
Grupo de Grado 3	Puntuación de Gleason $4+3=7$	Riesgo Intermedio Desfavorable
Grupo de Grado 4	Puntuación de Gleason $4+4=8$; $5+3=8$; $3+5=8$	Alto Riesgo
Grupo de Grado 5	Puntuación de Gleason $4+5=9$; $5+4=9$; $5+5=10$	Alto Riesgo

Debido a la facilidad de uso del ISUP grade, y que este retiene importante información clínica para el diagnóstico, en este trabajo se consolidó como la etiqueta de referencia sobre el cual se organizó el conjunto de datos y es la utilizada en esta investigación. Además del campo ISUP grade, el dataset incluye la columna `data_provider`, que señala la institución de procedencia de cada caso (Radboud University Medical Center o Karolinska Institute).

3.2. Métodos

3.2.1. Arquitecturas de redes neuronales - Clasificación

Para la tarea de clasificación se optó por utilizar las arquitecturas EfficientNet-B2 y DenseNet-121. Estas redes han demostrado resultados competitivos en diversos escenarios de clasificación de imágenes y constituyen modelos adecuados para las tareas requeridas en esta investigación (Rasheed et al. 2025).

EfficintNet-B2 se desarrolla a partir de técnicas de búsqueda automática de arquitecturas y optimización basada en compound scaling. Esta versión representa un escalamiento superior respecto a la arquitectura inicial EfficientNet-B0, aumentando la profundidad, ancho y resolución de entrada de la red de una forma equilibrada. Son estas mejoras las que le dan una mayor capacidad de representación y precisión que la hacían una arquitectura

deseable para esta tarea de graduación automática de cáncer de próstata.

Por otro lado, DenseNet-121 corresponde a una de las variantes más utilizadas dentro de la familia DenseNet, cuyo diseño se distingue por su uso de conexiones densas entre capas. Bajo este esquema, cada capa recibe como entrada la salida de todas las anteriores, lo que potencia la propagación de gradientes y favorece la reutilización de representaciones.

3.2.2. Métricas de evaluación para tareas de clasificación

La evaluación del desempeño de los modelos de clasificación se basa en distintas métricas que permiten cuantificar su precisión y capacidad predictiva. En clasificación, una métrica fundamental es la exactitud (accuracy), que mide la proporción de predicciones correctas sobre el total de observaciones:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Sin embargo, cuando las clases están desbalanceadas, la exactitud puede ser engañosa. Por esta razón se emplean métricas adicionales, como la precisión, que evalúa la proporción de verdaderos positivos entre todas las predicciones positivas:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

La matriz de confusión constituye otra herramienta esencial, pues muestra la distribución de aciertos y errores por clase. No obstante, estas métricas tradicionales presentan limitaciones en problemas donde los errores no son equivalentes, como ocurre en la clasificación ordinal.

Este trabajo se centra en la clasificación del grado de cáncer de próstata según el sistema ISUP, un problema de naturaleza ordinal donde las clases guardan una relación jerárquica (grado 0 es más cercano a 1 que a 5). En este contexto, confundir grados adyacentes resulta menos crítico que errores entre extremos de la escala. Para cuantificar adecuadamente este comportamiento, se emplea el coeficiente de concordancia cuadrática ponderada (Quadratic Weighted Kappa en inglés - QWK), que penaliza los errores en función de su magnitud. La QWK se define como:

$$\text{QWK} = 1 - \frac{\sum_{i,j} W_{i,j} O_{i,j}}{\sum_{i,j} W_{i,j} E_{i,j}} \tag{3.1}$$

donde:

- O representa la matriz de confusión observada

- E denota la matriz esperada por azar
- W es la matriz de ponderación cuadrática: $W_{i,j} = \frac{(i-j)^2}{(N-1)^2}$

El carácter cuadrático de la penalización asegura que errores entre clases distantes (por ejemplo grado 0 vs. 4) impacten más severamente en la métrica que desviaciones menores (por ejemplo grado 1 vs. 2). Como ilustra la Figura 5, dos modelos con idéntica exactitud pueden obtener QWK sustancialmente diferentes según la distribución de sus errores. Mientras la matriz izquierda muestra errores entre clases vecinas, la derecha presenta errores más graves entre clases distantes, lo que se traduce en una penalización QWK significativamente mayor.

Figura 5: *QWK penaliza más los errores entre clases distantes. Dos modelos con la misma exactitud pueden diferir en QWK: la matriz izquierda muestra errores leves (poca penalización), mientras que la derecha concentra errores graves (mayor penalización).*

Esta sensibilidad a la gravedad de los errores convierte a QWK en la métrica idónea para evaluar modelos de graduación histopatológica, donde las consecuencias clínicas de subestimar un grado alto son considerablemente más graves que errores entre grados adyacentes.

3.2.3. Predicción conforme para clasificación

En la predicción conforme aplicada a clasificación, el modelo no se limita a proporcionar una única clase basada en la mayor probabilidad del softmax, sino que genera un set de predicciones . Este set asegura, con un nivel de confianza especificado, que la clase verdadera se encuentra dentro de él. Como se ilustra en la Figura 3, el proceso de predicción conforme

sigue una serie de pasos bien definidos, desde el entrenamiento del modelo hasta la construcción del set de predicciones. En la Figura 6, se muestran tres ejemplos de predicciones conformes que evidencian cómo el tamaño del set conforme varía: a medida que el modelo tiene menos certeza, el set se expande, incluyendo más clases posibles; mientras que cuando el modelo está más seguro de su predicción, el set se reduce, conteniendo solo la clase más probable.

Esto refleja cómo el modelo ajusta sus predicciones dependiendo de la incertidumbre asociada a cada caso. Esta capacidad de los conjuntos conformes para reflejar la incertidumbre del modelo está estrechamente vinculada a las métricas que determinan qué tan bien se ajusta una nueva observación a los datos previamente calibrados. Entre estas métricas, el Non-Conformity Score (puntaje de no conformidad) ya que permite evaluar cuán atípica o diferente es una predicción en comparación con las observaciones del conjunto de calibración. En términos simples, refleja qué tan "diferente" es una nueva observación respecto a las observaciones anteriores en el conjunto de calibración (Shafer y Vovk 2008). Un valor alto en este puntaje indica que la predicción del modelo es menos confiable, ya que el nuevo ejemplo no se parece a los datos anteriores.

Figura 6: *Predicción Conforme en Clasificación. Adaptado de Angelopoulos y Bates (2021)*

Por otro lado, también es posible trabajar con el concepto complementario, conocido como puntaje de conformidad (conformity score), el cual puede interpretarse como una medida de similitud o familiaridad con los datos previos. La relación entre ambos puntajes puede expresarse de manera simple como:

$$C(x, y) = 1 - \alpha(x, y)$$

donde:

- $\alpha(x, y) \in [0, 1]$ es la función de no conformidad, y
- $C(x, y) \in [0, 1]$ es la función de conformidad, su complemento.

La predicción conforme es una técnica versátil que puede adaptarse a diferentes escenarios dependiendo de las características del conjunto de datos y los objetivos del análisis. Estas adaptaciones permiten optimizar el equilibrio entre eficiencia computacional, precisión y personalización de las predicciones. Entre las distintas variantes se encuentran la predicción conforme inductiva, la predicción conforme de Mondrian y conjuntos de predicciones adaptativos que fueron utilizadas en la investigación y se describen a continuación.

3.2.3.1. Predicción conforme inductiva. La predicción conforme inductiva (Inductive Conformal Prediction en inglés – ICP) surge como una solución a la principal limitación de la predicción conforme clásica o transductiva (TCP, por sus siglas en inglés): su ineficiencia computacional. En el enfoque transductivo, cada nueva predicción requiere volver a ejecutar el procedimiento completo, lo cual resulta impráctico para algoritmos con tiempos de entrenamiento prolongados, como las redes neuronales.

Para superar esta limitación, ICP reemplaza la inferencia transductiva por inferencia inductiva, dividiendo el conjunto de datos en dos subconjuntos: uno de entrenamiento y otro de calibración. El modelo se entrena únicamente sobre el subconjunto de entrenamiento, evitando la necesidad de reentrenarlo en cada predicción. Posteriormente, el subconjunto de calibración se utiliza para calcular los puntajes de conformidad (non-conformity scores) y, a partir de ellos, los valores p para cada posible clasificación o regresión (Papadopoulos 2008). Como se observa en el Algoritmo 1, los ejemplos en el conjunto de calibración permiten cuantificar la incertidumbre del modelo en nuevos datos de manera eficiente.

De este modo, ICP conserva las garantías de validez de la predicción conforme, pero con una eficiencia computacional comparable a la del algoritmo subyacente. Esto lo convierte en un enfoque especialmente adecuado para escenarios donde el entrenamiento es costoso, como en el caso de las redes neuronales, ampliando de manera significativa la aplicabilidad práctica de la predicción conforme (Vovk 2012).

Algoritmo 1: Predicción Conforme Inductiva (ICP)

Entrada: Conjunto de datos $\mathcal{D} = \{(X_i, y_i)\}_{i=1}^{N}$, nivel de confianza $1 - \alpha$, función de conformidad $S(y, \hat{y})$

Salida: Salida conforme para nuevas muestras

1. División del conjunto de datos:

Dividir $\mathcal{D}$ en tres subconjuntos:

Entrenamiento $\mathcal{D}_{\text{train}}$ para ajustar el modelo.

Calibración $\mathcal{D}_{\text{cal}}$ para calcular los puntajes de conformidad.

Prueba $\mathcal{D}_{\text{test}}$ donde se aplicará la predicción conforme.

2. Entrenamiento del modelo:

Entrenar el modelo f con $\mathcal{D}_{\text{train}}$.

3. Calibración:

Para cada *muestra* $(X_i, y_i) \in \mathcal{D}_{cal}$ **hacer**

  Obtener la predicción $\hat{y}_i = f(X_i)$.

  Calcular el puntaje de conformidad $S(y_i, \hat{y}_i)$.

fin

Ordenar los puntajes y calcular el cuantil $q_{1-\alpha}$.

4. Aplicación de la predicción conforme:

Para cada *nueva muestra* X_{new} *en* $\mathcal{D}_{test}$ **hacer**

  Obtener la predicción $\hat{y}_{\text{new}} = f(X_{\text{new}})$.

  Para cada clase $y' \in \mathcal{Y}$, calcular el puntaje de conformidad $S(y', \hat{y}_{\text{new}})$.

  Si $S(y', \hat{y}_{\text{new}}) \geq q_{1-\alpha}$ **entonces**

    Incluir la clase y' en el set conforme $\mathcal{O}_{\text{conforme}}$.

  Fin si

fin

3.2.3.2. Predicción conforme de Mondrian. Por otra parte, la predicción conforme de Mondrian (Mondrian Conformal Prediction en inglés - MCP), descrita en el Algoritmo 2, adapta los intervalos de predicción de acuerdo con grupos específicos dentro del conjunto de datos, conocidos como categorías Mondrianas. En lugar de aplicar un único puntaje de conformidad a todas las observaciones, este enfoque segmenta el conjunto de datos en subconjuntos basados en características específicas (por ejemplo, diferentes clases o regiones en el espacio de características). De acuerdo con la literatura, en la Predicción Conforme de Mondrian (MCP), cada etiqueta o clase se trata de forma separada, evaluando la confianza en la asignación de instancias de manera independiente. Esto permite controlar la tasa de error dentro de cada grupo, al dividir el conjunto de entrenamiento en categorías y asignar un nivel de significancia a cada una de ellas (Vazquez y Facelli 2022).

Esta segmentación posibilita que los intervalos de predicción se ajusten de manera más precisa para cada categoría, mejorando tanto la interpretación como la confiabilidad del modelo, especialmente en aplicaciones donde la heterogeneidad de los datos es significativa.

En escenarios con clases variadas o características diferenciadas, la MCP facilita predicciones más adaptadas y precisas para cada tipo de observación. Además, resulta particularmente útil en conjuntos de datos desbalanceados, ya que compara los puntajes de conformidad solo entre aquellos de la misma categoría y no a través de todo el conjunto de entrenamiento.

Algoritmo 2: Predicción Conforme Mondrian (MCP)

Entrada: Conjunto de datos $\mathcal{D} = \{(X_i, y_i)\}_{i=1}^{N}$, nivel de confianza $1 - \alpha$, función de conformidad $S(y, \hat{y})$, función de partición $g(X_i)$

Salida: Salida conforme para nuevas muestras

1. División del conjunto de datos:

Dividir $\mathcal{D}$ en tres subconjuntos:

- Entrenamiento $\mathcal{D}_{\text{train}}$ para ajustar el modelo.
- Calibración $\mathcal{D}_{\text{cal}}$ para calcular los puntajes de conformidad.
- Prueba $\mathcal{D}_{\text{test}}$ donde se aplicará la predicción conforme.

2. Entrenamiento del modelo:

Entrenar el modelo f con $\mathcal{D}_{\text{train}}$.

3. Agrupación por categorías (Mondrian):

Para cada *muestra* $(X_i, y_i) \in \mathcal{D}_{cal}$ **hacer**
  Asignar la muestra a un grupo G_k según $g(X_i)$.
fin

4. Calibración por grupo:

Para cada *grupo* G_k **hacer**
  Para cada *muestra* $(X_i, y_i) \in G_k$ **hacer**
    Obtener la predicción $\hat{y}_i = f(X_i)$.
    Calcular el puntaje de conformidad $S(y_i, \hat{y}_i)$.
  fin
  Ordenar los puntajes y calcular el cuantil $q_{1-\alpha}^{(k)}$.
fin

5. Aplicación de la predicción conforme:

Para cada *nueva muestra* X_{new} **hacer**
  Determinar su grupo G_k según $g(X_{\text{new}})$.
  Obtener la predicción $\hat{y}_{\text{new}} = f(X_{\text{new}})$.
  Para cada clase $y' \in \mathcal{Y}$, calcular el puntaje de conformidad $S(y', \hat{y}_{\text{new}})$.
  Si $S(y', \hat{y}_{\text{new}}) \geq q_{1-\alpha}^{(k)}$ **entonces**
    Incluir la clase y' en el set conforme $\mathcal{O}_{\text{conforme}}$.
  Fin si
fin

3.2.3.3. Predicción conforme de conjuntos adaptativos. La predicción conforme de conjuntos adaptativos (Adaptive Prediction Sets en inglés - APS) utiliza directamente los puntajes de salida del modelo (por ejemplo, las probabilidades de la función *softmax*) para cada clase. Dado un nivel de confianza $(1 - \alpha)$, el procedimiento consiste en ordenar las clases según la probabilidad predicha y, posteriormente, acumular dichas probabilidades hasta alcanzar o superar un umbral definido por el cuantil de calibración. El conjunto de predicción final está compuesto por todas las clases que fueron necesarias para superar este umbral.

Formalmente, el conjunto de predicción se define como:

$$C(x) = \{\pi_1(x), \ldots, \pi_k(x)\}, \quad \text{donde } k = \sup\left\{k' : \sum_{j=1}^{k'} \hat{f}(x)_{\pi_j(x)} < \hat{q}\right\} + 1.$$

En esta formulación, $C(x)$ representa el conjunto de predicción asociado a la instancia x. Cada elemento $\pi_j(x)$ corresponde a la clase en la posición j, luego de ordenar todas las clases según la probabilidad estimada en orden descendente. La probabilidad predicha por el modelo para la clase $\pi_j(x)$ se denota como $\hat{f}(x)_{\pi_j(x)}$. El parámetro $\hat{q}$ corresponde al cuantil de calibración obtenido a partir del conjunto de calibración y está asociado al nivel de error α. El operador $\sup\{k' : \ldots\}$ indica el mayor valor de k' tal que la suma acumulada de probabilidades se mantiene estrictamente por debajo de $\hat{q}$. Finalmente, el término $+1$ asegura que se incluya también la primera clase cuya incorporación permite que la suma acumulada alcance o supere dicho umbral.

Algoritmo 3: Adaptive Prediction Sets (APS)

Entrada: Conjunto de datos $\mathcal{D} = \{(X_i, y_i)\}_{i=1}^{N}$, nivel de confianza $1 - \alpha$, predicciones de probabilidad $\hat{p}(y|X)$

Salida: Conjuntos de predicción adaptativos para nuevas muestras

1. División del conjunto de datos:

Dividir $\mathcal{D}$ en tres subconjuntos:

Entrenamiento $\mathcal{D}_{\text{train}}$ para ajustar el modelo.

Calibración $\mathcal{D}_{\text{cal}}$ para calcular los umbrales adaptativos.

Prueba $\mathcal{D}_{\text{test}}$ donde se aplicará la predicción conforme.

2. Entrenamiento del modelo:

Entrenar el modelo f con $\mathcal{D}_{\text{train}}$.

3. Calibración adaptativa:

Para cada *muestra* $(X_i, y_i) \in \mathcal{D}_{cal}$ **hacer**

- Obtener el vector de probabilidades $\hat{p}(y|X_i)$.
- Ordenar las clases de mayor a menor probabilidad.
- Calcular la *función de cobertura acumulada* al ir sumando probabilidades hasta incluir y_i.
- Registrar el valor de corte adaptativo asociado a y_i.

fin

Calcular el cuantil $q_{1-\alpha}$ sobre los valores de corte registrados.

4. Aplicación de APS:

Para cada *nueva muestra* X_{new} *en* $\mathcal{D}_{test}$ **hacer**

- Obtener el vector de probabilidades $\hat{p}(y|X_{\text{new}})$.
- Ordenar las clases en orden descendente de probabilidad.
- Construir el conjunto de predicción sumando clases hasta alcanzar el umbral $q_{1-\alpha}$.

fin

3.2.4. Métricas de predicción conforme para la tarea de clasificación

En esta sección se describen las principales métricas utilizadas para evaluar la predicción conforme en los experimentos realizados. Las métricas consideradas son la cobertura, cobertura estratificada por características (Feature Stratified Coverage en inglés - FSC), cobertura estratificada por tamaños (Size stratified coverage en inglés - SSC) y el N-Criterion.

Iniciando con la cobertura, esta mide la proporción de veces que la predicción conforme incluye el valor verdadero. Se espera que la cobertura empírica sea cercana al nivel de confianza especificado $(1 - \alpha)$.

En clasificación, la cobertura se calcula como la proporción de veces que la clase verdadera (grado ISUP) está dentro del conjunto conforme generado:

$$\text{Cobertura} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{y_i \in C(X_i)\},$$

donde $C(X_i)$ es el conjunto conforme para la muestra i, y $\mathbb{1}$ es la función indicadora que vale 1 si la clase verdadera está dentro del conjunto conforme, y 0 en caso contrario. Cabe resaltar que esta función también se utiliza en FSC y SSC asumiendo el mismo comportamiento.

La métrica FSC (Feature Stratified Coverage) evalúa la uniformidad de la cobertura a través de subgrupos de datos predefinidos (por ejemplo, diferentes instituciones o categorías demográficas). Destaca los posibles sesgos o la falta de adaptabilidad cuando el rendimiento predictivo varía entre grupos. Específicamente, el FSC se define como la cobertura empírica mínima entre todos los subgrupos G. De esta manera se asegura que la predicción conforme no solo mantenga la cobertura global deseada, sino que también sea confiable de forma consistente a lo largo de todos los subgrupos de datos.:

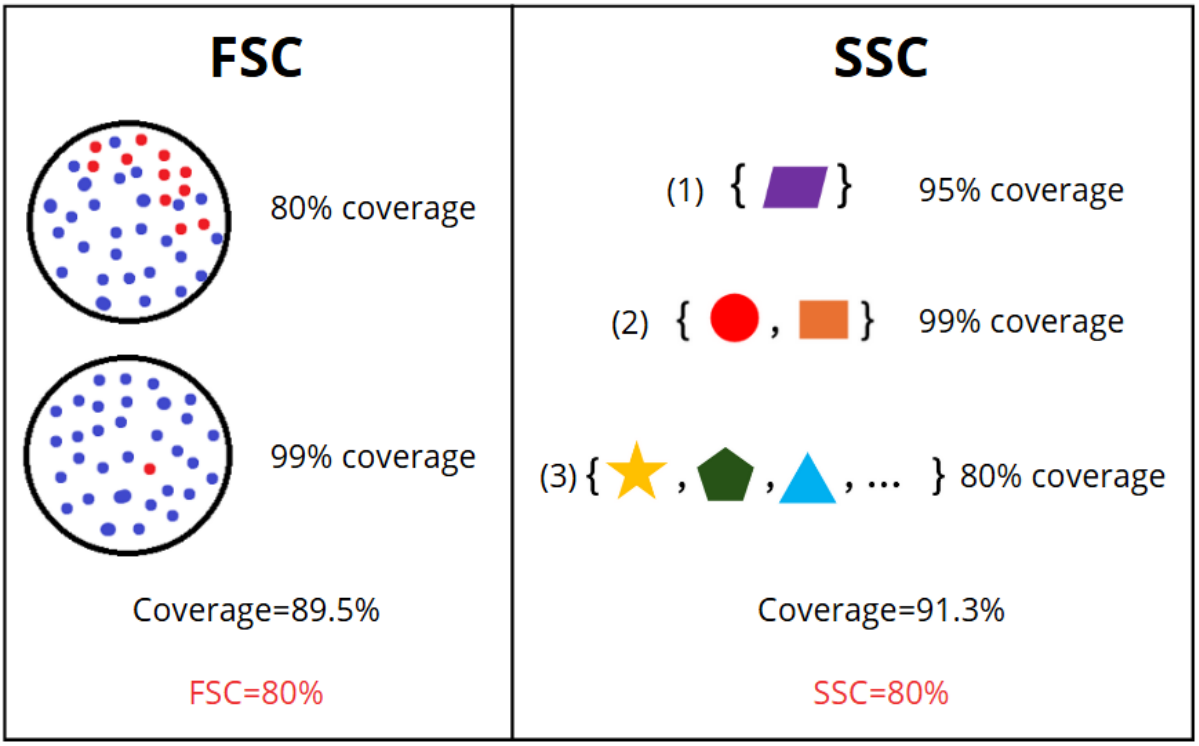
$$\text{Métrica FSC} \;=\; \min_{g\in\{1,\ldots,G\}} \frac{1}{|\mathcal{I}_g|}\sum_{i\in\mathcal{I}_g} \mathbb{1}\Big\{ Y_i^{(\text{val})} \in \mathcal{C}\Big(X_i^{(\text{val})}\Big)\Big\}. \tag{3.2}$$

Paralelamente, La métrica SSC (Size stratified Coverage) mide la consistencia de la cobertura a través de diferentes tamaños de conjuntos de predicción. Es decir, estratifica los ejemplos de validación basándose en el número de etiquetas en sus respectivos conjuntos de predicción (por ejemplo, predicciones de una sola etiqueta, dos etiquetas y múltiples etiquetas), y luego calcula la cobertura mínima entre estos estratos. Un SSC bajo indica que la cobertura del método se deteriora en ciertos regímenes de confianza:

$$\text{SSC} \;=\; \min_{g\in\{1,\ldots,G\}} \frac{1}{|\mathcal{I}_g|}\sum_{i\in\mathcal{I}_g} \mathbb{1}\Big\{ Y_i^{(\text{val})} \in \mathcal{C}\Big(X_i^{(\text{val})}\Big)\Big\}. \tag{3.3}$$

Tanto FSC como SSC se centran en la cobertura del peor caso dentro de particiones definidas, ofreciendo una perspectiva complementaria a la cobertura global y ayudando a evaluar la equidad y fiabilidad bajo la variabilidad de los subgrupos y se ilustran en la Figura 7.

Figura 7: *Comparación entre FSC (izquierda) y SSC (derecha). El FSC mide la cobertura mínima entre subgrupos, mientras que el SSC agrupa los conjuntos por tamaño (1, 2 o 3+ clases) y toma la cobertura mínima en cada grupo.*



N-criterion: Este mide el tamaño promedio de los sets generados:

$$\text{N-Criterion} = \frac{1}{N}\sum_{i=1}^{N} |C(X_i)|.$$

Un valor bajo de N-Criterion indica que el modelo es más informativo y refleja una alta confianza en sus predicciones.

3.3. Configuración Experimental

En esta sección se comenzará describiendo el preprocesamiento aplicado al conjunto de imágenes. Posteriormente, se explicará la configuración del entrenamiento de los modelos propuestos, para finalizar presentando la estructura de la experimentación, que se ha dividido en dos fases distintas

3.3.1. Preparación de los datos

En este estudio, las imágenes se sometieron a una serie de pasos diseñados para estandarizar el conjunto de datos, eliminar información irrelevante y resaltar las características de

interés. El flujo de trabajo se centró en la extracción de mosaicos (tiles) de alta calidad, que son representativos del tejido patológico, y en la creación de una representación compacta y uniforme de cada imagen.

3.3.2. Extracción y selección de mosaicos

El primer paso consistió en segmentar las imágenes de próstata en pequeñas secciones denominadas mosaicos o tiles. Este enfoque se justifica por la naturaleza de las imágenes de alta resolución, que a menudo superan la capacidad de memoria de las GPUs. Para ello, se implementó una estrategia de extracción que priorizaba el contenido del tejido, ya que grandes porciones de las imágenes originales consistían en áreas en blanco o de fondo sin valor diagnóstico.

- **Detección de tejido:** Se utilizó una máscara binaria para identificar y aislar las regiones con tejido. Esto se logró convirtiendo la imagen a escala de grises y aplicando un umbral (cv2.threshold) de 150. Los píxeles con una intensidad por debajo de este umbral se consideraron parte del tejido (negro), mientras que aquellos por encima se clasificaron como fondo (blanco). Esta máscara binaria se utilizó para diferenciar entre el tejido de interés y las áreas en blanco de la lámina. Posteriormente, se recortaron las áreas en blanco para optimizar el procesamiento.
- **Generación de mosaicos:** A partir de la región de tejido recortada, la imagen se dividió en una cuadrícula de mosaicos uniformes de 128x128 píxeles. Para garantizar una cobertura completa, la imagen se rellenó con bordes blancos (np.pad) si sus dimensiones no eran múltiplos del tamaño del mosaico.
- **Filtrado por contenido de tejido:** A diferencia de una división simple, se implementó un método de puntuación para cada mosaico. La puntuación se calculó basándose únicamente en el porcentaje de píxeles de tejido dentro de cada mosaico, es decir, el número de píxeles no blancos en la máscara correspondiente. Se seleccionaron los 36 mosaicos con la puntuación más alta, asegurando que cada muestra de datos contuviera las regiones más relevantes y ricas en información para el análisis posterior.

3.3.3. Composición de imágenes uniformes

Una vez seleccionados los 36 mosaicos más relevantes para cada muestra, se procedió a combinarlos para crear una nueva imagen única y de tamaño estandarizado. Este paso fue crucial para alimentar de manera eficiente la red neuronal, que requiere entradas con dimensiones uniformes.

- **Creación de la cuadrícula:** Los 36 mosaicos seleccionados, cada uno con un tamaño de 128x128 píxeles, se organizaron en una cuadrícula de 6x6.
- **Ensamblaje:** Los mosaicos se unieron en un lienzo en blanco para formar una imagen combinada final con una resolución de 768x768 píxeles. Este proceso garantiza que todas las imágenes del conjunto de datos tengan la misma estructura y dimensiones, lo que facilita el entrenamiento del modelo.

Después de organizar las imágenes compuestas en carpetas de acuerdo con su grado ISUP, se descartaron aquellas que no contaban con suficiente tejido para generar los 36 mosaicos requeridos, lo cual aseguró un conjunto de entrenamiento más consistente y de mejor calidad.

Para la aplicación de la Predicción Conforme de Mondrian y la evaluación detallada de la cobertura condicional FSC, fue necesario definir criterios de estratificación (o subgrupos) sobre el dataset. Si bien se disponía de la variable Data Provider, se buscó enriquecer la experimentación con un criterio adicional que reflejara las condiciones reales de calidad de las imágenes. Con este propósito, se definió una segunda variable de estratificación basada en el nivel de nitidez o desenfoque de cada imagen. Para este cálculo, se convirtió la imagen a la escala de grises y se aplicó el operador laplaciano, el cual se encarga de resaltar los bordes. La varianza de este Laplaciano otorga un valor escalar, donde valores altos indican imágenes más nítidas, mientras que valores más bajos indican imágenes más borrosas. A partir de la distribución de este puntaje, se definieron tres grupos de nitidez según los cuantiles Q_{25} y Q_{75}:

$$\text{Blurry} \quad (s \leq Q_{25}), \qquad \text{Moderate} \quad (Q_{25} < s < Q_{75}), \qquad \text{Sharp} \quad (s \geq Q_{75}),$$

donde s corresponde al puntaje de desenfoque calculado para cada imagen.

3.3.4. Configuración de los modelos de clasificación

El proceso de configuración de los modelos de clasificación se organizó en varias etapas con el fin de mantener una estructura clara. A continuación, se detallan los pasos seguidos.

Se seleccionaron las arquitecturas EfficientNet-B2 y DenseNet-121 por su reconocida capacidad para capturar patrones jerárquicos en imágenes médicas. Los modelos fueron inicializados con pesos preentrenados en ImageNet, garantizando una base sólida de representaciones visuales generales. Posteriormente, se reemplazó la capa de salida de cada red por un head de clasificación ajustado al número de clases de la tarea (seis grados ISUP).

El conjunto de datos se dividió en tres subconjuntos: 75 % para entrenamiento, 8.75 %

para validación (también utilizado como calibración en predicción conforme) y 16.25 % para prueba de la predicción conforme.Para mejorar la robustez del modelo y reducir el sobreajuste, se aplicaron aumentaciones con la librería Albumentations. Estas incluyeron volteo horizontal y vertical, rotaciones aleatorias de 90 grados y la transformación ShiftScaleRotate, que introduce variaciones en posición, escala y rotación. Todas las imágenes fueron normalizadas utilizando los valores estándar de media y desviación de ImageNet.

Dado el predominio de las clases 0 y 1 en el conjunto de datos, se implementó un muestreo ponderado mediante WeightedRandomSampler. Esta técnica asigna pesos inversamente proporcionales a la frecuencia de cada clase, asegurando que las clases minoritarias tengan mayor representación en cada época de entrenamiento.Los conjuntos fueron organizados en DataLoaders de PyTorch. El entrenamiento utilizó lotes de tamaño(batch size en inglés) 16 y 4 trabajadores (workers en inglés)

Para la función de perdida se utilizó la perdida de entropía cruzada (Cross-Entropy Loss en inglés) combinada con Label Smoothing ($\epsilon = 0.1$) como técnica de regularización para reducir la confianza excesiva del modelo y mejorar su generalización. La pérdida base para una muestra i con K clases es:

$$\ell_{\text{CE}} = -\sum_{k=1}^{K} y_{i,k} \log(\hat{p}_{i,k})$$

Tras aplicar suavizado, las etiquetas se transforman en:

$$\tilde{y}_{i,k} = \begin{cases} 1 - \epsilon + \dfrac{\epsilon}{K}, & \text{si } k = \text{clase verdadera} \\ \dfrac{\epsilon}{K}, & \text{si } k \neq \text{clase verdadera} \end{cases}$$

y la pérdida final sobre un lote de N muestras queda:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \left(-\sum_{k=1}^{K} \tilde{y}_{i,k} \log(\hat{p}_{i,k}) \right).$$

Para el optimizador se utilizó SGD con momentum de 0.9 y decaimiento de peso de 4×10^{-3} para EfficientNet y 5×10^{-2} para DenseNet. El aprendizaje se controló mediante el planificador (scheduler en inglés)ReduceLROnPlateau, configurado con factor 0.1 y paciencia de 4–5 épocas, según el modelo. En la Tabla 3 se visualizan las diferencias clave de las configuraciones para DenseNet y EfficientNet.

Tabla 3: *Configuración de Optimización para Modelos EfficientNet y DenseNet*

Arquitectura	Modelo	Optimizador	Scheduler
EfficientNet	EfficientNet-B2	SGD con momentum=0.9, weight 4e-3	ReduceLROnPlateau con factor=0.1, patience=4
DenseNet	DenseNet-121	SGD con momentum=0.9, weight 5e-2	ReduceLROnPlateau con factor=0.1, patience=5

Finalmente, entrenamiento se implementó con PyTorch Lightning, integrando EarlyStopping (paciencia de 10 épocas) y ModelCheckpoint para guardar el mejor modelo. La ejecución se realizó en GPU con precisión mixta (16 bits) para optimizar recursos computacionales, y mediante CSVLogger se registró las métricas accuracy y QWK que fueron clave para determinar el écito de los modelos.

3.3.5. Configuración de la predicción conforme

La configuración experimental se diseñó con el objetivo de evaluar distintas variantes de predicción conforme(ICP, APS y MCP) sobre el conjunto de datos del PANDA Challenge. La aplicación de cada una de estas técnicas se llevó a cabo siguiendo rigurosamente los pasos detallados en la Sección 3.2. En todas las técnicas, el puntaje de conformidad utilizado corresponde a las probabilidades predichas por el modelo a través de la función *softmax*, de acuerdo con la siguiente expresión:

$$s(x, y) = \hat{f}(x)_y,$$

donde $s(x, y)$ denota el puntaje de conformidad de la instancia x con respecto a la clase y, y $\hat{f}(x)_y$ corresponde a la probabilidad asignada por el modelo a dicha clase. Para el cálculo de SSC, los tamaños de conjuntos de predicción se agruparon en las categorías 1, 2, 3 y 4-5 (grados ISUP), consolidando los dos últimos debido a la baja frecuencia de muestras que generaban conjuntos de este tamaño. Adicionalmente, todos los casos el nivel de significancia (α) se estableció en 0.10, correspondiente a una cobertura objetivo del 90 %.

3.3.5.1. Experimento 1: Institución como Variable de Estratificación y Categoría Mondriana. En esta configuración, la institución de origen (`data_provider`) se empleó de manera unificada tanto como categoría Mondriana para MCP como variable de estratificación para el cálculo de FSC. Con ello se buscó evaluar simultáneamente la vali-

dez condicional y la equidad de los conjuntos de predicción en función de la procedencia institucional. La implementación metodológica fue:

- **ICP y APS:** Cálculo global del umbral de conformidad sobre el conjunto completo de calibración, con evaluación posterior de FSC condicionada por los subgrupos institucionales.
- **MCP:** Cálculo independiente de cuantiles conformes por institución, garantizando cobertura específica en cada grupo.

3.3.5.2. Experimento 2: Blurriness como Variable de Estratificación y Categoría Mondriana. En este segundo experimento, la variable blurriness_group (Blurry, Moderate, Sharp) se utilizó tanto como categoría Mondriana en MCP como variable de estratificación para FSC. Este diseño permitió analizar la validez de los conjuntos de predicción de acuerdo con la calidad de imagen, considerando de manera simultánea su impacto en la cobertura condicional y la equidad entre los distintos niveles de desenfoque.

3.4. Resultados

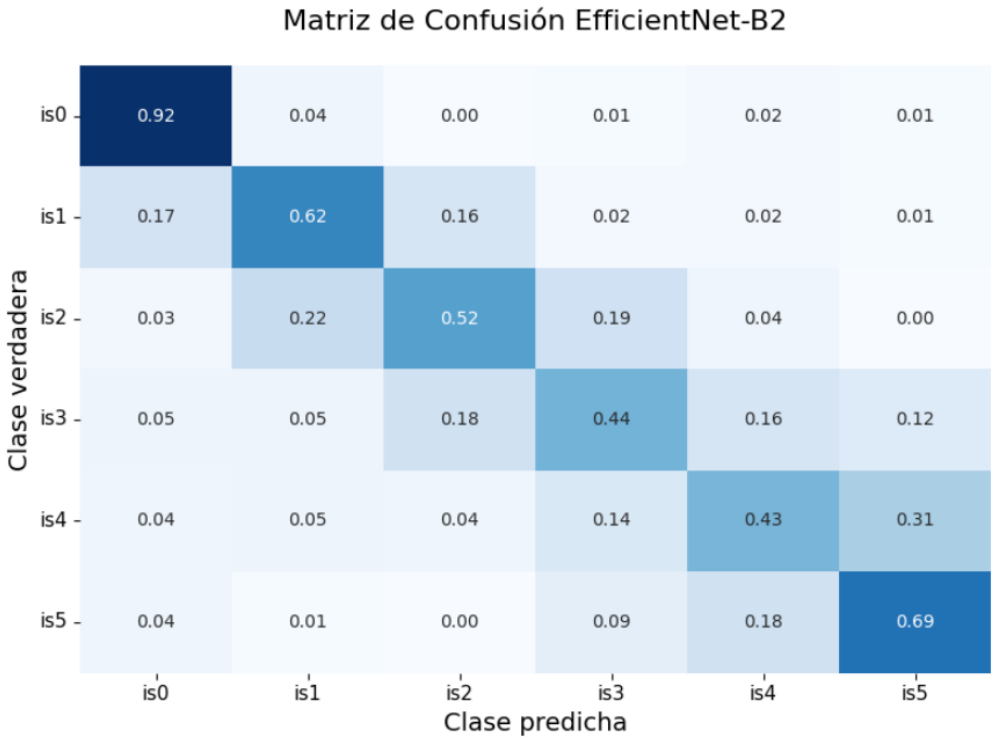
En esta sección se presentan los hallazgos obtenidos a lo largo del estudio. Para lograr una exposición más clara, los resultados se dividen en dos apartados principales: (i) el desempeño de los modelos base de redes neuronales antes de aplicar técnicas de predicción conforme, y (ii) la evaluación de dichos modelos bajo los distintos enfoques de predicción conforme considerados en este trabajo segmentados en los dos experimentos.

3.4.1. Estimación categórica puntual del grado ISUP

Antes de aplicar métodos de predicción conforme, se evaluó el rendimiento de las arquitecturas seleccionadas de Deep Learning sobre el conjunto de datos. Estos resultados permiten establecer una línea base frente a la cual se comparan posteriormente los efectos de la predicción conforme.

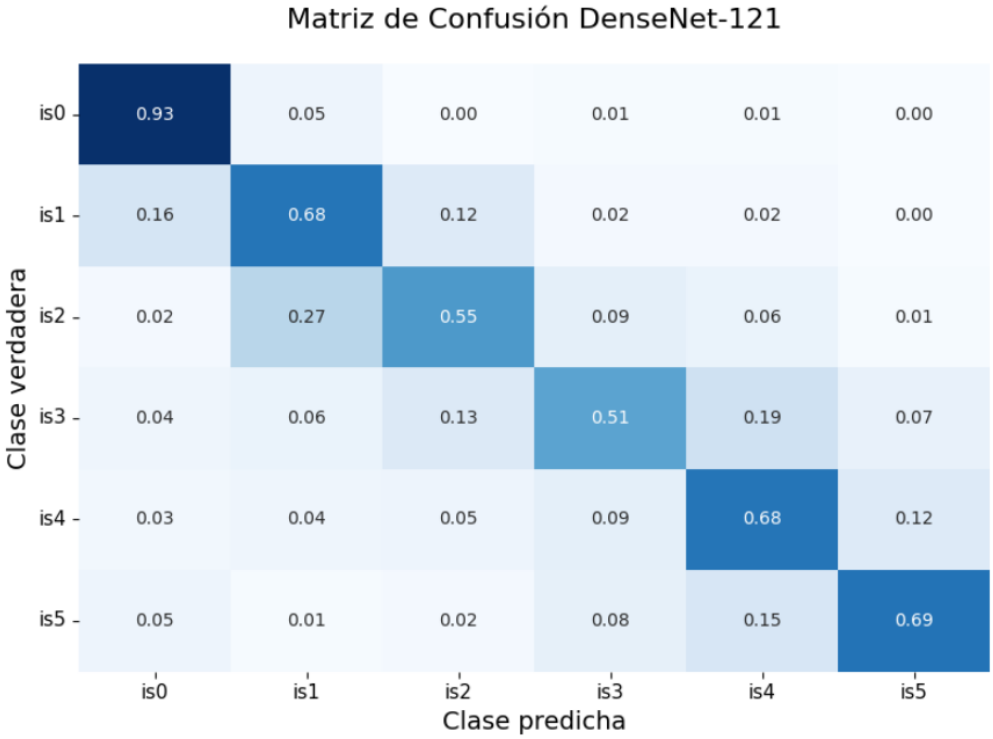
El modelo EfficientNet-B2 alcanzó una precisión de 0.6863 y un QWK de 0.8510. La matriz de confusión asociada, mostrada en la Figura 8, evidencia la distribución de los aciertos y errores de clasificación del modelo.

Figura 8: *Matriz de confusión del modelo EfficientNet-B2. Las etiquetas "isN" en los ejes corresponden a los grados ISUP N, donde N = 0,1,2,3,4,5*



Por su parte, el modelo DenseNet-121 obtuvo una precisión de 0.7071 y un QWK de 0.8485. La Figura 9 muestra la matriz de confusión correspondiente, que permite analizar de manera detallada el patrón de clasificaciones realizadas por este modelo.

Figura 9: *Matriz de confusión del modelo DenseNet-121. Las etiquetas "isN" en los ejes corresponden a los grados ISUP N, donde N = 0,1,2,3,4,5*



En síntesis, ambos modelos alcanzaron desempeños competitivos, aunque con ligeras diferencias: mientras que DenseNet-121 mostró una mayor precisión, EfficientNet-B2 presentó un QWK marginalmente superior. Estos resultados iniciales sirven como punto de partida para analizar cómo las técnicas de predicción conforme afectan y enriquecen la interpretación y confiabilidad de las predicciones.

3.4.2. Resultados con predicción conforme

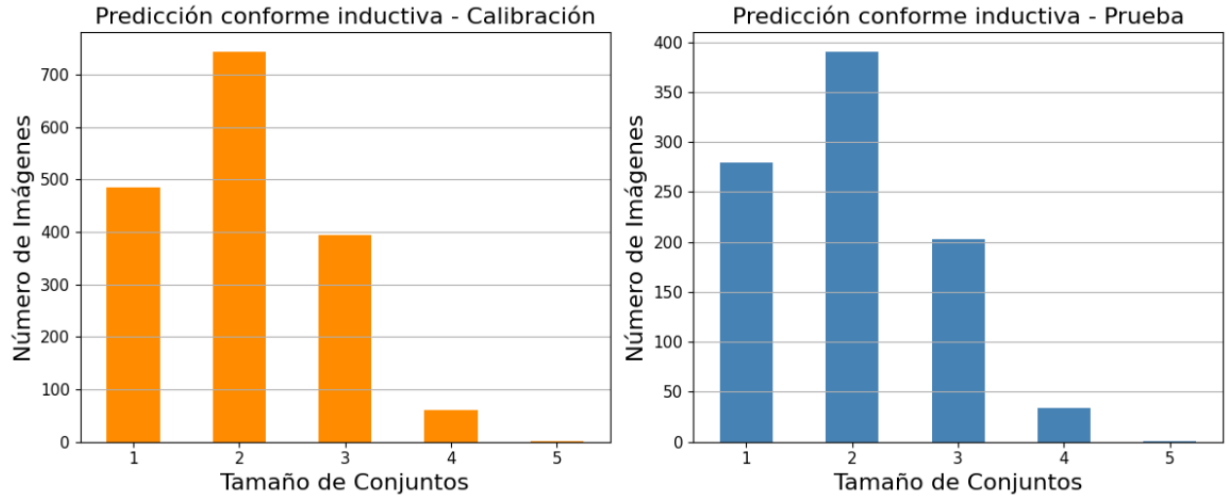
Una vez establecido el desempeño de las estimaciones puntuales de los modelos, se procedió a la aplicación de técnicas de predicción conforme con el objetivo de evaluar su impacto en la confiabilidad de las predicciones.

Como se indicó en la Sección 3.3, debido a que se plantearon dos configuraciones experimentales distintas —una basada en la variable institución y otra en blurriness_group—, los resultados se presentan de manera separada para cada experimento. En ambos casos se separaron las técnicas de Predicción Conforme Inductiva (ICP), Mondriana (MCP) y Conjuntos Adaptativos (APS), evaluando las métricas de cobertura, N-criterion, *Feature Stratified Coverage* (FSC) y *Size Stratified Coverage* (SSC) y a su vez comparando ambas arquitecturas

3.4.2.1. Experimento 1: Institución como Variable de Estratificación y Categoría Mondriana. En este experimento, la institución de origen se utilizó tanto como categoría Mondriana en MCP como variable de estratificación para FSC.

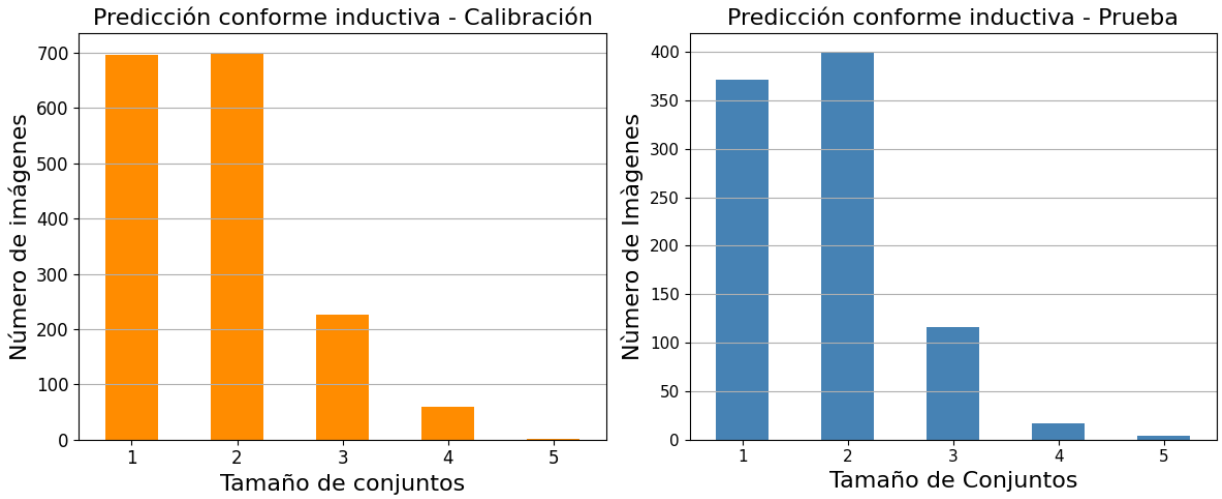
3.4.2.1.1. Predicción Conforme Inductiva. Para ICP, el modelo EfficientNet-B2 alcanzó una cobertura de 0.8831 con un N-criterion de 1.9945, lo que indica conjuntos conformes de tamaño moderado. La métrica FSC fue de 0.8393, mientras que la SSC obtuvo un valor de 0.8566 en el grupo de tamaño 1, reflejando un comportamiento consistente en subgrupos pequeños.

Figura 10: *Resultados de ICP en EfficientNet-B2*



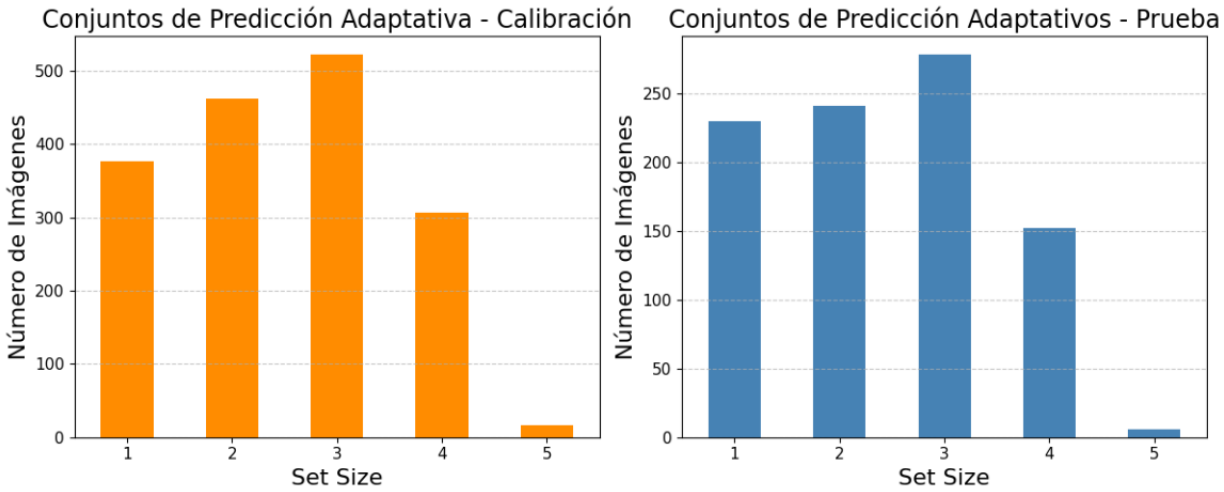
Por su parte, DenseNet-121 alcanzó una cobertura de 0.8953, más cercana al objetivo del 90 %, con un N-criterion de 1.8280 que indica mayor eficiencia. El FSC se ubicó en 0.8752 y la SSC en 0.8750 para el grupo 4-5, lo que sugiere que este modelo logra un mejor balance entre confiabilidad y tamaño de los conjuntos.

Figura 11: *Resultados de ICP en DenseNet-121*



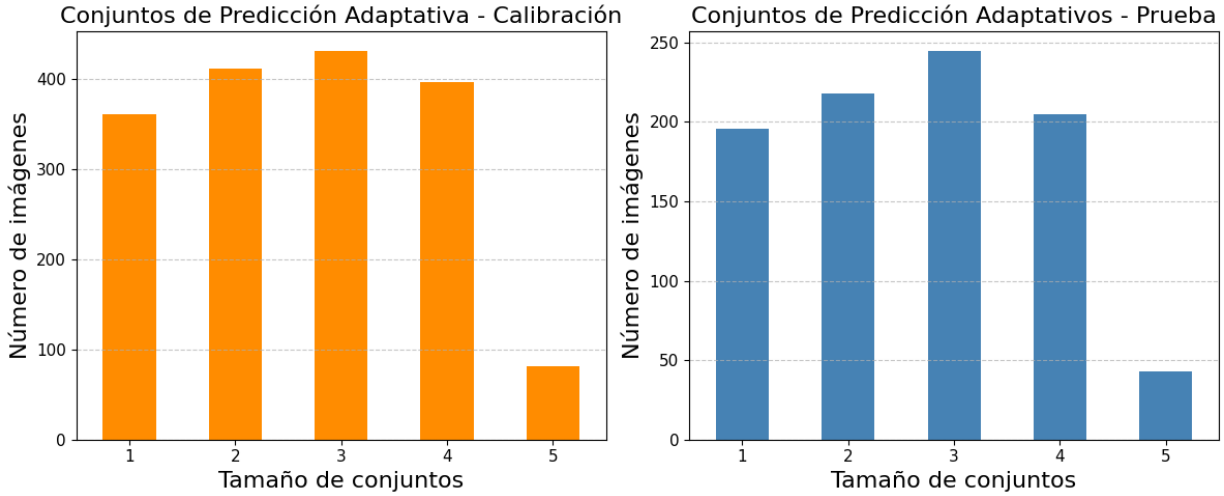
3.4.2.1.2. Predicción Conforme de Conjuntos Adaptativos. En APS, EfficientNet-B2 presentó la mayor cobertura entre las tres técnicas, alcanzando 0.9096. Este incremento se acompañó de un N-criterion más elevado (2.4079), lo que indica conjuntos conformes más grandes. El FSC se situó en 0.8777, mientras que la SSC alcanzó 0.8783 en el grupo de tamaño 1.

Figura 12: *Resultados de APS en EfficientNet-B2*



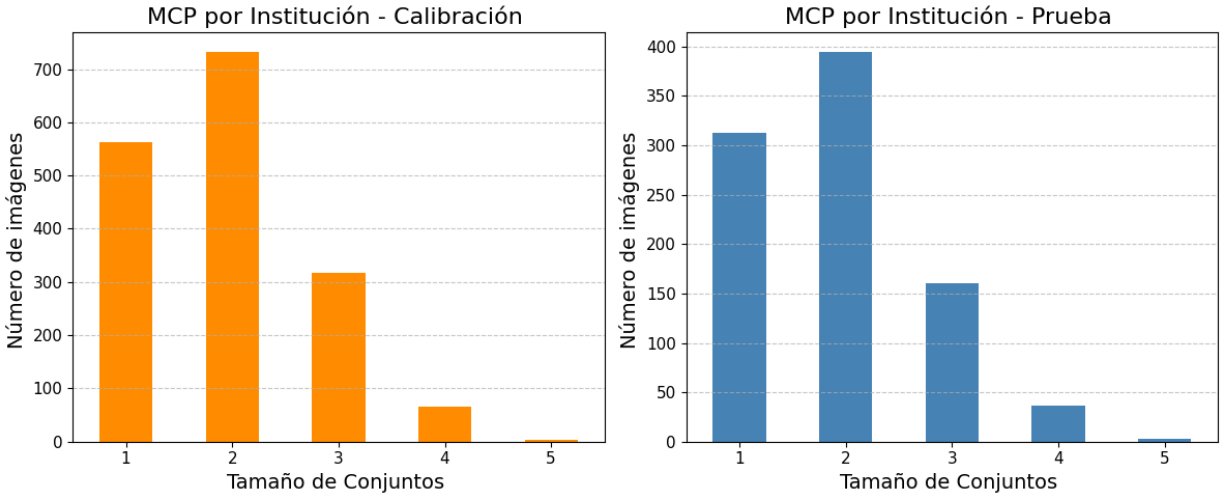
DenseNet-121 alcanzó la cobertura más alta de este experimento con 0.9526, aunque con un N-criterion de 2.6448, evidenciando conjuntos aún más amplios. El FSC fue de 0.9424 y la SSC de 0.9175 en el grupo 1. Esto confirma la tendencia de APS a garantizar alta cobertura a costa de menor N-criterion.

Figura 13: *Resultados de APS en DenseNet-121*



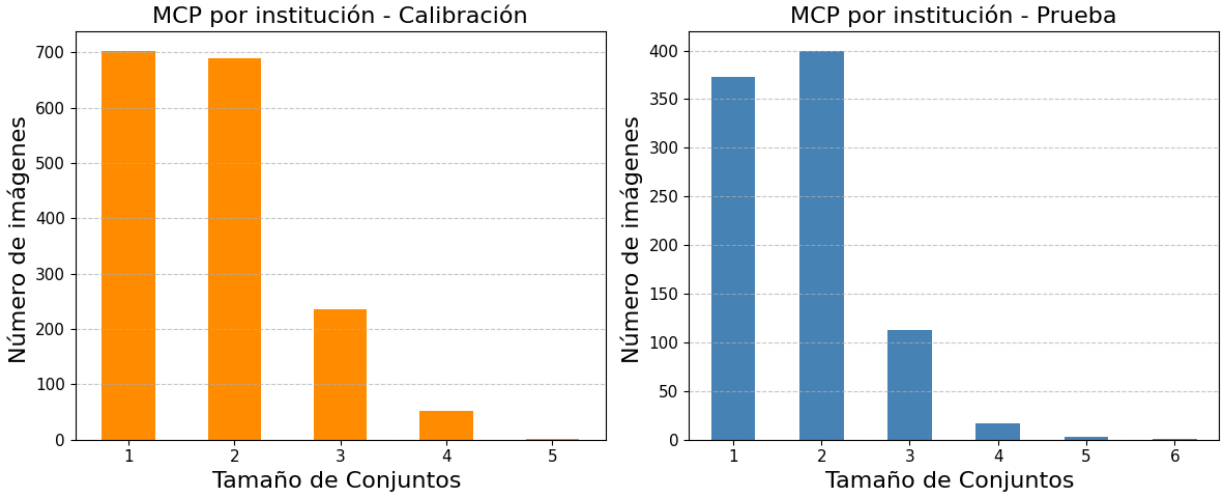
3.4.2.1.3. Predicción Conforme Mondriana. En MCP, EfficientNet-B2 alcanzó una cobertura de 0.8875 con un N-criterion de 1.9846. La métrica FSC se situó en 0.8537, mientras que la SSC fue de 0.8648 en el grupo de tamaño 1.

Figura 14: *Resultados de MCP por provider en EfficientNet*



DenseNet-121 obtuvo una cobertura de 0.8964, con un N-criterion de 1.8313, logrando nuevamente mayor N-criterion que EfficientNet-B2. El FSC fue de 0.8777 y la SSC de 0.8810 para el grupo 4-5. Estos resultados muestran un rendimiento similar al de ICP, con la ventaja de Mondrian de ofrecer cobertura específica por institución.

Figura 15: *Resultados de MCP por provider en DenseNet*



Finalmente, con el fin de consolidar los resultados presentados en este experimento, en la Tabla 4 y la Tabla 5 se resumen los valores de cobertura, N-criterion, FSC y SSC obtenidos para cada técnica y modelo. Estas tablas permiten observar de manera compacta las diferencias entre EfficientNet-B2 y DenseNet-121, así como entre las tres variantes de predicción conforme. En particular, se evidencia que DenseNet-121 tiende a producir conjuntos conformes más eficientes (menor N-criterion) y con mayor confiabilidad (mayores valores de FSC y SSC), mientras que APS, en ambos modelos, logra la cobertura más alta, aunque con el costo de conjuntos de predicción más amplios.

Tabla 4: *Cobertura y N-Criterion para métodos de predicción conforme del experimento 1*

Modelo	**Cobertura**			**N-criterion**		
	ICP	**APS**	**MCP**	**ICP**	**APS**	**MCP**
EfficientNet	0.8831	0.9096	0.8875	1.9945	2.4079	1.9846
DenseNet	0.8953	0.9526	0.8964	1.8280	2.6448	1.8313

Tabla 5: *FSC y SSC para métodos de predicción comforme del experimento 1*

Modelo	FSC			SSC		
	ICP	**APS**	**MCP**	**ICP**	**APS**	**MCP**
EfficientNet	0.8393	0.8777	0.8537	0.8566	0.8783	0.8648
DenseNet	0.8752	0.9424	0.8777	0.8750	0.9175	0.8810

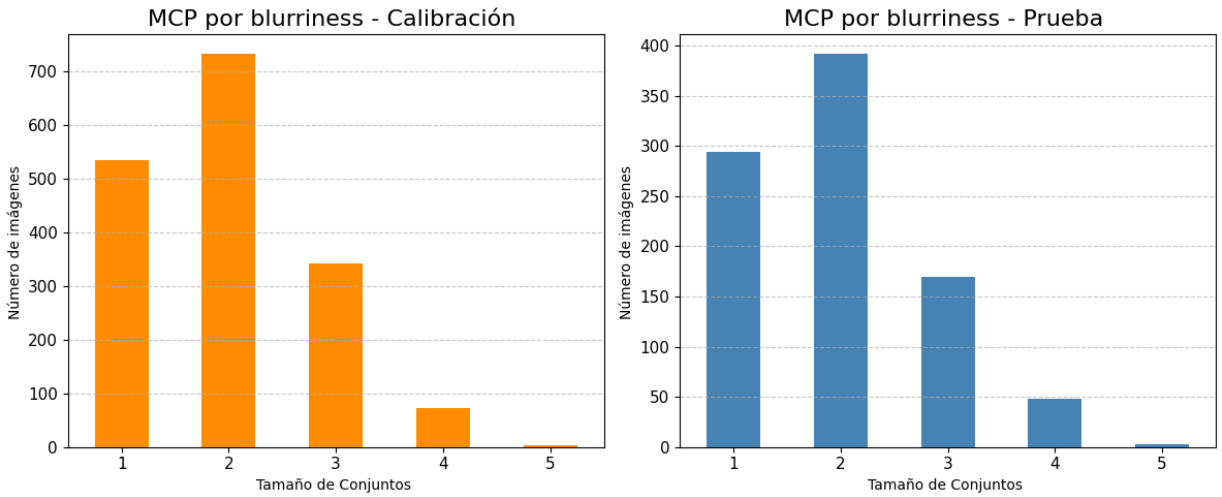
3.4.2.2. Experimento 2: Blurriness como Variable de Estratificación y Categoría Mondriana. En el segundo experimento, la variable `blurriness_group` se empleó de manera unificada como categoría Mondriana en MCP y como variable de estratificación en FSC. Cabe destacar que las métricas de cobertura global, SSC y N-criterion en ICP y APS permanecen inalteradas respecto al Experimento 1, dado que no dependen de la variable de segmentación.

3.4.2.2.1. Predicción Conforme Inductiva. Los resultados para ICP muestran que EfficientNet-B2 alcanzó un FSC de 0.8370, mientras que DenseNet-121 logró un FSC de 0.8634. Las métricas globales de cobertura y N-criterion se mantuvieron en 0.8831 y 1.9945 para EfficientNet-B2, y en 0.8953 y 1.8280 para DenseNet-121, respectivamente.

3.4.2.2.2. Predicción Conforme de Conjuntos Adaptativos. En APS, EfficientNet-B2 alcanzó un FSC de 0.8969, acompañado de sus métricas globales de cobertura (0.9096) y N-criterion (2.4079). DenseNet-121, en contraste, obtuvo un FSC de 0.9471, con cobertura global de 0.9526 y un N-criterion de 2.6448. Esto confirma que, aunque APS genera conjuntos más grandes, logra altos niveles de confiabilidad, especialmente en DenseNet-121.

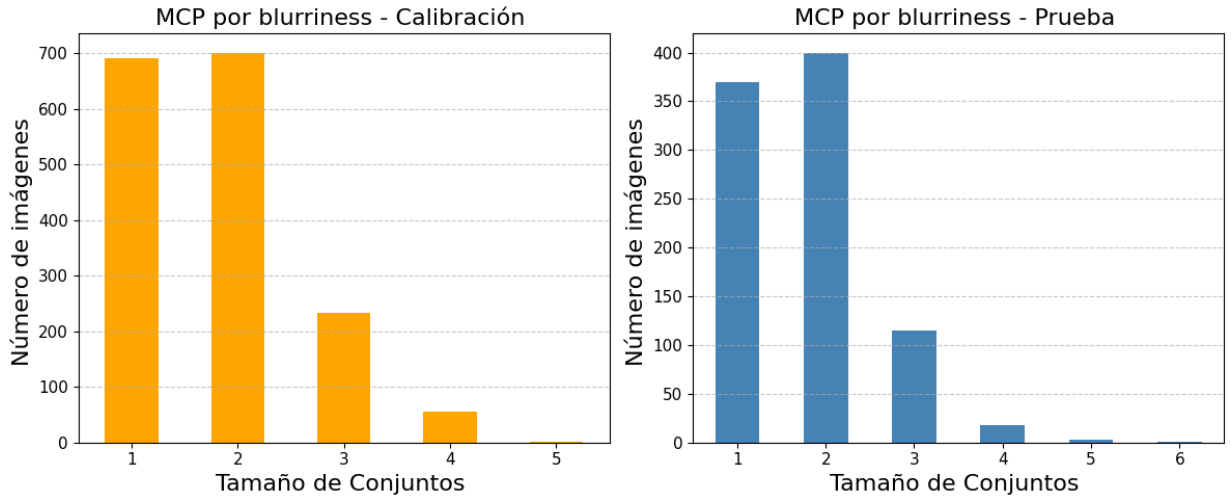
3.4.2.2.3. Predicción Conforme Mondriana. En MCP, EfficientNet-B2 alcanzó una cobertura de 0.8860, con un N-criterion de 2.0110. El FSC fue de 0.8634, mientras que la SSC obtuvo un valor de 0.8623 en el grupo 1.

Figura 16: *Resultados de MCP en EfficientNet-B2*



Por su parte, DenseNet-121 obtuvo una cobertura de 0.8920 con un N-criterion de 1.8015, lo que lo hace más eficiente. El FSC alcanzó un valor de 0.8502, mientras que la SSC fue de 0.8862 en el grupo 3. Esto evidencia un desempeño comparable entre ambos modelos, aunque con ligeras ventajas de N-criterion para DenseNet-121.

Figura 17: *Resultados de MCP en DenseNet-121*



De manera análoga, en la Tabla 6 y la Tabla 7 se presenta el resumen de los resultados del segundo experimento. Allí se observa que las métricas globales de cobertura y N-criterion se mantienen consistentes con el Experimento 1, mientras que las diferencias aparecen en los valores de FSC. En este caso, los resultados muestran que el uso de la variable blurriness_group como criterio de estratificación refuerza la robustez de DenseNet-121, que alcanza los valores más altos de FSC y SSC bajo la técnica APS, aunque nuevamente a costa de un incremento en el tamaño de los conjuntos conformes.

Tabla 6: *Cobertura y N-Criterion para métodos de predicción conforme del experimento 2*

Modelo	**Cobertura**			**N-criterion**		
	ICP	**APS**	**MCP**	**ICP**	**APS**	**MCP**
EfficientNet	0.8831	0.9096	0.8886	1.9945	2.4079	2.0110
DenseNet	0.8953	0.9526	0.8920	1.8280	2.6448	1.8015

Tabla 7: *FSC y SSC para métodos de predicción conforme del experimento 2*

Modelo	FSC			SSC		
	ICP	**APS**	**MCP**	**ICP**	**APS**	**MCP**
EfficientNet	0.8370	0.8969	0.8634	0.8566	0.8783	0.8623
DenseNet	0.8634	0.9471	0.8502	0.8750	0.9175	0.8862

3.5. Discusión de los resultados

En esta sección se analizan tanto los resultados cualitativos como cuantitativos obtenidos con las diferentes variantes de predicción conforme, así como las diferencias observadas entre las arquitecturas de redes convolucionales empleadas. El objetivo es comprender cómo estas técnicas y modelos balancean el compromiso entre cobertura, N-criterion y estabilidad en el problema de graduación histológica abordado.

Iniciando con los resultados de las arquitecturas, el modelo EfficientNet-B2 alcanzó una exactitud de 0.6863 y un QWK de 0.8510. La matriz de confusión asociada, mostrada en la Figura 8, evidencia la distribución de los aciertos y errores de clasificación del modelo. Se observa que los errores más frecuentes ocurren en clases adyacentes, un comportamiento esperado dado el carácter ordinal del problema.

Por su parte, el modelo DenseNet-121 obtuvo una exactitud de 0.7071 y un QWK de 0.8485. La Figura 9 muestra la matriz de confusión correspondiente, que permite analizar de manera detallada el patrón de clasificaciones realizadas por este modelo. A diferencia de EfficientNet, se evidencia una ligera mejora en la capacidad para identificar correctamente las clases intermedias, lo que sugiere una mayor robustez del modelo.

Al comparar cuantitativamente ambos resultados, DenseNet-121 superó a EfficientNet-B2 en exactitud por 2.07 puntos porcentuales (70.71 % frente a 68.63 %). En contraste, EfficientNet-B2 alcanzó un QWK mayor por 0.25 puntos porcentuales (85.10 % frente a 84.85 % en DenseNet).

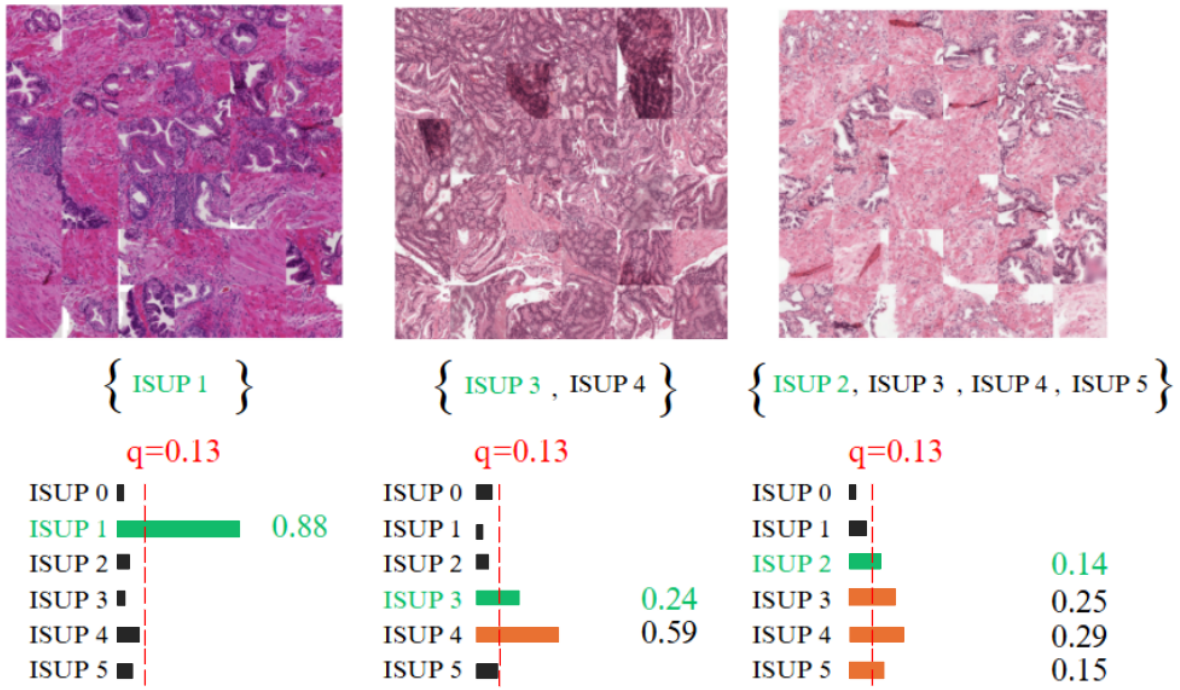
En síntesis, ambos modelos alcanzaron desempeños competitivos, aunque con ligeras diferencias: DenseNet-121 mostró una mejor capacidad de clasificación puntual, reflejada en su mayor exactitud, mientras que EfficientNet-B2 tendió a cometer errores más cercanos a la clase verdadera, lo que le permitió obtener un QWK ligeramente más alto. Este contraste refleja que DenseNet puede ofrecer predicciones más consistentes en términos absolutos, mientras que EfficientNet conserva una ventaja en la métrica ordinal. Estos resultados iniciales sirven como punto de partida para analizar cómo las técnicas de predicción conforme afectan y enriquecen la interpretación y confiabilidad de las predicciones.

Un primer aspecto cualitativo puede observarse en la Figura 18, donde se ilustran ejemplos concretos de la construcción de los conjuntos conformes. En cada caso, con las predicciones *softmax* del modelo se asignan los respectivos puntajes de conformidad a cada grado ISUP y, a partir de ellos, los sets conformes se forman aplicando un umbral q que garantiza la cobertura deseada.

Así, en la primera imagen el set conforme está compuesto únicamente por la clase verdadera (ISUP 1), reflejando una predicción altamente confiable. En contraste, en la segunda y tercera imagen se observa cómo el modelo incluye múltiples clases (por ejemplo,

$\{ISUP\,3, ISUP\,4\}$ o $\{ISUP\,2, ISUP\,3, ISUP\,4, ISUP\,5\}$), lo que indica un nivel de incertidumbre mayor. Estos ejemplos cualitativos permiten interpretar que, cuanto más amplio es el conjunto conforme, mayor es la duda del modelo respecto a la clase verdadera, lo que tendrá implicaciones en la utilidad clínica de las predicciones.

Figura 18: *Ejemplos de construcción de los sets conformes usando el umbral q. El tamaño del conjunto refleja el grado de incertidumbre del modelo*



3.5.1. Resultados del Experimento 1

A partir de esta base cualitativa, se analizan ahora los resultados cuantitativos globales obtenidos en los dos experimentos de predicción conforme. En términos de cobertura (Tabla 4), APS se destacó como el método con mejor, superando consistentemente el umbral objetivo del 90 %. En EfficientNet-B2, alcanzó un 90.96 %, lo que representa un incremento del 2.6 % respecto a ICP (88.31 %) y del 2.4 % respecto a MCP (88.75 %). En DenseNet-121, la diferencia fue aún más marcada: APS logró una cobertura del 95.26 %, superando en 5.9 % a ICP (89.53 %) y en 6.0 % a MCP (89.64 %). Estos resultados confirman que APS prioriza la inclusión de la clase verdadera, aunque a costa de conjuntos más amplios.

El N-criterion evidenció precisamente ese costo. En EfficientNet-B2, APS obtuvo un valor de 2.40, lo que significa que sus conjuntos de predicción fueron en promedio un 20.7 % menos eficientes(un N-criterion más grande) que los de ICP (1.99) y un 21.3 % menos que los de MCP (1.98). La misma tendencia se observó en DenseNet-121, donde APS alcanzó

2.64, siendo un 44.6 % menos eficiente que ICP (1.82) y un 44.4 % menos que MCP (1.83). Este comportamiento refleja el compromiso inherente entre cobertura y N-criterion: APS garantiza altos niveles de cobertura, pero genera intervalos menos informativos.

El análisis de las distribuciones de tamaño de conjuntos (Figuras 10, 11, 12, 13, 14, 15) respalda estas observaciones. En ICP y MCP, la mayoría de los conjuntos se concentraron en tamaños de 1 o 2 etiquetas, mientras que en APS se observó una mayor dispersión, con casos que alcanzaron hasta 5 etiquetas. Esta dispersión refleja la naturaleza adaptativa del método, que amplía el tamaño de los conjuntos en muestras donde la incertidumbre del modelo es mayor.

Finalmente, al considerar las métricas compuestas FSC y SSC (Tabla 5), APS nuevamente mostró ventajas claras en cobertura efectiva. En EfficientNet-B2, superó a ICP en 3.8 puntos porcentuales (87.77 % frente a 83.93 %), y en DenseNet-121 la diferencia fue de 6.7 puntos (94.21 % frente a 87.52 %). En SSC, la mejora fue más moderada: en EfficientNet-B2, APS alcanzó 87.83 %, solo 2.7 % por encima de ICP (85.66 %), y en DenseNet-121 llegó a 91.75 %, representando un incremento de 4.7 % frente a ICP (87.50 %).

En síntesis, los resultados del experimento 1 evidencian que APS se constituye como la alternativa más conservadora y robusta para garantizar cobertura, con incrementos de hasta 6 % frente a ICP y MCP. Sin embargo, este beneficio viene acompañado de una pérdida sustancial en N-criterion, con conjuntos que pueden contener hasta 5 clases, lo cual plantea un desafío para su utilidad clínica. Por el contrario, ICP y MCP mantienen conjuntos más compactos y eficientes, aunque sin alcanzar el nivel de cobertura de APS.

3.5.2. Resultados del Experimento 2

El segundo experimento de predicción conforme se diseñó a partir del uso de la variable *blurriness_group* en lugar de la institución . Este cambio impactó de manera directa a MCP, ya que sus conjuntos de calibración dependen de la variable empleada, modificando todos sus resultados. En contraste, los métodos ICP y APS mantuvieron coberturas, N-criterion y SSC idénticos al experimento anterior, dado que no dependen de dicha variable, observándose únicamente cambios en la métrica FSC.

En términos de cobertura, APS se mantuvo nuevamente como el método con mejor desempeño. En EfficientNet-B2 alcanzó 90.96 %, superando en 2.6 % a ICP (88.31 %) y en 2.9 % a MCP (88.86 %). En DenseNet-121, APS logró 95.26 %, con un incremento del 5.7 % frente a ICP (89.53 %) y del 6.3 % respecto a MCP (89.20 %). Estos resultados confirman la estabilidad de APS en ambos experimentos, priorizando la inclusión de la clase verdadera.

En cuanto a N-criterion, MCP mostró un desempeño diferenciado frente al experimento 1. En EfficientNet-B2, obtuvo un valor de 2.01, lo que implica que fue 0.8 % menos

eficiente que ICP (1.99) pero 19.6 % más eficiente que APS (2.40). En DenseNet-121, MCP alcanzó 1.80, superando en N-criterion a APS en un 31.8 % (2.64) y a ICP en un 1.5 % (1.82). Esto evidencia que, aunque MCP pierde en cobertura frente a APS, logra ofrecer conjuntos más compactos, lo que aumenta su potencial utilidad práctica.

Respecto a las métricas por subgrupos, el efecto del cambio de variable fue más evidente en FSC. En EfficientNet-B2, APS alcanzó 89.69 %, lo que representa una mejora de 6.0 puntos porcentuales frente a ICP (83.70 %) y de 3.3 puntos frente a MCP (86.34 %). En DenseNet-121, APS se ubicó en 94.71 %, superando en 8.4 puntos a ICP (86.34 %) y en 9.7 puntos a MCP (85.02 %). Por su parte, en SSC los resultados permanecieron similares al experimento 1: en EfficientNet-B2 las diferencias entre métodos fueron mínimas (86.23 % en MCP frente a 85.66 % en ICP y 87.83 % en APS), mientras que en DenseNet-121 APS mantuvo la ventaja con 91.75 %, superando en 4.3 % a ICP (87.50 %) y en 2.9 % a MCP (88.62 %).

En síntesis, el experimento 2 confirma las tendencias observadas previamente: APS garantiza consistentemente la mayor cobertura y la mejor combinación de FSC y SSC, aunque con el costo de conjuntos más amplios y menos eficientes. MCP, en este escenario, mostró una ligera ganancia de N-criterion frente al experimento 1 gracias al cambio de variable, pero sin alcanzar los niveles de cobertura de APS. Comparando ambos experimentos, se concluye que las variaciones en la variable Mondrian afectan principalmente a MCP, mientras que ICP y APS mantienen resultados estables, consolidando así a APS como el método más confiable en términos de cobertura y robustez frente a cambios en la partición.

3.6. Consideraciones finales

El análisis realizado permite extraer varias conclusiones sobre el desempeño de las arquitecturas EfficientNet-B2 y DenseNet-121 en nuestra tarea de clasificación, así como sobre la aplicación de métodos de predicción conforme en la tarea de graduación histológica categórica.

En primer lugar, al examinar los resultados de la estimación puntual se observa que DenseNet-121 alcanzó una exactitud de 70.71 %, superando en 2.1 puntos porcentuales a EfficientNet-B2 (68.63 %), lo que indica una mayor capacidad de clasificación correcta en términos absolutos. Sin embargo, EfficientNet-B2 presentó un QWK ligeramente superior (0.8510 frente a 0.8485), lo que sugiere que, aunque se equivoca más, sus errores tienden a concentrarse en clases adyacentes, lo cual es coherente con la naturaleza ordinal del problema.

Al incorporar métodos de predicción conforme, se observaron compromisos claros entre cobertura y N-criterion. APS se consolidó como la técnica más conservadora, alcanzando coberturas que superaron en hasta 6 % a ICP y MCP, pero a costa de intervalos más am-

plios, con tamaños de conjunto que en algunos casos llegaron hasta cinco etiquetas. Este comportamiento garantiza confiabilidad, pero limita la utilidad clínica directa al generar predicciones menos precisas. En contraste, ICP y MCP ofrecieron conjuntos más compactos y eficientes (con tamaños de 1 o 2 etiquetas en la mayoría de los casos), aunque con coberturas ligeramente inferiores al 90 %.

La comparación entre experimentos reveló que la variable utilizada en la calibración Mondrian incide de manera significativa en MCP. Cuando se usó la institución como variable (Experimento 1), MCP mostró coberturas más altas y un balance más cercano a ICP. En cambio, al emplear el *blurriness group* (Experimento 2), MCP redujo su cobertura en torno a 0.5 puntos porcentuales, aunque incrementó su N-criterion relativa frente a APS e ICP. Por su parte, ICP y APS se mantuvieron estables en sus métricas de cobertura y N-criterion entre ambos experimentos, evidenciando su mayor robustez frente a cambios en la partición Mondrian. El impacto se reflejó únicamente en la métrica FSC, donde APS mantuvo ventajas de entre 6 y 10 puntos porcentuales sobre ICP y MCP en DenseNet-121.

En conjunto, los hallazgos pueden sintetizarse en tres patrones principales: (i) DenseNet-121 supera consistentemente a EfficientNet-B2 en exactitud y en métricas derivadas, consolidándose como la arquitectura más robusta en esta tarea particular. (ii) APS garantiza la mayor cobertura y los mejores valores de FSC y SSC, aunque con pérdidas significativas en N-criterion debido a la ampliación de los conjuntos de predicción. (iii) MCP se muestra sensible a la variable Mondrian utilizada, con un desempeño más equilibrado bajo la partición por institución y más eficiente, aunque menos estable, bajo la partición por nivel de desenfoque.

Estas consideraciones finales refuerzan la idea de que la elección del modelo y de la técnica de predicción conforme debe responder al equilibrio deseado entre confiabilidad y eficiencia, así como a las características específicas de los datos y de la aplicación clínica prevista.

4. Predicción conforme para regresión: Celularidad en cáncer de mama

En este capítulo nos enfocamos en la predicción conforme aplicada a problemas de regresión, donde el objetivo del modelo no es asignar una etiqueta discreta, sino estimar un valor continuo dentro de un rango específico, en nuestro caso $[0, 1]$. La regresión plantea retos diferentes a la clasificación, tanto en la construcción de los modelos como en la evaluación de la incertidumbre de sus predicciones, lo que hace necesaria una adaptación de las técnicas de predicción conforme para este tipo de tareas.

Este capítulo se organiza de la siguiente manera. Primero, en la sección 4.1 se describe el dataset utilizado para la tarea de regresión y para la predicción conforme. Luego, en la sección 4.2 se detallan los modelos empleados, las métricas de evaluación y la manera en que se aplica la predicción conforme, incluyendo sus distintas variantes. A continuación, en la sección 4.3 se presentan las configuraciones del procedimiento experimental. La sección 4.4 muestra los resultados obtenidos, junto a una sección donde se discuten más a fondo los resultados (sección 4.5. Finalmente, la sección 4.6 recoge las conclusiones del capítulo.

4.1. Materiales

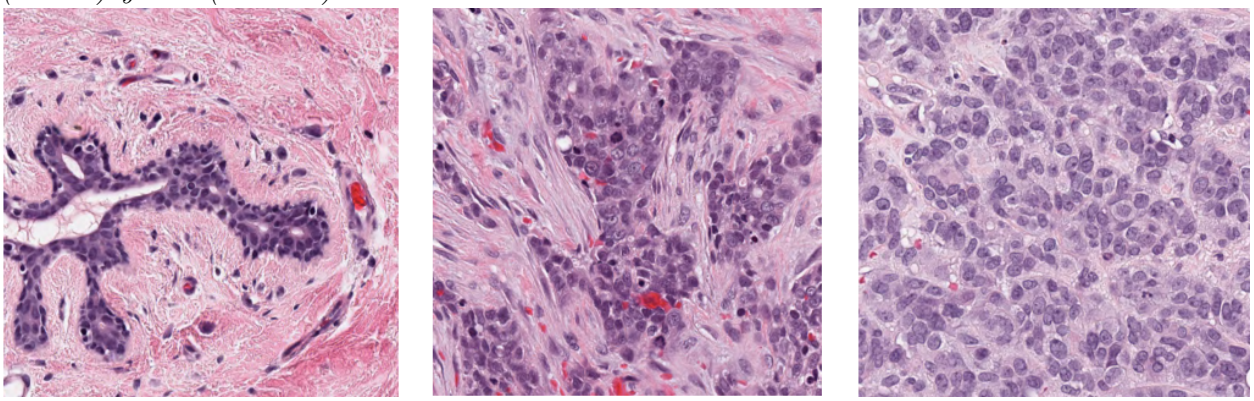
Para el caso de la predicción en tareas de regresión, utilizaremos un dataset diseñado para predecir el porcentaje de celularidad tumoral en imágenes histopatológicas de cáncer de mama. La mama es un órgano compuesto por unidades ducto–lobulillares (ductos y acinos glandulares) que constituyen la glandula mamaria, y que están embebidas en estroma fibroso y tejido adiposo. En cortes teñidos con hematoxilina y eosina (H&E), los núcleos celulares se tiñen predominantemente con hematoxilina (color azul-púrpura, basofilia), mientras que el citoplasma, el estroma y la matriz extracelular captan eosina (tonos rosados). Esta diferenciación cromática facilita distinguir la morfología celular versus el soporte tisular.

En el cancer de mama, la celularidad tumoral corresponde a la fracción de área ocupada por células neoplásicas respecto al campo histológico, y se constituye en un biomarcador útil para evaluar respuesta a terapia neoadyuvante y carga residual de enfermedad (Hossain et al. 2024). El dataset “BreastPathQ: Cancer Cellularity Challenge 2019” fue utilizado en el marco de la conferencia SPIE Medical Imaging 2019 en un desafío que congregó varios participantes. El objetivo principal de este reto era fomentar el desarrollo algoritmos automáticos para la estimación de celularidad tumoral en imágenes histopatológicas de cáncer de mama, teñidas de hematoxilina y eosina (Petrick et al. 2021). Este conjunto de datos se recolectó en el Sunnybrook Health Sciences Centre (Toronto, Canadá), y consta de un total de 96 whole slide images (WSI) obtenidas de 64 pacientes con cáncer de mama invasivo residual tras terapia neoadyuvante. Estas imágenes fueron digitalizadas a una resolución de 2X (0.5

μm/píxel).

A partir de estas imágenes, los organizadores extrajeron parches histológicos y los dividieron para el reto: se destinaron 2,579 parches provenientes de 69 WSI para entrenamiento y validación, cada uno con un puntaje de celularidad asignado por un patólogo experto; y 1,121 parches de 25 WSI para la prueba, cuyas etiquetas se mantuvieron ocultas.

Figura 19: *Ejemplos de parches extraídos BreastPathQ: Celularidad baja (izquierda), media (centro) y alta (derecha).*



La tarea que se propuso a los participantes consistió en desarrollar algoritmos para asignar de manera automática un puntaje de celularidad continuo entre 0 y 1, donde los valores cercanos a 0 representan una ausencia de tumor y los valores cercanos a 1 representan alta densidad tumoral.

4.2. Métodos

4.2.1. Arquitecturas de redes neuronales - Regresión

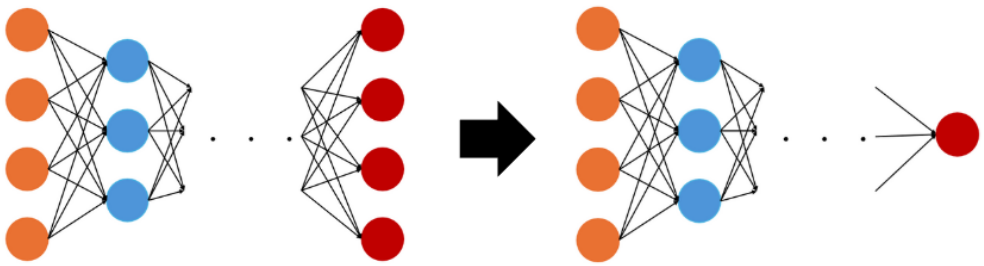
Para la tarea de regresión se emplearon arquitecturas de redes neuronales convolucionales previamente utilizados en clasificación, específicamente EfficientNet-B3 y DenseNet-169. Estas arquitecturas han destacado por su desempeño en visión por computador y originalmente fueron diseñadas para tareas de clasificación, ya que su capa final produce probabilidades para distintas clases.

EfficientNet-B3 corresponde a una versión más profunda y con mayor capacidad de representación que EfficientNet-B2, utilizada en la sección de clasificación. Esta familia de modelos se caracteriza por el uso de compound scaling, que ajusta de manera equilibrada la profundidad, el ancho y la resolución de la red para mejorar la eficiencia computacional y el rendimiento predictivo (Tan y Le 2019). De la misma manera, DenseNet-169 constituye una extensión más profunda respecto a DenseNet-121, lo que permite capturar representaciones más complejas a partir de las imágenes. Este conjunto de arquitecturas densenet se

caracteriza por el esquema de conexiones densas, en el cual cada capa recibe como entrada la salida de todas las capas anteriores, favoreciendo la reutilización de características y la propagación eficiente de gradientes durante el entrenamiento.

Para poder adaptar estas arquitecturas a nuestra tarea de regresión, lo único que debemos hacer es reemplazar nuestra capa de salida (originalmente diseñada para 1000 clases en imagenet) por un bloque denso de varias capas conectadas. Al utilizar este método para regresión, incluimos capas lineales intermedias con funciones de activación ReLU y capas de dropout para mejorar la capacidad de la generalización. La capa final es un perceptrón lineal con una única neurona, lo que nos produce un valor continuo que corresponde a la predicción de la celularidad.

Figura 20: *Representación del cambio en la última capa.* ***Izquierda:*** *Arquitectura convolucional estándar con salida* Linear $\rightarrow$ *Softmax.* ***Derecha:*** *Arquitectura adaptada a tareas de regresión con* Linear $\rightarrow 1$.



4.2.2. Métricas de evaluación para tareas de regresión

En los experimentos de regresión se consideraron métricas estándar ya usadas en la literatura, tales como el error cuadrático medio (MSE), el error absoluto medio (MAE) y el coeficiente de determinación (R^2) (Chicco; Warrens y Jurman 2021).

El error cuadrático medio (MSE) calcula el promedio de los cuadrados de los errores, penalizando desviaciones grandes:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

El error absoluto medio (MAE) mide el promedio de los errores absolutos y es menos sensible a valores atípicos en comparación con el MSE:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Finalmente, el coeficiente de determinación (R^2) evalúa qué tan bien el modelo explica la variabilidad de los datos observados:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Cada una de estas métricas ofrece información complementaria, por lo que su selección depende del contexto del problema y de la naturaleza de los datos.

Para nuestro trabajo solo nos enfocaremos en el coeficiente de determinación. Esta métrica fue seleccionada por su interpretación directa y ampliamente reconocida en la literatura, ya que cuantifica la proporción de la variabilidad de los datos de los modelos. En cuanto a los contextos biomédicos, donde es relevante evaluar la capacidad del modelo para reproducir patrones observados en las mediciones, esta métrica representa un indicador estándar y fácilmente comprensible.

4.2.3. Predicción conforme para regresión

Para aplicar la predicción conforme a tareas de regresión se sigue el mismo procedimiento general utilizado en clasificación, aunque con unas pequeñas diferencias clave.

En este contexto, la incertidumbre del modelo ya no se representa mediante conjuntos de predicción, sino mediante intervalos de predicción, los cuales buscan cubrir el valor real con un nivel de confianza preestablecido (Wisniewski; D. Lindsay y S. Lindsay 2020).

El procedimiento comienza con el entrenamiento de un modelo de regresión f a partir del conjunto de entrenamiento. Posteriormente, se calculan los scores de no conformidad sobre un conjunto de calibración, con el fin de medir la discrepancia entre las predicciones y los valores observados. Una función comúnmente empleada en regresión es el valor absoluto de los residuales:

$$S(y_i, \hat{y}_i) = |y_i - \hat{y}_i|,$$

donde y_i es el valor real y $\hat{y}_i = f(X_i)$ es la predicción.

A partir de los scores obtenidos, se ordenan y se calcula el cuantil $q_{1-\alpha}$, que corresponde al nivel de error que garantiza una cobertura mínima de $1-\alpha$. Finalmente, para cada nueva muestra X_{new} con predicción puntual $\hat{y}_{\text{new}}$, se construye el intervalo de predicción conforme como:

$$C(X_{\text{new}}) = \Big[\, \hat{y}_{\text{new}} - q_{1-\alpha},\ \hat{y}_{\text{new}} + q_{1-\alpha} \,\Big].$$

Este intervalo es simétrico y utiliza el mismo cuantil para todas las predicciones, lo cual asegura la validez del método de acuerdo con el nivel de confianza establecido. Para entenderlo de mejor manera, podemos ver como funciona la predicción conforme en este caso

en el Algoritmo 4.

Algoritmo 4: Predicción conforme en regresión (general)

Entrada: Conjunto de datos $\mathcal{D} = \{(X_i, y_i)\}_{i=1}^{N}$, nivel de confianza $1 - \alpha$, función de no conformidad $S(y, \hat{y})$

Salida: Intervalos de predicción conformes para nuevas muestras

1. Entrenamiento del modelo:

Entrenar el modelo f utilizando el conjunto de datos $\mathcal{D}$.

2. Cálculo de scores:

Para cada *muestra* $(X_i, y_i) \in \mathcal{D}$ **hacer**

Obtener la predicción $\hat{y}_i = f(X_i)$.

Calcular el puntaje de no conformidad $S(y_i, \hat{y}_i) = |y_i - \hat{y}_i|$.

fin

3. Obtención del cuantil:

Ordenar los puntajes $\{S(y_i, \hat{y}_i)\}$ y calcular el cuantil $q_{1-\alpha}$.

4. Construcción de intervalos:

Para cada *nueva muestra* X_{new} **hacer**

Obtener la predicción $\hat{y}_{\text{new}} = f(X_{\text{new}})$.

Construir el intervalo de predicción: $C(X_{\text{new}}) = [\hat{y}_{\text{new}} - q_{1-\alpha},\ \hat{y}_{\text{new}} + q_{1-\alpha}]$.

fin

En el marco de la tarea de regresión se emplearon distintos tipos de predicción conforme con el fin de generar intervalos de predicción válidos y evaluar qué tan bien desempeñan en la estimación de celularidad tumoral. Al igual que en la sección de clasificación, se reutilizaron la predicción conforme inductiva y la predicción conforme de Mondrian, siendo estas dos usadas ampliamente en la literatura. Adicional a estos dos métodos, se introdujo un tercer enfoque basado en la regresión por cuantiles, que permite estimar intervalos de predicción de manera directa a partir de los cuantiles de la distribución condicional de la variable de interés.

4.2.3.1. Predicción conforme inductiva en regresión. La predicción conforme inductiva (ICP) en regresión sigue la misma lógica general descrita previamente para el caso de clasificación. En ambos escenarios, el procedimiento consiste en dividir el conjunto de datos en dos partes: un subconjunto de entrenamiento, donde se ajusta el modelo de aprendizaje, y un subconjunto de calibración, empleado para calcular los puntajes de no conformidad. La diferencia fundamental es que, en lugar de obtener conjuntos de predicción sobre clases, en regresión se generan intervalos de predicción que buscan contener el valor real con un nivel de confianza predefinido.

El proceso comienza entrenando un modelo de regresión f sobre el conjunto $\mathcal{D}_{\text{train}}$. Posteriormente, se utiliza el conjunto de calibración $\mathcal{D}_{\text{cal}}$ para calcular los scores de no conformidad, definidos habitualmente como el error absoluto entre la predicción y el valor real:

$$S(y_i, \hat{y}_i) = |y_i - \hat{y}_i|,$$

donde y_i es el valor observado y $\hat{y}_i = f(X_i)$ es la salida del modelo.

A partir de estos scores, se ordenan y se calcula el cuantil $q_{1-\alpha}$:

$$\hat{q} \;=\; s\left(\left\lceil \frac{(n+1)(1-\alpha)}{n} \right\rceil \right)$$

donde n es el número de ejemplos. Este permite determinar un umbral de error compatible con la cobertura deseada. Finalmente, para cada nueva muestra X_{new} con predicciónbpuntual $\hat{y}_{\text{new}}$, se construye el intervalo de predicción conforme como:

$$C(X_{\text{new}}) = \Big[\hat{y}_{\text{new}} - q_{1-\alpha},\ \hat{y}_{\text{new}} + q_{1-\alpha}\Big].$$

Este enfoque mantiene la ventaja ya discutida en clasificación: el modelo se entrena una única vez, mientras que la calibración independiente permite estimar intervalos válidos para datos no vistos sin necesidad de reentrenar continuamente el algoritmo subyacente. Podemos ver este proceso de mejor manera en el Algoritmo 5.

Algoritmo 5: Predicción conforme inductiva (ICP) en regresión

Entrada: Conjunto de datos $\mathcal{D} = \{(X_i, y_i)\}_{i=1}^{N}$, nivel de confianza $1 - \alpha$, función de no conformidad $S(y, \hat{y})$

Salida: Intervalos de predicción conformes para nuevas muestras

1. División del conjunto de datos:

Dividir $\mathcal{D}$ en dos subconjuntos: Entrenamiento $\mathcal{D}_{\text{train}}$ y calibración $\mathcal{D}_{\text{cal}}$.

2. Entrenamiento:

Entrenar el modelo f con $\mathcal{D}_{\text{train}}$.

3. Calibración:

Para cada *muestra* $(X_i, y_i) \in \mathcal{D}_{cal}$ **hacer**

Obtener la predicción $\hat{y}_i = f(X_i)$.

Calcular el puntaje de no conformidad $S(y_i, \hat{y}_i) = |y_i - \hat{y}_i|$.

fin

Ordenar los puntajes y calcular el cuantil $q_{1-\alpha}$.

4. Construcción de intervalos:

Para cada *nueva muestra* X_{new} **hacer**

Obtener la predicción $\hat{y}_{\text{new}} = f(X_{\text{new}})$.

Construir el intervalo de predicción: $C(X_{\text{new}}) = [\hat{y}_{\text{new}} - q_{1-\alpha},\ \hat{y}_{\text{new}} + q_{1-\alpha}]$.

fin

4.2.3.2. Predicción conforme de Mondrian en regresión. Una vez entendido el funcionamiento general de la predicción conforme en regresión, resulta más sencillo extenderlo al caso de la predicción conforme de Mondrian (MCP). Al igual que en clasificación, este método introduce la idea de dividir el conjunto de datos en subgrupos o particiones definidos a partir de características específicas, como la edad del paciente, la institución de procedencia de la muestra o cualquier otra variable relevante del contexto.

Para cada subgrupo se calcula de manera independiente el cuantil $q_{1-\alpha}$ a partir de los puntajes de no conformidad obtenidos en ese subconjunto. De esta forma, al predecir para una nueva muestra X_{new}, el intervalo de predicción se construye utilizando el cuantil asociado al subgrupo al que dicha muestra pertenece.

El procedimiento es análogo al caso de clasificación, con la diferencia de que ahora se generan intervalos de predicción en lugar de conjuntos de etiquetas. El proceso completo puede observarse en el Algoritmo 6.

Algoritmo 6: Predicción conforme de Mondrian (MCP) en regresión

Entrada: Conjunto de datos $\mathcal{D} = \{(X_i, y_i)\}_{i=1}^{N}$, nivel de confianza $1-\alpha$, función de no conformidad $S(y, \hat{y})$, regla de partición en subgrupos Π

Salida: Intervalos de predicción conformes para nuevas muestras

1. División en subgrupos:

Asignar cada muestra (X_i, y_i) a un subgrupo $\Pi(X_i)$ de acuerdo con la regla de partición definida (ej. edad, institución, etc.).

2. Entrenamiento:

Entrenar el modelo f con el conjunto completo de entrenamiento.

3. Calibración por subgrupos:

Para cada *subgrupo* G **hacer**

  Para cada *muestra* $(X_i, y_i) \in G$ **hacer**

    Obtener la predicción $\hat{y}_i = f(X_i)$.

    Calcular el puntaje de no conformidad $S(y_i, \hat{y}_i) = |y_i - \hat{y}_i|$.

  fin

  Ordenar los puntajes de G y calcular el cuantil $q_{1-\alpha}^{(G)}$.

fin

4. Construcción de intervalos:

Para cada *nueva muestra* X_{new} **hacer**

  Determinar el subgrupo $G = \Pi(X_{\text{new}})$.

  Obtener la predicción $\hat{y}_{\text{new}} = f(X_{\text{new}})$.

  Construir el intervalo de predicción: $C(X_{\text{new}}) = [\hat{y}_{\text{new}} - q_{1-\alpha}^{(G)},\ \hat{y}_{\text{new}} + q_{1-\alpha}^{(G)}]$.

fin

4.2.3.3. Predicción conforme basada en regresión por cuantiles. En las variantes de predicción conforme vistas anteriormente, los cambios con respecto al método general fueron relativamente sencillos. Sin embargo, en la predicción conforme por cuantiles (Conformalized Quantile Regression, CQR) las modificaciones comienzan desde el propio entrenamiento del modelo (Angelopoulos y Bates 2021).

La idea central de CQR es combinar la *regresión de cuantiles*, que proporciona intervalos dependientes de x, con la *calibración conformal*, que ajusta de manera global para garantizar cobertura finito–muestral bajo suposiciones de intercambiabilidad.

En lugar de predecir un único valor, el modelo aprende a estimar dos cuantiles condicionales, uno inferior y otro superior. Para un nivel de confianza del 90 %, por ejemplo, se predicen los cuantiles $q_{0.05}(x)$ y $q_{0.95}(x)$, los cuales definen el intervalo base

$$I(x) \ = \ \Big[\hat{q}_{0.05}(x),\ \hat{q}_{0.95}(x)\Big].$$

Estos cuantiles se entrenan mediante la pérdida pinball, definida como

$$\mathcal{L}_\tau(y, \hat{q}) = \tau\,(y - \hat{q})\,\mathbf{1}\{y \geq \hat{q}\} + (1 - \tau)\,(\hat{q} - y)\,\mathbf{1}\{y < \hat{q}\},$$

para cada nivel $\tau \in \{0.05,\, 0.95\}$, más una penalización que evita el cruce de cuantiles.

Una vez entrenado el modelo, se pasa al conjunto de calibración. Para cada par (x_i, y_i) se define un *score de no conformidad* que mide qué tan lejos queda el valor real de los cuantiles predichos:

$$s_i = \text{máx}\Big(\,\hat{q}_{0.05}(x_i) - y_i,\;\; y_i - \hat{q}_{0.95}(x_i),\;\; 0\Big).$$

En otras palabras, $s_i = 0$ si y_i ya está dentro del intervalo base, y es positivo si el valor verdadero queda por fuera (ya sea por debajo o por encima).

A partir de los scores $\{s_i\}$ se ordenan los valores $s_{(1)} \leq s_{(2)} \leq \cdots \leq s_{(n)}$. El cuantil conforme no se calcula de forma arbitraria, sino siguiendo la fórmula finito-muestral estándar utilizada en la literatura de predicción conforme:

$$\hat{q} \;=\; s\left(\left\lceil \frac{(n+1)(1-\alpha)}{n} \right\rceil\right)$$

donde n es el número de ejemplos de calibración y $\alpha \in (0, 1)$ es el nivel de error predefinido. Este procedimiento garantiza que la cobertura final sea al menos $1 - \alpha$.

Finalmente, para una nueva muestra x, el intervalo de predicción se construye como

$$C(x) = \Big[\,\hat{q}_{0.05}(x) - \hat{q},\;\; \hat{q}_{0.95}(x) + \hat{q}\,\Big],$$

lo que significa que cada predicción tiene su propio intervalo adaptativo (depende de x), ampliado con un colchón global $\hat{q}$ proveniente de la calibración.

Algoritmo 7: Predicción conforme basada en regresión por cuantiles (CQR)

Entrada: Datos de entrenamiento $\mathcal{D}_{\text{train}}$, calibración $\mathcal{D}_{\text{cal}}$, nivel de confianza $1-\alpha$

Salida: Intervalos de predicción conformes basados en cuantiles

1. Entrenamiento de cuantiles:

Entrenar un modelo f que, para cada x, produzca dos salidas $\hat{q}_{\text{inf}}(x)$ y $\hat{q}_{\text{sup}}(x)$, usando *pinball loss* para $\tau \in \{\alpha/2,\ 1-\alpha/2\}$ y una penalización de no-cruce.

2. Calibración:

Para cada *muestra* $(x_i, y_i) \in \mathcal{D}_{cal}$ **hacer**

Obtener el intervalo base $I(x_i) = [\hat{q}_{\text{inf}}(x_i),\ \hat{q}_{\text{sup}}(x_i)]$.

Calcular el score $s_i = \max\Big(\hat{q}_{\text{inf}}(x_i) - y_i,\ \ y_i - \hat{q}_{\text{sup}}(x_i),\ \ 0\Big)$.

fin

Ordenar los scores $\{s_i\}$ y calcular el cuantil $\hat{q}$ correspondiente a $1-\alpha$.

3. Despliegue:

Para cada *nueva muestra* x **hacer**

Predecir $(\hat{q}_{\text{inf}}(x),\ \hat{q}_{\text{sup}}(x)) = f(x)$.

Construir el intervalo conforme: $C(x) = [\,\hat{q}_{\text{inf}}(x) - \hat{q},\ \ \hat{q}_{\text{sup}}(x) + \hat{q}\,]$.

fin

4.2.4. Métricas de predicción conforme para la tarea de regresión

En el contexto de regresión con predicción conforme, se emplean dos métricas fundamentales: la cobertura y el tamaño de los intervalos.

La cobertura mide la proporción de observaciones reales que caen dentro de los intervalos conformes generados. Sea $\{(x_i, y_i)\}_{i=1}^{n}$ el conjunto de prueba y $C(x_i)$ el intervalo asociado a x_i. La cobertura empírica se define como

$$\widehat{\text{Cov}} \ = \ \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\Big\{\, y_i \in C(x_i)\,\Big\},$$

donde $\mathbf{1}\{\cdot\}$ es la función indicadora que vale 1 si la condición se cumple y 0 en caso contrario.

El tamaño promedio de los intervalos se define como

$$\widehat{\text{Size}} \ = \ \frac{1}{n}\sum_{i=1}^{n} \Big(U(x_i) - L(x_i)\Big),$$

donde $L(x_i)$ y $U(x_i)$ representan, respectivamente, los extremos inferior y superior del intervalo $C(x_i)$.

En la variante inductiva (ICP) los intervalos se construyen como

$$C(x) = [\, f(x) - \hat{q},\ \ f(x) + \hat{q}\,],$$

por lo que el ancho de cada intervalo es constante y viene dado por

$$U(x) - L(x) = (f(x) + \hat{q}) - (f(x) - \hat{q}) = 2\hat{q}.$$

En contraste, en las variantes de Mondrian (MCP) y de cuantiles (CQR) los intervalos sí dependen de cada ejemplo, ya que se ajustan según el grupo (en MCP) o de manera directa a través de cuantiles (en CQR). En estos casos, el tamaño de los intervalos varía con x_i y la métrica relevante es el tamaño promedio de los intervalos en el conjunto de prueba.

4.3. Configuración Experimental

Para realizar la aplicación de prediccion conforme a nuestras tareas de regresión debemos hacer ciertos ajustes a nuestros datos y modelos. A continuación describiremos como fue que configuramos nuestro pryecto para cumplir con los objetivos trazados.

4.3.1. Preparación de los datos

Antes de entrenar los modelos fue necesario realizar una preparación exhaustiva de las imágenes del conjunto *BreastPathQ*. El dataset original estaba compuesto por 2,579 parches de WSI etiquetados para entrenamiento y validación, y 1,121 parches adicionales reservados como conjunto de prueba. Sin embargo, este último no incluía etiquetas reales, lo cual imposibilitaba su uso en el proceso de evaluación. En particular, métricas como la cobertura requieren conocer el valor real de la celularidad para cada imagen, de manera que se pueda verificar si este se encuentra dentro del intervalo de predicción. Por este motivo, se decidió descartar el conjunto de prueba oficial y trabajar únicamente con los parches de entrenamiento y validación provistos por los organizadores.

A partir de estos 2,579 parches disponibles se realizó una división en tres subconjuntos: entrenamiento, validación y calibración. Para ello, se aplicó un muestreo estratificado en función de la variable de celularidad, discretizada en deciles mediante la función q-cut de pandas. De esta forma se garantizó que la distribución de celularidad quedara representada de manera equilibrada en los tres conjuntos. El 70 % de los datos se destinó a entrenamiento, y el 30 % restante se dividió en partes iguales para validación y calibración.

Posteriormente, el conjunto de entrenamiento se sometió a un proceso de balanceo por sobre-muestreo (oversampling). Concretamente, se generó un muestreo uniforme de 300

imágenes por cada bin de celularidad, aplicando reemplazo en los bins con menos ejemplos. Esto permitió obtener un conjunto de entrenamiento balanceado en torno a la distribución de la variable de interés.

Adicionalmente, se incorporaron refuerzos moderados en regiones clave de la distribución: se añadieron 300 imágenes con celularidad $y = 0.0$, otras 300 en el rango intermedio ($0.3 \leq y \leq 0.7$), y 300 con celularidad alta ($y > 0.7$). Esta estrategia buscó reducir el sesgo del modelo hacia zonas específicas de la distribución y mejorar su capacidad de generalización.

De manera complementaria, se incorporó un ajuste manual orientado a incrementar la presencia de parches con menor nitidez dentro del conjunto de entrenamiento. La decisión se tomó de forma empírica, con el objetivo de asegurar que este tipo de variabilidad estuviera representado en el proceso de aprendizaje. Para evitar incrementar artificialmente el tamaño del conjunto, se reemplazó un número equivalente de imágenes seleccionadas aleatoriamente, manteniendo así el tamaño final constante.

Finalmente, tras todos estos ajustes, se obtuvo un conjunto de 3,900 imágenes para entrenamiento, 359 para validación y 360 para calibración. Estos subconjuntos se exportaron en carpetas separadas, cada una acompañada de su archivo CSV de etiquetas correspondiente, constituyendo la base definitiva para el entrenamiento y evaluación de los modelos desarrollados en este trabajo.

4.3.2. Construcción de dataset con subgrupos basados en blurriness

Definimos también subgrupos de datos (categorías) para calcular un cuantil independiente para cada uno de ellos. En nuestro caso, las etiquetas del dataset original no contenían información suficiente para derivar subgrupos relevantes, ya que únicamente se disponía del porcentaje de celularidad tumoral por parche. Por este motivo, se implementó el procedimiento adicional aplicado para clasificación en la sección 3.3.3, basado en la métrica visual de las imágenes: "blurriness". Con estas nuevas etiquetas "Blurry", "Moderate", "Sharp" (imágenes borrosas, moderadas y nítidas) haremos un nuevo split y así tendremos un dataset distinto. Una vez asignado cada parche a un grupo, se realizó una división estratificada 70 %–15 %–15 % (entrenamiento, validación y calibración) dentro de cada subgrupo, preservando la distribución de la variable de celularidad mediante binning en deciles. Esto garantizó que las tres particiones mantuvieran ejemplos representativos tanto en niveles de celularidad como en grados de desenfoque. Al igual que en la preparación general de datos en la sección 4.3.1, se aplicó un oversampling uniforme por deciles (300 imágenes por bin), seguido de eso se incorporó un ajuste manual orientado a incrementar la presencia de parches con menor nitidez dentro del conjunto de entrenamiento, y refuerzos específicos en zonas críticas de la distribución ($y = 0.0$, $0.3 \leq y \leq 0.7$ y $y > 0.7$). El resultado global para nues-

tros splits fueron: para entrenamiento quedaron configuradas 3900 imágenes, para validación 359 y calibración 360. Este mismo procedimiento de agrupación por desenfoque se aplicó a las imágenes del conjunto de prueba, empleando los cuantiles calculados en entrenamiento para asegurar consistencia. Así, cada parche de test quedó asignado a un grupo (Borrosa, Moderada o Nítida), permitiendo evaluar la cobertura de Mondrian conforme a subgrupos definidos objetivamente.

4.3.3. Configuración de los modelos de regresión

Como se mencionó anteriormente, se reutilizaron las arquitecturas de EfficientNet, y DenseNet para mantener consistencia metodológica con los experimentos de clasificación y facilitar la comparación bajo predicción conforme. Más especificamente, se utilizaron las arquitecturas de EfficientNet-B3 y DenseNet-169. Estos modelos se inicializaron con pesos preentrenados en ImageNet y se adaptaron a regresión reemplazando la capa de salida por un head completamente conectado (con capas intermedias y dropout) que produce un único valor continuo.

La adaptación del head en regresión se observa en la tabla 8.

Tabla 8: *Parámetros de los modelos de regresión adaptados de redes convolucionales*

Modelo	Backbone	Feat dim	Head MLP	Parámetros del head
EfficientNet-B3	ImageNet (IMAGENET1K_V1)	1536	Linear(1536,512) ReLU Dropout(0.4) Linear(512,128) ReLU Dropout(0.4) Linear(128,1)	852,737
DenseNet-169	ImageNet (IMAGENET1K_V1)	1664	Linear(1664,512) ReLU Dropout(0.4) Linear(512,128) ReLU Dropout(0.4) Linear(128,1)	918,273

4.3.3.1. Esquema de entrenamiento y validación. El entrenamiento de los modelos requirió definir no solo la arquitectura base, sino también un conjunto de transformaciones y estrategias de optimización que aseguran un buen desempeño y una adecuada capacidad de generalización. A continuación se detallan los componentes principales del proceso.

Preprocesamiento y aumentaciones. Todas las imágenes fueron redimensionadas a 380×380 píxeles para estandarizar la entrada a la red. Posteriormente se aplicó la normalización con medias y desviaciones estándar de ImageNet $(0.485, 0.456, 0.406)$ y $(0.229, 0.224, 0.225)$. En el conjunto de entrenamiento se incorporaron aumentaciones suaves —volteo horizontal, cambios de brillo y contraste, y transformaciones afines de rotación, escala y *shear*— con el objetivo de incrementar la diversidad de ejemplos y reducir el sobreajuste. Para validación se utilizó únicamente el redimensionamiento y la normalización, garantizando así una evaluación más estable.

Carga de datos. Los parches se organizaron en lotes de tamaño 16 y se utilizaron

cuatro workers para paralelizar la carga. Durante el entrenamiento los datos se mezclaron aleatoriamente, mientras que en validación se mantuvo el orden fijo para garantizar reproducibilidad.

Función de pérdida. Con el fin de manejar adecuadamente errores pequeños y a la vez ser robustos ante valores atípicos, se utilizó la pérdida de Huber, definida para un residuo $r_i = \hat{y}_i - y_i$ y un parámetro $\delta = 1.0$ como:

$$\ell_\delta(r_i) \;=\; \begin{cases} \dfrac{1}{2} r_i^2, & \text{si } |r_i| \le \delta, \\ \delta\Big(|r_i| - \tfrac{1}{2}\delta\Big), & \text{en otro caso.} \end{cases}$$

Para dar mayor importancia a las regiones de celularidad intermedia (0.3–0.7) y alta (> 0.7), se introdujo un esquema de ponderación:

$$w_i \;=\; \begin{cases} 2.0, & 0.3 \le y_i \le 0.7, \\ 1.5, & y_i > 0.7, \\ 1.0, & \text{en otro caso,} \end{cases} \qquad \mathcal{L} \;=\; \frac{1}{n} \sum_{i=1}^{n} w_i \, \ell_\delta(r_i).$$

Optimización y regularización. Para la optimización se empleó el algoritmo AdamW con una tasa de aprendizaje inicial de 3×10^{-4} y un weight decay de 10^{-4} para mitigar el sobreajuste. Adicionalmente se implementó un plan de aprendizaje Cosine Annealing, que reduce paulatinamente la tasa de aprendizaje siguiendo una curva coseno con $T_{\text{máx}} = 200$, favoreciendo la convergencia estable.

Criterios de parada. El entrenamiento se planificó para un máximo de 1000 épocas, pero se incluyó un esquema de early stopping con paciencia de 200 épocas, tomando como criterio principal la mejora del coeficiente de determinación R^2 en validación. Cada vez que el R^2_{val} alcanzaba un nuevo máximo, el modelo se almacenaba como checkpoint para evitar retrocesos en desempeño.

Métricas de evaluación. La evaluación de cada época se realizó mediante el cálculo del error cuadrático medio (MSE) y el coeficiente R^2. Este último se consideró la métrica principal para seleccionar los modelos, dado que refleja la proporción de varianza explicada en las predicciones.

Tabla 9: *Resumen de arquitecturas y configuración de entrenamiento*

Modelo	Feat dim	Paráms	Head MLP	Optimización
EfficientNet-B3	1536	852,737	Linear(1536,512) → ReLU → Dropout(0.4) → Linear(512,128) → ReLU → Dropout(0.4) → Linear(128,1)	AdamW (3e-4, 1e-4), CosineAnnealing ($T_{\text{máx}} = 200$), pérdida de Huber ponderada ($\delta = 1.0$)
DenseNet-169	1664	918,273	Linear(1664,512) → ReLU → Dropout(0.4) → Linear(512,128) → ReLU → Dropout(0.4) → Linear(128,1)	AdamW (3e-4, 1e-4), CosineAnnealing ($T_{\text{máx}} = 200$), pérdida de Huber ponderada ($\delta = 1.0$)

Resumen de arquitecturas y configuración de entrenamiento

Modelo	Feat dim	Paráms	Head MLP	Optimización
EfficientNet-B3	1536	852,737	Linear(1536,512) → ReLU → Dropout(0.4) → Linear(512,128) → ReLU → Dropout(0.4) → Linear(128,1)	AdamW (3e-4, 1e-4), CosineAnnealing ($T_{\text{máx}} = 200$), pérdida de Huber ponderada ($\delta = 1.0$)
DenseNet-169	1664	918,273	Linear(1664,512) → ReLU → Dropout(0.4) → Linear(512,128) → ReLU → Dropout(0.4) → Linear(128,1)	AdamW (3e-4, 1e-4), CosineAnnealing ($T_{\text{máx}} = 200$), pérdida de Huber ponderada ($\delta = 1.0$)

4.3.4. Configuración predicción conforme

Para cada variante de predicción conforme se utilizó una configuración especifica:

4.3.4.1. Predicción conforme inductiva. Fijamos el nivel de significancia en $\alpha = 0.10$ con cobertura objetivo $1 - \alpha = 90\,\%$. El flujo de ICP para regresión se estructuró en

dos etapas: calibración (para estimar el cuantil conforme) y predicción en validación (para medir cobertura y tamaño de intervalos).

Sobre el conjunto de calibración, se obtuvieron predicciones puntuales $\hat{y}_i$ y se calcularon scores de no conformidad como el error absoluto

$$s_i \;=\; \Big| y_i - \hat{y}_i \Big|.$$

El cuantil conforme se estimó como el percentil empírico al nivel $(1-\alpha)$:

$$\hat{q} \;=\; \text{Percentil}_{\,100\,(1-\alpha)}\Big(\{s_i\}_{i=1}^{n}\Big).$$

(Usamos percentil empírico directo; la corrección finito–muestral con índice $\lceil (n+1)(1-\alpha) \rceil$ es equivalente en la práctica para n moderado–grande). El valor $\hat{q}$ se persistió en disco como `q_value.npy` para su reutilización en inferencia.

En el conjunto de prueba, cada intervalo conforme se construyó como

$$C(x) \;=\; \Big[\, \hat{y}(x) - \hat{q},\;\; \hat{y}(x) + \hat{q} \,\Big],$$

con ancho constante $2\hat{q}$. Se reportó la cobertura empírica

$$\widehat{\text{Cov}} \;=\; \frac{1}{m} \sum_{j=1}^{m} \mathbf{1}\{\, y_j \in C(x_j) \,\},$$

y el tamaño promedio de los intervalos $\frac{1}{m}\sum_j \Big(U(x_j) - L(x_j)\Big) = 2\hat{q}$.

4.3.4.2. Predicción conforme de Mondrian. Fijamos el nivel de significancia en $\alpha = 0.10$ (cobertura objetivo $1-\alpha = 90\,\%$). MCP replica el flujo de ICP pero condicionado por subgrupos previamente definidos; en nuestro caso, el subgrupo es `blurriness_group` ("Blurry", "Moderate", "Sharp").

Sobre el conjunto de calibración, para cada muestra (x_i, y_i) se obtuvo la predicción puntual $\hat{y}_i$ y se calculó el score de no conformidad dentro de su grupo g:

$$s_i^{(g)} \;=\; \Big| y_i - \hat{y}_i \Big|, \qquad x_i \in g.$$

Para cada grupo g con n_g muestras de calibración, se ordenaron los scores y se aplicó la corrección finito–muestral:

$$\hat{q}^{(g)} \;=\; s^{(g)}_{(\,\lceil (n_g+1)(1-\alpha) \rceil)},$$

Adicionalmente, se calculó un $\hat{q}_{\text{global}}$ (a partir de todos los scores de calibración) como fallback para casos donde un grupo carezca de suficientes muestras en validación/test.

En el conjunto de prueba, el intervalo conforme para una muestra x perteneciente al grupo g se construyó como

$$C(x) \;=\; \Big[\, \hat{y}(x) - \hat{q}^{(g)},\;\; \hat{y}(x) + \hat{q}^{(g)} \,\Big],$$

es decir, con ancho nominal $2\hat{q}^{(g)}$ dentro de cada subgrupo. Si $\hat{q}^{(g)}$ no estuviera disponible (p. ej., grupo ausente), se usó $\hat{q}_{\text{global}}$ como sustituto.

No obstante, el tamaño promedio que se reporta por grupo se calculó de forma empírica sobre el conjunto de prueba: para cada grupo g con m_g muestras de test definimos

$$\widehat{\text{Size}}^{(g)} \;=\; \frac{1}{m_g} \sum_{i:x_i \in g} \Big(U(x_i) - L(x_i) \Big),$$

donde, cuando procedía, se recortaron los extremos a $L(x_i) = \text{máx}(0, \hat{y}(x_i) - \hat{q}^{(g)})$ y $U(x_i) = \text{mín}(1, \hat{y}(x_i) + \hat{q}^{(g)})$ para asegurar que los intervalos permanezcan en $[0, 1]$.

4.3.4.3. Predicción conforme basada en regresión por cuantiles. Fijamos $\alpha = 0.10$ (cobertura objetivo $1 - \alpha = 90\,\%$). Entrenamos un modelo que, para cada imagen x, predice dos cuantiles condicionales en $[0, 1]$: $\hat{q}_{0.05}(x)$ y $\hat{q}_{0.95}(x)$. En calibración, para cada (x_i, y_i) se calculó el score de no conformidad como el exceso del valor real por fuera del intervalo base dependiente de x:

$$s_i \;=\; \text{máx}\Big(\hat{q}_{0.05}(x_i) - y_i,\;\; y_i - \hat{q}_{0.95}(x_i),\;\; 0 \Big).$$

Con $\{s_i\}_{i=1}^{n}$ ordenados, se estimó el cuantil conforme con corrección finito–muestral:

$$\hat{q} \;=\; s_{(\,\lceil (n+1)(1-\alpha) \rceil)},$$

y se persistió para inferencia.

En la prueba, el intervalo conforme final para una muestra x se construyó como

$$C(x) \;=\; \Big[\, \hat{q}_{0.05}(x) - \hat{q},\;\; \hat{q}_{0.95}(x) + \hat{q} \,\Big],$$

esto es, un intervalo adaptativo dependiente de x ampliado por el ajuste global $\hat{q}$. Se reportó la cobertura empírica, y para el tamaño promedio de los intervalos se realizó de la misma manera que para la predicción conforme de Mondrian en la sección 4.3.4.2.

4.4. Resultados

En esta sección se presentan los resultados de los modelos de regresión aplicando ICP, MCP y CQR. En lugar de dividir por arquitectura, los resultados se organizan por técnica de predicción conforme, lo que permite comparar directamente el desempeño de EfficientNet-B3 y DenseNet-169 dentro de cada enfoque.

Se presentan los elementos mínimos para evaluar la predicción conforme: cobertura y tamaño de intervalos, como se describe en la sección 4.2.4.

4.4.1. Estimación puntual de la celularidad

Antes de entrar a predicción conforme, resumimos el rendimiento base en validación. La tabla 10 reporta tres puntajes R^2 por arquitectura, uno por cada técnica (ICP, MCP y CQR). Las diferencias entre ellos se explican por dos factores: el dataset, porque MCP usó un particionado distinto al incorporar el atributo de nitidez (*blurriness*); y el objetivo de entrenamiento, porque en CQR se optimizaron cuantiles con *pinball loss* y el R^2 se calculó sobre la predicción central $\left(\hat{q}_{0.05} + \hat{q}_{0.95}\right)/2$. Así, ICP y CQR comparten split pero difieren por el objetivo, mientras que MCP difiere por el split; estos dos elementos justifican los tres R^2 de la tabla. En particular, las diferencias entre ICP y MCP son mínimas: 0.0032 para EfficientNet-B3 y 0.0067 para DenseNet-169.

Tabla 10: *R^2 en validación por técnica de predicción conforme y arquitectura*

Modelo	**ICP**	**MCP**	**CQR**
EfficientNet-B3	0.9287	0.9319	0.9403
DenseNet-169	0.9046	0.8979	0.9341

4.4.2. Predicción Conforme Inductiva

Una vez teniendo los modelos entrenados, se realizó el proceso completo de predicción conforme y obtuvimos los siguientes resultados.

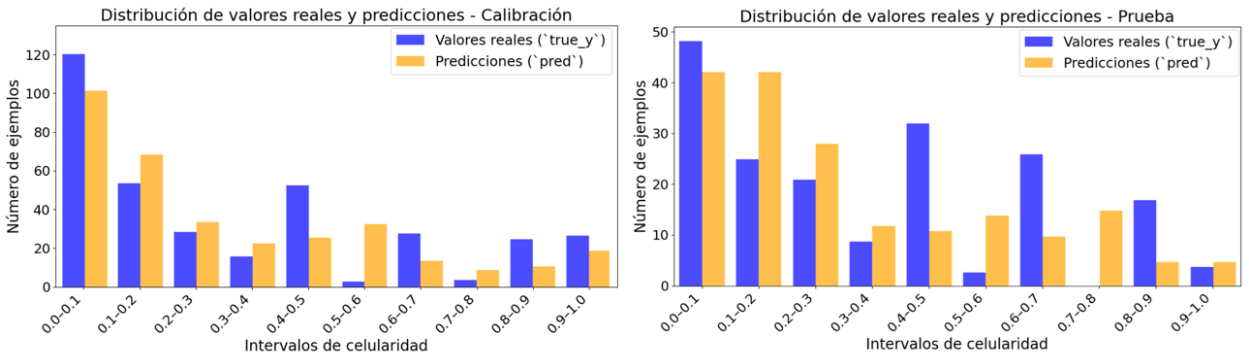
Para la predicción conforme inductiva se configuró para alcanzar un 90 % de confianza en las predicciones, como se mencionó anteriormente, y una vez se realizó la calibración, obtuvimos estos resultados para EfficientNet-B3 y DenseNet-169.

EfficientNet-B3. Para EfficientNet-B3 realizamos la calibración y obtuvimos los resultados de las predicciones frente a los valores reales. Tras calcular el cuantil, se obtuvo un valor de $q = 0.1379$. Este cuantil permitió construir, para nuestro conjunto de calibración, un primer acercamiento de los intervalos de predicción conforme con un nivel de confianza aproximado del 90 %.

De manera similar, en el conjunto de prueba se aplicó el mismo procedimiento: se compararon las predicciones con los valores reales y se construyeron los intervalos de predicción conforme usando el cuantil obtenido en calibración.

En la figura 21 podemos observar la diferencia entre predicciones y valores reales en los conjuntos de calibración y prueba. En calibración, las predicciones se ajustan razonablemente a los valores reales, con errores más notorios en los bins intermedios, lo que indica un desempeño aceptable. En el conjunto de prueba, se aprecia una mayor diferencia en los bins $[0.4 - 0.5]$, $[0.5 - 0.6]$ y $[0.6 - 0.7]$, así como predicciones en el bin $[0.7 - 0.8]$ a pesar de que no existían valores reales en ese rango.

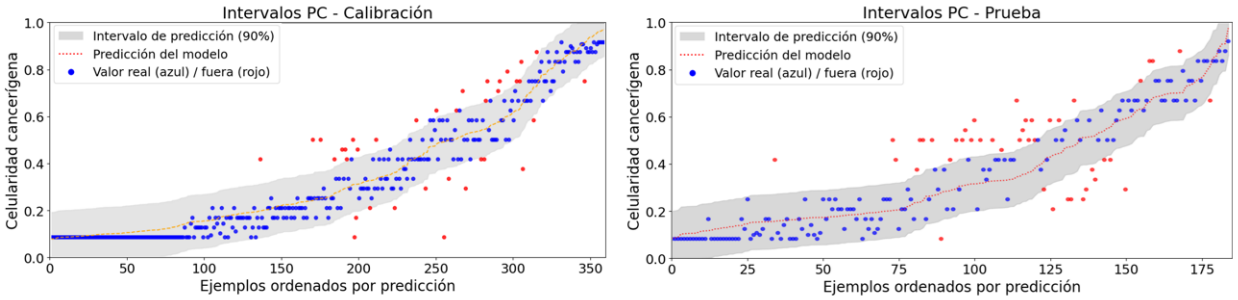
Figura 21: *Predicciones frente a valores reales para los conjuntos de calibración y prueba para EfficientNet-B3 en ICP*



Al construir los intervalos de predicción conforme, podemos evaluar qué tan alejados se encuentran los valores reales de las predicciones y estimar la cobertura lograda.

En la figura 22 se muestran los puntos azules, que representan valores dentro del intervalo de confianza, y los puntos rojos, que quedan fuera. En calibración, algunos valores se encuentran fuera de los intervalos, como era de esperar, dado que el objetivo era aproximar una cobertura del 90 %. Para el conjunto de prueba, la cobertura obtenida fue del 76.76 %, inferior al valor esperado. Esto se debe principalmente a los valores que quedan fuera del intervalo en los bins centrales, lo que reduce la cobertura global.

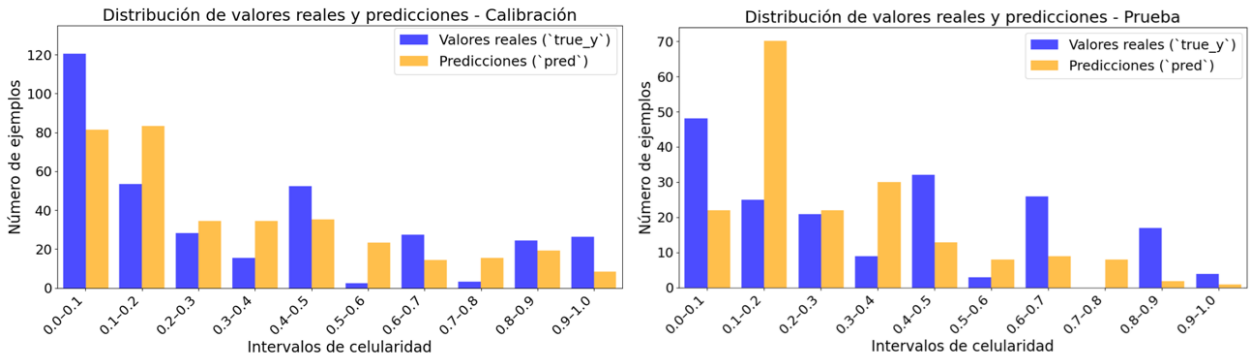
Figura 22: *Intervalos de predicción conforme para calibración y prueba para EfficientNet-B3 en ICP.* ***Azul*** *= predicción dentro del intervalo;* ***rojo*** *= predicción fuera.*



DenseNet-169. Para la arquitectura DenseNet-169 realizamos nuevamente la comparación entre valores predichos y reales, así como el análisis de los intervalos de predicción conforme, tanto para calibración como para prueba.

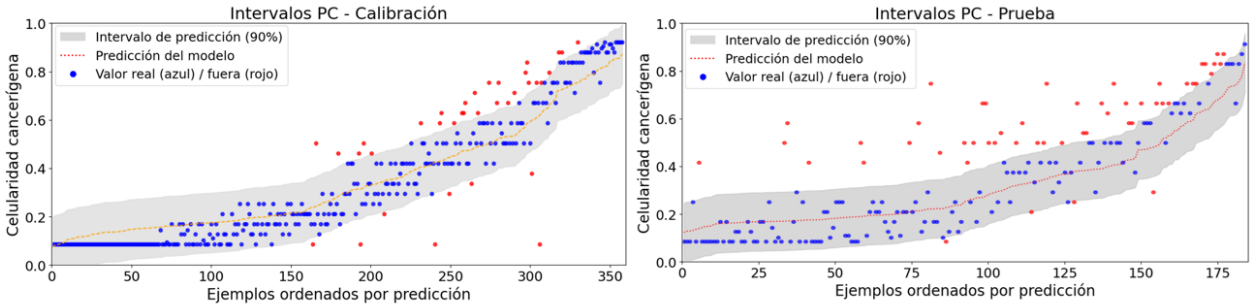
En calibración, los resultados fueron aceptables con errores más pronunciados en el primer bin y en algunos bins finales. En el caso de prueba, los errores son más notorios en los bins iniciales y finales, como se muestra en la figura 23. Al igual que en EfficientNet, se predijeron valores en el bin $[0.7 - 0.8]$, aunque no había valores reales en ese rango.

Figura 23: *Predicciones frente a valores reales para calibración y prueba para DenseNet-169 en ICP*



Tras calcular el cuantil para calibración, se obtuvo $q = 0.1483$. Con este valor se construyeron los intervalos para ambos conjuntos, observándose en la figura 24 que algunos puntos en calibración quedan fuera del intervalo, como era de esperar, dado que se utilizó este conjunto para determinar el cuantil. En prueba, una mayor cantidad de predicciones se encuentra fuera del intervalo, lo que explica la cobertura obtenida del 71.89 %.

Figura 24: *Intervalos de predicción conforme para calibración y prueba para DenseNet-169 en ICP.* ***Azul*** *= predicción dentro del intervalo;* ***rojo*** *= predicción fuera*



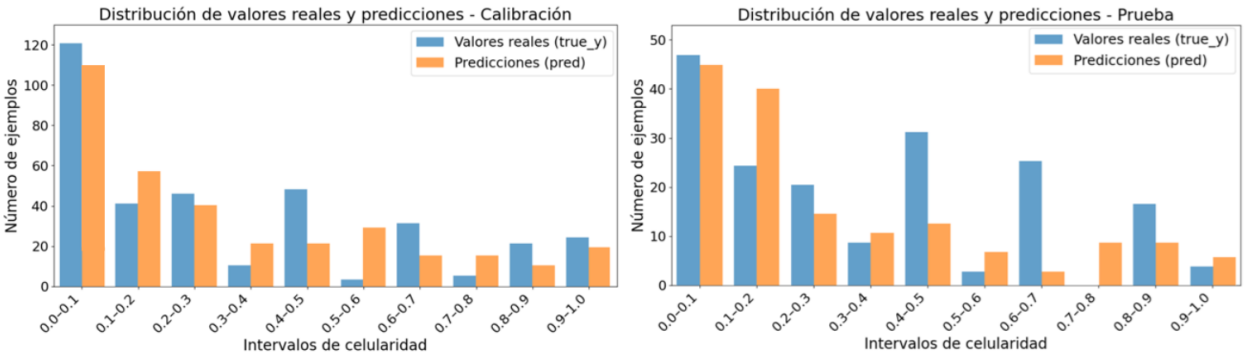
4.4.3. Predicción conforme de Mondrian

Para la sección MCP una vez entrenamos nuestros modelos procedemos a realizar el proceso de predicción conforme y plasmar los resultados. Es importante recordar que se fijó el nivel de confianza a 90 %.

Para esta sección vamos a tener tres cuantiles distintos, uno para cada uno de los subgrupos definidos en la sección 4.3.2.

EfficientNet-B3. Siguiendo la misma dinámica de los resultados, en la figura 25 se muestran primero las diferencias entre valores predichos y reales en los datos de calibración y prueba. En calibración las predicciones resultaron relativamente acertadas, mientras que para nuestro conjunto de prueba obtuvimos en los bins $[0.1 - 0.2]$, $[0.4 - 0.5]$ y $[0.6 - 0.7]$ una alta diferencia entre predicciones y valores reales. Nuevamente el bin $[0.7 - 0.8]$ tiene predicciones sin tener valores reales incluidos.

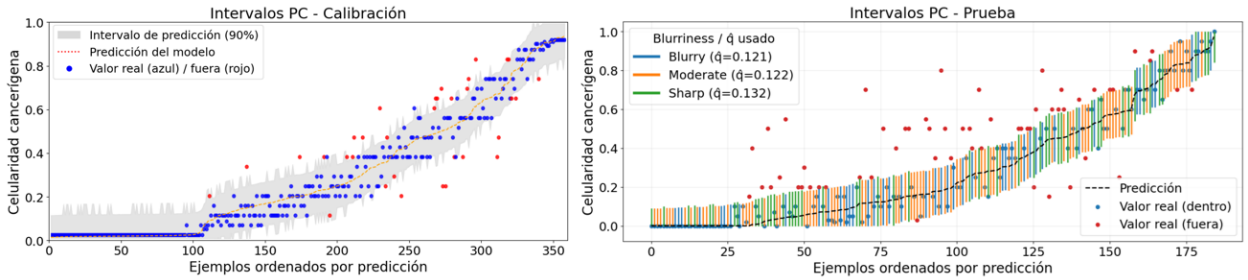
Figura 25: *Predicciones frente a valores reales para calibración y prueba para EfficientNet-B3 en MCP*



En la variante anterior de predicción conforme se obtenía un único cuantil; aquí, al considerar tres subgrupos por calidad de imagen, se estimó un cuantil por subgrupo: borrosas $= 0.1115$, moderadas $= 0.1122$ y nítidas $= 0.1671$.

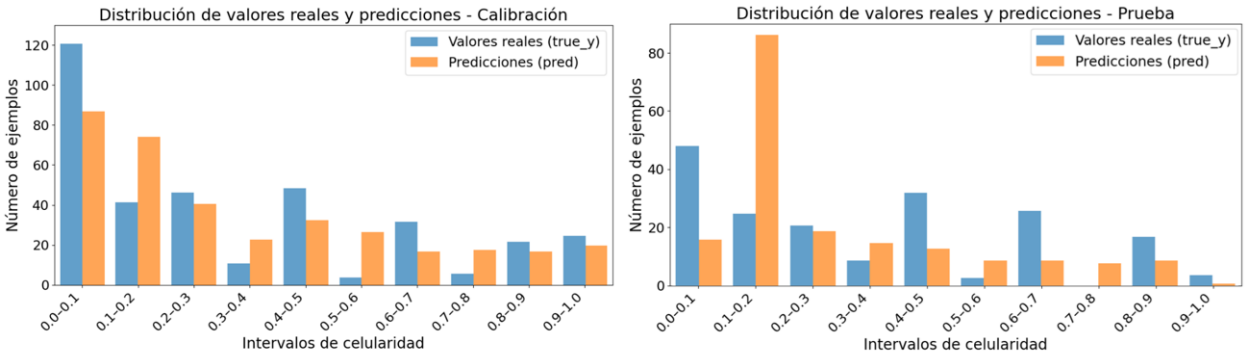
Cada valor permitió construir intervalos distintos. En la figura 26 se aprecia primero el comportamiento en calibración, donde los intervalos se adaptan para alcanzar aproximadamente el 90 % de cobertura. A la derecha se presentan los resultados sobre las imágenes de prueba. Los colores distinguen los subgrupos: azul para borrosas, amarillo para moderadas y verde para nítidas. En esta gráfica se evidencia la variación entre intervalos, aunque los cuantiles de los dos primeros grupos sean bastante similares. Para el set de prueba, la cobertura alcanzada fue de 67.03 % y se puede evidenciar observando la cantidad de puntos rojos (predicciones del modelo) por fuera del intervalo.

Figura 26: *Intervalos de predicción conforme para calibración y prueba para EfficientNet-B3 en MCP. Dentro del intervalo: color por subgrupo de calidad (**borrosas** = azul, **moderadas** = amarillo, **nítidas** = verde)*



DenseNet-169. Para DenseNet, las predicciones obtenidas en los conjuntos de calibración y prueba se ilustran en la figura 27. En el primero, los resultados fueron en general aceptables dentro del margen de error, con discrepancias más notorias en los bins iniciales e intermedios. Por el otro lado, en el conjunto de prueba, se observó un aumento de predicciones en el segundo bin y mayores diferencias en los rangos altos. El intervalo $[0.7 - 0.8]$ volvió a ser el más problemático.

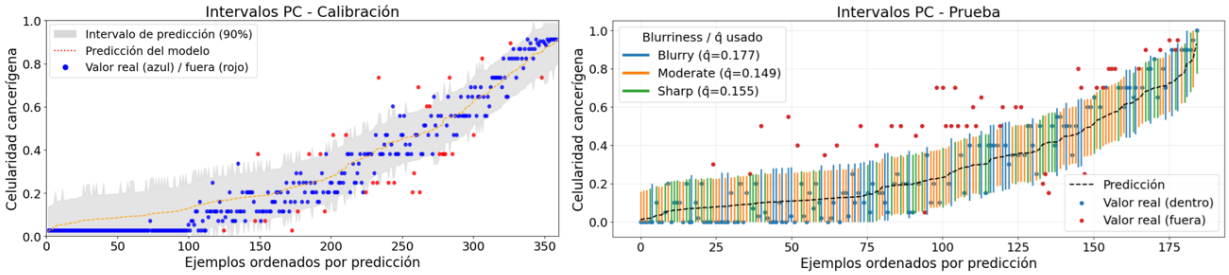
Figura 27: *Predicciones frente a valores reales para calibración y prueba para DenseNet-169 en MCP*



En este caso, los tres subgrupos de calidad de imagen produjeron cuantiles distintos: borrosas = 0.1287, moderadas = 0.1264 y nítidas = 0.1865

A partir de estos valores se construyeron los intervalos de predicción conforme, mostrados en la figura 28. Se mantuvo la misma paletade colores utilizada con EfficientNet (azul para borrosas, amarillo para moderadas y verde para nítidas). En calibración, el comportamiento fue el esperado, con pocas observaciones fuera del rango; sin embargo, en prueba el número de puntos excluidos aumentó considerablemente, reduciendo la cobertura a 63.24 %.

Figura 28: *Intervalos de predicción conforme para calibración y prueba para DenseNet-169 en MCP. Dentro del intervalo: color por subgrupo de calidad (**borrosas** = azul, **moderadas** = amarillo, **nítidas** = verde)*

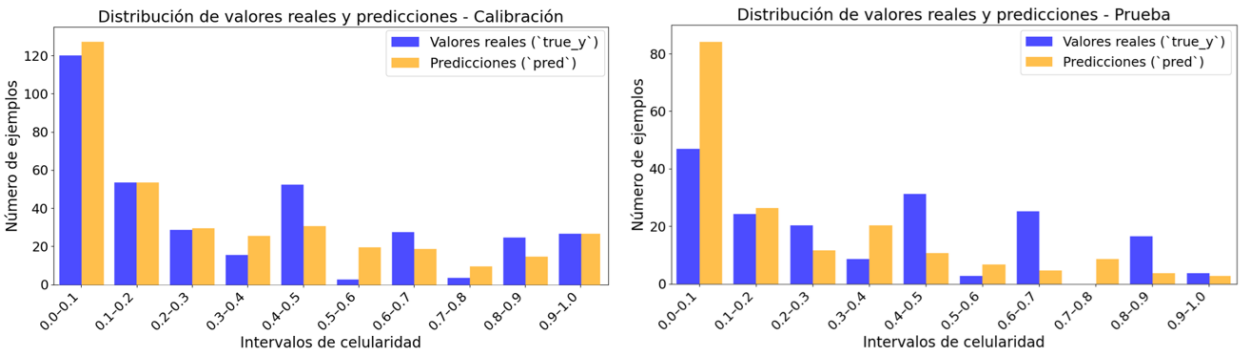


4.4.4. Predicción conforme basada en regresión por cuantiles

Para la predicción conforme basada en regresión por cuantiles tenemos la misma dinámica que en ICP. Presentaremos diferencias en predicciones, intervalos y pacientes para nuestros datasets de calibración y prueba. Nuevamente se recuerda que se apuntó por un 90 % de nivel de confianza.

EfficientNet-B3. En la figura 29 se presentan las predicciones obtenidas en los conjuntos de calibración y prueba. En el primero, los resultados fueron en general adecuados, con una mínima diferencia frente a los valores reales. En contraste, en las muestras de prueba se observa una desigualdad alta marcada en el primer bin, lo que repercute en el desempeño global. Al igual que en experimentos anteriores, el intervalo $[0.7 - 0.8]$ registró predicciones sin que existieran valores reales en ese rango.

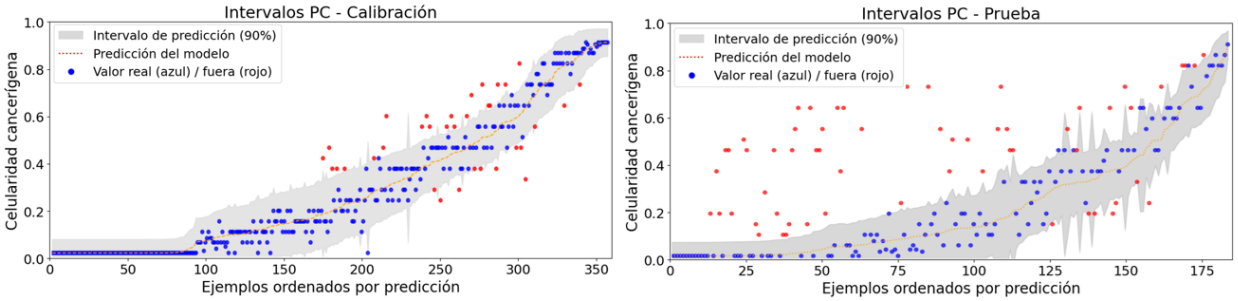
Figura 29: *Predicciones frente a valores reales para calibración y prueba para EfficientNet-B3 en CQR*



A partir de las predicciones y los scores, se calculó el cuantil utilizado para ajustar los límites inferior y superior de cada intervalo, cuyo valor fue $q = 0.0655$. Con este ajuste, los intervalos resultantes se muestran en la figura 30, alcanzando en el conjunto de prueba una cobertura del 66.49 %.

Los gráficos evidencian claramente cómo los intervalos se adaptan a cada predicción. En calibración, los puntos fuera del rango corresponden a lo esperado al buscar una cobertura del 90 %. En prueba, en cambio, se aprecia un número considerable de valores que quedan fuera, lo que explica la disminución de la cobertura total.

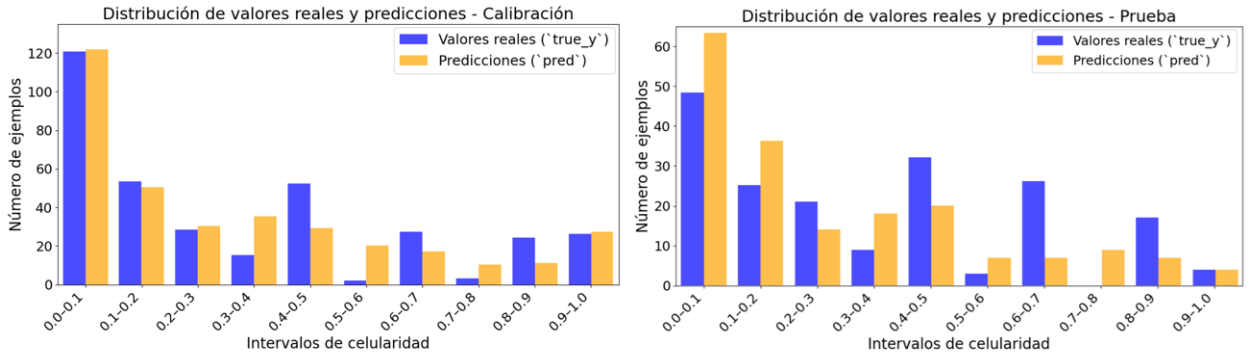
Figura 30: *Intervalos de predicción conforme para calibración y prueba para EfficientNet-B3 en CQR.* ***Azul*** *= predicción dentro del intervalo;* ***rojo*** *= predicción fuera*



DenseNet-169. Para nuestro modelo DenseNet dentro de la variante de regresión CQR se siguió el mismo procedimiento de presentación de resultados. En la figura 31 se ilustran las diferencias en las predicciones para los dos conjuntos de datos.

En el conjunto de calibración, las predicciones pueden considerarse relativamente buenas, con la mayoría de los errores concentrados en los bins intermedios. En cambio, para el conjunto de prueba se observa mayor discrepancia en el primer bin y en los bins finales. Como en experimentos anteriores, el bin $[0.7 - 0.8]$ vuelve a mostrar predicciones sin valores reales por predecir.

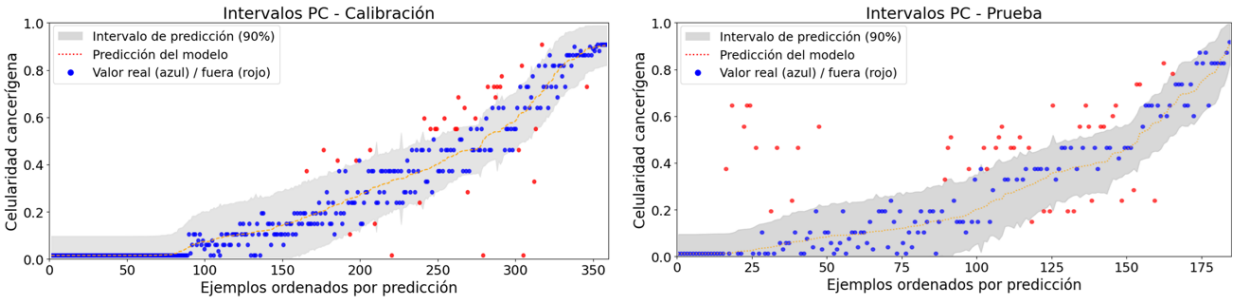
Figura 31: *Resultados de predicciones para calibración y prueba para DenseNet-169 en CQR*



En la figura 32 se presentan los intervalos de predicción conforme para ambos conjun-

tos. A partir de las predicciones y scores de calibración se calculó un cuantil de $q = 0.0921$. Para calibración, los errores son razonables dentro de lo esperado. Sin embargo, en el conjunto de prueba se observa un número considerablemente mayor de predicciones fuera del intervalo, lo que repercute en una cobertura inferior a la deseada: 74.59 %.

Figura 32: *Intervalos de predicción conforme para calibración y prueba para DenseNet-169 en CQR.* ***Azul*** *= predicción dentro del intervalo;* ***rojo*** *= predicción fuera*



4.5. Discusión de los resultados

En esta sección se analizan los resultados obtenidos con cada variante de predicción conforme, apoyándose en las tablas de coberturas y tamaños promedio, así como en las figuras que muestran los intervalos por paciente.

Las tablas 11 y 12 presentan un resumen global de la cobertura alcanzada y del tamaño promedio de los intervalos por técnica y arquitectura. A partir de ellas se observa que ninguna combinación alcanzó la cobertura objetivo del 90 %, y que existen diferencias tanto entre modelos como entre variantes de predicción conforme.

Tabla 11: *Cobertura en el conjunto de prueba por técnica de predicción conforme y arquitectura*

Modelo	ICP	MCP	CQR
EfficientNet-B3	76.76 %	67.03 %	66.49 %
DenseNet-169	71.89 %	63.24 %	74.59 %

Tabla 12: *Tamaño promedio de intervalos por técnica de predicción conforme y arquitectura*

Modelo	ICP	MCP	CQR
EfficientNet-B3	0.2758	0.2079	0.2634
DenseNet-169	0.2966	0.2860	0.2822

Para complementar la información, en la tabla 13 se observan los cuantiles obtenidos para las tres variantes de predicción conforme.

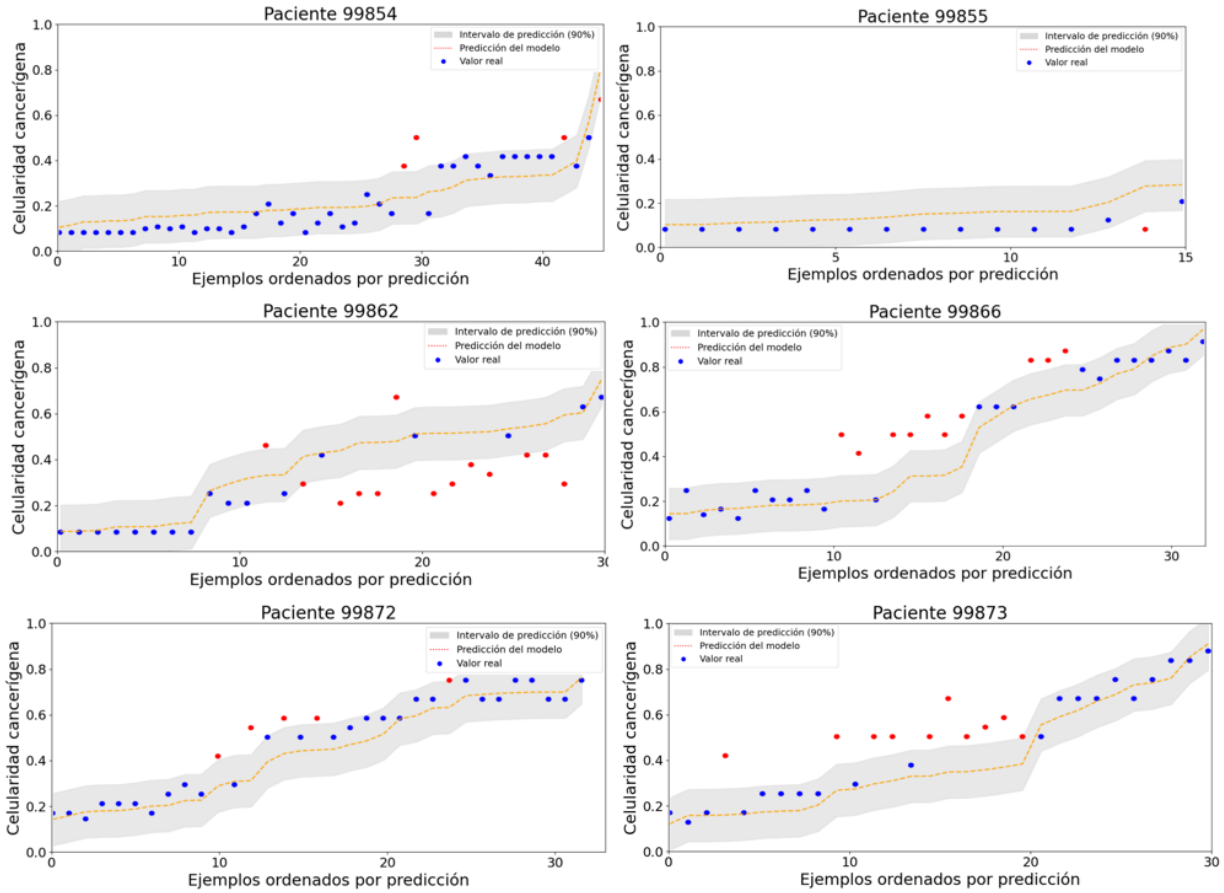
Tabla 13: *Cuantiles usados para construir intervalos por técnica de predicción conforme y arquitectura*

		MCP (subgrupos)			
Modelo	**ICP**	**Borrosas.**	**Moderadas.**	**Nítidas.**	**CQR**
EfficientNet-B3	0.1379	0.1115	0.1122	0.1671	0.0655
DenseNet-169	0.1483	0.1287	0.1264	0.1865	0.0921

Predicción conforme inductiva

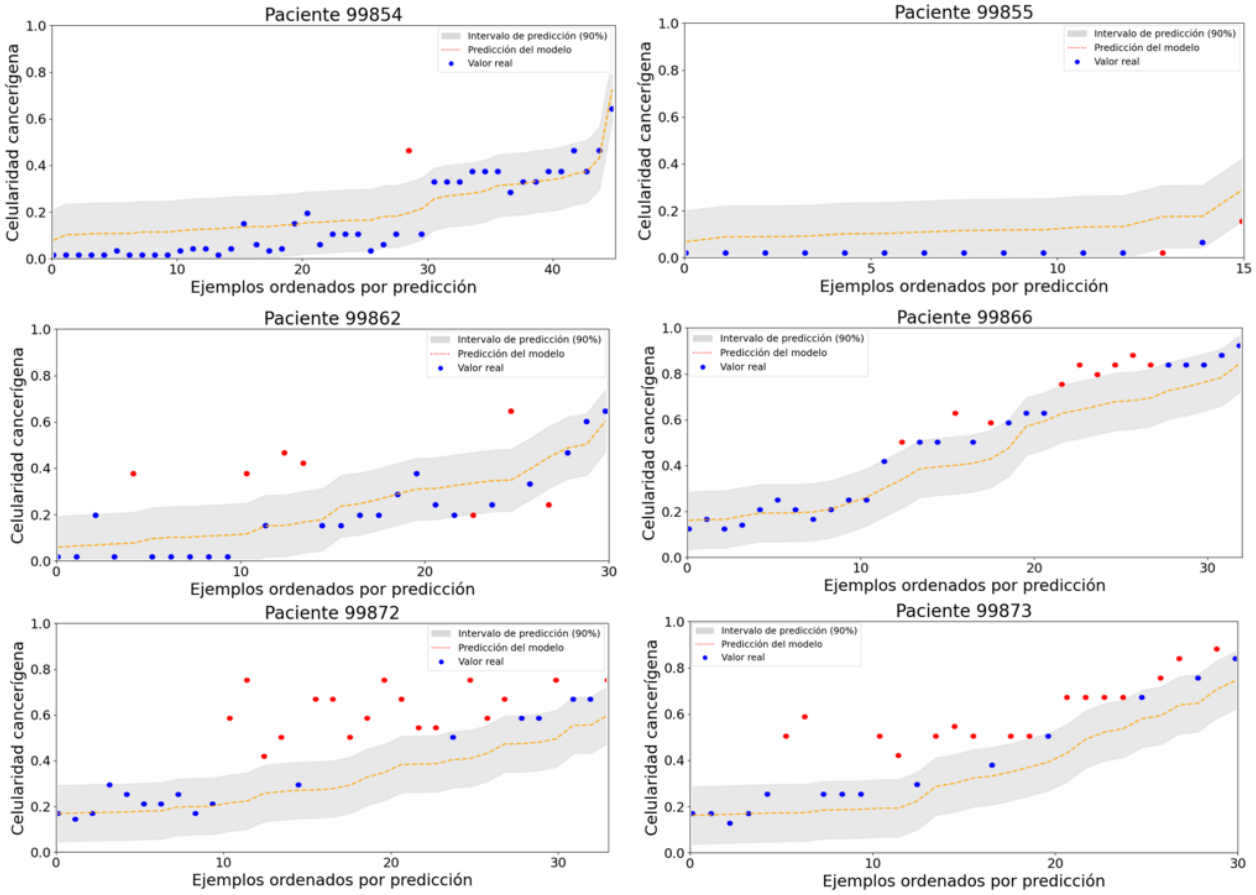
En la figura 33 se presentan los intervalos de ICP por paciente con EfficientNet-B3. Allí se observa que algunos pacientes, como **99854**, **99855** y **99872**, muestran intervalos relativamente estables con pocas predicciones fuera, mientras que en otros casos, como **99862**, **99866** y **99873**, aparecen varias etiquetas fuera del rango.

Figura 33: *Intervalos de predicción conforme por paciente para la arquitectura EfficientNet-B3 en ICP.* ***Azul*** *= predicción dentro del intervalo;* ***rojo*** *= predicción fuera*



En la figura 34, con DenseNet-169, se observa un patrón similar: la mayoría de pacientes permanecen contenidos en los intervalos, con excepciones en casos concretos como los pacientes **99872** y **99873**, donde se evidencian varias predicciones fuera del rango esperado.

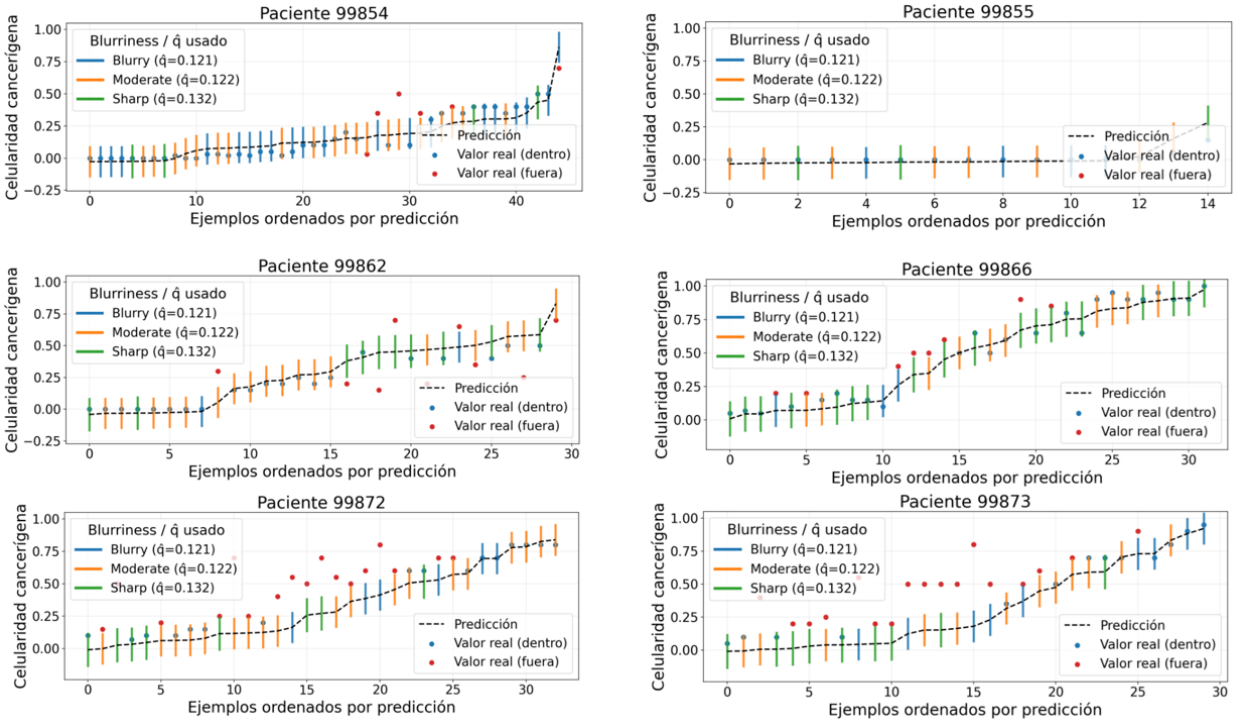
Figura 34: *Intervalos de predicción conforme por paciente para la arquitectura DenseNet-169 en ICP.* ***Azul*** *= predicción dentro del intervalo;* ***rojo*** *= predicción fuera.*



Predicción conforme de Mondrian

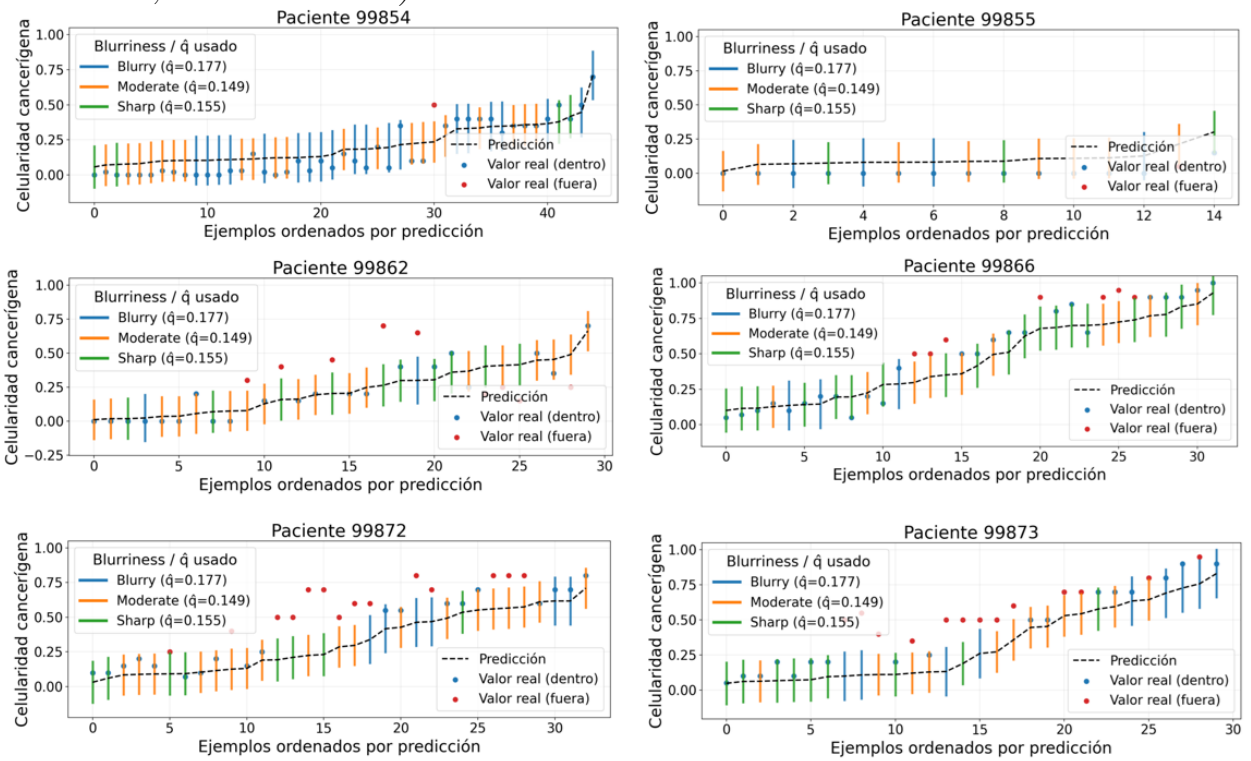
En la figura 35 se muestran los resultados por paciente con EfficientNet-B3. Cada barra está coloreada según el grupo de borrosidad. Se observa que, en la mayoría de los pacientes, las observaciones se concentran dentro de los intervalos, aunque algunos, como **99872** y **99873**, tienen más casos que quedaron fuera.

Figura 35: *Intervalos de predicción conforme por paciente para la arquitectura EfficientNet-B3 en MCP. Dentro del intervalo: color por subgrupo de calidad (**borrosas** = azul, **moderadas** = amarillo, **nítidas** = verde)*



La figura 36 muestra el mismo análisis para DenseNet-169. En este caso, la tendencia general es similar: la mayoría de los pacientes presentan intervalos adecuados, mientras que en algunos específicos aparecen varias predicciones fuera, especialmente en los mismos pacientes **99872** y **99873**.

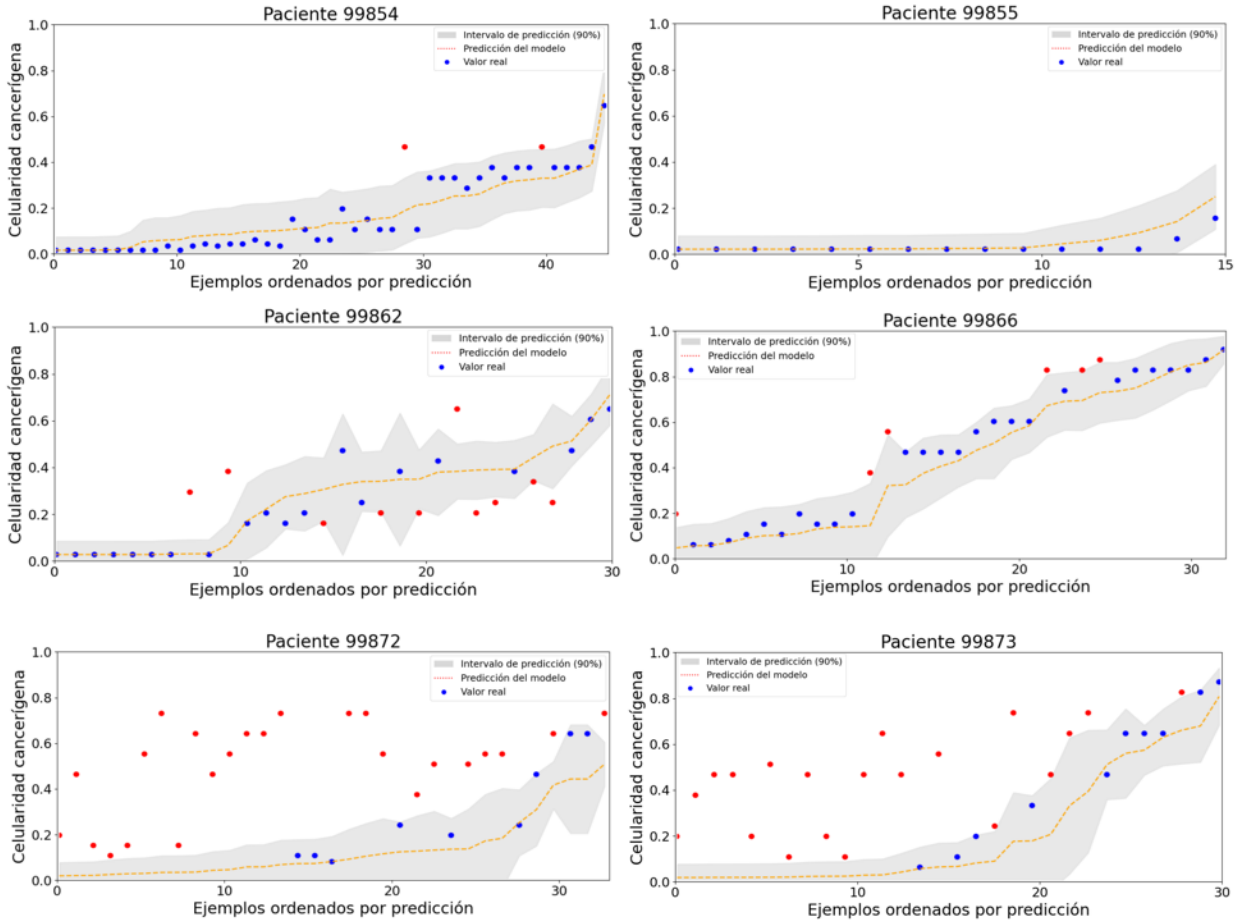
Figura 36: *Intervalos de predicción conforme por paciente para la arquitectura DenseNet-169 en MCP. Dentro del intervalo: color por subgrupo de calidad (**borrosas** = azul, **moderadas** = amarillo, **nítidas** = verde)*



Predicción conforme basada en regresión por cuantiles

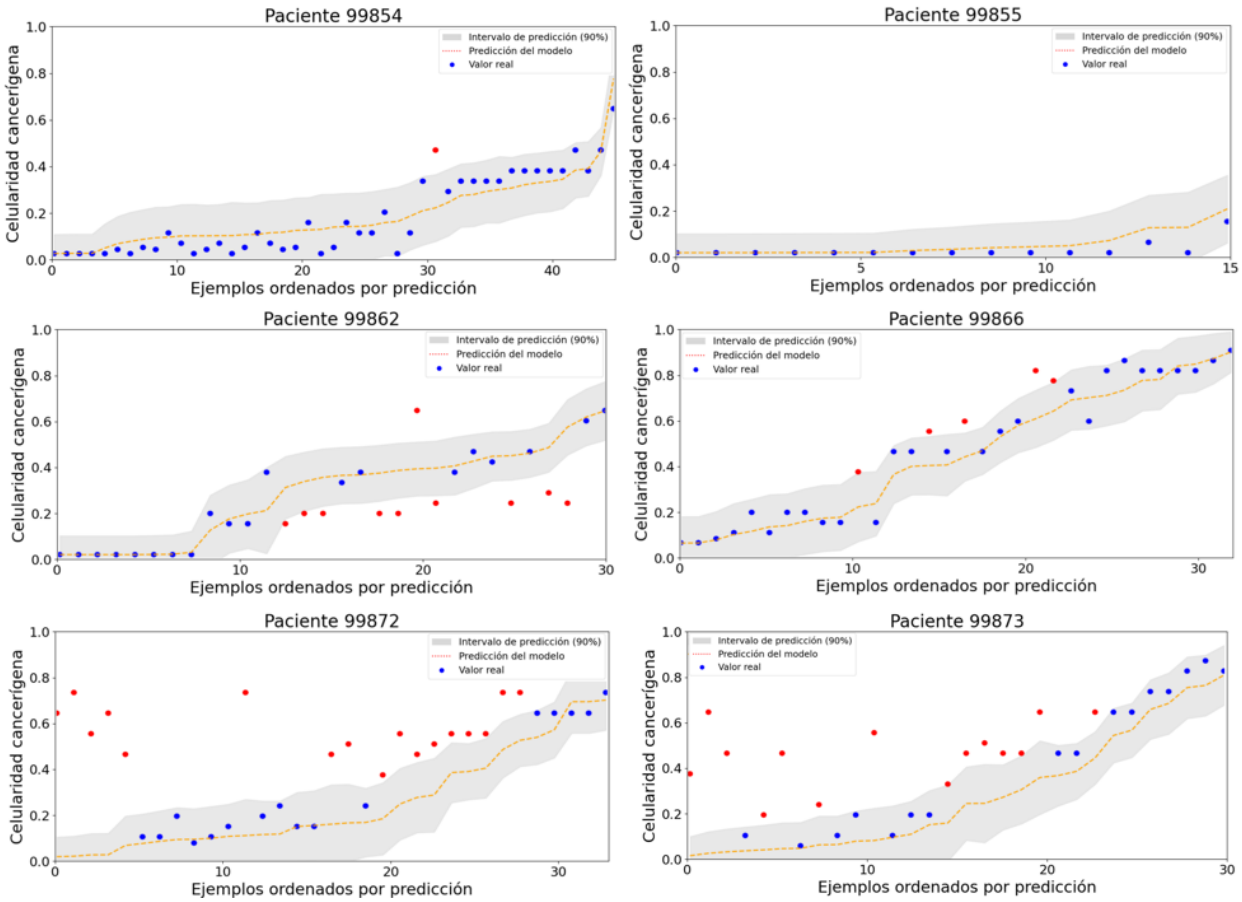
En la figura 37 se presentan los resultados por paciente con EfficientNet-B3. Se observa que los intervalos son adaptativos y varían de acuerdo con las predicciones individuales. Sin embargo, en varios pacientes aparecen observaciones fuera del rango, lo que reduce la cobertura global.

Figura 37: *Intervalos de predicción conforme por paciente para la arquitectura EfficientNet-B3 en CQR.* ***Azul*** *= predicción dentro del intervalo;* ***rojo*** *= predicción fuera*



La figura 38 corresponde al mismo análisis con DenseNet-169. En este caso, los intervalos muestran una mayor contención en varios pacientes, aunque persisten ciertos casos con valores fuera, en particular en los mismos pacientes ya mencionados en las variantes anteriores.

Figura 38: *Intervalos de predicción conforme por paciente para la arquitectura DenseNet-169 en CQR.* ***Azul*** *= predicción dentro del intervalo;* ***rojo*** *= predicción fuera*



Este análisis paciente por paciente aportó información clave. Casos como los pacientes **99872** y **99873** influyeron de manera determinante en la caída de la cobertura, debido a que gran parte de sus parches correspondían a imágenes borrosas o moderadas, con pocas imágenes nítidas. Esto contrasta con pacientes como **99854**, que pese a tener también una baja proporción de imágenes nítidas, mantuvo un buen desempeño general. Estos hallazgos confirman que existen características específicas de ciertos pacientes que deterioran la confiabilidad de los intervalos, más allá de la calidad técnica de las imágenes.

4.6. Consideracionales finales

Se evidenció que las variantes de predicción conforme para regresión brindan intervalos que permiten cuantificar la certidumbre de las predicciones. En calibración, las técnicas estimaron cuantiles adecuados y, por construcción, se logró una cobertura cercana al 90 % (dejando fuera en torno al 10 %). En prueba, sin embargo, ninguna combinación alcanzó el 90 % de cobertura. La mejor fue ICP con EfficientNet-B3 (cobertura = 76.76 %, tamaño

promedio del intervalo = 0.2758), seguida por ICP con DenseNet-169 (71.89 %, 0.2966). En MCP, EfficientNet-B3 obtuvo 67.03 % con tamaño 0.2079, mientras que DenseNet-169 llegó a 63.24 % con 0.2860. En CQR ocurrió lo contrario a ICP/MCP: DenseNet-169 logró mayor cobertura (74.59 %) que EfficientNet-B3 (66.49 %), a costa de intervalos algo más amplios (0.2822 vs. 0.2634). Dado que la variable objetivo está acotada en $[0, 1]$, un intervalo de tamaño 0.28 cubre aproximadamente el 28 % del rango; aun así, un intervalo más ancho no garantiza mejor cobertura en presencia de heterogeneidad.

El análisis por paciente reveló que unos pocos casos difíciles (predicciones de pacientes 99872 y 99873) condicionan la cobertura global debido a la calidad de sus datos. Esto subraya que la validez práctica de los intervalos depende también de la calidad de los datos y de rasgos específicos de cada paciente, y que un buen desempeño puntual no implica necesariamente incertidumbre bien calibrada.

5. Conclusiones

En conclusión, la predicción conforme sí entrega un nivel de certidumbre útil sobre las predicciones (intervalos o conjuntos que indican cuándo confiar y cuándo ser cautos). En este trabajo, además de reproducir tareas de los challenges usados como base (graduación de cáncer de próstata en PANDA y estimación de celularidad tumoral en BreastPathQ), implementamos sobre ellos el procedimiento completo de predicción conforme para ambos tipos de problema. No solo alcanzamos los objetivos de desempeño propios de los conjuntos, sino que añadimos una capa explícita de cuantificación de incertidumbre y lo hicimos con un proceso replicable y transparente: definimos los scores de no conformidad, separamos con claridad entrenamiento–calibración–prueba y evaluamos con métricas de validez y eficiencia. Este protocolo deja una guía práctica que puede servir de base para estudios multicéntricos con requisitos de transparencia y auditoría. Para interpretar adecuadamente la incertidumbre, combinamos métricas complementarias (cobertura, tamaño, N-criterion, FSC/SSC) con análisis por paciente: esta combinación ofrece una lectura más amplia del por qué de los resultados y ayuda a evitar decisiones basadas en una sola métrica o en promedios que ocultan heterogeneidad.

En clasificación, el modelo DenseNet-121 se establece como la arquitectura de mejor desempeño para la graduación del cáncer de próstata en los dos experimentos realizados, demostrando un equilibrio óptimo entre la confiabilidad y la utilidad práctica en el diagnóstico. Aunque el modelo EfficientNet logra niveles de cobertura comparables, DenseNet-121, particularmente al emplear las térnicas ICP y MCP produce consistentemente conjuntos de predicción notablemente más compactos, lo que se traduce en un diagnóstico menos ambiguo y de mayor valor clínico. Esta ventaja se complementa con una robustez y consistencia interinstitucional (FSC) marcadamente superiores, evidenciando que DenseNet-121 es intrínsecamente menos sensible a las variaciones en la partición de los datos y, por ende, tiene mayor generalización. Finalmente, la observación de que MCP es sensible a la estrategia de calibración subraya la importancia de definir criterios clínicos sólidos, confirmando que la elección exitosa del modelo y la técnica de predicción conforme debe priorizar la consistencia y robustez para una aplicación clínica fiable.

En regresión construimos intervalos de predicción conforme de manera efectiva: primero estimamos correctamente el cuantil en la fase de calibración y luego ajustamos los límites de cada predicción sumando y restando ese cuantil. Según la variante, cada arquitectura mostró ventajas distintas. En ICP y MCP, intervalos más compactos se asociaron con mejores coberturas, siendo EfficientNet-B3 la arquitectura con mejor desempeño; en CQR, en cambio, intervalos algo más anchos resultaron convenientes, ahora siendo DenseNet-169 la arquitectura sobresaliente. Aun así, el tamaño del intervalo por sí solo no garantiza una

mejor cobertura. Como no se alcanzó la cobertura objetivo, analizamos el desempeño por paciente y encontramos que la calidad de los datos de algunos casos puede arrastrar la cobertura global. Por ello, mirar los resultados a nivel individual aporta más información que una métrica promedio: la cobertura y el tamaño de los intervalos dependen de rasgos del caso. Además, en escenarios con múltiples parches por paciente, un patrón "raro" o atípico en ese paciente puede propagarse a muchos parches y arrastrar la cobertura global del conjunto de prueba. En la práctica, hay pacientes en los que el modelo se mantiene estable y otros en los que pierde fiabilidad; esa heterogeneidad debe guiar la toma de decisiones y el uso clínico de los resultados.

Como trabajo futuro, el tema nos resulta lo bastante interesante como para seguirlo trabajando y llevarlo un paso más allá. En concreto, queremos (i) extender el marco de predicción conforme a otras patologías y conjuntos de datos, (ii) diseñar y validar un procedimiento propio de extracción de scores (funciones de no conformidad) adaptado a histopatología, e (iii) incorporar retroalimentación clínica en la evaluación —con criterios y umbrales de derivación acordados con expertos— para cerrar el ciclo y mejorar la utilidad en práctica real. Además, nos planteamos explorar robustez frente a cambios de dominio (desenfoque, variaciones de tinción) con recalibración cuando sea necesario, y evaluar reglas de agregación a nivel paciente para reportar incertidumbre de forma más cercana a la toma de decisiones clínica.

Bibliografía

AHMED, Tashin y SABAB, Noor Hossain Nuri. «Classification and Understanding of Cloud Structures via Satellite Images with EfficientUNet». En: *Journal of Cloud Computing: Advances, Systems and Applications* 9.1 (2021), págs. 1-12. DOI: 10.1007/s42979-021-00981-2.

ANGELOPOULOS, Anastasios N. y BATES, Stephen. «A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification». En: *arXiv preprint arXiv:2107.07511* (2021). A practical introduction to conformal prediction with Python examples. DOI: 10.48550/arXiv.2107.07511.

BERA, Kaustav; SCHALPER, Kurt A.; RIMM, David L.; VELCHETI, Vamsidhar y MADABHUSHI, Anant. «Artificial intelligence in digital pathology — new tools for diagnosis and precision oncology». En: *Nature Reviews Clinical Oncology* 16.11 (2019), págs. 703-715. DOI: 10.1038/s41571-019-0252-y.

CHICCO, Davide; WARRENS, Matthijs J. y JURMAN, Giuseppe. «The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation». En: *PeerJ Computer Science* 7 (2021). Publisher: PeerJ Inc. ISSN: 2376-5992. DOI: 10.7717/peerj-cs.623. URL: https://peerj.com/articles/cs-623.

CRESSWELL, Jesse C.; SUI, Yi; KUMAR, Bhargava y VOUITSIS, Noël. *Conformal Prediction Sets Improve Human Decision Making*. arXiv:2401.13744. 2024. URL: http://arxiv.org/abs/2401.13744.

FONTANA, Matteo; ZENI, Gianluca y VANTINI, Simone. «Conformal Prediction: A Unified Review of Theory and New Challenges». En: *Bernoulli* 29.1 (2023). Publisher: Bernoulli Society for Mathematical Statistics and Probability, págs. 1-23. DOI: 10.3150/21-BEJ1447.

HE, Kaiming; ZHANG, Xiangyu; REN, Shaoqing y SUN, Jian. «Deep Residual Learning for Image Recognition». En: *arXiv:1512.03385 [cs.CV]* (dic. de 2016). arXiv: 1512.03385. DOI: 10.48550/arXiv.1512.03385.

HOSSAIN, MD Shakhawat; RAHMAN, MD. Sahilur; AHMED, Munim; ALFAZ, Nazia; MUNIRA SHIFAT, Sirajum; MAHBUBUL SYEED, M. M.; HUSSEN, Mohammad Anowar y UDDIN, Mohammad Faisal. «Residual Tumor Cellularity Assessment of Breast Cancer After Neoadjuvant Therapy Using Image Transformer». En: *IEEE Access* 12 (2024),

págs. 86083-86095. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2024.3415665. URL: https://ieeexplore.ieee.org/abstract/document/10559819.

HUANG, Gao; LIU, Zhuang; PLEISS, Geoff; MAATEN, Laurens van der y WEINBERGER, Kilian Q. «Convolutional Networks with Dense Connectivity». En: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.12 (dic. de 2022), págs. 8704-8716. DOI: 10.1109/TPAMI.2019.2918284.

HUANG, Gao; LIU, Zhuang; VAN DER MAATEN, Laurens y WEINBERGER, Kilian Q. «Densely Connected Convolutional Networks». En: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI: IEEE, 2017, págs. 2261-2269. DOI: 10.1109/CVPR.2017.243.

KAGGLE. *Prostate cANcer graDe Assessment (PANDA) Challenge Dataset*. 2020. URL: https://www.kaggle.com/competitions/prostate-cancer-grade-assessment.

KHANNA, Mukul. *Paper Review: DenseNet – Densely Connected Convolutional Networks*. https://medium.com/data-science/paper-review-densenet-densely-connected-convolutional-networks-acf9065dfefb. Sep. de 2020.

KOTSIANTIS, S. B.; ZAHARAKIS, I. D. y PINTELAS, P. E. «Machine learning: a review of classification and combining techniques». en. En: *Artificial Intelligence Review* 26.3 (nov. de 2006), págs. 159-190. ISSN: 1573-7462. DOI: 10.1007/s10462-007-9052-3.

KUMAR, Sunil y BHATNAGAR, Vaibhav. «A Review of Regression Models in Machine Learning». en. En: *JOURNAL OF INTELLIGENT SYSTEMS AND COMPUTING* 3.1 (dic. de 2022), págs. 40-47. ISSN: 2976-8098. DOI: 10.51682/jiscom.v3i1.30.

MADABHUSHI, Anant y LEE, George. «Image Analysis and Machine Learning in Digital Pathology: Challenges and Opportunities». en. En: *Medical Image Analysis* 33 (jul. de 2016), págs. 170-175. DOI: 10.1016/j.media.2016.06.037.

MUKHERJEE, Suvaditya. *The Annotated ResNet-50*. en. https://towardsdatascience.com/the-annotated-resnet-50-a6c536034758. Ago. de 2022.

NIAZI, Muhammad Khalid Khan; PARWANI, Anil V. y GURCAN, Metin N. «Digital Pathology and Artificial Intelligence». En: *The Lancet Oncology* 20.5 (2019), e253-e261. DOI: 10.1016/S1470-2045(19)30154-8.

OLSSON, Henrik et al. «Estimating diagnostic uncertainty in artificial intelligence assisted pathology using conformal prediction». en. En: *Nature Communications* 13.1 (dic. de 2022).

Publisher: Nature Publishing Group, pág. 7761. ISSN: 2041-1723. DOI: 10.1038/s41467-022-34945-8.

ORCHARD, Edited by Guy y NATION, Brian, eds. *Histopathology*. Second Edition, Second Edition. Fundamentals of Biomedical Science. Oxford, New York: Oxford University Press, ene. de 2018. ISBN: 978-0-19-871733-1.

PAPADOPOULOS, Harris. «Inductive Conformal Prediction: Theory and Application to Neural Networks». en. En: *Tools in Artificial Intelligence*. Ed. por Paula FRITZSCHE. InTech, ago. de 2008. ISBN: 978-953-7619-03-9. DOI: 10.5772/6078.

PETRICK, Nicholas; AKBAR, Shazia; CHA, Kenny H.; NOFECH-MOZES, Sharon; SAHINER, Berkman; GAVRIELIDES, Marios A.; KALPATHY-CRAMER, Jayashree; DRUKKER, Karen y MARTEL, Anne L. «SPIE-AAPM-NCI BreastPathQ Challenge: an image analysis challenge for quantitative tumor cellularity assessment in breast cancer histology images following neoadjuvant treatment». En: *Journal of Medical Imaging* 8 (mayo de 2021), pág. 034501. ISSN: 2329-4302. DOI: 10.1117/1.JMI.8.3.034501.

PINTAWONG, Surayuth; SHUANGSHOTI, Shanop; JITPASUTHAM, Tikamporn; SHUANGSHOTI, Somruethai; WIWATWARAYOS, Kulachet; KOBCHAISAWAT, Thananop y CHALIDABHONGSE, Thanarat H. «Conformal Prediction for Uncertainty Quantification and Reliable HER2 Status Classification in Breast Cancer IHC Images». En: *IEEE Access* (2025).

RASHEED, Mehwish; JAFFAR, Muhammad Arfan; AKRAM, Arslan; RASHID, Javed; ALSHALALI, Tagrid Abdullah N; IRSHAD, Asma y SARWAR, Nadeem. «Improved brain tumor classification through DenseNet121 based transfer learning». En: *Discovery Oncology* 27.1 (2025), pág. 16. DOI: 10.1007/s12672-025-03501-3.

SHAFER, Glenn y VOVK, Vladimir. «A tutorial on conformal prediction». En: *Journal of Machine Learning Research* 9 (2008), págs. 371-421.

SONG, Andrew H.; JAUME, Guillaume; WILLIAMSON, Drew F. K.; LU, Ming Y.; VAIDYA, Anurag; MILLER, Tiffany R. y MAHMOOD, Faisal. «Artificial intelligence for digital and computational pathology». en. En: *Nature Reviews Bioengineering* 1.12 (dic. de 2023). Publisher: Nature Publishing Group, págs. 930-949. ISSN: 2731-6092. DOI: 10.1038/s44222-023-00096-8.

TAN, Mingxing y LE, Quoc V. «EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks». En: *arXiv preprint arXiv:1905.11946* (mayo de 2019). https://doi.org/10.48550/arXiv.1905.11946.

TIZHOOSH, Hamid Reza y PANTANOWITZ, Liron. «Artificial Intelligence and Digital Pathology: Challenges and Opportunities». En: *Journal of Pathology Informatics* 9.1 (2018), pág. 38. DOI: 10.4103/jpi.jpi_53_18.

VAZQUEZ, Janette y FACELLI, Julio C. «Conformal Prediction in Clinical Medical Sciences». En: *Journal of Healthcare Informatics Research* 6.3 (2022), págs. 241-252. DOI: 10.1007/s41666-021-00113-8.

VOVK, Vladimir. «Conditional Validity of Inductive Conformal Predictors». en. En: *Proceedings of the Asian Conference on Machine Learning*. ISSN: 1938-7228. PMLR, nov. de 2012, págs. 475-490. URL: https://proceedings.mlr.press/v25/vovk12.html.

WISNIEWSKI, Wojciech; LINDSAY, David y LINDSAY, Sian. «Application of conformal prediction interval estimations to market makers' net positions». en. En: *Proceedings of the Ninth Symposium on Conformal and Probabilistic Prediction and Applications*. PMLR, 2020.

Apéndices

Productos científicos

Abstract:

Deep learning has shown significant results in histopathology image analysis. However, despite their high predictive accuracy, conventional machine learning (ML) models often lack mechanisms to quantify uncertainty, which is necessary for building trust in clinical scenarios. This limitation hinders the integration of these models into real-world diagnostic workflows.

Conformal Prediction is a relatively recent statistical framework that enhances ML models by generating prediction sets according to a defined confidence level. These sets indicate the range of plausible outcomes for each input, thus allowing clinicians to assess model confidence.

In this study, we explore Conformal Prediction methods for multi-class classification of histopathological images using state-of-the-art convolutional neural networks, specifically DenseNet and EfficientNet. We implement three Conformal Prediction approaches: Inductive Conformal Prediction (ICP) with and without Adaptive Prediction Sets (APS), and Mondrian Conformal Prediction (MCP). Each method is evaluated on prostate cancer whole-slide images from two different institutions.

Evaluation metrics include the obtained coverage, N-criterion, Fairness in Subgroup Coverage (FSC), and Stratified Size Coverage (SSC), which assess adaptability across institutional and prediction set variations. Our results show that all the methods were able to generate prediction sets with coverage close to the desired confidence levels ($\alpha = 0.9$). ICP and MCP obtained similar coverage levels (around 0.88–0.89) while maintaining more compact prediction sets when compared to APS.

Additionally, the FSC and SSC metrics show that the conditional coverage was approximately equal, with the lowest result corresponding to the DenseNet architecture (0.86). Contrary to our expectations, MCP did not outperform ICP, suggesting that additional information associated with clinical site did not translate into clear improvements.

Regarding the convolutional architectures, DenseNet was able on average to achieve a better N-criterion than EfficientNet. This is notable for both ICP (1.76 vs. 2.01) and MCP (1.78 vs. 2.01) methods. Our results show that the APS conformal method outperforms ICP and MCP for this classification task. Additionally, the selection of the architecture might introduce important differences in the uncertainty of the predictions as reflected in the N-criterion values of the resulting conformal prediction sets.

Referencias del producto científico

GELVEZ, Santiago; CLAVIJO, Diego y ROMO-BUCHELI, David. «Conformal Prediction for Deep Learning Classification Model in Histopathological Images». En: *2025 XXV Symposium of Image, Signal Processing, and Artificial Vision (STSIVA)*. 2025, págs. 1-5. DOI: 10.1109/STSIVA66383.2025.11156416.