Pontificia Universidad Javeriana Cali
Facultad de Ingeniería y Ciencias
Maestría en Ciencia de Datos

Proyecto de Grado

Detección de Tejido Canceroso en Glándulas Mamarias Basado en Aprendizaje Automático Supervisado con Múltiples Expertos

Adrián Sebastián Muñoz Hoyos
Jean Cristopher Martínez Reyes

Director: Dr. Julián Gil González

14 de julio de 2024

Índice general

Índice de figuras

Índice de cuadros

Introducción

La creciente demanda de especialistas médicos y la alta carga laboral de los patólogos son factores que intensifican los desafíos en el diagnóstico de cáncer, esta problemática es un punto interesante por abordar pues cualquier tiempo que se gane en un diagnóstico de cáncer puede significar la vida de un paciente. Si bien existen diversos tipos de cáncer, clasificados según la ubicación y el tipo de células involucradas, hay uno en particular que afecta a las mujeres independientemente de su región o grupo étnico: el cáncer de glándula mamaria. Esta investigación se enfoca en proporcionar herramientas que contribuyan a la lucha contra este tipo de enfermedad.

En este proyecto se desarrolló un modelo de inteligencia artificial que permite detectar tejido canceroso en imágenes histológicas de glándulas mamarias. Las imágenes empleadas en entrenamiento y validación provienen de bases de datos públicas y cuentan con la anotación de múltiples expertos. Con esta información, se entrenó un modelo de aprendizaje automático basado en múltiples anotadores para mejorar la precisión de las clasificaciones. La implementación de este modelo permite aumentar la velocidad de diagnóstico reduciendo los costos operacionales. Este sistema podría convertirse en un importante apoyo para el personal médico, proporcionando una herramienta que les permita optimizar su flujo de trabajo al priorizar los casos clasificados como cáncer.

Capítulo 1

Definición Del Problema

1.1. Planteamiento Del Problema

El cáncer de mama es una enfermedad de alta prevalencia y mortalidad en la población femenina a nivel global [22] y nacional, ya que el 25.7 % de las colombianas con padecimientos cancerígenos son por este tipo [9]. A pesar de los avances en la tecnología de diagnóstico, la detección temprana y precisa de esta patología sigue siendo un desafío significativo. La complejidad y heterogeneidad de las estructuras mamarias dificultan la identificación oportuna de tejido canceroso, lo que resulta en una tarea que requiere mucho tiempo y especialización por parte de los profesionales médicos.

La supervisión constante y la revisión de casos por parte de especialistas médicos no solo es costosa y consume mucho tiempo, sino que también puede ser susceptible a errores humanos ya que este proceso requiere no solo de la destreza del profesional de la salud sino también su capacidad de distinguir los rasgos propios de las células cancerígenas, con esto, surge la necesidad de implementar modelos de inteligencia artificial que permitan automatizar las lecturas de imágenes histológicas, permitiendo filtrar casos sin presencia de cáncer para que los patólogos puedan enfocarse en los casos sospechosos y así mejorar el flujo de trabajo, la carga laboral y los costos de diagnóstico [14].

Dado el significativo impacto en la morbilidad que conlleva el cáncer de seno y las múltiples variables involucradas en la clasificación de imágenes histológicas por parte del personal médico, es crucial investigar y avanzar en el campo del Machine Learning para la creación de modelos diagnósticos. La combinación del conocimiento médico con el aprendizaje automático puede mejorar significativamente la precisión y rapidez en el diagnóstico de esta patología.

Sin embargo, uno de los desafíos más importantes en este ámbito es la disponibilidad limitada de personal experto para el etiquetado de las fuentes de entrenamiento. Este proceso es fundamental para el desarrollo de modelos competentes, pero puede ser complejo y demandante.

En este contexto, la implementación del etiquetado a través de crowdsourcing se presenta como una solución interesante. Aprovechar la colaboración de una amplia base de personal médico para la tarea de etiquetado puede reducir la carga sobre los médicos expertos, acelerar el proceso y facilitar la construcción de modelos eficaces. De esta manera, podemos desarrollar herramientas que, mediante el entrenamiento con modelos de Machine Learning, permitan realizar la clasificación de imágenes de manera sistemática, ordenada y precisa, aportando la rapidez necesaria en el diagnós-

tico del cáncer de seno.

1.2. Formulación del Problema

De acuerdo con las problemáticas expuestas, como son la carga laboral, la oferta de especialistas, el costo de las horas laborales de los especialistas, el riesgo de falsos positivos o negativos por error humano y la disponibilidad de da tos etiquetado por expertos para el entrenamiento de modelos diagnósticos, se propuso la siguiente pregunta de investigación:

- ¿Cómo construir un modelo de aprendizaje automático supervisado con múltiples expertos que permita clasificar de forma precisa y rápida imágenes histológicas positivas o negativas a cáncer?

Para dar respuesta a nuestra pregunta central de investigación e implementar el modelo propuesto debemos resolver preguntas más específicas:

- ¿Cómo procesar imágenes histológicas priorizando regiones de interés que contengan patrones de tejido canceroso?
- ¿Como entrenar modelos de aprendizaje automático basado en múltiples anotadores que permita clasificar imágenes histológicas con patrones de tejido canceroso?
- ¿Cómo evaluar el rendimiento de los modelos creados y compararlos con otros modelos utilizados para este tipo de tareas?
- ¿Cómo se pueden integrar estos modelos en el proceso de diagnóstico actual para aliviar la carga laboral y reducir la tasa de falsos positivos y falsos negativos?

Capítulo 2

Objetivos Del Proyecto

2.1. Objetivo General

Desarrollar un modelo de aprendizaje automático supervisado basado en múltiples expertos que permita identificar la presencia de cáncer en imágenes histológicas de tejido mamario.

2.2. Objetivos Específicos

- Implementar un esquema de procesamiento y extracción de características en imágenes histológicas que permita la identificación de tejido cancerígeno.
- Desarrollar modelos de aprendizaje automático que utilicen los diagnósticos de múltiples expertos para la identificación de patrones tumorales en tejido mamario.
- Evaluar el rendimiento de los modelos entrenados mediante métricas como la precisión (*accuracy*), la puntuación F1 (*F1-score*) y la sensibilidad (*recall*), y analizar los resultados basado en las diferentes fuentes de crowdsourcing empleadas.
- Desarrollar una herramienta web utilizando Python que permita aplicar el modelo en la clasificación de imágenes histológicas de forma interactiva.

Capítulo 3

Marco Teórico y Antecedentes

3.1. Marco Teórico

3.1.1. Cáncer de seno

El cáncer de mama es definido científicamente como "*proliferación acelerada, desordenada y no controlada de células con genes mutados, los cuales actúan normalmente suprimiendo o estimulando la continuidad del ciclo celular pertenecientes a distintos tejidos de una glándula mamaria*" [17]. En mujeres, el cáncer de mama es uno de los principales tipos de cáncer diagnosticados en regiones de alto y bajo desarrollo, siendo una de las principales causas de muerte en mujeres a nivel mundial [8].

El cáncer de mama es la segunda causa más frecuente de muerte por cáncer entre las mujeres estadounidenses, representando el 15 % de todas las muertes por cáncer en mujeres. Solo el cáncer de pulmón causa más muertes por cáncer. Según datos del programa de Vigilancia, Epidemiología y Resultados Finales del Instituto Nacional del Cáncer (SEER) [12].

Las predicciones actuales y las estadísticas sugieren un aumento tanto en la incidencia mundial de cáncer de mama como en la mortalidad relacionada. Según las estadísticas de GLOBOCAN de 2012, casi 1.7 millones de mujeres fueron diagnosticadas con cáncer de mama y se registraron 522,000 muertes relacionadas, lo que representa un aumento del 18 % desde 2008. Estas predicciones señalan que la incidencia mundial de cáncer de mama femenino alcanzará aproximadamente 3.2 millones de nuevos casos por año para 2050, reflejando la magnitud de la incidencia de cáncer de mama, su impacto en la sociedad a nivel mundial y la necesidad urgente de medidas preventivas y de tratamiento [28].

3.1.2. Histología e imágenes histológicas

La histología es una ciencia que se centra en el estudio de los detalles microscópicos de células y tejidos biológicos, utilizando microscopios de luz, fluorescencia o electrónicos. Para ello, se preparan previamente láminas histológicas mediante procesos de extracción, corte y preparación de regiones de tejido especificas y posteriormente, la aplicación de técnicas de tinción, siendo la hematoxilina y eosina (H&E) la más utilizada.

El proceso de tinción permite resaltar diferentes estructuras biológicas como núcleos celulares, citoplasma y otros elementos. La interpretación de las imágenes histológicas es crucial para determinar la salud o enfermedad de las células y tejidos [21].

3.1.3. Digitalización de imágenes histológicas

Los laboratorios de patología están pasando por una transformación hacia un flujo de trabajo completamente digital. Además de la gestión digital de muestras de tejidos, órdenes de patología e informes, esto incluye la digitalización de láminas de histopatología y el uso de monitores y computadoras para su visualización, lo que busca reemplazar al microscopio óptico como la herramienta principal utilizada por los patólogos. Esta transformación ha sido posible recientemente gracias a la introducción de escáneres de imágenes de toda la muestra (WSI, por sus siglas en inglés), que son más eficientes en costos y tiempo en comparación con las cámaras digitales montadas en microscopios. Uno de los principales beneficios de la migración digital en histología es que se permite el uso de métodos automáticos de análisis de imágenes cuantitativos, estos métodos tienen el potencial de abordar los problemas derivados de la interpretación subjetiva por parte de los patólogos y, al mismo tiempo, reducir su carga de trabajo en contraste con el método manual convencional [30].

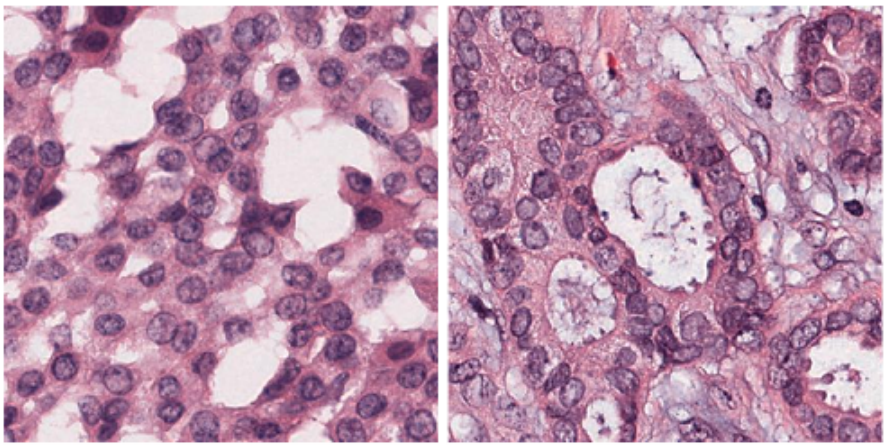


Figura 3.1: Imagen histológica de tejido mamario digitalizada [30].

3.1.4. Aprendizaje automático

El aprendizaje automático conocido también como Machine Learning hace parte de los procesos de inteligencia artificial en la cual el lenguaje de máquina a través de algoritmos y la utilización de sus recursos informáticos permite homologar procesos de inteligencia, realizando así acciones de manera automatizada, rápida y con mejora en términos de rendimiento y precisión [15].

Para poder llevar a cabo este proceso automático es primordial que la máquina cuente con las herramientas para realizar el cálculo de aprendizaje (conjunto de entrenamiento) y validar la información del modelo adaptado verificando los errores en la predicción o clasificación (datos de prueba) [19].

El aprendizaje automático generalmente se clasifica dependiendo de la técnica utilizada en el entrenamiento (supervisado, no supervisado). Las técnicas de Aprendizaje Supervisados se basan en datos que se encuentran agrupados y debidamente etiquetados, permitiendo identificar características específicas en los conjuntos de datos para realizar las actividades de entrenamiento y predicción. Estas técnicas son usadas principalmente en modelos de clasificación y regresión. Por otro lado, en las técnicas de aprendizaje no supervisado, los conjuntos de datos no cuentan con un etiquetado y se requiere implementar técnicas para agrupar características o generar clústers, de esta manera se pueden resaltar características relevantes para el entrenamiento de los modelos [10].

3.1.5. Redes Neuronales Convolucionales - CNN

El reconocimiento y la detección de imágenes son problemas clásicos en el ámbito del aprendizaje automático. Identificar un objeto o reconocer una imagen a partir de una imagen digital o un vídeo es una tarea muy compleja espacialmente para los algoritmos de Machine Learning Tradicional. Los algoritmos de aprendizaje profundo han logrado avances significativos en el ámbito de la visión por computadora. El aprendizaje profundo utiliza redes neuronales artificiales con varias capas ocultas para imitar las funciones de la corteza cerebral humana. Las capas de una red neuronal profunda extraen diversas características y, por lo tanto, proporcionan varios niveles de abstracción. A diferencia de las redes superficiales, que no pueden trabajar con múltiples características, las redes neuronales convolucionales son un potente algoritmo de aprendizaje profundo que puede manejar millones de parámetros y reducir el costo computacional al procesar una imagen 2D, convolucionarla con filtros/kernels y producir volúmenes de salida [5].

Las redes neuronales convolucionales son un tipo de red neuronal especialmente apta para la clasificación de imágenes. Estas redes están formadas por una serie de capas, cada una de las cuales lleva a cabo una operación de convolución sobre la imagen de entrada. La salida de la capa final proporciona las probabilidades de clase para la imagen de entrada. Las redes neuronales convolucionales han demostrado ser muy exitosas en diversas tareas, como la detección de objetos, el reconocimiento facial y la clasificación de texto [18].

3.1.6. Red neuronal - VGG16

La red VGG16 se desarrolló en el año 2014, publicándose en el artículo "Very Deep Convolutional Networks for Large-Scale Image Recognition", esta red surgió con el objetivo de explorar cómo la profundidad de las CNNs afecta su rendimiento. Los autores propusieron una arquitectura basada en el uso de capas convolucionales pequeñas (3×3), argumentando que estas podrían capturar características visuales más complejas sin incrementar significativamente el costo computacional.

Esta simplicidad estructural, combinada con una gran profundidad (16 capas con pesos), permitió a la VGG16 alcanzar excelentes rendimientos en el conjunto de datos ImageNet, compuesto por millones de imágenes etiquetadas en miles de categorías [26].

3.1.7. Extracción de características

Las imágenes histopatológicas son de gran tamaño, lo que hace que su procesamiento sea una tarea que lleva mucho tiempo. Por eso, se dividen en partes más pequeñas para acelerar el proceso. Después, se extraen características de cada parte de la imagen. Estas características descriptivas se utilizan luego en algoritmos de clasificación para categorizar las imágenes [23].

La extracción de características con redes neuronales convolucionales (CNNs) se refiere al proceso de aprender y extraer atributos relevantes de las imágenes para usarlos en tareas de clasificación y reconocimiento. Este proceso es fundamental porque permite a las CNNs capturar patrones complejos y representaciones de alto nivel de las imágenes, lo que mejora significativamente el rendimiento de los modelos y la capacidad de computo necesaria para los entrenamiento. La extracción de características además permite flexibilidad de combinar la representación 2d de las características de alto valor producidas por las CNNs con el uso de otros tipos de modelos como Máquinas de Vectores de Soporte (SVM), arboles de decisión, regresión lineal, regresión logística, entre otros [27].

3.1.8. Crowdsourcing

El crowdsourcing es una metodología que consiste en la recolección de información, datos y soluciones a problemas a través de la contribución de una gran cantidad de personas, generalmente mediante plataformas en línea. Esta técnica aprovecha la inteligencia colectiva de la multitud para realizar tareas que serían difíciles o costosas de ejecutar de manera tradicional. En el contexto de la ciencia de datos y el Machine Learning, el crowdsourcing permite la recolección de grandes volúmenes de datos etiquetados necesarios para entrenar modelos, mejorando la diversidad y calidad de los datos.

Los beneficios del crowdsourcing en la ciencia de datos son múltiples. En primer lugar, reduce significativamente los costos y el tiempo necesarios para obtener conjuntos de datos grandes y diversos. Esto es especialmente útil en proyectos de Machine Learning donde la cantidad y calidad de los datos son cruciales para el rendimiento del modelo. Además, al involucrar a una multitud, se pueden obtener diferentes perspectivas y enfoques para la resolución de problemas, enriqueciendo el conjunto de datos con una variedad de patrones y características. La aplicabilidad del crowdsourcing se extiende a numerosas áreas, incluyendo la recolección de datos para el entrenamiento de modelos de reconocimiento de imágenes, procesamiento del lenguaje natural y análisis de sentimientos, entre otros. Este enfoque permite a los investigadores y empresas construir modelos más robustos y generalizables, aprovechando la variabilidad y riqueza de los datos obtenidos mediante la colaboración masiva [13].

3.1.9. Algoritmos de Crowdsourcing

Se han desarrollado algunos métodos para aplicar a los etiquetados generados por crowdsourcing, estos algoritmos son fundamentales para mejorar la calidad de los datos etiquetados utilizados en el entrenamiento de modelos de Machine Learning. Al elegir y aplicar el algoritmo adecuado, se pueden obtener etiquetas de alta calidad que mejoran el rendimiento de los modelos predictivos y de clasificación [29].

3.1.9.1. Majority Vote

El algoritmo Majority Vote es una técnica simple para la agregación de respuestas en tareas de crowdsourcing. Cada trabajador proporciona una etiqueta para una tarea específica y la etiqueta final asignada es la que recibe la mayoría de los votos. Se cuenta con la ecuación basica para el cálculo 3.1, donde $\mathbb{I}$ es una función indicadora que vale 1 si la etiqueta del trabajador y_i es igual a la clase c, y 0 en caso contrario. n es el número total de trabajadores y c representa cada posible clase.

$$\text{Etiqueta final} = \arg\max_{c} \sum_{i=1}^{n} \mathbb{I}(y_i = c) \tag{3.1}$$

El voto mayoritario se ha aplicado en numerosos campos, incluyendo la salud, el medio ambiente, la detección de intrusiones, el reconocimiento de expresiones faciales, la minería de textos y la ingeniería de software. Esta técnica ha sido útil para clasificar señales EEG en la detección de convulsiones epilépticas, identificar automáticamente contaminantes en el entorno, mejorar la capacidad predictiva en el análisis de sentimientos y estimar posibles errores en proyectos de software [7].

3.1.9.2. WAWA (Worker Agreement with Aggregate)

WAWA es un algoritmo que calcula el acuerdo entre los trabajadores y la etiqueta agregada, ponderando las respuestas de los trabajadores según su historial de precisión. Teniendo su ecuación 3.2 donde w_i es el peso asignado al trabajador i, basado en su acuerdo pasado con la etiqueta agregada, y $\mathbb{I}$ es la función indicadora. El peso w_i puede ser calculado como la proporción de respuestas correctas del trabajador en tareas previas.

$$\text{WAWA}(c) = \sum_{i=1}^{n} w_i \mathbb{I}(y_i = c) \tag{3.2}$$

3.1.9.3. DS (Dawid-Skene)

El modelo Dawid-Skene utiliza un enfoque basado en la máxima verosimilitud para estimar las tasas de error de los observadores y corregir las etiquetas. El algoritmo emplea el método de

Expectation-Maximization (EM) para iterar entre la estimación de las habilidades de los trabajadores y las etiquetas verdaderas. Las ecuaciones clave del modelo Dawid-Skene son en primera instancia, paso de Expectativa (E-step) 3.3, donde π^c es la proporción de la clase c, θ_i^c es la probabilidad de que el trabajador i clasifique correctamente una etiqueta de la clase c, y y_{ij} es la etiqueta proporcionada por el trabajador i para la tarea j. Seguidamente otra ecuación importante del modelo es Paso de Maximización (M-step) 3.4 y 3.5, donde m es el número total de tareas y γ_{ij}^c es la probabilidad de que la tarea j pertenezca a la clase c dado el trabajador i [29].

$$\gamma_{ij}^c = \frac{\pi^c \prod_{i=1}^{n} (\theta_i^c)^{y_{ij}} (1-\theta_i^c)^{1-y_{ij}}}{\sum_{k=1}^{C} \pi^k \prod_{i=1}^{n} (\theta_i^k)^{y_{ij}} (1-\theta_i^k)^{1-y_{ij}}} \tag{3.3}$$

$$\pi^c = \frac{1}{m} \sum_{j=1}^{m} \gamma_{ij}^c \tag{3.4}$$

$$\theta_i^c = \frac{\sum_{j=1}^{m} \gamma_{ij}^c y_{ij}}{\sum_{j=1}^{m} \gamma_{ij}^c} \tag{3.5}$$

3.1.10. Aprendizaje supervisado con múltiples expertos

El aprendizaje automático con múltiples expertos se basa en la creación de modelos a partir de conjuntos de datos etiquetados por diferentes profesionales o expertos en un campo de estudio especifico, de esta manera, los modelos se destacan por su capacidad para imitar el razonamiento humano debido a la heterogeneidad de los conjuntos de datos etiquetados por expertos [3].

Rossin, resalta las diferencias entre los sistemas de programas convencionales (CPS) y los sistemas expertos (ES). Mantiene que los CPS implican el deseo del investigador de crear un sistema que aborde tareas interesantes y difíciles sin considerar si estas son similares a las utilizadas por los humanos. Por otro lado, los ES intentan comprender cómo los humanos resuelven problemas y luego utilizan la computadora para explicar y predecir su comportamiento. En la práctica, muchos sistemas contienen elementos de ambos. Las ventajas de los ES incluyen la capacidad de proporcionar asesoramiento experto a no expertos, asistir a expertos para resolver problemas y actuar como una herramienta de enseñanza para no expertos [25].

Un problema común en la formación de conjuntos de datos para entrenamiento de modelos en el campo de la medicina es la disponibilidad de etiquetas de expertos médicos, esto se explica por lo complejo que puede resultar la tarea de etiquetar bases de datos gigantescas para un experto o médico o incluso un grupo de expertos médicos. Ante esta dificultad se han propuestos programas de crowdsourcing que permiten que individuos, instituciones u organizaciones se encarguen de generar etiquetas en bases de datos abiertas y en línea. Este proceso de crowdsourcing se ha tornado valioso ya que permite plasmar conocimientos y criterios de grupos heterogéneos de expertos [1].

Se ha investigado si el crowdsourcing sería adecuado para el etiquetado y delimitación de imágenes histopatológicas , una tarea de considerable complejidad que generalmente se asigna a expertos. Los experimentos cubrieron un amplio rango de complejidad de clases y tamaño de imagen, encontrando que el etiquetado de multitudes tiene una mayor efectividad en imágenes pequeñas, pocas clases y tamaños y apariencias constantes, por ejemplo, estructuras del tejido renal como el glomérulo. De acuerdo con esto, la tasa de error aumenta con la creciente complejidad de la imagen histológica [11].

El desafío de los Sistemas Expertos radica en su capacidad para utilizar el conocimiento de expertos, en este caso, expertos en clasificación histológica. Después de condensar sus conocimientos médicos, habilidades de observación y razonamiento de clasificación en datos etiquetados, estos se trasladan a una base de datos que el algoritmo puede usar para ajustar los parámetros del proceso y minimizar los errores, así se debe garantizar que la base de datos tenga robustez y un etiquetado de buena calidad para que el modelo tenga éxito [4].

Además, en el aprendizaje supervisado existen modelos conocidos como "múltiples expertos", en los que cada conjunto de datos posee anotaciones para cada ejemplo y clase realizadas por anotadores expertos [24]. Dentro del proceso de aplicación y uso del aprendizaje con múltiples expertos se establecen algunos retos: La estimación subjetiva que tiene cada uno de los expertos frente a la etiqueta del conjunto de datos, generando diferencias entre ellas y ruido en la estimación de la verdad que buscará el modelo de aprendizaje automático; por el otro lado está la importancia que tiene el ruido, la calibración de la fiabilidad y el sesgo que tienen las etiquetas expertas dentro del proceso de aprendizaje; para buscar existen algoritmos de esperanza-maximización (algoritmo EM) que ayudan a encontrar estimadores de mayor similitud en el proceso [20].

No obstante, uno de los principales desafíos en la implementación de la IA para la detección del cáncer de mama es la amenaza de falsos positivos y negativos. Un falso positivo puede conducir a un tratamiento innecesario y a una ansiedad innecesaria para el paciente, mientras que un falso negativo puede retrasar el tratamiento necesario, empeorando el pronóstico del paciente. Para minimizar estos riesgos, se propone un enfoque de aprendizaje automático basado en múltiples expertos, donde se emplea el conocimiento de varios expertos para entrenar el algoritmo, lo cual puede mejorar la precisión del modelo [14].

3.2. Antecedentes

Yun Liu y colaboradores desarrollaron un algoritmo de inteligencia artificial (LYNA) para el diagnóstico del cáncer, utilizando el conjunto de evaluación Camelyon16. LYNA demostró capacidad para detectar tanto macro metástasis como micro metástasis como se muestra en la Ilustración 2. En las pruebas de evaluación, LYNA logró un área bajo la curva (AUC) del 99.3 % a nivel de diapositiva (presencia o ausencia de metástasis nodal), superior al 96.6 % obtenido por un patólogo

practicante. LYNA también demostró ser capaz de detectar todas las 40 macro metástasis, esto demuestra que los algoritmos de inteligencia artificial como LYNA pueden ser herramientas valiosas en el diagnóstico de cáncer en términos de eficacia y velocidad de diagnóstico [14].

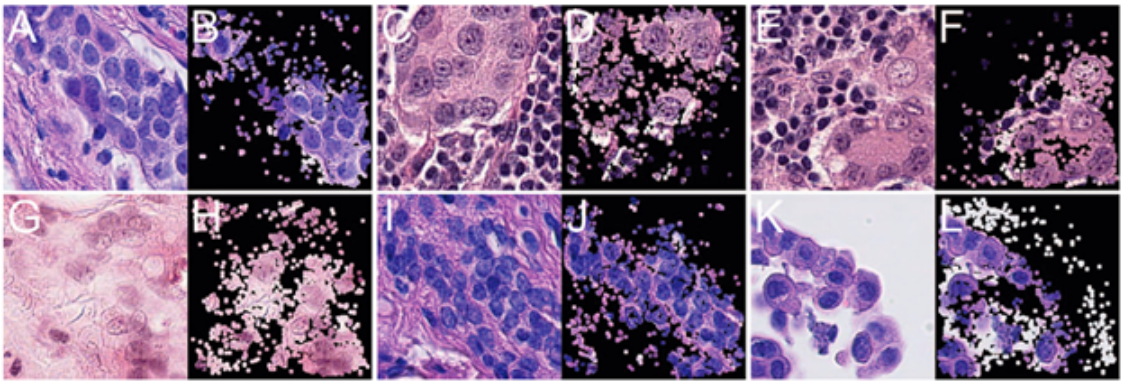


Figura 3.2: Visualización de los píxeles en imágenes teñidas con hematoxilina-eosina por el Asistente de Ganglios Linfáticos (LYNA) [14].

Amgad y colaboradores en su trabajo "*NuCLS: A scalable crowdsourcing approach and dataset for nucleus classification and segmentation in breast cancer*" presentaron un marco de colaboración novedoso para la participación de multitudes de estudiantes de medicina y patólogos en la generación de etiquetas de calidad para núcleos celulares, utilizando aprendizaje profundo para mapear con precisión estructuras celulares en cáncer de mama. Se aborda la barrera crítica de la generación de etiquetas al involucrar a no expertos en el proceso, resultando en el conjunto de datos NuCLS con más de 220,000 anotaciones de núcleos celulares en cánceres de mama. Se presenta un enfoque innovador que utiliza sugerencias algorítmicas para recopilar datos de segmentación precisa sin la necesidad de un laborioso trazado manual de los núcleos. Los resultados indican que las sugerencias algorítmicas, aunque ruidosas, no afectan negativamente la precisión de los patólogos y pueden mejorar la calidad de las anotaciones de no expertos [2].

La investigación evidencia la importancia del uso de múltiples anotadores en este tipo de modelos, un enfoque conocido como crowdsourcing. Este proceso permite la recopilación de una amplia gama de anotaciones, aunque estas puedan ser ruidosas o imprecisas. A pesar de la variabilidad y la presencia de ruido", la metodología de agregación propuesta por Albarqouni y colaboradores, AggNet, demostró ser robusta frente a estas condiciones, e incluso llegó a mejorar el rendimiento de la red convolucional. Este enfoque de múltiples anotadores y agregación de etiquetas contribuye a la variación codificada dentro del modelo, lo que podría ser útil en escenarios reales donde los datos no siempre son claros o precisos. Sin embargo, es crucial planificar y diseñar cuidadosamente estos procesos de crowdsourcing para garantizar que la tarea de anotación no genere más ruido"que el beneficio que se obtiene de la diversidad de opiniones. Esta investigación resalta la utilidad del crowdsourcing y proporciona una metodología robusta para lidiar con las incertidumbres inherentes

a la recopilación de datos en masa [1].

Con estos antecedentes en el uso de machine learning para entrenar modelos en detección de imágenes, se evidencian bases para la identificación de características específicas de células neoplásicas en imágenes de tejidos de histología. La disponibilidad de datasets públicos de imágenes histológicas de tejidos positivos y negativos a cáncer con etiquetado de múltiples expertos nos permite generar una ruta de trabajo para este proyecto que además de desarrollar un modelo para predicción de células tumorales basado en extracción de características, utiliza el método de múltiples expertos para mejorar el rendimiento del modelo, aportando valor académico al etiquetado de múltiples expertos como una valiosa fuente de información para entrenamiento de modelos de machine learning.

Capítulo 4

METODOLOGÍA

4.1. Análisis exploratorio

Para la realización del presente estudio, se empleó una base de datos de acceso público que ya ha sido sometida a un proceso de preprocesamiento de imágenes, conforme a lo descrito en el estudio "*Learning from crowds in digital pathology using scalable variational Gaussian processes*"[16]. Esta base de datos comprende imágenes histológicas teñidas con hematoxilina y eosina, correspondientes a pacientes diagnosticados con cáncer de glándula mamaria. Estas imágenes se digitalizaron y se segmentaron en secciones cuadradas de 224 x 224 píxeles, denominadas *patch*. La naturaleza de alta resolución y el volumen significativo de píxeles de una imagen histológica digitalizada permite la generación de múltiples secciones por caso clínico, abarcando áreas con células tumorales, tejido de soporte normal e infiltrados inflamatorios. Este enfoque facilita la creación de un conjunto de datos extenso y diverso, idóneo para el entrenamiento y resolución de problemas de clasificación múltiple en el ámbito de la patología digital.

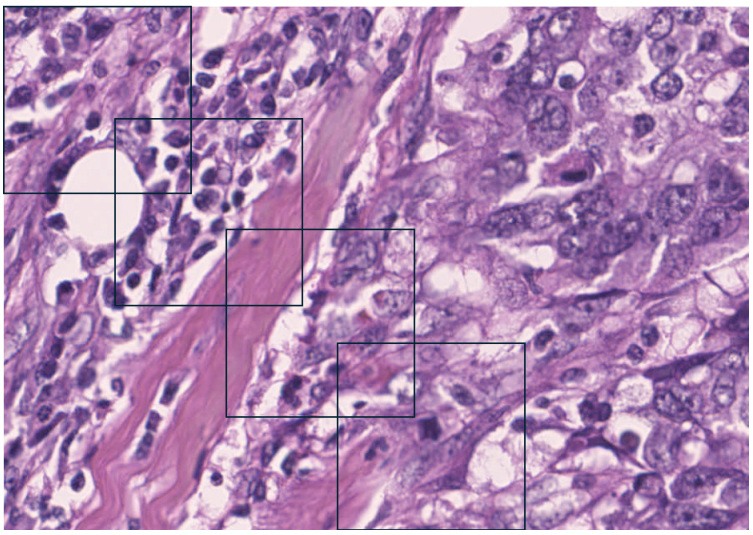


Figura 4.1: Representación del recorte de imágenes de 224 X 224 píxeles que componen la base datos. Cada recorte nuevo se genera a partir de la suma de 134 píxeles en las coordenadas.

La base de datos alberga un dataframe guardado en formato pkl '*train_crowdsourced_labels*' que es una representación vectorizada de las 75243 secciones de imágenes en formato png. Esta

tabla incorpora la variable "*annotations*", que registra las anotaciones y clasificaciones realizadas por diversos anotadores en un array de numpy. Cada subarray de esta estructura incluye, en su primera posición, el código identificador del anotador, seguido de la clasificación numérica asignada, que varía entre 1 y 3 (por ejemplo, [23, 1] corresponde al anotador número 23 asignando la clasificación 1). Adicionalmente, la variable "*label*" refleja la categorización definitiva establecida por un profesional experto, que oscila entre 1 (célula tumoral), 2 (estroma) y 3 (infiltrado inflamatorio). Por último, la variable "*patch*" almacena el nombre del archivo de imagen correspondiente.

Annotations	Label	Patch
[[13, 1], [23, 1], [17, 2], [7, 1], [5, 1], [11, 1], [12, 1], [21, 1], [19, 1], [8, 1], [10, 1], [22, 1], [14, 1], [24, 1], [20, 1], [9, 1], [15, 1]]	1	TCGA-A1-A0SK-DX1_xmin47201_ymin23382 _A0SK_A1_x_ini_0_y_ini_0.png
[[13, 1], [23, 1], [17, 2], [7, 1], [5, 1], [11, 1], [12, 2], [21, 1], [19, 1], [8, 1], [10, 1], [22, 2], [14, 1], [24, 2], [20, 1], [9, 1], [15, 1]]	1	TCGA-A1-A0SK-DX1_xmin47201_ymin23382 _A0SK_A1_x_ini_134_y_ini_0.png
[[13, 1], [23, 1], [17, 2], [7, 1], [5, 1], [11, 1], [12, 1], [21, 1], [19, 1], [8, 1], [10, 1], [22, 1], [14, 1], [24, 1], [20, 1], [9, 1], [15, 1]]	1	TCGA-A1-A0SK-DX1_xmin47201_ymin23382 _A0SK_A1_x_ini_0_y_ini_134.png

Cuadro 4.1: Etiquetas y anotaciones del conjunto de datos.

Realizamos un estudio de frecuencias para determinar la distribución de las etiquetas asignadas por los expertos, encontrando que la mayoría de estás categorizaciones hacen referencia a campos en los que sobresalen las células tumorales (tumor 49.5 %, estroma 36.8 %, inflamación benigna 13.7 %).

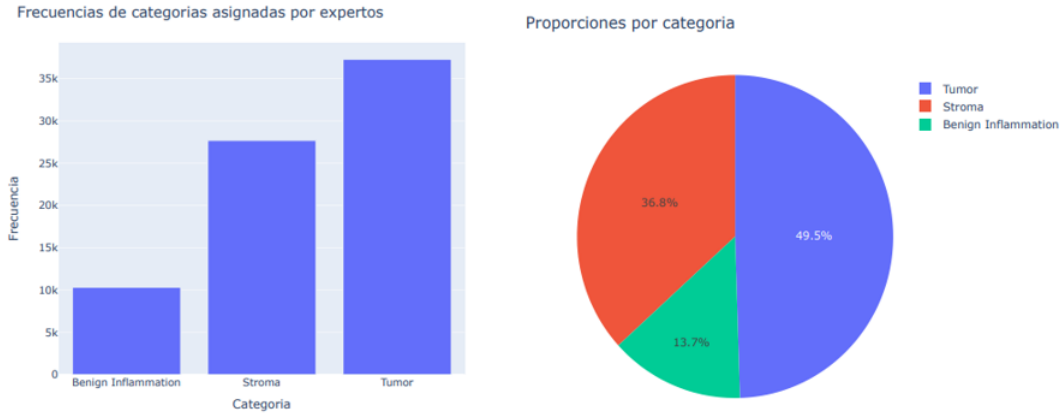
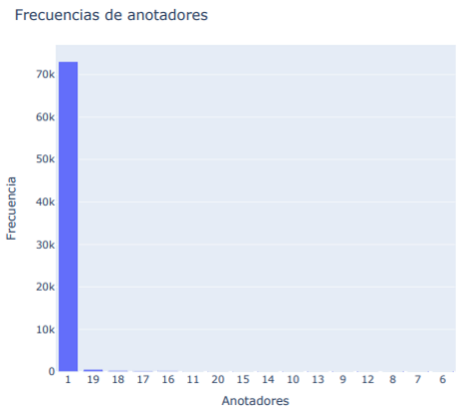


Figura 4.2: Distribución de las etiquetas.

Otro resultado interesante del análisis fue la cantidad de anotaciones por imagen, pues al mostrar las primeras filas del dataframe, se podía observar algunas filas con varias anotaciones, pero al generar el conteo se pudo observar que estas correspondían a una minoría, y que la mayor cantidad de imágenes empleadas en el dataset contienen una única anotación por crowdsourcing.



annotators	
1	73103
19	615
18	345
17	297
16	254
11	188
20	110
15	105
14	60
10	51
13	35
9	30
12	21
8	15
7	10
6	4

Figura 4.3: Frecuencia de las anotaciones.

Finalmente generamos una inspección visual de los "*patches*", que componen la base de datos. Al estudiar las imágenes de las 3 categorías vemos que se han clasificado de acuerdo a la predominancia celular que se encuentra en cada sección, esto es debido a que las imágenes son extraídas casos positivos a cáncer, debido a la complejidad del arreglo celular y la naturaleza de los infiltrados se pueden encontrar secciones.

Además, se realizó una inspección visual de algunas imágenes de muestra para comprender su clasificación. Las imágenes originales son muestras de tejido mamario de pacientes positivos para cáncer de seno. Sin embargo, debido a la naturaleza de los arreglos celulares, es imposible obtener muestras que contengan únicamente células tumorales. Por lo tanto, el banco de imágenes se creó generando recortes de 224 x 224 píxeles de las imágenes histológicas originales. Posteriormente, cada recorte fue observado por un experto, quien asignó la etiqueta correcta ("true label") basada en el tipo celular predominante en cada recorte: 1. tumor, 2. estroma, 3. inflamación.

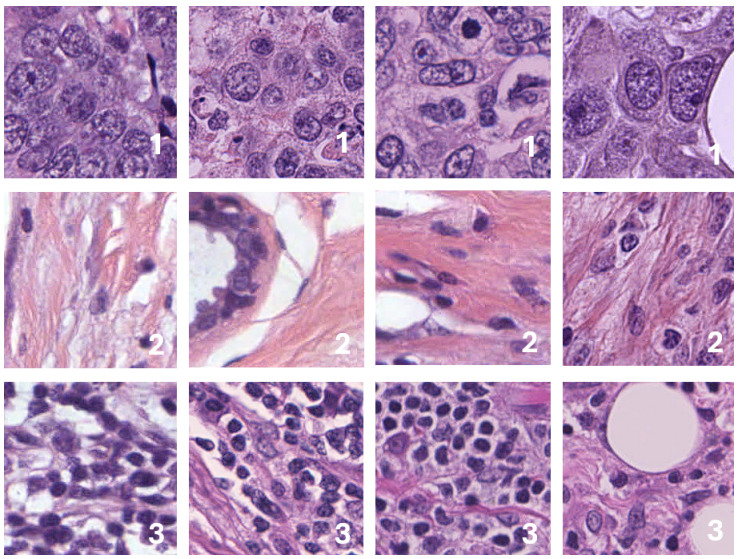


Figura 4.4: Ejemplos de *patches* contenidos en el conjunto de entrenamiento. Label 1 (tumor). Se observa un predominio de células tumorales en las secciones, estas células tienen núcleos aumentados de tamaños, núcleos vacualizados, migración de la cromatina, mitosis activas, entre otras características comunes de células neoplásicas. Label 2 (estroma) Se observa predominio de células de sostén y tejido conectivo en los campos, este tejido seguramente está próximo a los infiltrados tumorales y se caracteriza por tener células con núcleos normales y la presencia de fibras eosinofílicas que componen el tejido de sostén. Label 3 (infiltrado inflamatorio) se observa un predominio de células inflamatoria (linfocitos y células plasmáticas) que se caracterizan por tener núcleos de color basófilo intenso.

4.2. Aplicación de algoritmos de crowdsowrcing

Las anotaciones están inicialmente almacenadas en un DataFrame, donde cada fila representa una imagen y contiene una lista de pares [anotador, anotación] que representan el identificador del anotador y la etiqueta asignada por este. Las listas de anotaciones se expanden para que cada par [anotador, anotación] tenga su propia fila, lo que facilita la manipulación, análisis y aplicación de los métodos de crowdsourcing.

Se reformatean los datos para que se ajusten a los requisitos de los algoritmos de agregación, identificando task, worker, y label. Se eligieron los métodos de Majority Vote (MV), Worker Agreement with Aggregate (WAWA) y Dawid Skeene (DS) para procesar las etiquetas de crowdsourcing, por medio de la biblioteca crowd-kit para consolidar las anotaciones de múltiples anotadores en una única etiqueta consensuada por imagen.

Para este punto se definieron dos enfoques de estudio, el primero con una parte del dataframe (2141 imágenes) que en su totalidad tenían mas de 1 anotaciones de crowdsourcing y el segundo con la totalidad del dataframe (75243 imágenes) que pueden contener múltiples anotaciones o una única anotación por imagen, esto con el fin de estudiar con mas profundidad la variabilidad del trabajo de los anotadores.

4.3. Procesamiento y extracción de características para clasificación GOLD y crowdsourcing

Se utilizó la biblioteca de redes neuronales **Keras** bajo el entorno de Pyhton para cargar librerías, utilidades y aplicaciones necesarias para el procesamiento y entrenamiento de los modelos. Adicionalmente, otras bibliotecas esenciales como *os* para la manipulación de archivos y directorios, *numpy* para operaciones numéricas.

Se realizó la carga del conjunto de imágenes en el entorno de Python, un total de 75243 imágenes de entrenamiento contenidas en el directorio `/Train_non_experts_simple/` y 4364 imágenes de prueba contenidas en el directorio `/Test/`.

Directorio	Cantidad de imágenes
Train/1	37260
Train/2	27668
Train/3	10315
Test/1	2692
Test/2	1196
Test/3	476

Cuadro 4.2: composición de los directorios de entrenamiento y test.

Se configuró un tensor de entrada para recibir imágenes de tamaño 224x224 con 3 canales de color, que es el formato que contienen todas las imágenes de nuestro conjunto de datos. Se carga el modelo VGG16 pre-entrenado con los pesos obtenidos al entrenar en ImageNet, un gran conjunto de datos de imágenes utilizado para el entrenamiento de modelos de visión por computadora.

Se llamó la librería *AveragePooling2D* que añade una capa de agrupamiento promedio (average pooling) a la arquitectura del modelo. El agrupamiento promedio es una operación que reduce la dimensionalidad de las entradas calculando el promedio de los valores dentro de una ventana para cada canal de características. Esto reduce significativamente el tamaño de la salida, lo cual es útil para reducir el número de parámetros y la cantidad de cómputo necesario en las capas siguientes, así como para ayudar a evitar el sobreajuste.

4.3.1. Construcción y compilación del modelo

Para los entrenamientos con la base de anotaciones sencillas + anotaciones múltiples (75243 imágenes), se empleó la biblioteca de TensorFlow para el entrenamiento del modelo, configurando

una capa de entrada de 512, correspondiente a las características de cada imagen, siguiendo con una capa densa (completamente conectada) de 128 neuronas con la función de activación ReLU (Rectified Linear Unit). Esta capa transforma los datos de entrada en una representación de mayor nivel. Se aplicó una capa de Dropout con una tasa de 0.5. Esto significa que durante el entrenamiento, el 50 % de las características de la capa anterior se pondrán a cero aleatoriamente en cada paso de entrenamiento. Esta técnica ayuda a prevenir el sobreajuste. Posteriormente, se aplicó Una segunda capa densa de 64 neuronas, también con la función de activación ReLU. Finalmente, una capa de salida densa de 3 neuronas con la función de activación softmax, esta capa produce la probabilidad de que la entrada pertenezca a cada una de las 3 clases.

Para los entrenamientos con la base de múltiples anotaciones (2141 imágenes) se emplearon las mismas configuraciones de capa de entrada, activación ReLU y softmax. En la capa densa 512 neuronas con regularización L2, una capa de dropout 0.5, segunda capa densa de 1024 neuronas con regularización L2 y una segunda capa de dropout 0.5, esta configuración fue la que produjo mejor rendimiento considerando que estos modelos se entrenaron con numero bastante inferior de imágenes, de esta manera el aumento de neuronas en las capas permite un mejor aprendizaje con pocas imágenes.

Capítulo 5

Resultados y Discusión

5.1. Resultados del etiquetado por crowdsowrcing

Los resultados obtenidos al aplicar los métodos de agregación se añaden al DataFrame, proporcionando una comparación directa entre los diferentes métodos de agregación para cada imagen.

Patch	True Label	Majority Vote	WAWA	Dawid Skene	Annotations
TCGA-C8-A26Y-DX1_xmin14088_ymin37052_A26Y_C8_x_ini_1072_y_ini_1608.png	3	2	2	2	[[16,2],[11,2],[17,2],[13,2],[23,3],[22,2],[14,2],[9,2],[10,2],[18,2],[15,2],[7,2],[6,2],[12,2],[5,3]]
TCGA-C8-A26Y-DX1_xmin14088_ymin37052_A26Y_C8_x_ini_536_y_ini_1742.png	3	3	3	3	[[16,3],[11,3],[13,3],[23,3],[8,3],[14,3],[10,3],[18,3],[7,3],[6,3],[12,3],[24,3],[5,3]]
TCGA-C8-A26Y-DX1_xmin14088_ymin37052_A26Y_C8_x_ini_670_y_ini_1742.png	3	3	3	3	[[19,3],[16,3],[11,3],[13,3],[23,3],[8,3],[22,3],[10,3],[18,3],[7,3],[6,3],[12,3],[24,3],[5,3]]
TCGA-C8-A26Y-DX1_xmin14088_ymin37052_A26Y_C8_x_ini_1072_y_ini_1742.png	3	2	2	2	[[16,2],[17,2],[13,2],[23,3],[8,3],[22,2],[14,2],[9,2],[18,2],[15,2],[7,2],[6,3],[12,2],[5,3]]

Cuadro 5.1: Print del dataframe con la implementación de los métodos de crowdsourcing

5.2. Análisis comparativo de crowdsourcning/etiquetado verdadero

Iniciamos expandiendo el dataframe a partir de la columna "*annotations*", de esta manera obtenemos el listado de anotaciones individuales por anotador. El fundamento detrás de los métodos de crowdsourcing es que existen anotadores que tienen buenos rendimientos y otros pueden tener rendimientos regulares, esto se visualiza al calcular la precisión general de cada anotador vs la etiqueta verdadera.

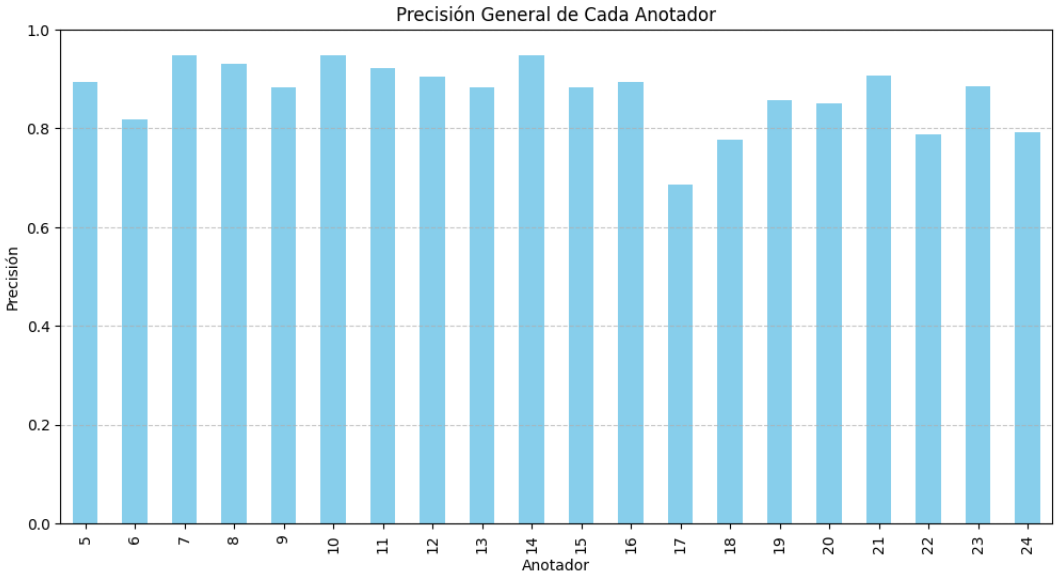


Figura 5.1: Gráfico de precisión de los anotadores.

En análisis del rendimiento de los anotadores se destacan los anotadores 7, 10 y 14 con precisiones por encima de 94 %, por otro lado, anotadores como 18 y 22 no pasan del 79 % y el peor rendimiento se registra en el anotador 17 con una precisión del 68.6 %. Las diferencias notables en los rendimientos del grupo son lo que fundamenta la metodología del crwodsorcing, donde en escenarios reales hay etiquetas ruidosas y etiquetas mas robustas en las que se pueden aplicar métodos como WAWA o Dawid Skeene para dar mas peso a unas tareas o unos trabajadores específicos.

Se aplicaron los métodos de crowdsourcing seleccionados al DataFrame y se analizó la distribución de las etiquetas generadas versus las etiquetas reales. Observamos que los conteos para la clasificación de la etiqueta 1 son similares en todos los métodos. Sin embargo, para las etiquetas 2 y 3 se observa una mayor variabilidad entre los métodos. Esto sugiere que las tareas 2 y 3, al ser tareas más difíciles de realizar, los métodos WAWA y Dawid-Skene (DS) dan más peso a las etiquetas de ciertos trabajadores o a etiquetas específicas, diferenciándose del Majority Vote (MV), que simplemente representa la moda (la etiqueta más repetida por imagen), considerando que todos los anotadores tienen el mismo peso.

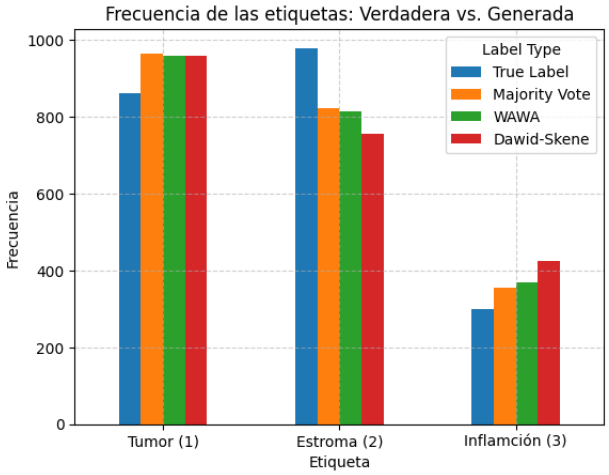


Figura 5.2: Distribución de las etiquetas generadas vs etiqueta real

Al revisar los reportes de precisión en detalle, observamos que, en general, los tres métodos tienen resultados similares, siendo el método de Dawid Skene el que presenta una precisión más baja en términos generales. Analizando la precisión por etiqueta, notamos que la tarea 2 (clasificación de estroma) es la más sencilla, mostrando la mayor precisión en todos los métodos. Por otro lado, la tarea 3 (clasificación de infiltrados inflamatorios) resulta ser especialmente confusa o difícil, lo que provoca que el método de Dawid Skene cometa más errores, reduciendo significativamente su precisión en esta categoría.

Método	**Precisión Etiqueta 1**	**Precisión Etiqueta 2**	**Precisión Etiqueta 3**	**Precisión General**
Majority Vote (MV)	87.34 %	92.58 %	69.86 %	86.45 %
WAWA	87.60 %	93.36 %	69.29 %	86.64 %
Dawid Skene	87.50 %	94.05 %	61.88 %	84.73 %

Cuadro 5.2: Precisión general y por categoría para los métodos de crowdsourcing aplicados al dataframe de múltiples anotaciones.

Se finaliza el análisis de los métodos de corowdsowrcing, concluyendo que existen diferencias notables entre las etiquetas verdaderas y las etiquetas generadas. La adaptación del dataframe para agregar los métodos de crowdsourcing permitió producir un entorno de trabajo organizado y útil para generar las extracciones de características, arrays de etiquetas y modelados comparativo de los próximos capítulos.

5.3. Caracteristicas para entrenamiento

En la impresión de algunas características extraídas con este método, vemos como la aplicación del Average Pooling produce un total de 512 características por imagen, clasificadas en 3 labels con One Hot Encoding para 75243 registros del conjunto de entrenamiento.

Nombre	Características (512)	Etiqueta verdadera
TCGA-BH-A0BL-DX1_xmin25219_ymin42847_A0BL_BH_x_ini_5226_y_ini_1340.png	[2.7226715e+00 1.6715584e-02 9.8814144e+00 ... 0.0000000e+00 5.6466312e+00 1.2513410e-01]	1
TCGA-OL-A5D7-DX1_xmin114443_ymin22490_A5D7_OL_x_ini_4288_y_ini_4020.png	[3.1049452e-03 0.0000000e+00 2.4595180e+00 ... 0.0000000e+00 6.0292780e-01 0.0000000e+00]	1
TCGA-EW-A1P1-DX1_xmin50567_ymin38988_A1P1_EW_x_ini_2680_y_ini_938.png	[5.6900997e+00 0.0000000e+00 1.1595472e+01 ... 0.0000000e+00 2.3865326e-01 5.6102210e-01]	2

Cuadro 5.3: Ejemplos de características extraídas con la red VGG16

Este método de extracción de características se implementó en una función que produce 6 arrays por cada imagen procesada, uno con el nombre, un segundo array que almacena una lista de 512 características extraídas y los 4 arrays restantes con el numero de la etiqueta correspondiente (etiqueta verdadera, Majority Vote, WWAWA y DS). De esta manera, se obtuvieron todas las informaciones necesarias para el entrenamiento de los modelos y el método *AveragePooling2D* produjo características con beneficios como la reducción de la dimensionalidad y mejora en tiempos de entrenamiento y capacidad de computo necesaria.

5.4. Entrenamiento del modelo

Se plantean hasta 30 épocas procesando lotes de a 8 entradas para el entrenamiento de los modelos, se utiliza la funciona callback para detener el entrenamiento en la época con mejor valor de perdida. En el proceso de entrenamiento se observa una tendencia general hacia la disminución de la pérdida (tanto en entrenamiento como en validación) y el aumento de la precisión, lo cual indica que los modelos aprendieron de manera efectiva.

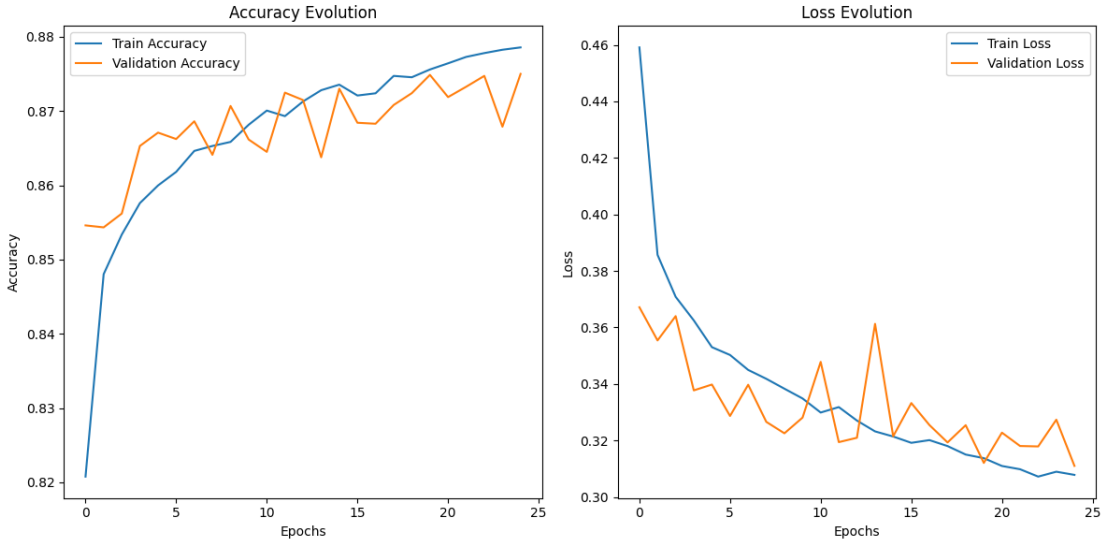


Figura 5.3: Evolución del entrenamiento en el modelo Gold

5.5. Implementación

Con el objetivo de demostrar cómo un modelo de Machine Learning para la clasificación del cáncer de mama puede integrarse en el flujo de trabajo de un profesional de la salud, se desarrolló una herramienta que permite al usuario utilizar el modelo para clasificar múltiples imágenes y diagnosticar a un gran número de pacientes en el menor tiempo posible. La rapidez es crucial en el diagnóstico de enfermedades como el cáncer, donde el tiempo es una variable crítica.

Para el desarrollo de la aplicación se utilizó la librería Dash de Python, un framework que permite a los programadores de Python desarrollar aplicaciones web completas centradas en datos y análisis, así como paneles de control interactivos, sin necesidad de conocimientos en Javascript. Dash utiliza Flask para el backend, y por defecto, emplea la librería Plotly para la creación de gráficos, aunque su uso no es estrictamente necesario. React se utiliza para el manejo de todos los componentes, de modo que una aplicación de Dash se renderiza como una aplicación de una sola página en React [6].

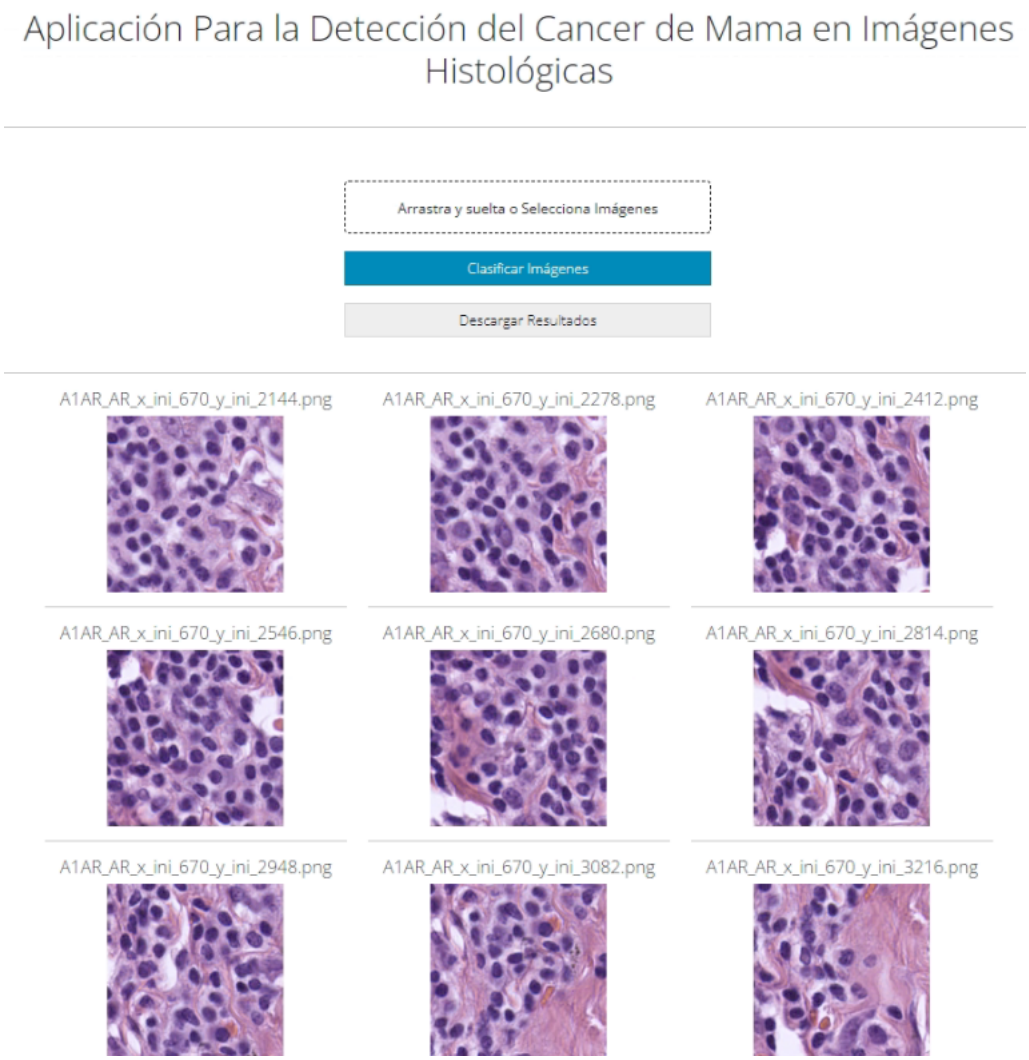


Figura 5.4: Ejemplo de algunas imágenes histológicas cargadas en la interfaz para su clasificación.

La aplicación permite al usuario cargar múltiples imágenes en una interfaz sencilla e intuitiva. Seguidamente el botón "*Clasificar Imágenes*", que es el disparador principal de la aplicación, aplica la tarea de predicción del modelo que asigna las probabilidades a cada categoría (tumor, estroma o inflamación) y genera la etiqueta final basado en la probabilidad más alta. Posteriormente, los resultados de cada imagen se presentan en un dataframe, con una fila por imagen y las probabilidades y categorización en las columnas correspondientes. Finalmente, el usuario puede descargar el dataframe con los resultados en una hoja de cálculo.

Imagen	Predicción	Prob Célula Tumoral	Prob Estromal	Prob Infiltrado Inflamatorio
A1AR_AR_x_ini_670_y_ini_2144.png	Infiltrado Inflamatorio	0	0	1
A1AR_AR_x_ini_670_y_ini_2278.png	Infiltrado Inflamatorio	0	0	1
A1AR_AR_x_ini_670_y_ini_2412.png	Infiltrado Inflamatorio	0	0	1
A1AR_AR_x_ini_670_y_ini_2546.png	Infiltrado Inflamatorio	0	0	1
A1AR_AR_x_ini_670_y_ini_2680.png	Infiltrado Inflamatorio	0	0	1
A1AR_AR_x_ini_670_y_ini_2814.png	Infiltrado Inflamatorio	0	0	1
A1AR_AR_x_ini_670_y_ini_2948.png	Infiltrado Inflamatorio	0	0	1
A1AR_AR_x_ini_670_y_ini_3082.png	Infiltrado Inflamatorio	0	0.03	0.97
A1AR_AR_x_ini_670_y_ini_3216.png	Infiltrado Inflamatorio	0.02	0.39	0.59
A1AR_AR_x_ini_670_y_ini_3350.png	Infiltrado Inflamatorio	0.01	0.2	0.79
A1AR_AR_x_ini_670_y_ini_3484.png	Infiltrado Inflamatorio	0.01	0.38	0.61
A1AR_AR_x_ini_804_y_ini_2010.png	Infiltrado Inflamatorio	0	0	1
A1AR_AR_x_ini_804_y_ini_2144.png	Infiltrado Inflamatorio	0	0	1
A1EW_BH_x_ini_2278_y_ini_268.png	Estromal	0	1	0
A1EW_BH_x_ini_2278_y_ini_402.png	Estromal	0.01	0.99	0
A1EW_BH_x_ini_2278_y_ini_536.png	Estromal	0	0.99	0.01
A1EW_BH_x_ini_2278_y_ini_670.png	Estromal	0.01	0.99	0
A1EW_BH_x_ini_2278_y_ini_134.png	Estromal	0	1	0
A1EW_BH_x_ini_2278_y_ini_804.png	Estromal	0	0.99	0.01

Figura 5.5: Dataframe generado con las categorizaciones y las probabilidad de cada categoría para cada imagen

Esta herramienta es un ejemplo de como se pueden emplear modelos producidos por crowdsourcing para que el personal administrativo de los laboratorios de patología pueda optimizar el flujo de casos y dar prioridad al diagnostico profesional de pacientes que tengan mas probabilidad de presencia de células tumorales en sus estudios de histopatología.

5.6. Validación del modelo

Se carga el modelo y se ejecuta una tarea de predicción en el conjunto de datos “test”, un conjunto compuesto por 4364 imágenes histológicas clasificadas en las mismas 3 clases, pero que no fueron empleadas en el entrenamiento del modelo.

Se utiliza la librería “clasification_report” de la biblioteca “sklearn.metrics” para generar los reportes de Precisión, recall y f1-score para las 3 clases. Esta validación se emplea para todos para el etiquetado gold y los métodos de crowdsourcing para su posterior análisis comparativo.

5.7. Resultados de la validación

Se obtuvieron resultados por dos métodos diferentes, el primero por el entrenamiento y validación del conjunto completo de 75243 imágenes (múltiples anotaciones + única anotación), y un segundo método de entrenamiento y validación con 2141 imágenes (únicamente múltiples anotaciones), este segundo enfoque pensamos que es el mas valioso para estudiar el impacto del crowdsourcing.

El primer método obtuvo nuestra referencia Gold estándar con una precisión de 85.6 % y un F1-score de 85.3 %, estos resultados muestran un modelo muy eficiente para la tarea de clasificación de imágenes histológicas de de tejido mamario en las categorías de región tumoral, estroma o infiltrado inflamatorio, esta eficiencia se debe al amplio conjunto de entrenamiento (75243 imágenes) que además estaba etiquetado en su totalidad por un experto. Los métodos de crowdsourcing evaluados mostraron resultados similares siendo el WAWA (Worker Agreement with Aggregate) el mas cercano presentando una precisión 84.8 % y un F1-score 84.4 %, seguidamente el DS (Dawid-Skeene) presentó una precisión 83.6 % y un F1-score 83.6 %, y por ultimo el método MV (Mayority Vote) obtuvo una precisión 83.4 % y un F1-score 83.3 %. Considerando que el método Mayority vote obtuvo los resultados mas bajos continua siendo un modelo competente para la clasificación de imágenes histológicas de tejido mamario y en comparación con el modelo Gold obtuvo diferencias mínimas (Precisión -2.2 %, Sensibilidad -2.8 % y F1 score -2.0 %).

Single annotation + Multiple annotations (75243)	**Gold**	**Mayority vote**	**WAWA**	**DS**
Accuracy	85,6	83,4	84,8	83,6
Recall	81,4	78,6	80,1	79,9
F1 score	85,3	83,3	84,4	83,6

Cuadro 5.4: Resultados de la validación, método Single annotation + Multiple annotations (75243 imágenes)

El segundo método evaluado nos mostró resultados muy interesantes en cuanto a la aplicación de métodos de crowdsourcing, filtramos y utilizamos para el entrenamiento únicamente 2141 imágenes que en su totalidad contenían múltiples anotaciones, de esta manera, sacrificando una parte importante del conjunto inicial logramos dar prioridad al estudio comparativo de los múltiples anotadores. Los resultados mostraron un buen rendimiento en general para los modelos siendo el mejor método el MV con una precisión 83.1 % y un F1 Score 82.7 %, seguidamente el modelo Gold con una precisión 82.2 % y un F1-score 81.9 %, el método DS fue muy cercano al Gold mostrando una precisión 82.2 % y un F1-score 81.8 % y finalmente el método WAWA obtuvo el rendimiento mas bajo del experimento, sin embargo, continua siendo muy cercano al modelo Gold con una precisión 81.6 % y un F1-score 81.3 %.

Multiple annotations only (2141)	Gold	Mayority vote	WAWA	DS
Accuracy	82,2	83,1	81,6	82,2
Recall	76,4	78,2	76,9	76,9
F1 score	81,9	82,7	81,3	81,8

Cuadro 5.5: Resultados de la validación, método Múltiples anotaciones (2141 imágenes)

Estos resultados han sido muy interesantes en comparación con el primer experimento, ya que al filtrar y utilizar únicamente las imágenes con múltiples anotaciones redujimos la variabilidad de rendimiento entre los métodos, mostrando que el etiquetado por crowdsourcing puede producir resultados similares a los obtenidos por medio de etiquetado de expertos (Precisión +0.9/-0.6 %, Sensibilidad +0.5/+1.8 % y F1 score +0.8/-0.6 %).

5.8. Discusión y conclusiones

La tarea de clasificar imágenes histológicas es muy compleja debido a la naturaleza de los tejido y sus arreglos celulares, los tejidos de glándulas mamarias normalmente se componen de diferentes tipos celulares como células alveolares, mioepiteliales, adipocitos, células plasmáticas, tejido conjuntivo, entre otros; con esto podemos asumir que cuando se presentan crecimientos tumorales estos van a estar acompañados de una variedad celular normal del tejido. La figura 5.1 muestra un ejemplo de la composición de las imágenes del dataset que empleamos en el entrenamiento de los modelos, vemos como a medida que los *patchs* se desplazan a secciones diferentes de la imagen la etiqueta cambia de acuerdo al criterio del experto que clasificó basado en la predominancia de las células que componían cada patch.

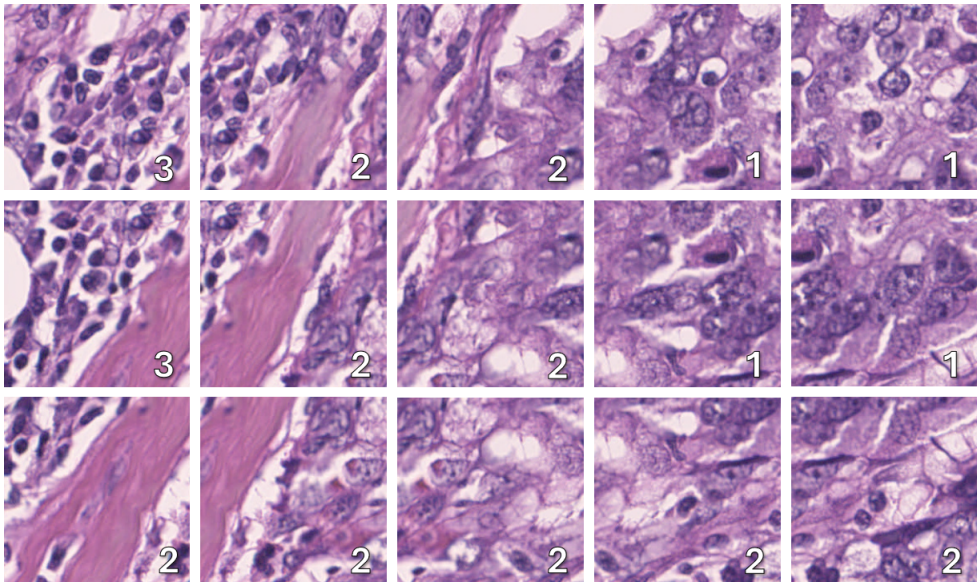


Figura 5.6: Composición del etiquetado. Vemos como en los recortes de la izquierda predomina el infiltrado inflamatorio (3), este se compone de células pequeñas con tinción basofila intensa. En el centro predomina el estroma (2), un arreglo fibrilar de color eosinofilico intenso que se compone de tejido conjuntivo y proteínas. En los recortes de la derecha predominan las células tumorales (1), células de tamaños aumentados y desuniformes con núcleos irregulares y mitosis activas.

La variación de tipos celulares que hay en cada recorte se traduce en problemas de sensibilidad y precisión del modelo, la tabla 5.3 muestra las métricas detalladas por clase para los modelos de múltiples anotaciones, vemos como los modelos en general son competentes para la tarea de clasificar células de tumorales (clase 1), sin embargo, la tarea de clasificar el estroma (clase 2) y el infiltrado inflamatorio (clase 3) genera mas inconvenientes para los anotadores, produciendo un rendimiento mas bajo.

Clase	Precisión	Recall	F1-score
MV - Clase 1	0.847	0.921	0.883
MV - Clase 2	0.818	0.682	0.744
MV - Clase 3	0.75	0.693	0.721
MV General	**0.831**	**0.782**	**0.827**
WAWA - Clase 1	0.844	0.9010	0.872
WAWA - Clase 2	0.757	0.681	0.717
WAWA - Clase 3	0.775	0.672	0.720
WAWA General	**0.816**	**0.769**	**0.813**
DS - Clase 1	0.840	0.919	0.878
DS - Clase 2	0.780	0.684	0.729
DS - Clase 3	0.798	0.624	0.700
DS General	**0.822**	**0.769**	**0.818**

Cuadro 5.6: Resultados de los modelos entrenados con el método múltiples anotaciones detallado por clases. MV (Majority Vote), WAWA (Worker Agreement with Aggregate), DS (dawid-Skeene).

Los problemas de clasificación para las clases 2 y 3 se sustentan en la complejidad natural de los tejidos que mencionábamos anteriormente, un modelo de salida binaria que solo tenga que clasificar si hay presencia o no de células tumorales tendrá rendimientos mas altos debido a la simplicidad de la tarea, cuando generamos un problema de clasificación de múltiples clases al considerar el estroma y la inflamación se añade dificultad a la tarea del etiquetado debido a la naturaleza de los infiltrados inflamatorios, que normalmente, se pueden ver junto al estroma, en este caso los anotadores deben estimar bajo su criterio que grupo celular predomina para asignar un etiquetado, este proceso termina generando etiquetas ruidosas que disminuyen el rendimiento. Incluso, un caso hipotético en el que se agregue una cuarta clase como tejido adiposo, generará mas complejidad a las tareas de clasificación ya que es normal encontrar adipocitos, celulas inflamatorias y tejido conjuntivo cercanos. Estos desafíos de variedad celular y etiquetado ruidoso son abordados mediante la creación de *patchs* o recortes de menor tamaño, que permiten entrenar modelos con menos capacidad de computo y disminuir la probabilidad de mezclar muchos grupos celulares en una misma imagen de entrenamiento, sin embargo, se requiere mas trabajos e investigación encaminada a mejorar los rendimientos en problemas de clasificación multiclase.

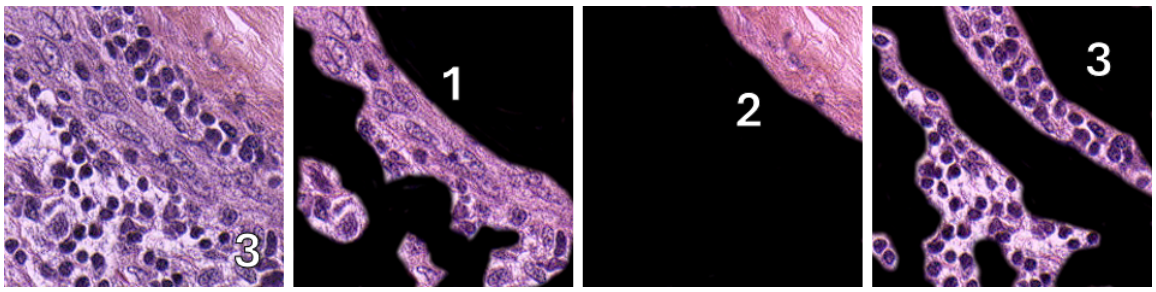


Figura 5.7: Ejemplo de una imagen etiquetada como clase 3 por el experto. Vemos como a pesar de manejar recortes pequeños de 224 x 224 píxeles, algunas imágenes pueden contener características evidentes de diferentes clases, esto puede generar ruido en el proceso de etiquetado y de entrenamiento de los modelos.

Finalmente, al comparar los modelos de crowdsorcing de múltiples anotaciones entrenados con 2.141 imágenes contra el modelo Gold estándar entrenado con 75.243 imágenes con etiquetado real, encontramos que los modelos de crowdsourcing continúan siendo competentes obteniendo rendimientos similares, lo que también nos permite abordar planteamientos interesantes de disminución del trabajo de etiquetado, disminución del personal experto en el etiquetado, reducción de los conjuntos de entrenamiento y de la capacidad de computo necesaria para producir modelos competentes en la predicción de células tumorales.

Capítulo 6

Conclusiones y trabajos futuros

6.1. Conclusiones y trabajos futuros

6.1.1. Conclusiones

La integración entre la inteligencia artificial y la medicina está creciendo a pasos agigantados, pero aún tiene mucha tela por cortar. La investigación científica dirigida a generar modelos de Machine Learning para diagnostico por imágenes ha mostrado avances interesantes que abren puertas para la integración de modelos de inteligencia artificial y diagnósticos de los profesionales de la medicina.

En este trabajo desarrollamos el enfoque de construcción de modelos de Machine Learning para la identificación de patrones tumorales en imágenes histológicas demostrando la eficacia de los métodos de crowdsourcing para la anotación de imágenes histológicas y su potencial para reducir la carga de trabajo en el etiquetado sin comprometer significativamente la precisión del modelo.

Hemos demostrado que la red neuronal VGG16 es efectiva para aplicaciones de visión por computadora, particularmente en la extracción de características de imágenes histológicas.

La implementación de métodos de crowdsourcing como MV (Majority Vote), WAWA (Worker Agreement with Aggregate) y DS (Dawid-Skene) demostró ser efectiva para consolidar las anotaciones de múltiples anotadores en una única etiqueta consensuada por imagen. Los métodos WAWA y DS mostraron una capacidad superior para manejar las variabilidades en la calidad de las anotaciones, destacándose en tareas más difíciles como la clasificación de infiltrados inflamatorios.

La evaluación de los métodos de crowdsourcing en un conjunto de datos con múltiples anotaciones (2141 imágenes) mostró que los modelos entrenados con estas etiquetas alcanzaron rendimientos muy cercanos al modelo Gold estándar (75243 imágenes). Esto sugiere que, aunque el etiquetado de expertos proporciona una referencia sólida, los métodos de crowdsourcing pueden ser una alternativa viable para reducir el trabajo de etiquetado y la dependencia de personal altamente especializado, sin comprometer significativamente la precisión del modelo.

La clasificación de imágenes histológicas en múltiples clases (tumor, estroma e infiltrado inflamatorio) presentó desafíos significativos debido a la complejidad y la variabilidad natural de los tejidos. Las imágenes que contenían múltiples tipos celulares dificultaron el proceso de anotación y,

en consecuencia, afectaron el rendimiento del modelo. La tarea de clasificar el estroma y los infiltrados inflamatorios resultó ser especialmente complicada, como se refleja en las métricas de precisión y recall más bajas para estas categorías.

6.1.2. Trabajos futuros

Los resultados destacan la importancia de abordar la variabilidad y complejidad natural de los tejidos en la clasificación multiclase y abren la puerta a futuras investigaciones para optimizar estos procesos en patología digital.

Para mejorar la precisión en problemas de clasificación multiclase en imágenes de histología, se recomienda explorar técnicas adicionales de preprocesamiento y segmentación de imágenes. La creación de conjuntos de datos más balanceados y la mejora en la calidad y cantidad de las anotaciones podría contribuir a mejorar los rendimientos de los modelos de clasificación.

Bibliografía

[1] Shadi Albarqouni, Christoph Baur, Felix Achilles, Vasileios Belagiannis, Stefanie Demirci, and Nassir Navab. Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images. *IEEE transactions on medical imaging*, 35(5):1313–1321, 2016.

[2] Mohamed Amgad, Lamees A Atteya, Hagar Hussein, Kareem Hosny Mohammed, Ehab Hafiz, Maha AT Elsebaie, Ahmed M Alhusseiny, Mohamed Atef AlMoslemany, Abdelmagid M Elmatboly, Philip A Pappalardo, et al. Nucls: A scalable crowdsourcing approach and dataset for nucleus classification and segmentation in breast cancer. *GigaScience*, 11:giac037, 2022.

[3] Sebastián Badaró, Leonardo Javier Ibañez, and Martín Jorge Agüero. Sistemas expertos: fundamentos, metodologías y aplicaciones. *Ciencia y tecnología*, (13):349–364, 2013.

[4] Eucario Parra Castrillón. Tecnología de sistemas expertos para el análisis del comportamiento humano de acuerdo con el modelo del cerebro triádico. *Revista Virtual Universidad Católica del Norte*, (11), 2003.

[5] Rahul Chauhan, Kamal Kumar Ghanshala, and R.C Joshi. Convolutional neural network (cnn) for image detection and recognition. In *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)*, pages 278–282, 2018.

[6] Elias Dabbas. *Interactive Dashboards and Data Apps with Plotly and Dash: Harness the power of a fully fledged frontend web framework in Python–no JavaScript required*. Packt Publishing Ltd, 2021.

[7] Alican Dogan and Derya Birant. A weighted majority voting ensemble approach for classification. In *2019 4th International Conference on Computer Science and Engineering (UBMK)*, pages 1–6, 2019.

[8] Jacques Ferlay, Isabelle Soerjomataram, Rajesh Dikshit, Sultan Eser, Colin Mathers, Marise Rebelo, Donald Maxwell Parkin, David Forman, and Freddie Bray. Cancer incidence and mortality worldwide: sources, methods and major patterns in globocan 2012. *International journal of cancer*, 136(5):E359–E386, 2015.

[9] International Agency for Research on Cancer. Global cancer observatory, 2023.

[10] Javier Galarza Hernández. *Reducción de dimensionalidad en Machine Learning. Diagnóstico de cáncer de mama bsado en datos genómicos y de imagen*. PhD thesis, Universitat Politècnica de València, 2017.

[11] Anne Grote, Nadine S Schaadt, Germain Forestier, Cédric Wemmert, and Friedrich Feuerhake. Crowdsourcing of histological image labeling and object delineation by medical students. *IEEE transactions on medical imaging*, 38(5):1284–1294, 2018.

[12] James V Lacey Jr, Susan S Devesa, and Louise A Brinton. Recent trends in breast cancer incidence and mortality. *Environmental and molecular mutagenesis*, 39(2-3):82–88, 2002.

[13] Yunhui Li, Liang Chang, Long Li, Xuguang Bao, and Tianlong Gu. Key research issues and related technologies in crowdsourcing data collection. *Wireless Communications and Mobile Computing*, 2021:1–13, 2021.

[14] Yun Liu, Timo Kohlberger, Mohammad Norouzi, George E Dahl, Jenny L Smith, Arash Mohtashamian, Niels Olson, Lily H Peng, Jason D Hipp, and Martin C Stumpe. Artificial intelligence–based breast cancer nodal metastasis detection: Insights into the black box for pathologists. *Archives of pathology & laboratory medicine*, 143(7):859–868, 2019.

[15] Ana López Díaz. Fundamentos matemáticos de los métodos kernel para aprendizaje supervisado. 2018.

[16] Miguel López-Pérez, Mohamed Amgad, Pablo Morales-Álvarez, Pablo Ruiz, Lee AD Cooper, Rafael Molina, and Aggelos K Katsaggelos. Learning from crowds in digital pathology using scalable variational gaussian processes. *Scientific reports*, 11(1):11612, 2021.

[17] Miguel Lugones Botell and Marieta Ramírez Bermúdez. Aspectos históricos y culturales sobre el cáncer de mama. *Revista cubana de medicina general integral*, 25(3):0–0, 2009.

[18] Nachaat Mohamed. Importance of artificial intelligence in neural network through using mediapipe. In *2022 6th International Conference on Electronics, Communication and Aerospace Technology*, pages 1207–1215, 2022.

[19] María Navelonga Moreno García, Francisco J García-Peñalvo, et al. Modelos de estimación del software basados en técnicas de aprendizaje automático. 2007.

[20] Santiago Murillo Rendón. *Metodología para el aprendizaje de máquina a partir de múltiples expertos en procesos de clasificación de bioseñales*. PhD thesis, 2013.

[21] Giuseppe Musumeci. Past, present and future: overview on histology and histopathology. *J Histol Histopathol*, 1(5):1–3, 2014.

[22] Organizacion Mundial de la Salud (OMS). Cáncer de mama. World Health Organization, March 2021. [En línea]. Available: `https://www.who.int/es/news-room/fact-sheets/detail/breast-cancer`. [Último acceso: 16 Mayo 2023].

[23] Şaban Öztürk and Bayram Akdemir. Application of feature extraction and classification methods for histopathological image using glcm, lbp, lbglcm, glrlm and sfta. *Procedia computer science*, 132:40–46, 2018.

[24] Enrique G Rodrigo, Juan A Aledo, and José A Gámez. Cglad: Glad en problemas de big crowdsourced data. In *XVIII Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA 2018): avances en Inteligencia Artificial. 23-26 de octubre de 2018 Granada, España*, pages 1111–1116. Asociación Española para la Inteligencia Artificial (AEPIA), 2018.

[25] Peter Rossini et al. Using expert systems and artificial intelligence for real estate forecasting. In *Sixth Annual Pacific-Rim Real Estate Society Conference, Sydney, Australia*, pages 24–27. Citeseer, 2000.

[26] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[27] Shuqiu Tan, Jiahao Pan, Jianxun Zhang, and Yahui Liu. Casvm: An efficient deep learning image classification method combined with svm. *Applied Sciences*, 12(22):11690, 2022.

[28] ZiQi Tao, Aimin Shi, Cuntao Lu, Tao Song, Zhengguo Zhang, and Jing Zhao. Breast cancer: epidemiology and etiology. *Cell biochemistry and biophysics*, 72:333–338, 2015.

[29] Dmitry Ustalov, Nikita Pavlichenko, and Boris Tseitlin. Learning from crowds with crowd-kit. *arXiv preprint arXiv:2109.08584*, 2021.

[30] Mitko Veta, Josien PW Pluim, Paul J Van Diest, and Max A Viergever. Breast cancer histopathology image analysis: A review. *IEEE transactions on biomedical engineering*, 61(5):1400–1411, 2014.