Reconocimiento de células sanguíneas por medio de aprendizaje profundo y generación de datos sintéticos

Presenta:

Nohemí Sánchez Medel

Tesis sometida como requisito parcial para obtener el grado de

Maestra en Ciencias en el área de Ciencias y Tecnologías Biomédicas

en el:

Instituto Nacional de Astrofísica, Óptica y Electrónica

17 de Febrero de 2023
Santa María Tonantzintla, Puebla

Dirigida Por:

Dra. Raquel Díaz Hernández
INAOE

Resumen

El aumento de datos (Data Augmentation) es una técnica que adopta el enfoque de generar más datos de entrenamiento a partir de los datos disponibles. El aumento de datos también es útil para mejorar el rendimiento y la precisión de los modelos de aprendizaje profundo, mediante la creación de ejemplos nuevos y diferentes para entrenar conjuntos de datos.

En este trabajo de tesis, se experimenta con algoritmos de aumento de datos para la generación de imágenes sintéticas tradicionales y poligonales, con el fin de contar con un mayor número de imágenes para realizar el entrenamiento de los modelos, reconociendo enfermedades hematológicas mediante imágenes de células sanguíneas. Esto con la finalidad de resolver el problema de la escasez de datos y clases desbalanceadas al momento de entrenar redes neuronales. Se generaron imágenes sintéticas tradicionales mediante procedimientos clásicos como efecto espejo, rotaciones, contraste, brillo, entre otros; y se generaron imágenes sintéticas poligonales mediante la aplicación de máscaras. Se realizó el reconocimiento de células sanguíneas por medio del aprendizaje profundo.

Se utilizó el conjunto de datos ALL-IDB2, el cual contiene 260 imágenes originales segmentadas de células normales y blásticas. Se generaron 12000 imágenes sintéticas tradicionales mediante transformaciones geométricas, y se generaron 12000 imágenes sintéticas poligonales mediante máscaras. Se probaron 5 modelos diferentes: *Compact, MobileNetV2, AlexNet, ResNet-50 y Enhanced.* Se comprobó que con la utilización de imágenes sintéticas se aumenta la precisión para el reconocimiento de enfermedades hematológicas.

Los beneficios de este trabajo se pueden resumir en la obtención de más datos para el entrenamiento, se reducen los costos de recopilación y etiquetado de datos, se mejora el desempeño de los modelos y se elimina el sobreajuste al tener variabilidad en los datos. Como beneficio adicional, con la técnica del aumento de datos se elimina el desequilibrio de clases.

Se muestra la precisión de cada una de las pruebas realizadas, y se presenta la gráfica comparativa de los resultados obtenidos.

Por último, se observó que con las imágenes sintéticas poligonales los porcentajes de precisión se sitúan por encima del 98% y con base en los resultados obtenidos se deduce que el número ideal de imágenes para entrenar los 5 modelos de aprendizaje profundo se sitúa entre 2500 y 5000.

Palabras reservadas: Aumento de datos, aprendizaje profundo, diseño de algoritmos, enfermedades hematológicas, análisis de imágenes.

Abstract

Data augmentation is a technique that takes the approach of generating more training data from the available data. Data augmentation is also useful for improving the performance and accuracy of deep learning models by creating new and different examples to train datasets.

In this thesis work, we experiment with data augmentation algorithms for the generation of traditional and polygonal synthetic images, to have a larger number of images to perform model training, recognizing hematological diseases using blood cell images. This is to solve the problem of data scarcity and unbalanced classes when training neural networks. Traditional synthetic images were generated employing classical procedures such as mirror effect, rotations, contrast, and brightness, among others; and polygonal synthetic images were generated by applying masks. Blood cell recognition was performed utilizing deep learning.

The ALL-IDB2 dataset was used, which contains 260 original segmented images of normal and blast cells. 12000 traditional synthetic images were generated by geometric transformations, and 12000 polygonal synthetic images were generated by masks. Five different models were tested: Compact, MobileNetV2, AlexNet, ResNet-50, and Enhanced. It was found that the use of synthetic images increases the accuracy of the recognition of hematologic diseases.

The benefits of this work can be summarized as obtaining more data for training, reducing data collection and labeling costs, improving model performance, and eliminating over-fitting by having variability in the data. As

an additional benefit, with the data augmentation technique, the class imbalance is eliminated.

The precision of each of the tests performed is shown, and the comparative graph of the results obtained is presented.

Finally, it was observed that with the polygonal synthetic images, the accuracy percentages are above 98%, and based on the results obtained it is deduced that the ideal number of images to train the 5 deep learning models is between 2500 and 5000.

Keywords: data augmentation, deep learning, algorithm design, hematological diseases, image analysis.

Agradecimientos

Quiero agradecer a todas y cada una de las personas que me han brindado su apoyo en todo momento para lograr este objetivo. No tengo palabras para expresarles lo importante y necesario que es para mí contar con todos ustedes, mil gracias, en particular mencionaré a quienes fueron pilares para la realización de este trabajo de tesis.

Primeramente, a mi asesora, la Dra. Raquel Díaz Hernández, toda mi admiración para ella, gracias por motivarme a seguir adelante, por creer en mí en todo momento, y porque con una sola frase me dijo tantas cosas “solo tú puedes hacer que las cosas pasen”, gracias por todo.

También quiero agradecer especialmente al Dr. Leopoldo Altamirano Robles, por compartir sus conocimientos conmigo y porque con su visión me permite encontrar una salida a las dificultades que se me presentan, gracias.

Agradezco al MC. José de Jesús Velázquez Arreola, (próximo Doctor), por compartirme sus conocimientos, por dedicarme su tiempo y por tener la paciencia para trabajar conmigo, sin ti no lo hubiera logrado, gracias mil.

Agradezco también a mis compañeros de maestría, porque hoy en día ya somos más que compañeros, somos buenos amigos, en especial al equipo de trabajo que hemos formado con Edgar platas, Nancy López y Estefanía Ruíz, también a Gloria Conde y Alejandra López, gracias.

Y, por último, pero no menos importante agradezco al Instituto Nacional de Astrofísica, Óptica y Electrónica por aceptarme en el posgrado, gracias.

Dedicatorias

Dedico este y todos los proyectos de mi vida a Dios.

Y en especial, dedico este trabajo al amor de mi vida, al motor que me impulsa a seguir en esta tierra, a Mateo de Jesús, te amo hijo. Y a Matías también, porque es un ángel que me prestará sus alas para volar hacia él.

No podría dejar de mencionar a mi madre, que, aunque ya no está con nosotros físicamente su recuerdo nunca se aleja de mí, “te amo má”.

También dedico este trabajo de tesis a mis hermanos y a toda mi familia, tanto de sangre como la que me ha adoptado, familia “Tecuatl”, gracias a Marina y Elenita por ser mis compañeras de vida y por estar conmigo, no quisiera poner más nombres porque no terminaría de mencionar a tantas personas, pero sé que conocen mis sentimientos, saben que están en mi corazón y siento su apoyo siempre y en todo momento.

En fin, a todos los que me han brindado la oportunidad de conocerlos gracias, por todas esas vivencias y experiencias. A mis amigos de toda la vida, ellos saben quiénes son, gracias; porque a pesar de los años siempre seguimos en contacto, y cuando nos necesitamos siempre nos hacemos presentes, mil gracias a todos. También a los “nuevos” amigos, los que empiezo a conocer, a mis equipos de futbol, y a todos los que se van agregando a mi vida, porque me demuestran que siempre hay espacio para alguien más.

Gracias.

“Cuando algo es lo suficientemente importante, hazlo. Incluso cuando todo esté en tu contra”.

Elon Musk.

Índice de figuras

Pág.

Figura 1. Los datos sintéticos se convertirán en la principal forma de datos utilizados por la inteligencia artificial 5

Figura 2. La sangre y sus componentes [7] 11

Figura 3. Anemia microcítica [10] 13

Figura 4. Anemia macrocítica [11] 13

Figura 5. Anemia normocítica [12] 14

Figura 6. Variabilidad morfológica asociada a las células blásticas según la clasificación FAB: (a) célula sana, (b-d) células enfermas, donde (b), (c) y (d) son L1, L2 y L3 respectivamente. 19

Figura 7. Trombocitopenia [14] 20

Figura 8. (Izq.) Célula normal, (Der) célula enferma [2] 21

Figura 9. Estructura de la inteligencia artificial [4] 23

Figura 10. Estructura de una neurona biológica [15] 28

Figura 11. Estructura de una neurona artificial [15] 28

Figura 12. Ejemplo de red con 4 capas [15] 29

Figura 13. Estructura de una Red Neuronal Profunda [15] 30

Figura 14. Estructura de una Red Neuronal Convolucional [15] 31

Figura 15. Estructura de una Red Neuronal Recurrente [15] 31

Figura 16. Ejemplo de red con 4 capas para identificar una imagen [15] 32

Figura 17. Metodología para la generación de datos sintéticos 34

Figura 18. Imágenes sintéticas tradicionales de células enfermas 36

Figura 19. Región de interés (ROI) 37

Figura 20. Imágenes sintéticas poligonales de células enfermas 37

Figura 21. Imágenes sintéticas poligonales de células enfermas 38

Figura 22. Imágenes sintéticas poligonales de células enfermas 39

Figura 23. Imágenes sintéticas poligonales de células enfermas 40

Figura 24. Arquitectura del modelo Compact [16] 41

Figura 25. Arquitectura del modelo Enhanced [16] 41

Figura 26. Arquitectura del modelo ResNet [16] 42

Figura 27. Arquitectura del modelo MobileNet V2 [16] 42

Figura 28. Arquitectura del modelo AlexNet [16] 43

Figura 29. Ejemplos de las imágenes contenidas en ALL-IDB2:las células de la A a la D están etiquetadas como células sanas, las células de la E a la H están etiquetadas como probables linfoblastos o células enfermas. 49

Figura 30. . Variabilidad morfológica asociada a las células blásticas según la clasificación FAB: (a) célula sana, (b-d) células enfermas, donde (b), (c) y (d) son L1, L2 y L3 respectivamente. 51

Figura 31. Ejemplos de imágenes segmentadas. (izq.); célula enferma, (der.); célula sana 52

Figura 32. Esquema de la metodología 53

Figura 33. Fórmula para obtener la métrica *F1-Score* 57

Figura 34. Fórmula para obtener la métrica *Mean Precision* 57

Figura 35. Fórmula para obtener la métrica *Mean Recall* 58

Figura 36. Ejemplos de imágenes etiquetadas como células enfermas. 61

Figura 37. Creación de los parámetros del Split. 62

Figura 38. Parámetros e hiperparámetros que se asignaron para realizar el entrenamiento. 62

Figura 39. Función de pérdida 63

Figura 40. Matriz de confusión 64

Figura 41. Mapa de calor 64

Figura 42. Número de imágenes que utilizó el método para realizar el entrenamiento, la validación y la prueba. 65

Figura 43. Parámetros utilizados para el entrenamiento 65

Figura 44. Falsos negativos del entrenamiento 66

Figura 45. Resultados de las métricas de evaluación 66

Figura 46. Gráfica del resultado de la precisión alcanzada con el entrenamiento de las imágenes sintéticas tradicionales 68

Figura 47. Gráfica del resultado de la precisión alcanzada con el entrenamiento de las imágenes sintéticas poligonales 69

Figura 48. Gráfica comparativa de la precisión de los 5 modelos 69

Índice de tablas

Tabla 1. Clasificación morfológica de la LLA según la FAB 22

Tabla 2. Procedimiento de imágenes poligonales 36

Tabla 3. Tabla comparativa de trabajos relacionados 47

Tabla 4. Conjunto de datos 53

Tabla 5. Precisión obtenida por el entrenamiento de las imágenes sintéticas tradicionales 67

Tabla 6. Precisión obtenida por el entrenamiento de las imágenes sintéticas poligonales 68

ÍNDICE GENERAL

Pág.

Resumen II

Abstract III

Agradecimientos IV

Dedicatorias V

Índice de figuras VI

Índice de tablas VII

CAPÍTULO 1. INTRODUCCIÓN 1

1.1 Motivación 3

1.2 Antecedentes 4

1.3 Planteamiento del problema 5

1.4 Justificación 6

1.5 Hipótesis 6

1.6 Preguntas de investigación 6

1.7 Objetivos 7

1.7.1 Objetivo general 7

1.7.2 Objetivos específicos 7

1.8 Guía del documento 7

CAPÍTULO 2. MARCO TEÓRICO 9

2.1 Conceptos hematológicos generales 9

2.1.1 Componentes de la sangre 10

2.2 Enfermedades de la sangre 11
2.2.1 Anemia 12
2.2.1.1 Clasificación de la anemia 12
2.2.2 Leucemia 15
2.2.2.1 Tipos de leucemias 17
2.2.2.2 Características morfológicas de las células leucémicas 17
2.2.3 Trombocitopenia 18
2.2.3.1 Clasificación de los trastornos plaquetarios 18
2.3 Diagnóstico por medio de un examen 20
2.4 Clasificación FAB 21
2.5 Inteligencia Artificial (IA) 22
2.6 Evolución de la Inteligencia Artificial (IA) 22
2.7 Aprendizaje Automático (Machine Learning) 23
2.7.1 Entrenamiento 24
2.7.1.1 Aprendizaje supervisado 24
2.7.1.2 Aprendizaje no supervisado 25
2.7.1.3 Aprendizaje por refuerzo 25
2.7.2. Predicción 26
2.8 Aprendizaje Profundo (*Deep Learning*) 26
2.9 Redes neuronales 27
2.10 Tipos de redes neuronales 29
2.10.1 Red Neuronal Profunda (DNN) 29
2.10.2 Red Neuronal Convolucional (CNN) 30

2.10.3 Red Neuronal Recurrente (CNN) 30
2.11 ¿Cómo se entrena una red neuronal? 31
2.12 Generación de imágenes sintéticas *(Data Augmentation)* 32
2.12.1 Algoritmo para la generación de Imágenes sintéticas tradicionales 34
2.12.2 Algoritmo propuesto para la generación de imágenes sintéticas poligonales 35
2.13 Modelos utilizados 40

CAPÍTULO 3. REVISIÓN DE TRABAJOS PREVIOS 44

CAPÍTULO 4. METODOLOGIA 48
4.1 Base de datos 51
4.2 Esquema general de la metodología 52
4.3 Generación de imágenes sintéticas tradicionales 54
4.4 Generación de imágenes sintéticas poligonales 54
4.5 Entrenamiento y prueba de los modelos de aprendizaje 55

CAPÍTULO 5. ANÁLISIS DE RESULTADOS DEL TRABAJO DESARROLLADO 59
5.1 Imágenes sintéticas tradicionales 59
5.2 Imágenes sintéticas poligonales 59
5.3 Análisis de los modelos de aprendizaje 60

5.3.1 Fase de entrenamiento 60

5.4 Interpretación de los resultados del entrenamiento 67

5.5 Comparación de los resultados 69

5.6 Análisis 70

CAPÍTULO 6. CONCLUSIONES 71

6.1 Conclusión 71

6.2 Trabajo futuro 72

REFERENCIAS 73

ANEXOS

A. Ejemplos de imágenes sintéticas tradicionales 76

B. Ejemplos de imágenes sintéticas poligonales 77

C. Glosario 78

CAPÍTULO 1. INTRODUCCIÓN

El aumento de datos (Data Augmentation) es una técnica que adopta el enfoque de generar más datos de entrenamiento a partir de los datos disponibles. Cuando existen datos muy limitados para entrenar, existen menos posibilidades de obtener predicciones precisas para los modelos. En el caso de imágenes, esto se consigue aplicando una serie de transformaciones aleatorias a la imagen que producen nuevas imágenes. El objetivo es que, en el momento del entrenamiento, los modelos nunca verán exactamente la misma imagen [1]. Esta idea simple, pero potente, ayuda a exponer el modelo a más aspectos de los datos y a generalizar mejor.

En otras palabras, la técnica de aumento de datos es una forma eficaz de mejorar el rendimiento de los modelos con datos de imágenes, ya que aplica transformaciones sencillas para generar nuevas imágenes. Sin embargo, es importante tener precaución al elegir las técnicas específicas de aumento de datos, considerando el contexto del conjunto de datos de entrenamiento y el conocimiento del dominio del problema [3]. De esta manera, se evita generar imágenes que nunca podrían ocurrir en la realidad, lo cual podría empeorar el entrenamiento del modelo.

El aumento de datos es una técnica que consiste en aplicar una serie de transformaciones a las imágenes existentes para generar nuevos datos. Estas transformaciones se realizan mediante un procesamiento previo y se añaden al conjunto de datos original, lo que permite un preprocesamiento más rápido y evita tener que realizar las transformaciones en tiempo de ejecución [15]. El objetivo del aumento de datos es mejorar la calidad o hacer más representativo un conjunto de datos con relación al mundo real. Por ejemplo, añadiendo imágenes de perros a un conjunto de datos que sólo contiene imágenes de gatos, se hace más representativo para entrenar un modelo de aprendizaje profundo, ya que incluye una variedad de animales.

Además, el aumento de datos es una técnica comúnmente utilizada en el entrenamiento de modelos de aprendizaje profundo, ya que ayuda a mejorar la precisión de estos modelos. También es útil en situaciones en las que hay una cantidad limitada de datos disponibles. El aumento de datos consiste en aumentar la cantidad y diversidad de los datos existentes mediante la aplicación de transformaciones, en lugar de recopilar nuevos datos. Esto permite un preprocesamiento más rápido y ayuda al modelo a generalizar mejor lo que también puede ayudar a reducir el sobreajuste [15]. Por lo tanto, se considera una técnica viable para mitigar el sobreajuste en los modelos de redes neuronales.

El aumento de datos es un proceso esencial en el aprendizaje profundo debido a la necesidad de grandes cantidades de datos para entrenar los modelos. En algunos casos, recopilar miles o millones de imágenes puede ser difícil, por lo que el aumento de datos se convierte en una técnica valiosa para aumentar el tamaño y diversidad del conjunto de datos existente. Las operaciones de transformación más utilizadas para realizar el aumento de datos son: rotación, espejo, ruido sal y pimienta, contraste, brillo, iluminación, polígonos, mascaras, entre otras.

Por otra parte, una enfermedad hematológica es un trastorno que afecta a algún componente de la sangre, lo que impide su correcto funcionamiento y puede causar problemas en otros órganos y tejidos del cuerpo. La sangre está compuesta por una parte líquida, el plasma, que contiene agua, sales y proteínas y permite su flujo por los vasos sanguíneos, y una parte sólida compuesta por células sanguíneas: glóbulos rojos (que transportan oxígeno), glóbulos blancos (del sistema inmune) y plaquetas (que coagulan la sangre en caso de herida). Estas enfermedades pueden ser causadas por factores genéticos, déficit en la dieta, problemas de absorción de nutrientes, déficit de vitaminas o problemas respiratorios o alérgicos [2].

Cualquiera de los componentes de la sangre es posible que no estén en condiciones óptimas, que pueden ser causados por errores genéticos, déficit de minerales esenciales como el hierro, problemas de absorción de vitaminas y nutrientes, déficit de vitaminas como la B12, anticuerpos contra las células sanguíneas del propio cuerpo o problemas respiratorios o alergias [2]. Estos factores impiden que la sangre funcione adecuadamente, dando lugar a una enfermedad hematológica.

En este trabajo de tesis se realizaron experimentos con imágenes de células leucémicas, ya que esta enfermedad hematológica es una de las más comunes en la infancia en México, representando el 80% de todas las leucemias agudas en edad pediátrica. Se mencionan también otras enfermedades hematológicas comunes como anemia y trombocitopenia.

1.1 Motivación

Los programadores necesitan conjuntos de datos grandes y cuidadosamente etiquetados para entrenar redes neuronales, ya que los datos de entrenamiento más diversos generalmente hacen que los modelos de IA (Inteligencia Artificial) sean más precisos [3]. Sin embargo, recopilar y etiquetar estos conjuntos de datos puede ser un desafío, ya que pueden contener miles o incluso millones de elementos y puede ser altamente costoso.

Es aquí donde entra en juego el aumento de datos, ya que permite generar datos artificialmente, reduciendo los costos y garantizando una diversidad de datos que represente el mundo real [4]. Además, ya que los datos sintéticos se etiquetan automáticamente, a veces son incluso mejores que los datos del mundo real. Es importante destacar que los datos sintéticos son clave para lidiar con los problemas de privacidad y reducir el sesgo[7].

1.2 Antecedentes

Los datos sintéticos son una herramienta valiosa en la resolución de problemas, ya sea en juegos de computadora, simulaciones científicas o estudios estadísticos. El término fue acuñado por el profesor de estadística Donald B. Rubin en 1993, en el contexto de ayudar a las ramas gubernamentales a resolver problemas de recuento insuficiente en censo de personas pobres [28]. Los datos sintéticos son conjuntos de datos simulados que parecen haber sido generados por el mismo proceso que generó el conjunto de datos real, pero sin revelar datos reales, lo que los hace ideales para estudiar conjuntos de datos personales y confidenciales.

La competencia ImageNet de 2012 [29], en la que una red neuronal reconoció objetos más rápido de lo que un humano podría, fue el detonante para que los investigadores comenzaran a buscar datos sintéticos con más interés. En poco tiempo, dio inicio la utilización de imágenes renderizadas en experimentos, lo cual dio resultados suficientemente prometedores para que empresas como NVIDIA invirtieran en productos y herramientas para generar datos sintéticos a través de motores 3D y pipelines de contenido. Hoy en día, datos sintéticos son utilizados en una amplia variedad de industrias, incluyendo bancos, fabricantes de automóviles, drones, fábricas, hospitales, y robots científicos [28].

Los investigadores de Ford recientemente han descubierto cómo combinar motores de juego y redes generativas adversarias (GANs) para generar datos sintéticos para el entrenamiento de IA. Por su parte, BMW ha creado una fábrica virtual utilizando NVIDIA Omniverse, una plataforma de simulación que permite a las empresas colaborar con diferentes herramientas. Los datos generados por BMW ayudan a mejorar la eficiencia en el proceso de fabricación de automóviles, a través del trabajo en equipo entre trabajadores y robots [29].

La investigación actual [3] destaca el uso de datos sintéticos como una de las técnicas más prometedoras en el aprendizaje profundo moderno, especialmente en el campo de la visión por computadora, ya que permite trabajar con datos no estructurados como imágenes y video. La creciente importancia de los datos sintéticos se ve reflejada en la perspectiva de líderes en IA como Andrew Ng [3], quien está abogando por un enfoque más centrado en los datos en el aprendizaje profundo y promoviendo una competencia para evaluar la calidad de los datos, ya que se considera que representa el 80% del trabajo en IA.

1.3 Planteamiento del problema

Uno de los desafíos fundamentales en el aprendizaje automático es el acceso a conjuntos de datos de alta calidad y suficiente cantidad. El rendimiento de los modelos de aprendizaje profundo está estrechamente relacionado con la cantidad y calidad de los datos de entrenamiento [1]. En este sentido, la generación de datos sintéticos se presenta como una alternativa para superar la escasez de datos y permitir la viabilidad de proyectos que requieren grandes conjuntos de datos, como vemos en la figura 1.

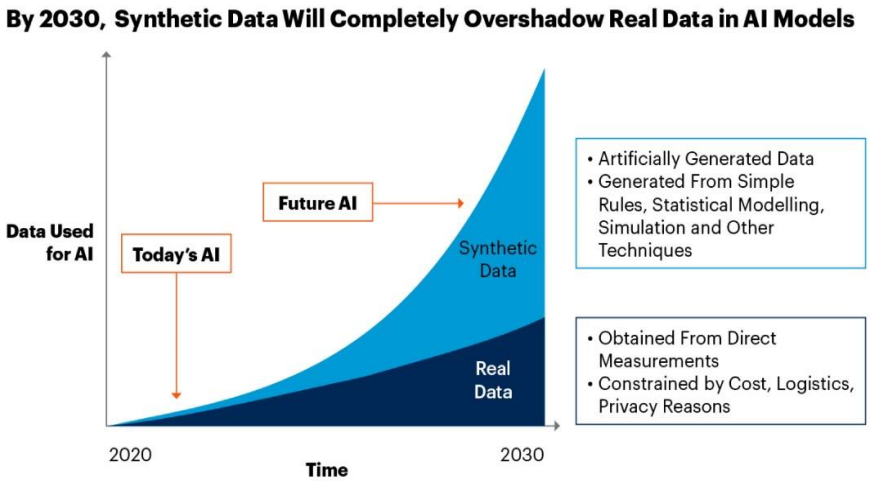


Figura 1. Los datos sintéticos se convertirán en la principal forma de datos utilizados por la inteligencia artificial [1].

1.4 Justificación

El rendimiento de la mayoría de los modelos de aprendizaje profundo depende de la calidad, cantidad y relevancia de los datos de entrenamiento [22]. Sin embargo, la insuficiencia de datos es uno de los desafíos más comunes en la implementación del aprendizaje profundo, un ejemplo de esto se puede ver cuando se trabaja con imágenes médicas de algunas enfermedades tales como anemia, retinopatía leucémica, tomosíntesis. Por lo que, se puede aprovechar el aumento de datos para reducir la dependencia de la recopilación y preparación de datos y para crear modelos de aprendizaje profundo más precisos.

1.5 Hipótesis

Utilizar imágenes sintéticas, generadas por medio de algoritmos para aumento de datos ayuda a mejorar el desempeño de los modelos de entrenamiento, en particular; para el problema de la clasificación de imágenes médicas, el caso de estudio para el desarrollo de este trabajo considera imágenes de células sanguíneas.

1.6 Preguntas de investigación

¿Se puede aumentar el desempeño de los modelos de Deep Learning utilizando imágenes sintéticas en el área hematológica?

¿En qué porcentaje se mejora el desempeño de los modelos con el uso de imágenes sintéticas?

¿Cuál es el número máximo de imágenes sintéticas para lograr un desempeño óptimo?

1.7 Objetivos

1.7.1 Objetivo general

Evaluar algoritmos de aumento de datos para la generación de imágenes sintéticas, con el fin de mejorar el desempeño de los modelos de aprendizaje profundo.

1.7.2 Objetivos específicos

- Investigar los métodos más utilizados para la generación de imágenes sintéticas por medio del aumento de datos.
- Generar un conjunto de datos de imágenes sintéticas, con técnicas de aumento de datos, por medio de transformaciones a las imágenes originales las cuales serán tomadas del conjunto de datos ALL-IDB2.
- Utilizar las imágenes sintéticas obtenidas para realizar el entrenamiento de los modelos.
- Evaluar los resultados contrastando las imágenes originales con las imágenes sintéticas.

1.8 Guía del documento

Los capítulos de este documento están organizados de la siguiente manera:

El capítulo 1 contiene los antecedentes, el planteamiento del problema, la justificación, así como los objetivos de este trabajo de tesis. En el capítulo 2 se describen los conceptos importantes sobre la hematología, componentes de la sangre y sus enfermedades, así como inteligencia artificial, redes neuronales y aprendizaje profundo, aumento de datos, entre otros. En el capítulo 3 se exponen los trabajos relacionados para este proyecto de investigación, cómo aprendizaje profundo, aumento de datos, y las principales enfermedades de la sangre. El capítulo 4 detalla la metodología de este proyecto de tesis, es decir, el proceso que se siguió

desde la adquisición de los datos hasta la fase de entrenamiento y prueba del algoritmo. En el capítulo 5 se presentan los resultados y los análisis de los resultados del trabajo realizado. Finalmente, en el capítulo 6 se muestran las conclusiones y el trabajo a futuro que se planea realizar.

CAPÍTULO 2. MARCO TEÓRICO

En este capítulo se abordan conceptos hematológicos generales para el desarrollo de este proyecto. Se describe también la leucemia linfoblástica aguda y sus conceptos generales, así como también se explica la implementación de los algoritmos para la generación del aumento de datos y la obtención de imágenes sintéticas describiendo los conceptos significativos relacionados con el aprendizaje profundo.

2.1 Conceptos hematológicos generales

La hematología es una rama de la medicina que estudia la morfología de la sangre y los tejidos que la producen, permite generar diagnósticos, y trata las enfermedades de la sangre y de sus componentes celulares. Cubre la composición celular y sérica de la sangre, el proceso de coagulación, la formación de células sanguíneas, la síntesis de la hemoglobina y todos los trastornos relacionados [4].

Por otro lado, se encarga de estudiar los hematíes, leucocitos y plaquetas, así como analizar sus proporciones relativas, su estado general y las enfermedades provocadas por los desequilibrios entre ellas.

Entre las enfermedades más comunes derivadas del desequilibrio celular, tenemos: la anemia, la cual consiste en la falta de hematíes, un trastorno provocado por diversos factores; la leucemia, es considerada como una enfermedad en la que la médula ósea produce demasiados leucocitos anormales, de modo que reemplazan la eritropoyesis y la trombopoyesis y provocan síntomas peligrosos; finalmente la trombocitopenia, consiste en una enfermedad adquirida grave o el síntoma de una enfermedad subyacente.

2.1.1 Componentes de la sangre

La sangre es el líquido que mantiene la vida y circula a través del corazón, las arterias, las venas, los capilares, el cerebro y el resto del cuerpo. La sangre se encarga de transportar nutrientes, electrólitos, hormonas, vitaminas, anticuerpos, calor, oxígeno y células inmunológicas hacia los tejidos del cuerpo, también transporta desperdicios y dióxido de carbono.

La sangre está constituida por una parte líquida y una parte sólida. La parte líquida se denomina plasma, la cual está compuesta por agua, sales y proteínas. Por otro lado, la parte sólida está formada por el grupo de células sanguíneas de los glóbulos rojos, glóbulos blancos y plaquetas [7].

Los Glóbulos rojos o eritrocitos se encargan de transportar oxígeno desde los pulmones al resto del cuerpo, a través de una proteína denominada hemoglobina (Hb), y también se encarga de liminar de los tejidos el dióxido de carbono como sustancia residual, para redirigirlos a los pulmones.

Los Glóbulos blancos o leucocitos contribuyen a combatir infecciones y asisten al proceso inmunológico. Existen varios tipos de glóbulos blancos y cada uno cumple un papel distinto en el combate contra infecciones bacterianas, virales, fúngicas y parasitarias, estos glóbulos son: linfocitos, monocitos, eosinófilos, basófilos y neutrófilos.

Finalmente, se tienen a las plaquetas o trombocitos, y su función principal es la de colaborar en la coagulación sanguínea. Las plaquetas tienen un tamaño más pequeño que el resto de las células y se agrupan para formar una acumulación, o tapón, en el orificio de un vaso sanguíneo para detener las hemorragias.

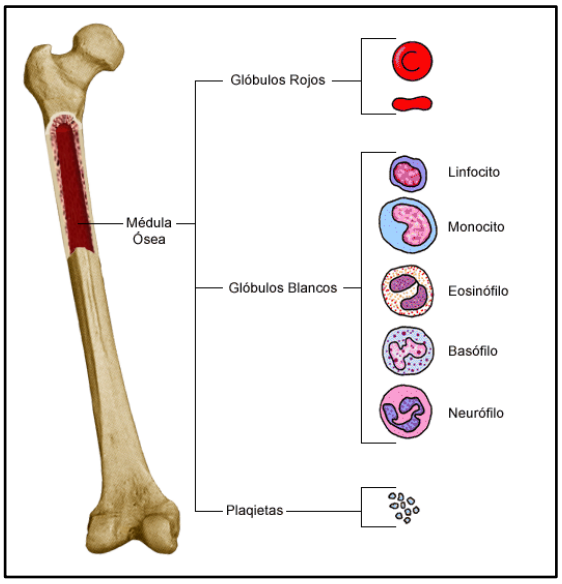


Figura 2. La sangre y sus componentes [7].

Las células sanguíneas o glóbulos rojos, glóbulos blancos y plaquetas se producen en la médula ósea, la cual es un material esponjoso que se localiza en el centro de los huesos. Las células sanguíneas que se producen en la médula ósea se forman como células madre. Una célula madre (o célula hematopoyética) es la primera etapa de todas las células sanguíneas, a medida que las células madre van madurando, se crean distintas células. Las células que no llegan a madurar se denominan blastos.

2.2 Enfermedades de la sangre

Las funciones de la sangre vienen determinadas por las células que viajan en ella. Los glóbulos rojos, los blancos y las plaquetas pueden generar una serie de problemas de salud (enfermedades hematológicas) cuando su calidad o su cantidad se encuentran alteradas.

Así, pueden predisponer a situaciones de mayor viscosidad y peligro de trombos, mayor vulnerabilidad para infecciones o un mayor riesgo de sangrados incontrolados. Los análisis de sangre ayudarán a determinar estos

trastornos sanguíneos. Algunas enfermedades de la sangre pueden ser: anemia, leucemia, trombocitopenia, talasemia, hemofilia, leucopenia, hemocromatosis y trombosis venosa.

Para los propósitos de esta investigación, se va a realizar el análisis de imágenes a un conjunto de datos de leucemia linfoblástica aguda.

2.2.1 Anemia

La anemia se define como una reducción de la concentración de la hemoglobina o de la masa global de hematíes en la sangre periférica por debajo de los niveles considerados normales para una determinada edad, sexo y altura sobre el nivel del mar [8].

Existen también factores fisiológicos que alteran los valores normales de hemoglobina, como son la altitud, la gestación, la raza y ser fumador [9].

2.2.1.1 Clasificación de la Anemia

Las anemias pueden clasificarse según sus criterios morfológicos o por su velocidad de propagación.

En cuanto a la clasificación morfológica se tienen tres categorías generales: anemia microcítica, macrocítica y normocítica.

La anemia microcítica o anemia ferropénica, el cual, se caracteriza por la deficiencia de hierro. En este tipo de padecimiento los pacientes cuentan con glóbulos rojos más pequeños de lo normal [10].

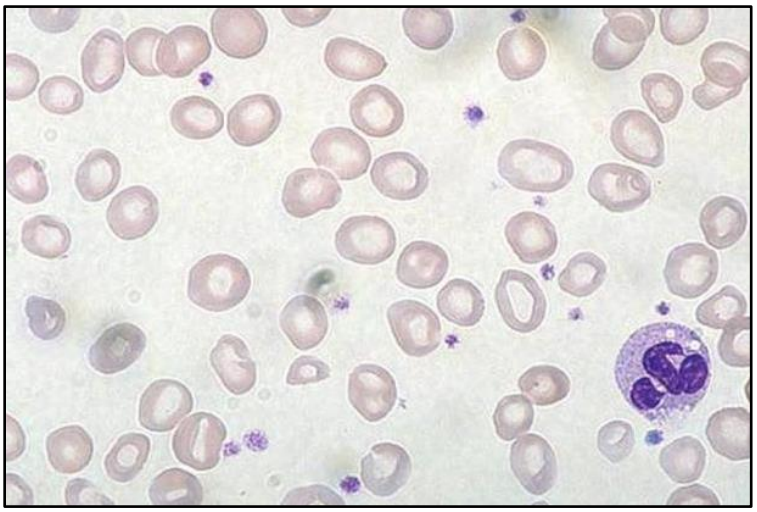


Figura 3. Anemia microcítica [10].

La anemia macrocítica que está relacionada con deficiencia de ácido fólico o vitamina B12. En la siguiente figura se presenta una muestra de frotis de Anemia macrocítica, se puede observar la morfología redonda de los eritrocitos, con exceso de membrana de los eritrocitos y presencia de células diana [11].

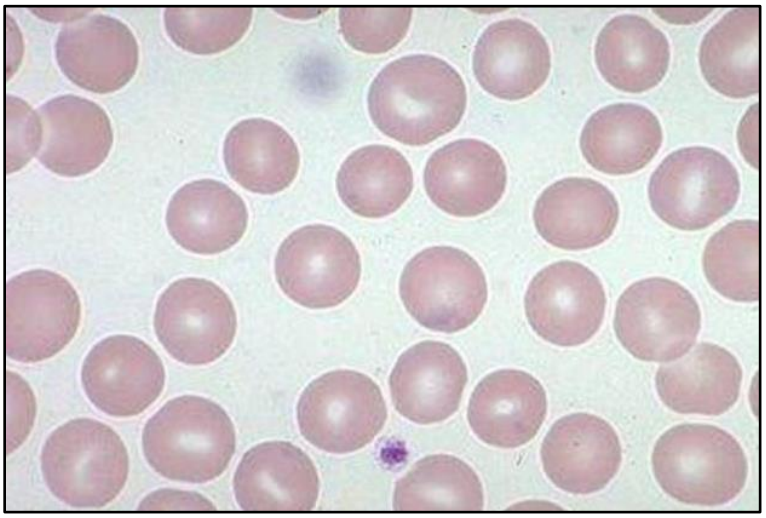


Figura 4. Anemia macrocítica [11].

Finalmente, la anemia normocítica es un padecimiento que se caracteriza por la ausencia de glóbulos rojos, si bien la apariencia y tamaño de los eritrocitos son normales, la cantidad es inferior a la necesaria. Este tipo de anemia está asociada a una gran variedad de trastornos, como son: hepatopatías, insuficiencia renal, enfermedades autoinmunes, neoplasias, endocrinopatías o infecciones crónicas [12].

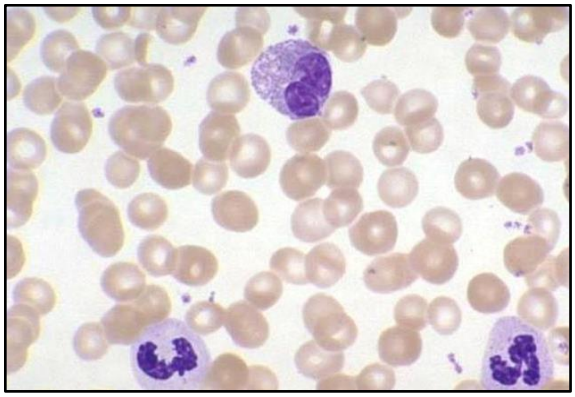


Figura 5. Anemia normocítica [12].

Con respecto a la velocidad de propagación, se divide en anemia aguda y anemia crónica.

En la anemia aguda los valores de hemoglobina y hematíes descienden por debajo de los niveles normales. Este tipo de anemia se presenta en dos situaciones: por hemorragias y por un aumento en la desintegración de los hematíes.

La anemia crónica se propaga de manera lenta y progresiva, provocando diversas enfermedades como insuficiencia en la generación de hematíes por la médula ósea o limitación en la síntesis de la hemoglobina de carácter hereditario o adquirido [5].

2.2.2 Leucemia

Las leucemias son un tipo de cáncer, de origen desconocido en la mayoría de los casos, que afecta a las células sanguíneas, generalmente a los glóbulos blancos. La enfermedad se produce a consecuencia de un error en el proceso de maduración de una célula madre a glóbulo blanco, que supone una alteración cromosómica que provoca que las células afectadas se vuelvan cancerosas y se multipliquen sin cesar, infiltrándose en la médula ósea, donde sustituyen a las células que producen las células sanguíneas normales [8].

Estas células cancerosas se diseminan por la sangre, y además pueden invadir otros órganos, como el hígado, los riñones, los ganglios linfáticos, el bazo y el cerebro.

A medida que la enfermedad progresa, las células malignas interfieren en la producción de otro tipo de células sanguíneas, como los glóbulos rojos y las plaquetas, lo que tiene como consecuencia el desarrollo de anemia.

Las leucemias tienen una incidencia aproximada de dos o tres casos por cada 100.000 habitantes. Son las neoplasias más frecuentes en la infancia (alrededor del 25% de los cánceres infantiles son leucemias), y afectan con más frecuencia a los varones.

No parece haber diferencias sustanciales en la prevalencia de leucemia entre las distintas razas o áreas geográficas, el entorno rural o urbano, ni entre las distintas clases sociales.

Sin embargo, dependiendo del tipo de leucemia, es más frecuente su aparición a determinadas edades. Por ejemplo, en el caso de la leucemia linfoblástica aguda (linfocítica), suele presentarse en niños de entre tres y cinco años, y aunque también afecta a los adolescentes, es poco común.

Aunque la causa de las leucemias no se conoce exactamente, se sabe que hay diversos factores que pueden provocar la aparición de esta enfermedad.

- Genéticos
- Inmunodeficiencias
- Factores ambientales

En cuanto a la relación de los factores genéticos con el desarrollo de leucemia, se sabe que la enfermedad es más frecuente en gemelos que en el resto de la población, y padecer trastornos genéticos como el síndrome de *Down* y el síndrome de Fanconi supone un factor de riesgo asociado a la aparición de leucemia. Las personas con el sistema inmunitario debilitado por la administración de quimioterapia o fármacos inmunosupresores (que se suministran a pacientes que han sufrido un trasplante de órganos), también son más susceptibles de desarrollar leucemia.

Uno de los factores más estudiados son los factores ambientales, sobre todo la exposición a radiaciones ionizantes, sustancias químicas como el benceno y ciertos fármacos, y los virus.

La relación entre las radiaciones ionizantes y la leucemia se descubrió a partir de accidentes nucleares (explosiones o incidentes en centrales nucleares). Diversos productos químicos también están relacionados con la aparición de la enfermedad, sobre todo algunos pesticidas, y otras sustancias como los gases mostaza utilizados en la I Guerra Mundial.

También ciertos virus están asociados con el desarrollo de las leucemias, en especial el virus de Epstein-Barr, relacionado con el linfoma de Burkitt africano o los linfomas en pacientes inmunodeprimidos.

2.2.2.1 Tipos de leucemias

Existen varios criterios para clasificar las leucemias. Una forma de clasificación se basa en su historia natural:

- De novo: cuando ocurren sin que exista un proceso previo que desencadene la enfermedad.

- Secundarias: cuando existe un proceso previo que desemboca en leucemia, como por ejemplo una enfermedad sanguínea.

Otra forma de clasificarlas se basa en el tipo de célula sanguínea en la que empieza la transformación maligna, y en la velocidad de progresión de la enfermedad. En el caso de las leucemias agudas, su desarrollo es muy rápido, mientras que las leucemias crónicas progresan lentamente [21]. Además, las leucemias pueden ser:

1. Linfoblásticas o linfocíticas: cuando afectan a los linfocitos (variedad de leucocitos de la médula ósea). Dan lugar a linfocitos T (células T), linfocitos B (células B) o linfocitos citolíticos naturales (NK).

2. Mieloblásticas o mielocíticas: que afectan a la célula precursora de la serie mieloide o serie roja (de los glóbulos rojos y plaquetas).

2.2.2.2 Características morfológicas de las células leucémicas

La clasificación de los linfocitos en las imágenes es bastante compleja, ya que incluso un operador experto puede tener problemas para clasificar algunas células de linfocitos viéndolos directamente por el microscopio. En realidad, las diferencias morfológicas entre las imágenes blásticas y los linfocitos normales son muy pequeñas.

De acuerdo con el análisis morfológico visual más común para la enfermedad de LLA (clasificación FAB) [30], las características que los técnicos de laboratorio capacitados consideran durante la observación de imágenes con células enfermas son las mencionadas en el apartado 2.4.

La siguiente figura muestra la gran variabilidad morfológica de las células blásticas según la clasificación FAB para la enfermedad de LLA. El objetivo principal es detectar sin diferenciación la presencia de los tres subtipos de leucemias linfoblásticas en las imágenes.

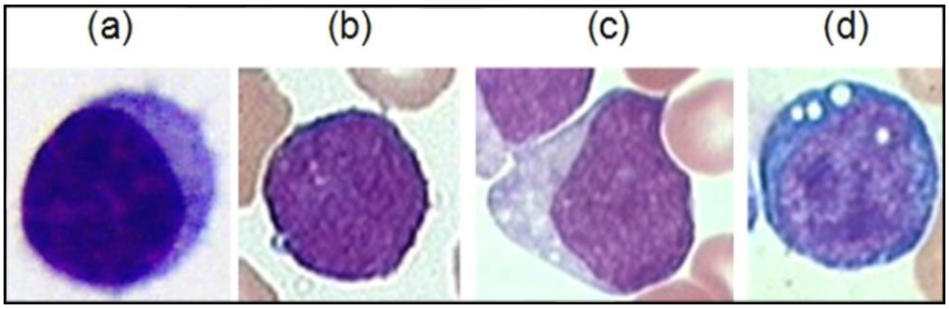


Figura 6. Variabilidad morfológica asociada a las células blásticas según la clasificación FAB: (a) célula sana, (b-d) células enfermas, donde (b), (c) y (d) son L1, L2 y L3 respectivamente [30].

2.2.3 Trombocitopenia

Las plaquetas (trombocitos) son células sanguíneas diminutas e incoloras que se forman en la medula ósea y tienen como función principal intervenir en la coagulación de la sangre, esto es, cuando se sufre una lesión, las plaquetas se agrupan para sellar la herida, también se denomina coágulo de sangre [13].

2.2.3.1 Clasificación de los trastornos plaquetarios

Los trastornos plaquetarios son padecimientos que aparecen cuando el recuento de plaquetas en la sangre es muy alto o bajo.

Un recuento de plaquetas alto se llama trombocitosis o trombocitemia, si el recuento de plaquetas es más bajo de lo normal se llama trombocitopenia.

La trombocitemia está relacionada con un alto número de plaquetas y suele llamarse trombocitemia primaria o esencial. Por otro lado, la trombocitosis es un recuento alto de plaquetas el cual se debe a otra afección de salud, suele denominarse trombocitosis secundaria o reactiva. La trombocitosis es más frecuente que la trombocitemia [13].

La trombocitopenia es una afección que aparece cuando el recuento de plaquetas de la sangre es demasiado bajo. En un adulto sano, podemos encontrar entre 150,000 y 450,000 plaquetas por microlitro (µl) de sangre. Se estipula que se está en presencia de una trombocitopenia cuando la cantidad de plaquetas circulantes en sangre es menor a 100.000 unidades por µl.

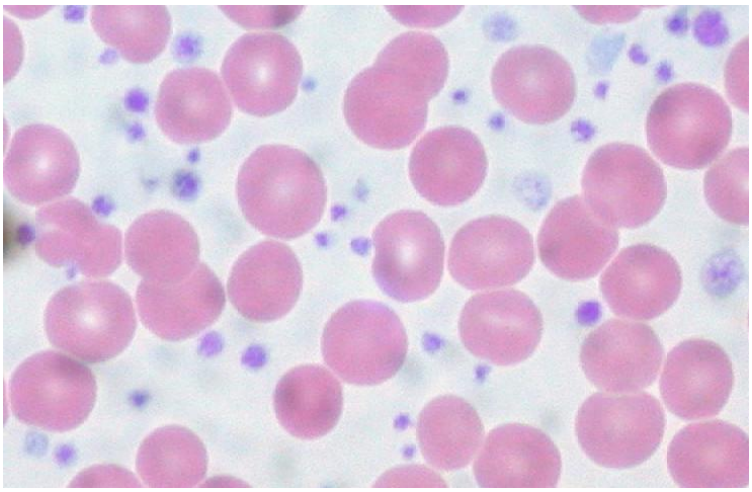


Figura 7. Trombocitopenia [13].

La trombocitopenia puede presentarse a partir de un trastorno en la médula ósea, como la leucemia o un problema del sistema inmunitario. Afecta tanto a niños como a adultos. Algunos tipos de trombocitopenia son la trombocitopenia inmunitaria y la púrpura trombótica trombocitopénica.

2.3 Diagnóstico por medio de un examen

La hematología identifica dichos desequilibrios. Una de las pruebas de laboratorio más importantes, es el hemograma completo, un análisis de sangre con un recuento y análisis de los diferentes tipos de células que forman la sangre. Un hemograma puede contribuir al diagnóstico de estos trastornos para facilitar la prescripción de tratamientos adecuados. Cuando se examina una muestra de sangre del paciente al microscopio, se observan glóbulos rojos, glóbulos blancos muy inmaduros (blastos) y plaquetas. Una biopsia de médula ósea servirá para confirmar el diagnóstico y determinar qué tipo de enfermedad sanguínea sufre el paciente [2].

Los trastornos bioquímicos que se registran con mayor frecuencia son hiperuricemia (ácido úrico elevado), hipocalcemia (calcio bajo), hiperfosfatemia (fósforo elevado), hiperpotasemia (potasio elevado) e incremento de la actividad sérica de la láctico-deshidrogenasa (LDH). Estas alteraciones se observan sobre todo en los casos con leucocitosis, grandes visceromegalias o adenopatías, y reflejan el elevado recambio celular. En algunos enfermos se detecta hipogammaglobulinemia (bajo nivel de gammaglobulinas). Las pruebas médicas realizadas para detectar enfermedades sanguíneas son: biopsia de médula ósea, análisis de sangre, citometría de flujo, frotis de sangre periférica, punción lumbar, tomografía computarizada, resonancia magnética, ecografía, entre otras.

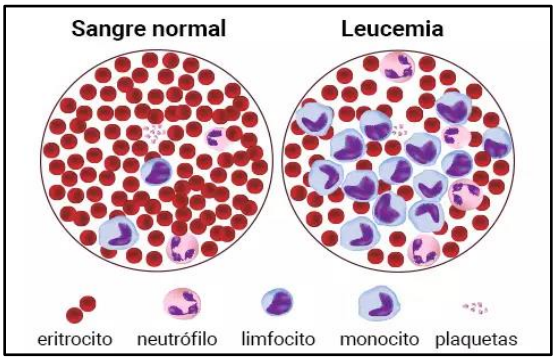


Figura 8. (Izq.) Célula normal, (Der) célula enferma [2].

2.4 Clasificación FAB

La clasificación FAB (Francesa- Americana- Británica) [30], fue creada por un grupo de franceses, americanos y británicos con el objetivo de clasificar la leucemia según el tipo de célula que origina la enfermedad y el grado de madurez que presentaba la célula en ese momento. En la siguiente tabla se puede apreciar la clasificación morfológica de los 3 subtipos de leucemia linfoblástica aguda.

Clasificación morfológica de la Leucemia Linfoblástica Aguda (LLA) según la FAB.			
Características	**L1** Predominio de células pequeñas	**L2** Predominio de células grandes	**L3** Población homogénea de células hiperbasófilas
Tamaño celular	Pequeño, hasta 2 veces el linfocito	Grande, Heterogéneo en tamaño	Grande y homogéneo
Cromatina nuclear	Homogénea en cada caso	Heterogénea	Finamente punteada y homogénea
Forma del núcleo	Regular, identaciones ocasionales	Irregular. Identación	Regular, oval o redondo
Nucleolos	No visible o pequeño	Una o más a veces grande	Prominente o más vesiculoso
Extensión del citoplasma	Escaso	Variable, a veces moderadamente abundante	Moderadamente abundante con múltiples vacuolas que se superponen incluso al núcleo
Basofilia del citoplasma	Ligera o moderadamente intensa	Variable	Muy intensa
Vacuolización del citoplasma	Variables	Variables	A menudo prominentes

Tabla 1. Clasificación morfológica de la LLA según la FAB, modificado de [30].

2.5 Inteligencia Artificial (IA)

La inteligencia artificial (IA) es la combinación de algoritmos planteados con el propósito de crear máquinas que presenten las mismas capacidades que el ser humano. Una tecnología que todavía nos resulta lejana y misteriosa, pero que desde hace algunos años está presente en nuestro día a día a todas horas [4].

La IA es probablemente una de las ramas de las Ciencias de la Computación que más crecimiento está teniendo en la actualidad. Pese a haber nacido hace más de 70 años, se encuentra en el periodo de su historia en el que mayor interés ha generado debido a la revolución que está provocando en el mercado actual.

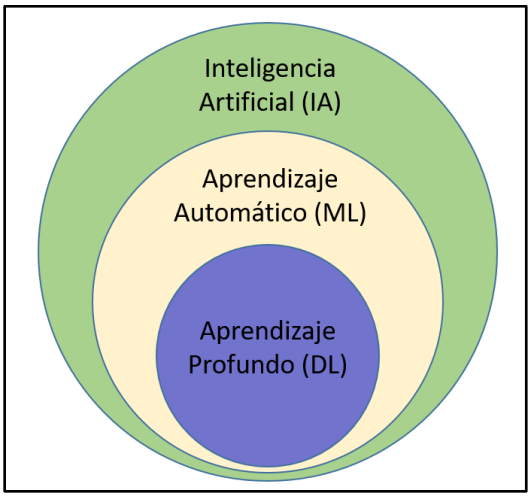


Figura 9. Estructura de la inteligencia artificial [4].

2.6 Evolución de la Inteligencia Artificial

Hasta hace poco, existía una limitación de capacidad de cómputo que hacía que la inteligencia artificial mostrara resultados muy pobres en los problemas a los que se aplicaba, lo que produjo varios periodos históricos de descontento en la industria y una considerable reducción tanto del interés en esta disciplina como del número de investigadores dedicados.

Sin embargo, en los últimos años la inteligencia artificial está tomando un gran impulso, al ser capaz de resolver problemas utilizando las computadoras, llegando a niveles a los que nunca se había llegado. Incluso los dispositivos móviles se benefician de investigaciones en este campo, por ejemplo, a través del texto predictivo del teclado, del desbloqueo de pantalla con huella dactilar o de la detección de rostros captados por la cámara. Se podrían enumerar varias razones que han servido de motor de despegue de la IA, pero destaca con especial énfasis la democratización de la capacidad de cómputo, en especial a partir del año 2009 con la publicación del primer artículo científico sobre la paralelización masiva de cómputo de IA usando GPU's y otro en 2010, por la demostración de su uso en el reconocimiento automático de dígitos escritos a mano, superando por primera vez la capacidad humana en esta tarea.

De esta forma, la inteligencia artificial adquiere su nombre debido a la capacidad que tiene un dispositivo electrónico de resolver problemas para los que se requiere inteligencia y que, de forma tradicional (programando), no podría resolverse en un ordenador.

2.7 Aprendizaje automático *(Machine Learning)*

El aprendizaje automático (*Machine Learning)* es una sub-rama de la inteligencia artificial, que tiene como objetivo la resolución de problemas sin que sea necesario programar explícitamente el algoritmo que los soluciona.

Para llevar esto a cabo con éxito, es necesario contar con datos que permitan al sistema ser capaz de inferir patrones [23].

El concepto de aprendizaje profundo está muy influenciado por la capacidad que tenemos los seres humanos para aprender. Por poner un ejemplo, si a una persona que nunca ha visto una pitaya, le indicamos que es una fruta de piel rosácea y que por dentro es blanca con puntos negros, le sería muy fácil detectarla en una cesta llena de diferentes frutas.

Es más, incluso sin decirle ninguna descripción, una persona que conociera todas las frutas de la cesta excepto la pitaya, sería capaz de concluir que esa fruta que desconoce es la que estamos buscando. Este tipo de deducciones, que para el ser humano pueden parecer muy sencillas, son extremadamente complejas de llevar a cabo por una computadora [23].

Gracias a la increíble evolución de las computadoras en cuanto a velocidad de cálculo y almacenamiento, el aprendizaje profundo está hoy en día en su mejor momento. Esto, aunado a que cada vez son más los datos de los que disponemos, hace que los campos de aplicación de esta sub-rama de la inteligencia artificial estén aumentando exponencialmente.

Todo proceso de aprendizaje profundo cuenta con al menos dos etapas muy diferenciadas: entrenamiento y predicción.

2.7.1 Entrenamiento

En esta fase el sistema trata de aprender tendencias, comportamientos y/o patrones que se ajusten a los datos que forman el conjunto de entrenamiento.

Desde un punto de vista matemático, el entrenamiento de un sistema inteligente corresponde con un proceso de optimización de los parámetros de una función para que su salida sea lo más parecida posible al resultado que queremos obtener.

Dependiendo de los datos con los que contamos para entrenar nuestro sistema, existen dos grandes categorías de aprendizaje: supervisado y no supervisado.

2.7.1.1 Aprendizaje supervisado

El aprendizaje supervisado comienza típicamente con un conjunto establecido de datos y una cierta comprensión de cómo se clasifican estos datos.

El aprendizaje supervisado tiene la intención de encontrar patrones en datos que se pueden aplicar a un proceso de analítica. Estos datos tienen características etiquetadas que definen el significado de los datos. Por ejemplo, se puede crear una aplicación de *Machine Learning* con base en imágenes y descripciones escritas que distinga entre millones de animales.

2.7.1.2 Aprendizaje no supervisado

El aprendizaje no supervisado se utiliza cuando el problema requiere una cantidad masiva de datos sin etiquetar. Por ejemplo, las aplicaciones de redes sociales, tales como *Twitter, Instagram y Snapchat*, tienen grandes cantidades de datos sin etiquetar. La comprensión del significado detrás de estos datos requiere algoritmos que clasifican los datos con base en los patrones o clústeres que encuentran. El aprendizaje no supervisado lleva a cabo un proceso iterativo, analizando los datos sin intervención humana. Se utiliza con la tecnología de detección de spam en e-mails.

Existen demasiadas variables en los *emails* legítimos y de *spam* para que un analista etiquete una cantidad masiva de *email* no solicitado.

En su lugar, los clasificadores de aprendizaje automático, basados en *clustering* y asociación, se aplican para identificar *emails* no deseados.

2.7.1.3 Aprendizaje por refuerzo

El aprendizaje por refuerzo es un modelo de aprendizaje conductual.

El algoritmo recibe retroalimentación del análisis de datos, conduciendo al usuario hacia el mejor resultado. El aprendizaje por refuerzo difiere de otros tipos de aprendizaje supervisado, porque el sistema no está entrenado con el conjunto de datos de ejemplo.

Más bien, el sistema aprende a través de la prueba y el error. Por lo tanto, una secuencia de decisiones exitosas conduce al fortalecimiento del proceso, porque es el que resuelve el problema de manera más efectiva.

2.7.2 Predicción

La etapa de predicción se realiza una vez se ha completado el entrenamiento y consiste en evaluar datos nuevos que no se han tenido en cuenta para el entrenamiento utilizando la función optimizada para obtener un resultado.

Si el resultado que se obtiene es una categoría, estamos ante un problema de clasificación. Por ejemplo, el resultado de un partido de la quiniela (1 X 2), el grupo sanguíneo al que pertenece una persona o si en una foto aparece un perro o un gato.

Si, por el contrario, el valor que devuelve la función es un número, estamos ante un problema de regresión. La predicción del precio del barril de *Brent*, o la estimación del peso de una persona, son ejemplos de regresión.

2.8 Aprendizaje Profundo (*Deep Learning)*

El aprendizaje profundo (*Deep Learning)* es un método específico de aprendizaje automático (Machine Learning) que incorpora las redes neuronales en capas sucesivas para aprender de los datos de manera iterativa. El aprendizaje profundo es especialmente útil cuando se trata de aprender patrones de datos no estructurados. Las redes neuronales complejas de aprendizaje profundo están diseñadas para emular cómo funciona el cerebro humano, así que las computadoras pueden ser entrenadas para lidiar con abstracciones y problemas mal definidos. Las redes neuronales y el aprendizaje profundo se utilizan a menudo en el reconocimiento de imágenes, voz y aplicaciones de visión por computadora.

El concepto de aprendizaje profundo nace a raíz de utilizar un gran número de capas ocultas en las redes.

2.9 Redes neuronales

Las redes neuronales artificiales están basadas en el funcionamiento de las redes de neuronas biológicas. Las neuronas que todos tenemos en nuestro cerebro están compuestas de dendritas, el soma y el axón: Las dendritas se encargan de captar los impulsos nerviosos que emiten otras neuronas. Estos impulsos, se procesan en el soma y se transmiten a través del axón que emite un impulso nervioso hacia las neuronas contiguas.

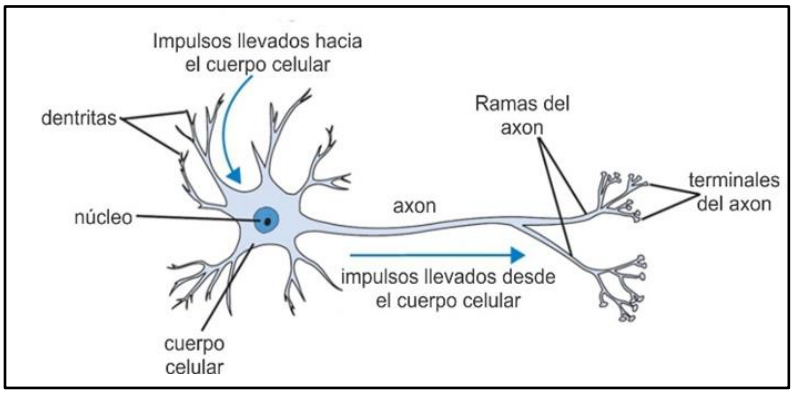


Figura 10. Estructura de una neurona biológica [15].

A nivel esquemático, una neurona artificial se representa del siguiente modo:

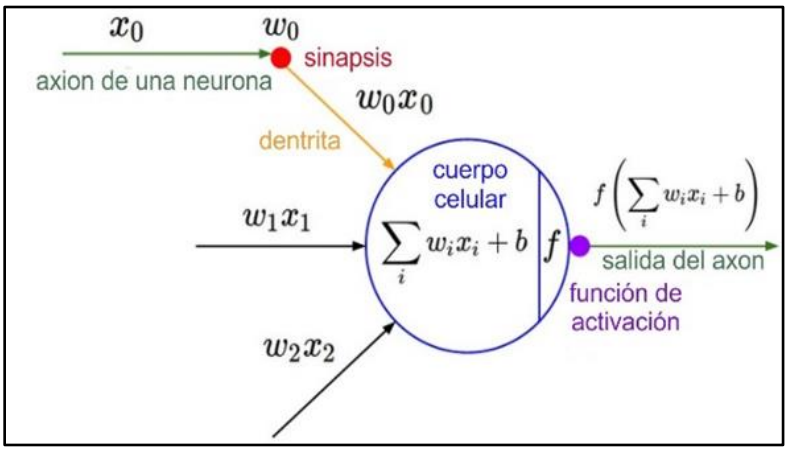


Figura 11. Estructura de una neurona artificial [15].

En el caso de las neuronas artificiales, la suma de las entradas multiplicadas por sus pesos asociados determina el “impulso nervioso” que recibe la neurona. Este valor se procesa en el interior de la célula mediante una función de activación que devuelve un valor que se envía como salida de la neurona.

Del mismo modo que nuestro cerebro está compuesto por neuronas interconectadas entre sí, una red neuronal artificial está formada por neuronas artificiales conectadas entre sí y agrupadas en diferentes niveles que denominamos capas.

Una capa es un conjunto de neuronas cuyas entradas provienen de una capa anterior (o de los datos de entrada en el caso de la primera capa) y cuyas salidas son la entrada de una capa posterior.

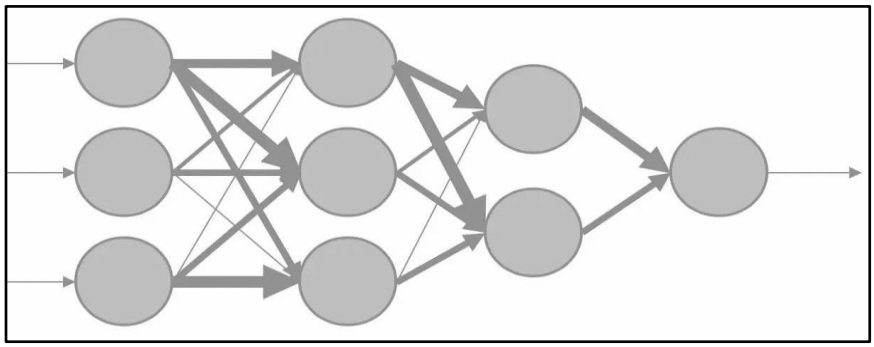


Figura 12. Ejemplo de red con 4 capas [15].

Las neuronas de la primera capa reciben como entrada los datos reales que alimentan a la red neuronal. Es por eso por lo que la primera capa se conoce como capa de entrada. La salida de la última capa es el resultado visible de la red, por lo que la última capa se conoce como la capa de salida. Las capas que se sitúan entre la capa de entrada y la capa de salida se conocen como capas ocultas ya que desconocemos tanto los valores de entrada como los de salida.

Una red neuronal, por lo tanto, siempre está compuesta por una capa de entrada, una capa de salida (si solo hay una capa en la red neuronal, la capa de entrada coincide con la capa de salida) y puede contener 0 o más capas ocultas.

2.10 Tipos de redes neuronales

Existen varios tipos de arquitecturas de redes neuronales, sin embargo, las más utilizadas se pueden clasificar en Redes Neuronales Profundas (DNN), Redes Neuronales Convolucionales (CNN) y Redes Neuronales Recurrentes (RNN) cada una de estas redes se describen a continuación:

2.10.1 Red Neuronal Profunda (DNN).

Con este tipo de red se puede procesar texto, imágenes pequeñas o datos. Sin embargo, esta red tiene una limitante, debido a la existencia de tantas conexiones en cada una de las capas que cuando se requiere procesar datos grandes, por ejemplo, una imagen de tamaño 300 x 300 píxeles se tienen un total de 90,000 datos en la capa de entrada, realizar los cálculos para cada uno de los pixeles es demasiado y requeriría mayor poder computacional con este tipo de red neuronal. Para darle solución a este problema con las imágenes, se crea la red neuronal convolucional. La estructura del DNN se muestra en la figura 13.

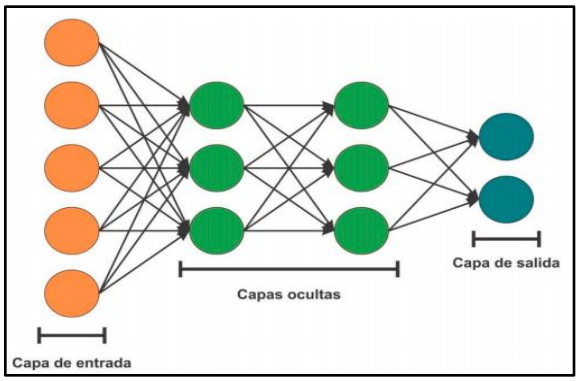


Figura 13. Estructura de una Red Neuronal Profunda (DNN) [15].

2.10.2 Red Neuronal Convolucional (CNN).

El uso común de esta red neuronal es para el procesamiento de imágenes, sin embargo, se ha estado implementando también para el procesamiento de texto. Generalmente esta red neuronal en la última capa oculta tiene una función *Softmax*, que le permite conectar todas las neuronas de las convoluciones y *max pooling* que emplea la red. La estructura de esta red se muestra en la figura 14.

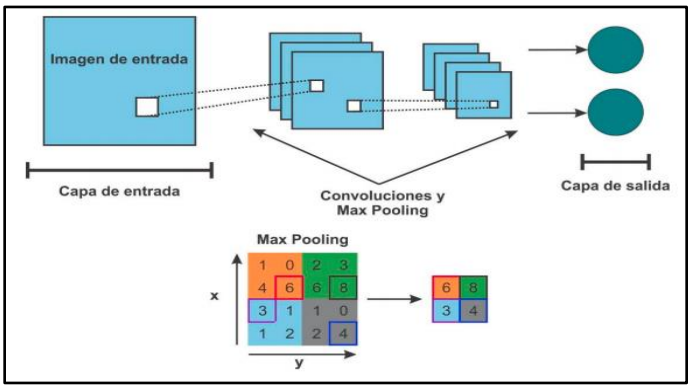


Figura 14. Estructura de una Red Neuronal Convolucional (CNN) [15].

2.10.3 Red Neuronal Recurrente (RNN).

Este tipo de redes se usan para tipos de datos que son secuenciales, es decir, datos en el que el valor de una variable en particular dependerá de el o los valores que se tuvo previamente. Por ejemplo, datos de tipo texto. A diferencia de las redes anteriores, esta red dentro de sus capas ocultas cuenta con capas recurrentes con celdas de *Long Short-Term Memory* (LSTM), que le permite saber el valor que tenía anteriormente. La estructura de esta red la podemos ver en la figura 15.

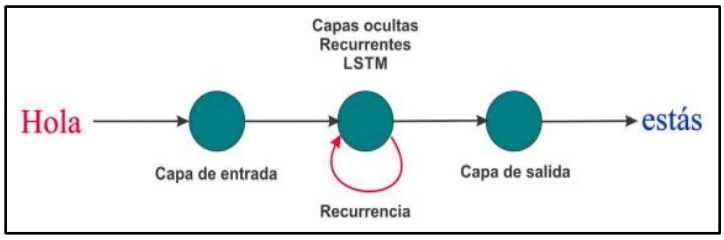


Figura 15. Estructura de una Red Neuronal Recurrente (RNN) [15].

2.11 ¿Cómo se entrena una red neuronal?

Entrenar una red neuronal consiste en ajustar cada uno de los pesos de las entradas de todas las neuronas que forman parte de la red neuronal, para que las respuestas de la capa de salida se ajusten lo más posible a los datos que conocemos.

En la siguiente figura, podemos ver un ejemplo muy simplificado del proceso de entrenamiento de una red para la detección de, en este caso, un gato en una imagen.

El grosor de cada flecha representa el peso que tiene esa entrada en la red neuronal, como se puede observar en la figura 16.

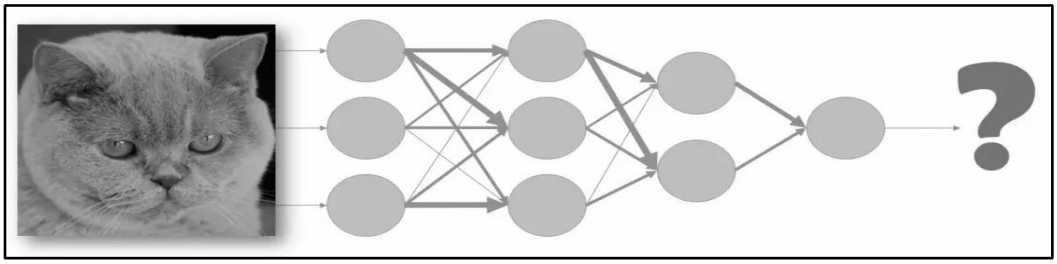


Figura 16. Ejemplo de red con 4 capas para identificar una imagen [15].

Si queremos conseguir que la red neuronal sea capaz de generalizar e identificar gatos en cualquier imagen, es importante utilizar un elevado número de imágenes para realizar el entrenamiento, tanto de imágenes que son gatos

(etiquetadas como 1) como de imágenes que no son gatos (etiquetadas como 0), incluyendo la mayor variabilidad posible. Con esto, la red será capaz de ajustar sus parámetros para satisfacer en la medida de lo posible a todas las imágenes.

2.12 Generación de imágenes sintéticas (*Data Augmentation*)

El rendimiento de la mayoría de los modelos de aprendizaje automático y, en particular, de los modelos de aprendizaje profundo, dependen de la calidad, la cantidad y la relevancia de los datos de entrenamiento. Sin embargo, la insuficiencia de datos es uno de los desafíos más comunes en la implementación del aprendizaje automático. Esto se debe a que, en muchos casos, la recopilación de dichos datos puede resultar costosa.

El aprendizaje profundo requiere de un gran número de datos para que las redes neuronales aprendan las características más relevantes de los *inputs* y después puedan realizar el proceso de inferencia de forma correcta, ya que cuando los modelos se entrenan con ejemplos limitados no son capaces de generalizar a los datos no vistos. Incluso si se utilizan modelos pre-entrenados *(Transfer Learning)*, muchas veces las imágenes para los casos particulares siguen siendo insuficientes y el modelo no se entrena correctamente.

Ante este contratiempo surge la necesidad de buscar métodos para generar imágenes sintéticas *(Data Augmentation*) con el objetivo de hacer viables proyectos con un conjunto de datos reducido de imágenes.

El aumento de datos es un conjunto de técnicas para aumentar artificialmente la cantidad de datos mediante la generación de nuevos datos a partir de datos existentes. Esto incluye realizar pequeños cambios en los datos o utilizar modelos de aprendizaje profundo para generar nuevos datos.

Para los modelos de aprendizaje profundo, la recopilación y el etiquetado de datos pueden ser procesos agotadores y costosos. Las transformaciones en conjuntos de datos mediante el uso de técnicas de aumento de datos permiten reducir estos costos operativos.

Uno de los pasos en un modelo de datos es la limpieza de esos datos, la cual es necesaria para modelos de alta precisión. Sin embargo, si la limpieza reduce la representabilidad de los datos, entonces el modelo no puede proporcionar buenas predicciones para las entradas del mundo real.

Las técnicas de aumento de datos pueden permitir que los modelos de aprendizaje automático sean más sólidos al crear variaciones que el modelo puede ver en el mundo real. En la siguiente figura mostramos las 2 técnicas utilizadas.

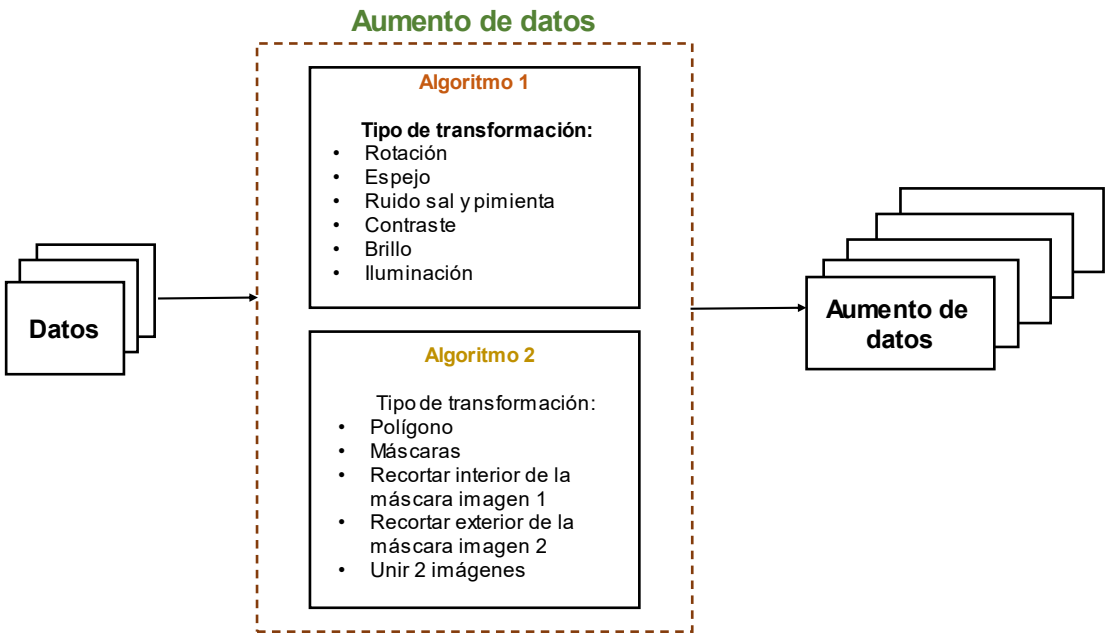


Figura 17. Metodología para la generación de datos sintéticos.

Se investigaron varias técnicas de generación de datos sintéticos y se escogieron 2 debido a su eficacia y facilidad de implementación:

- Generación de imágenes sintéticas tradicionales, mediante procedimientos clásicos como efecto espejo, rotaciones, contraste, brillo, etc.
- Generación de imágenes sintéticas poligonales, aplicando máscaras a partir de un algoritmo, para así generar nuevas imágenes.

La generación de imágenes o de cualquier otro tipo de dato está a la orden del día en un gran número de proyectos donde los datos son escasos [25]. El aumento de la variabilidad de los datos de entrenamiento permite una mayor generalización de los modelos.

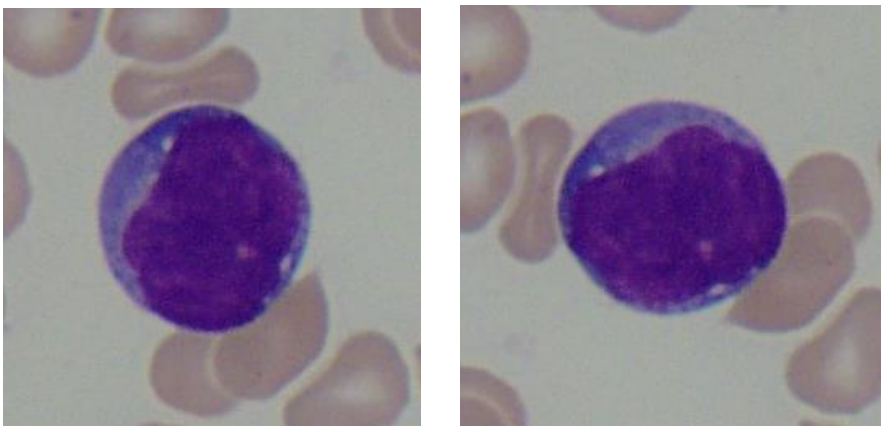
Los beneficios del aumento de datos incluyen: mejora de la precisión de la predicción del modelo, agregar más datos de entrenamiento a los modelos, prevención de la escasez de datos para mejores modelos [23].

Otro beneficio es reducir el sobreajuste de datos (es decir, un error en las estadísticas significa que una función corresponde demasiado a un conjunto limitado de puntos de datos), crear variabilidad en los datos, aumentar la capacidad de generalización de los modelos, ayudando a resolver problemas de desequilibrio de clases en la clasificación, reducir los costos de recopilación y etiquetado de datos, habilitar la predicción de eventos raros, previniendo problemas de privacidad de datos [24].

2.12.1 Algoritmo para la generación de imágenes sintéticas tradicionales

Cómo ya se mencionó antes, el objetivo de generar imágenes sintéticas es el de hacer viables proyectos con un conjunto de datos reducido de imágenes. Para este proyecto se realizó un programa el cual consiste en aplicar transformaciones geométricas a las imágenes cómo son, rotación y efecto espejo, y se modificaron también mediante ruido, contraste, brillo e iluminación.

Se obtuvieron en total 12,000 imágenes en formato TIF de tamaño 257 x 257 x 3. La obtención de estas imágenes permitió probar los 5 diferentes modelos. En la siguiente figura se muestra el ejemplo de 2 imágenes sintéticas tradicionales de células enfermas.



(Izq.) *(Der.)*

Figura 18. Imágenes sintéticas tradicionales de células enfermas. A la izquierda tenemos una imagen rotada a 90°, y a la derecha una imagen rotada a 180°.

2.12.2 Algoritmo propuesto para la generación de imágenes sintéticas poligonales

Para obtener las imágenes sintéticas poligonales se creó un procedimiento que consiste en obtener formas poligonales, es decir, convierte un polígono, generado de forma aleatoria, de una región de interés (ROI) en una máscara binaria.

A continuación, en la tabla 2, se muestra el procedimiento para realizar la generación de imágenes sintéticas poligonales.

Generación de imágenes sintéticas poligonales (Ms-Mix)
1.Generar las coordenadas polares
2.Cambiar de coordenadas polares a cartesianas
Función ***"pol2cart (theta, radius)"***
3.Convertir los valores de -1 a 1 a valores entre 0 y 256
4.Convertir el polígono a máscara
Función ***"poly2mask (x, y, tamaño-imag, tamaño-imag)***
5. Invertir la matriz
6. Duplicar la máscara aumentando 2 canales para imágenes RGB
7. Unir las 2 imágenes recortadas
8. Guardar las máscaras generadas

Tabla 2. Procedimiento para la generación de imágenes poligonales.

La siguiente figura muestra como se ve la máscara generada por los polígonos.

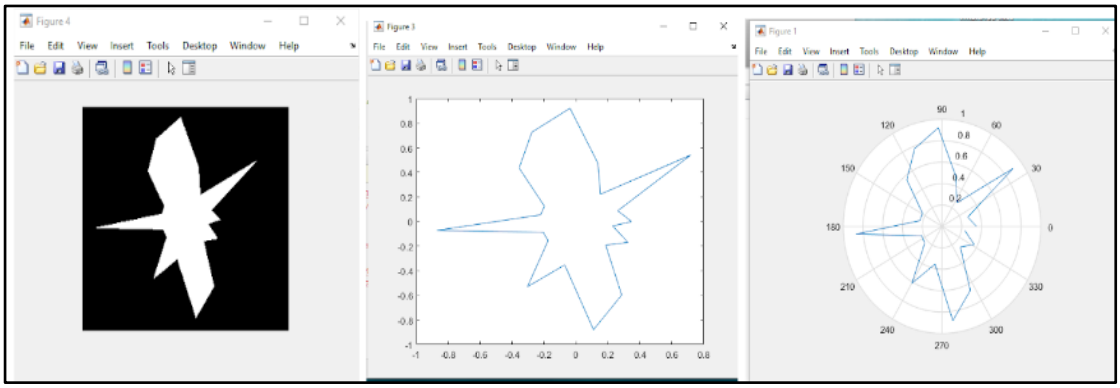


Figura 19. Región de interés (ROI).

Utilizando las máscaras se obtuvieron en total 12,000 imágenes en formato TIF de tamaño 257 x 257 x 3. Con este nuevo conjunto de datos se probaron los 5 diferentes modelos.

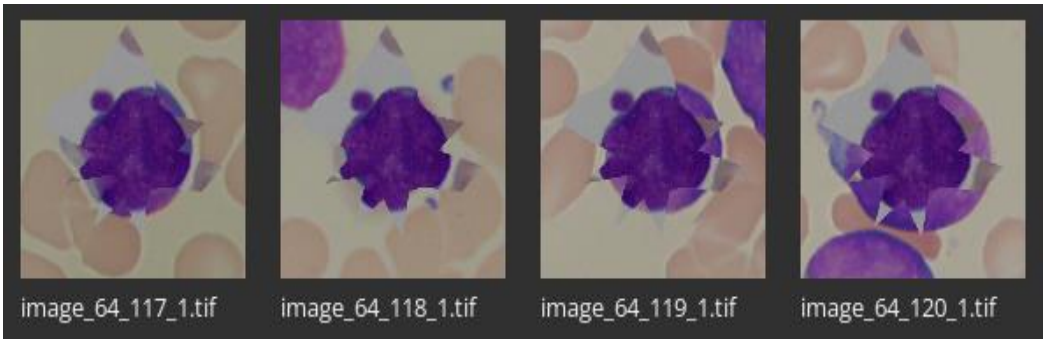


Figura 20. Imágenes sintéticas poligonales de células enfermas.

La idea de proponer un algoritmo de este tipo surge después de revisar en la literatura los métodos que utilizan algunos de los autores [23, 24, y 25] para generar los conjuntos de datos de imágenes sintéticas.

Como ejemplo tenemos el método de **CutMix** [23]. Ellos utilizaron métodos simples y efectivos para mejorar los clasificadores recortando al azar un parche de una imagen y pegándolo en otra imagen.
Con el fin de mejorar aún más el rendimiento, utilizaron información destacada de la imagen para guiar sus mezclas. Su implementación contiene las siguientes fórmulas:

$$\tilde{x} = \mathbf{M} \odot x_A + (\mathbf{1} - \mathbf{M}) \odot x_B$$
$$\tilde{y} = \lambda y_A + (1 - \lambda) y_B,$$

Donde **M** es la máscara binaria que indica las regiones recortadas y de relleno de las dos imágenes dibujadas al azar y λ (en [0, 1]) se extrae de una Beta(α, α) distribución. Las coordenadas de los cuadros delimitadores están dadas por: B = (rx, ry, rw, rh); y el muestreo del cuadro delimitador está representado por:

$$r_x \sim \text{Unif}\ (0, W)\,, \quad r_w = W\sqrt{1-\lambda},$$
$$r_y \sim \text{Unif}\ (0, H)\,, \quad r_h = H\sqrt{1-\lambda}$$

Donde r_x y r_y se extraen aleatoriamente de una distribución uniforme.

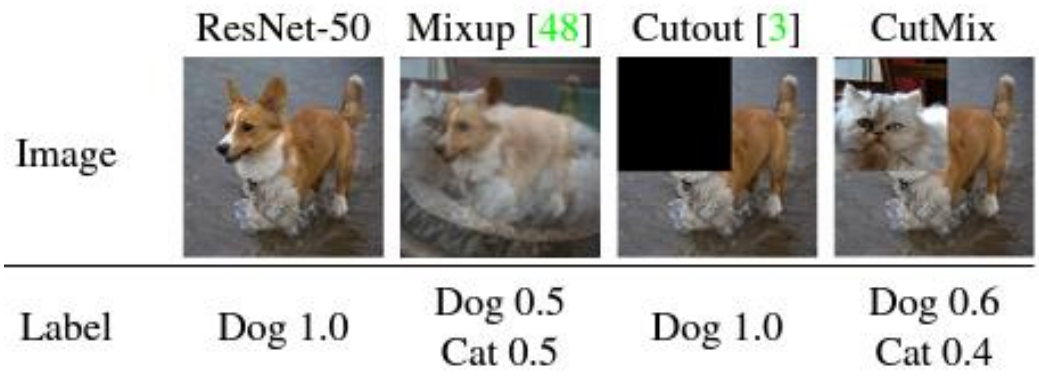


Figura 21. Método CutMix

Otro ejemplo es **ResizeMix** [24]. Ellos mezclaron los datos cambiando directamente el tamaño de la imagen de origen a un pequeño parche y pegándolo en otra imagen. El parche obtenido conserva información más sustancial del objeto en comparación con los métodos convencionales basados en cortes.

Este método muestra una ventaja evidente sobre CutMix e incluso supera a los métodos de aumento automático basados en búsquedas más costosas.
Las fórmulas que utilizaron son las siguientes:

$$I_s \in R^{W \times H} \quad y \quad I_t \in R^{W \times H}$$

Donde se designa la imagen de origen y destino respectivamente. También se designa el parche con:

$$P \in R^{WP \times HP}.$$

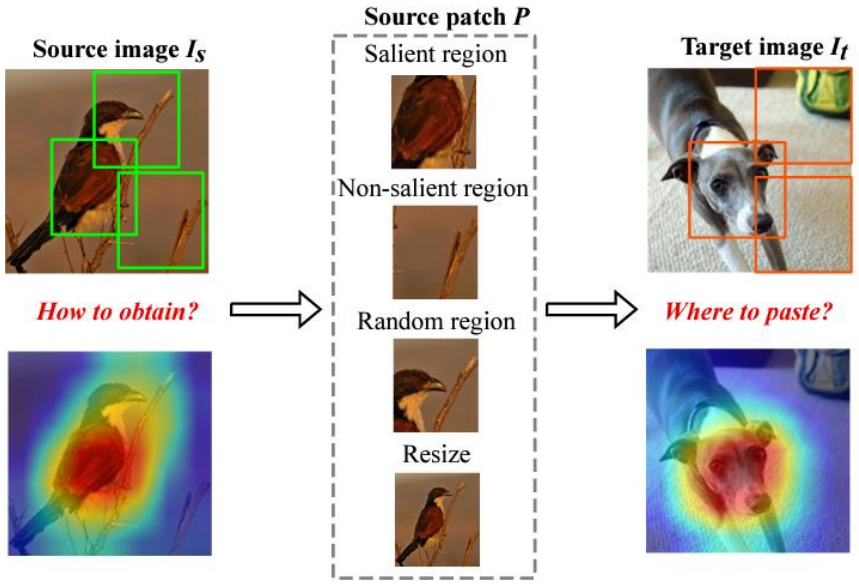


Figura 22. Método ResizeMix

Por último, tenemos a **F-Mix** [25]. Ellos trabajan con MSDA un método que utiliza máscaras binarias aleatorias obtenidas aplicando un umbral a imágenes de baja frecuencia muestreadas del espacio de Fourier. Estas máscaras aleatorias pueden adoptar una amplia gama de formas y pueden generarse para uso con datos de una, dos y tres dimensiones. FMix mejora el rendimiento sobre MixUp [26] y CutMix [23], sin aumento en el tiempo de entrenamiento, para una cantidad de modelos en una variedad de conjuntos de datos y configuraciones de problemas. Su procedimiento fue:

1. Recortar una parte de cualquier forma de una imagen aleatoria y pegarla en la imagen relacionada.

2. Es diferente de cortar y pegar en general, que requiere una máscara para definir qué partes de la imagen deben considerarse.

3. La máscara se obtiene mediante el umbral de la imagen de baja frecuencia muestreada en el espacio de Fourier.

Figura 23. Método F-Mix

2.13 Modelos utilizados

Revisando la literatura y considerando una mezcla de modelos complejos y modelos compactos, se escogieron los siguientes: *Compact, Enhanced, ResNet-50, MobileNetV2 y AlexNet.*

Los modelos utilizados se describen a continuación:

Compact: Esta red neuronal está diseñada para ser eficiente en cuanto a memoria y tiempo de ejecución. No contiene ninguna capa completamente conectada. La altura de una imagen no debe ser inferior a 15 píxeles.

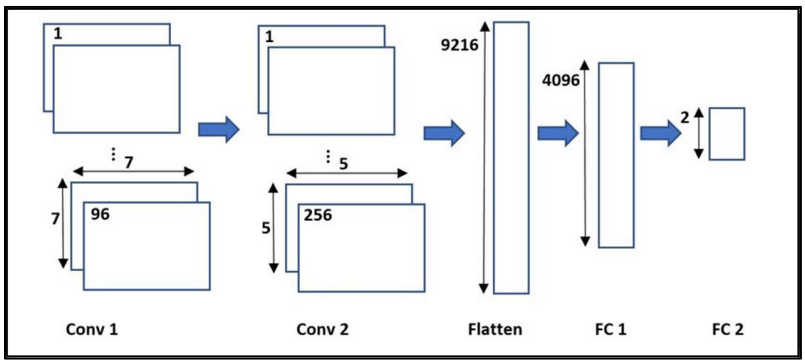


Figura 24. Arquitectura del modelo Compact [16].

Enhanced: Esta red neuronal tiene más capas ocultas que "*Compact"* y, por lo tanto, se supone que es más adecuada para tareas de clasificación más complejas. Esto tiene el costo de ser más exigente en cuanto a tiempo y memoria. El ancho y la altura de la imagen no deben ser inferiores a 47 píxeles. No hay un tamaño máximo, pero un tamaño de imagen grande aumentará significativamente la demanda de memoria y el tiempo de ejecución.

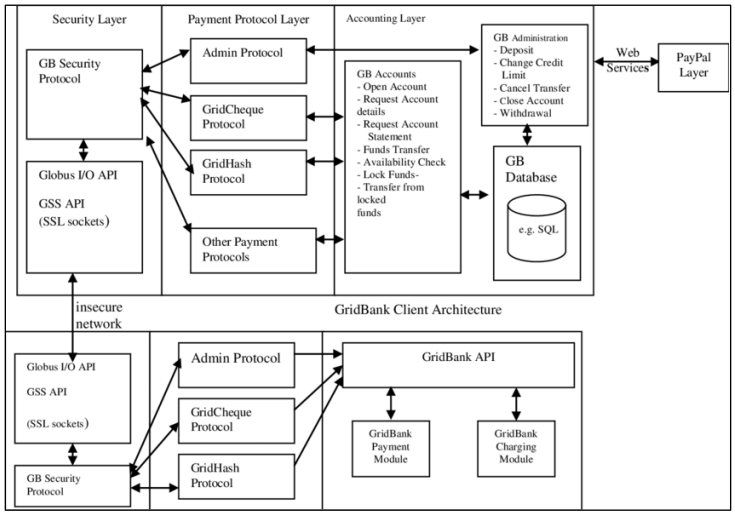


Figura 25. Arquitectura del modelo Enhanced [16].

ResNet-50: Similar a la red "*Enhanced"*, este clasificador es adecuado para tareas más complejas. Sin embargo, su estructura es diferente, lo que trae la ventaja de hacer que el entrenamiento sea más estable y robusto internamente. El ancho y la altura de la imagen no deben ser inferiores a 32 píxeles.

No hay un tamaño máximo, pero un tamaño de imagen grande aumentará significativamente la demanda de memoria y el tiempo de ejecución.

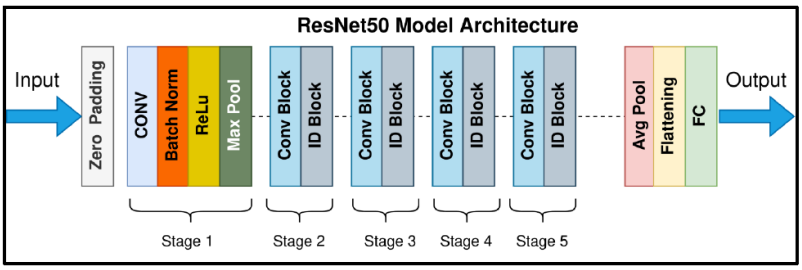


Figura 26. Arquitectura del modelo ResNet [16].

MobileNetV2: Este clasificador es un modelo pequeño y de bajo consumo, adecuado para aplicaciones de visión móviles e integradas. El ancho y la altura de la imagen no deben ser inferiores a 32 píxeles. No hay un tamaño máximo, pero un tamaño de imagen grande aumentará significativamente la demanda de memoria y el tiempo de ejecución. En GPU, la arquitectura de red puede beneficiarse enormemente de optimizaciones especiales; en la CPU, el tiempo de ejecución puede ser significativamente más lento.

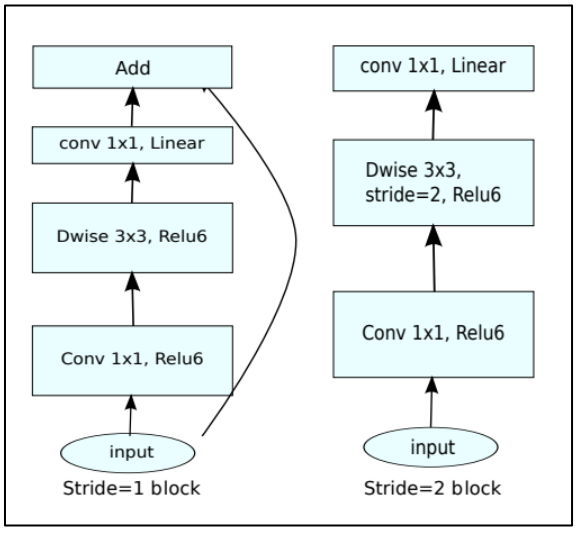


Figura 27. Arquitectura del modelo MobileNet V2 [16].

AlexNet: Esta red neuronal está diseñada para tareas de clasificación simples. Se caracteriza por sus núcleos de convolución en las primeras capas de

convolución, que son más grandes que en otras redes con un rendimiento de clasificación comparable (por ejemplo, la red "*Compact*").

El ancho y la altura de la imagen no deben ser inferiores a 29 píxeles. No hay un tamaño máximo, pero un tamaño de imagen grande aumentará significativamente la demanda de memoria y el tiempo de ejecución.

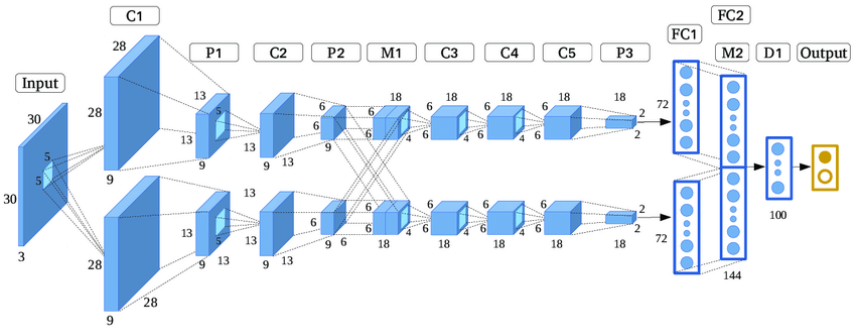


Figura 28. Arquitectura del modelo AlexNet [16].

CAPÍTULO 3. REVISIÓN DE TRABAJOS PREVIOS

En este capítulo comentaremos sobre la revisión que se realizó en relación con las enfermedades sanguíneas más comunes, las diferentes técnicas para aplicar el aumento de datos, así como también mencionaremos los modelos utilizados para entrenar el conjunto de datos.

En el año 2018 Gunčar et al., [17] pertenecientes al Smart Blood Analytics Swiss, en Suiza; diseñaron un modelo predictivo de aprendizaje automático para predecir enfermedades hematológicas, obteniendo precisiones de predicción de 0.88 y 0.86 al considerar las cinco enfermedades sanguíneas más probables.

Por su parte, Molina [18] de la Universidad Abierta de Cataluña, España; realizó un clasificador para pronosticar linfomas y leucemias agudas mediante el reconocimiento morfológico de células linfoides y blásticas anormales, los clasificadores empleados fueron Random Forest, SVM y LDA obteniendo en los tres casos una exactitud cercana al 80%.

Posteriormente, en el año 2019, Tambe et al. [19] del Instituto de Tecnología Informática de SCTR, India; desarrollaron un algoritmo de clasificación para identificar leucemia linfocítica crónica, linfoma folicular y linfoma de células del manto; empleando una red Inception-V3, obteniendo un porcentaje de exactitud de 97.33 %.

Por su parte, Kamal & Hassan [20] de la universidad de Beni-Suef, Egipto, utilizaron diferentes clasificadores de aprendizaje profundo para detectar diferentes enfermedades sanguienas, obteniendo los siguientes resultados: LogitBoost 98.16%, Random forests 97.12%, Decision Tree 97.00%, Regression analysis 96.54%, K-Nearest Neighbor 92.97%, Bayesian network

92.86%, MultilayerPerceptron 91.80%, NaiveBayes 81.60%, Support Vector Machine 71.20%.

En el año 2020, Font [21] de la Universidad Politécnica de Catalunya, España, diseño un sistema para detectar Linfoma de Burkitt; Linfoma Folicular; Tricoleucemia (HCL); Linfoma de Células del Manto (LCM) y Leucemia Linfocítica Crónica (LLC); utilizando Redes Neuronales Siamesas (SNN) y Few-Shot Learning (FSL). Los clasificadores probados fueron ResNet-34 y ResNet-50 obteniendo un porcentaje de exactitud de 93% y del 98% respectivamente.

Finalmente, en el año 2022, Wang, Et. Al., [22] de la Universidad Médica Shanxi en China, diseñaron un modelo de aprendizaje profundo para el reconocimiento automático de anemia aplásica (AA), síndromes mielodisplásicos (MDS) y leucemia mieloide aguda (AML) basada en frotis de médula ósea; los resultados en cuanto a la precisión y la sensibilidad del modelo fue: 0,968 para AA, 0,929 para MDS y 0,857 para AML.

Por otro lado, en la parte del aumento de datos, hay estrategias recientes basadas en la mezcla de datos, por ejemplo, tenemos a CutMix, Sangdoo Yun, Dongyoon Han, Et. al [23]. Ellos utilizaron métodos simples y efectivos para mejorar los clasificadores recortando al azar un parche de una imagen y pegándolo en otra imagen. Con el fin de mejorar aún más el rendimiento de CutMix, exploraron información destacada de la imagen para guiar sus mezclas.

También tenemos a ResizeMix, en el 2020 Jie Qin, Jiemin Fang, Et al. [24] Mezclaron los datos cambiando directamente el tamaño de la imagen de origen a un pequeño parche y pegándolo en otra imagen. El parche obtenido conserva información más sustancial del objeto en comparación con los métodos convencionales basados en cortes. Este método muestra una ventaja evidente

sobre CutMix e incluso supera a los métodos de aumento automático basados en búsquedas más costosas.

Y por último encontramos a Ethan Harris, Antonia Marcu, Et. al. Con el desarrollo de FMix [25], en 2020. Utilizaron MSDA un método que utiliza máscaras binarias aleatorias obtenidas aplicando un umbral a imágenes de baja frecuencia muestreadas del espacio de Fourier. Estas máscaras aleatorias pueden adoptar una amplia gama de formas y pueden generarse para uso con datos de una, dos y tres dimensiones. FMix mejora el rendimiento sobre MixUp [26] y CutMix [23], sin aumento en el tiempo de entrenamiento, para una cantidad de modelos en una variedad de conjuntos de datos y configuraciones de problemas.

Después de realizar una exhaustiva revisión de la literatura sobre estos temas de interés, se concluye que la información sobre el aumento de datos, (data augmentation) para la detección de Leucemias Linfoblásticas Agudas aún es escasa. Considerando el material encontrado, podemos observar que los autores están utilizando los mismos tipos de clasificadores, por lo que en este trabajo de tesis se plantea utilizar nuevos modelos de entrenamiento combinados con las técnicas de aumento de datos.

En la tabla 3, se muestra la comparación de los trabajos relacionados con este trabajo de tesis.

Autor / año	Reconocimiento de imágenes con aprendizaje profundo	Utilización de datos sintéticos	Imágenes de objetos cotidianos	Imágenes médicas	Algoritmos tradicionales	Algoritmos Poligonales
Mix-up, Hongyi Zhang (2018)	✓	✓	✓		✓	
Jordi Torres, Deep Learning (2019)	✓	✓	✓		✓	
FMix, Ethan Harris (2020)	✓	✓	✓			✓
ResizeMix, Jie Qin (2020)	✓	✓	✓		✓	
CutMix, Sangdoo Yun (2022)	✓	✓	✓		✓	
Mas-Mix, Sánchez (2023)	✓	✓		✓	✓	✓

Tabla 3. Tabla comparativa de trabajos relacionados. Al final de la tabla se compara el trabajo propuesto para esta tesis.

CAPITULO 4. METODOLOGÍA

En esta sección se describe la metodología utilizada para abordar el problema planteado en esta investigación. En primer lugar, se selecciona el conjunto de datos que se utilizará como entrada del proceso. En este caso, se ha elegido el conjunto de imágenes de células leucémicas ALL_IDB2, que consta de 260 imágenes segmentadas y etiquetadas: 130 imágenes de células sanas y 130 imágenes de células enfermas, todas en formato TIF con un tamaño de 257 x 257 x 3 píxeles.

Se utilizó la base de datos ALL-IDB2 *"Acute Lymphoblastic Leukemia Image Database for Image Processing"* [27]. La base es un conjunto de datos público y gratuito de imágenes microscópicas de muestras de sangre, diseñado específicamente para la evaluación y comparación de algoritmos de segmentación y clasificación de imágenes.

Estas imágenes se centran en la leucemia linfoblástica aguda (LLA). Para cada imagen en el conjunto de datos, oncólogos expertos proporcionan la clasificación/posición de todos los linfoblastos. El esfuerzo que realizó el conjunto de investigadores italianos dio como resultado una nueva herramienta de prueba a las comunidades de procesamiento de imágenes y reconocimiento de patrones, con el objetivo de estimular nuevos estudios en las áreas antes mencionadas.

Las imágenes del conjunto de datos se capturaron con un microscopio óptico de laboratorio acoplado con una cámara *Canon PowerShot G5*.
Todas las imágenes están en formato TIF con una profundidad de color de 24 bits, y una resolución de 2592 x 1944. La base de imágenes se compone de dos conjuntos: ALL-IDB1 y ALL-IDB2.

El primer conjunto de imágenes contiene imágenes de células sin segmentar y en el segundo conjunto se encuentran las células ya segmentadas. En ambos conjuntos se localizan las áreas de interés de células normales y blásticas.

El conjunto de datos ALL-DB2 consta de 260 imágenes recortadas y etiquetadas, 130 de células sanas y 130 de células enfermas. Las imágenes ALL-IDB2 tienen propiedades similares a las imágenes de ALL-IDB1, excepto que son más pequeñas y están etiquetadas.

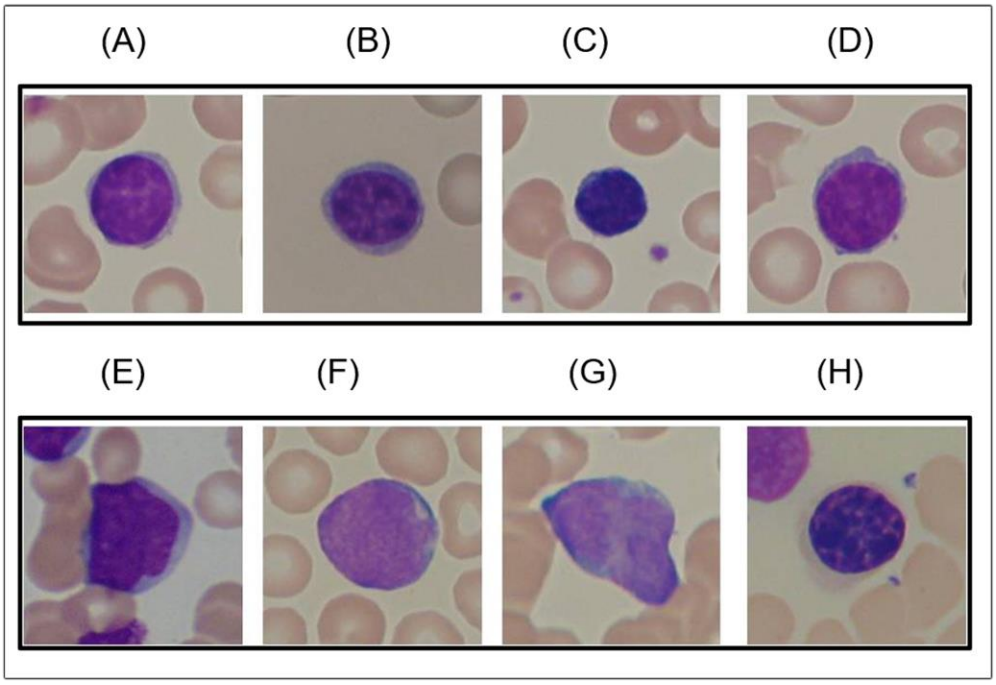


Figura 29. Ejemplos de las imágenes contenidas en ALL-IDB2: las células de la A a la D están etiquetadas como células sanas, las células de la E a la H están etiquetadas como probables linfoblastos o células enfermas.

Características morfológicas de las células blásticas:

La clasificación de los linfocitos en las imágenes es bastante compleja, ya que incluso un operador experto puede tener problemas para clasificar algunas células de linfocitos viéndolos directamente por el microscopio.

En realidad, las diferencias morfológicas entre las imágenes blásticas y los linfocitos normales son muy pequeñas.

De acuerdo con el análisis morfológico visual más común para la enfermedad de LLA (clasificación FAB), las características que los técnicos de laboratorio capacitados consideran durante la observación de imágenes con células enfermas son las siguientes:

Clasificación FAB, tipo L1, L2 y L3.

- **Para el tipo L1:** todos los blastos son pequeños y homogéneos. Los núcleos son redondos y regulares con pocas hendiduras y nucléolos discretos. El citoplasma es escaso y generalmente sin vacuolas.

- **Para el tipo L2:** todos los blastos son grandes y heterogéneos. Los núcleos son irregulares y a menudo hendidos. Están presentes uno o más nucléolos, generalmente grandes. El volumen del citoplasma es variable, pero a menudo abundante y puede contener vacuolas.

- **Para el tipo L3:** todos los blastos son de tamaño moderado-grande y homogéneos. Los núcleos son regulares y de forma redondeada-ovalada. Uno o más nucléolos prominentes están presentes. El volumen del citoplasma es moderado y contiene vacuolas prominentes.

La siguiente figura muestra la gran variabilidad en forma y patrón de las células blásticas según la clasificación FAB para la enfermedad de LLA. El objetivo principal es detectar sin diferenciación la presencia de los tres subtipos de leucemia en las imágenes.

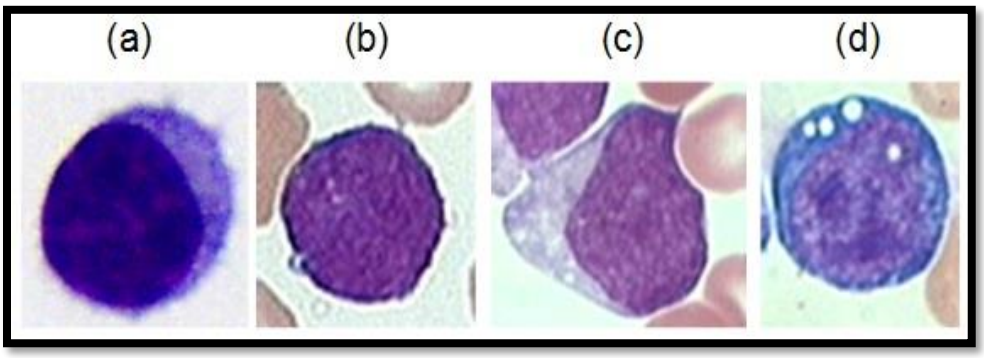


Figura 30. Variabilidad morfológica asociada a las células blásticas según la clasificación FAB: (a) célula sana, (b-d) células enfermas, donde (b), (c) y (d) son L1, L2 y L3 respectivamente.

4.1 Base de datos

Se utilizó la base de datos ALL-IDB2 "Acute Lymphoblastic Leukemia Image Database for Image Processing" creada por Fabio Scotti del Departamento de Ciencias de la Computación de la Universidad de Milán.

Es un conjunto de datos público y gratuito de imágenes microscópicas de muestras de sangre, diseñado específicamente para evaluar y comparar algoritmos de segmentación y clasificación de imágenes relacionadas con enfermedades hematológicas comunes como anemia, leucemia y trombocitopenia.

Los oncólogos expertos proporcionan la clasificación y posición de todos los linfoblastos en cada imagen del conjunto de datos. Además, se procesaron cifras de mérito específicas para comparar de manera justa diferentes algoritmos con el conjunto de datos.

Esta iniciativa proporciona una herramienta de prueba valiosa para las comunidades de procesamiento de imágenes y coincidencia de patrones, con el objetivo de estimular nuevos estudios en este campo de investigación importante.

Las imágenes se capturaron con un microscopio óptico de laboratorio y una cámara Canon PowerShot G5, con un formato TIF y una profundidad de color de 24 bits y resolución 2592 x 1944.

El conjunto de datos consta de 260 imágenes recortadas y etiquetadas, 130 de células sanas y 130 de células enfermas, como se observa en la figura 26, y es una versión actualizada de ALL-IDB1.

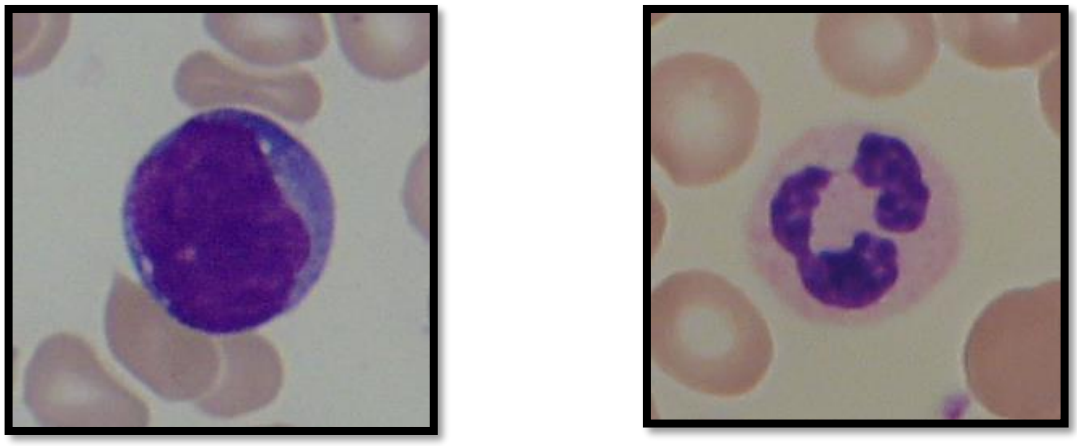


Figura 31. Ejemplos de imágenes segmentadas. (izq.); célula enferma, (der.); célula sana

4.2 Esquema general de la metodología

La figura 32 representa el esquema general de la metodología que se llevó a cabo durante la investigación.

Metodología

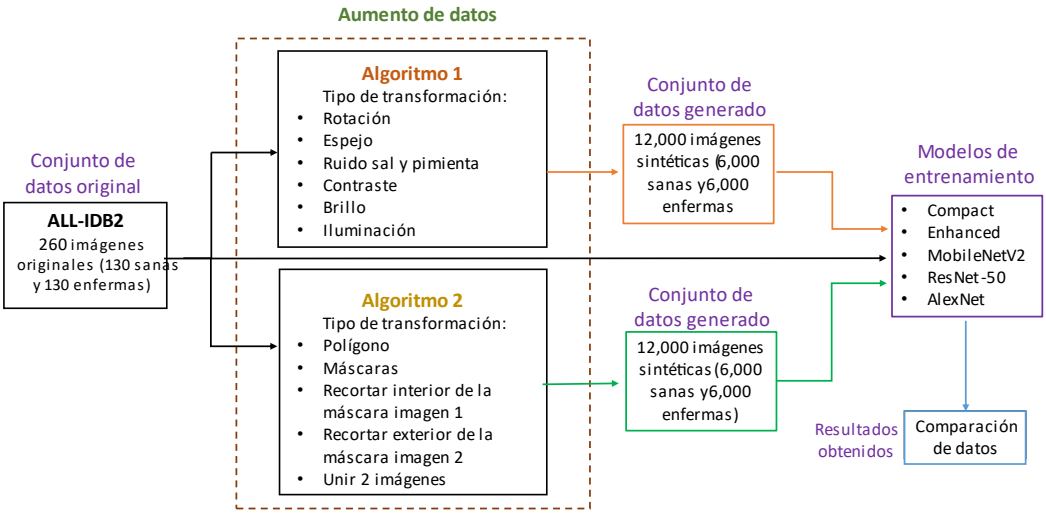


Figura 32. Esquema general de la metodología.

La siguiente tabla muestra el conjunto de datos generado. Se ha optado por un gran número de imágenes para poder responder adecuadamente a las preguntas de investigación planteadas.

Tipo de imagen	Total
Originales	260
Algoritmo 1	12,000
Algoritmo 2	12000
.	.
.	.

Tabla 4. Se muestran los conjuntos de datos utilizados. Primeramente, el conjunto de datos original, que contiene 260 imágenes, y cuando se aplicó el algoritmo 1 y 2 se obtuvieron 12,000 imágenes por cada algoritmo.

Por último, se experimenta con varios algoritmos de redes neuronales para entrenar los modelos, entre los cuales se incluyen: Compact, Enhanced, ResNet-50, MobileNetV2 y AlexNet.

4.3 Generación de imágenes sintéticas tradicionales

Las aplicaciones de aprendizaje automático, especialmente en el dominio del aprendizaje profundo, continúan diversificándose y aumentando rápidamente. Los enfoques centrados en datos para el desarrollo de modelos, así como las técnicas de aumento de datos, pueden ser una buena herramienta contra los desafíos que enfrenta el mundo de la inteligencia artificial, hablando de conjuntos de datos pequeños.

El aumento de datos es útil para mejorar el rendimiento y los resultados de los modelos de aprendizaje automático, mediante la creación de ejemplos nuevos y diferentes para entrenar conjuntos de datos. Si el conjunto de datos en un modelo de aprendizaje automático es grande, variable y suficiente, el modelo funciona mejor y con mayor precisión.

Se creó un procedimiento para poder obtener ejemplos de imágenes sintéticas tradicionales. Se aplicaron 4 transformaciones geométricas a la imagen original, las cuales fueron: rotación a 90°, rotación a 180°, rotación a 270° y efecto espejo. Adicionalmente se modificó el contenido de las imágenes sumando el cambio de contraste, cambio de brillo, y ruido sal y pimienta, obteniendo un total de 12000 imágenes sintéticas.

4.4 Generación de imágenes sintéticas poligonales

Ya mencionamos la importancia de la generación de imágenes sintéticas, y todas las aplicaciones que se le pueden dar dentro del aprendizaje automático. Sabemos también que pueden ser una buena herramienta contra los desafíos que enfrenta el mundo de la inteligencia artificial, ya que mediante estás

técnicas de aumento de datos se podrá mejorar el rendimiento y por ende los resultados obtenidos tendrán mayor precisión.

Para obtener las imágenes poligonales se creó un procedimiento y así se obtuvieron las imágenes sintéticas poligonales. Este algoritmo consiste precisamente en obtener formas poligonales, es decir convierte un polígono de una región de interés (ROI, por sus siglas en inglés) en una máscara binaria.

Primeramente, se genera un polígono en coordenadas polares, posteriormente el polígono se extrapola al plano cartesiano generando con ello la máscara binaria.

Empleando esta máscara se recorta primeramente el interior de una imagen, este procedimiento se aplica a 2 imágenes distintas y ambos resultados se combinan, el polígono puede variar en cada imagen.

Lo mismo sucede con la parte exterior de la máscara. En total se obtuvieron 12,0000 imágenes, con este nuevo conjunto de datos se entrenaron los 5 diferentes modelos, los cuales se abordaron en la sección 2.

4.5 Entrenamiento y prueba de los modelos de aprendizaje

Cómo ya se mencionó anteriormente entre los modelos que podemos entrenar con estas herramientas se encuentran: *Compact, Enhanced, ResNet-50, MobileNetV2 y AlexNet.* El entrenamiento se realizó utilizando estos 5 modelos, los modelos son redes pre-entrenadas que tienen algunas diferencias entre ellas.

Fase de entrenamiento: Para entrenar los diferentes modelos de redes neuronales se realizó los siguiente:

1. Primero se definen las imágenes que se van a usar para el entrenamiento.
2. Posteriormente se etiquetan las imágenes, para nuestro caso; se colocaron dos etiquetas una para células sanas y otra para células enfermas.
3. Posteriormente se definen los parámetros de entrenamiento de la red.
4. Se deben ajustar los valores predeterminados comunes utilizados para entrenar a los clasificadores.
5. Ejecutar el entrenamiento.
6. El último paso es evaluar los resultados, esto se puede hacer mediante las medidas de calidad y precisión.
7. Se obtendrá una matriz de confusión y un mapa de calor mostrando los detalles del rendimiento de los modelos.

Cabe recalcar que, para llevar a cabo el entrenamiento, se utilizaron varios parámetros específicos. Se seleccionaron conjuntos de 1000, 2500, 5000, 8000 y 12000 imágenes, de las cuales el 70% se utilizó para el entrenamiento, el 15% para la validación y el 15% para la prueba, en cada caso. El número de épocas se estableció en 2 para evitar el sobreajuste, y el número de iteraciones quedó en 145. La tasa de aprendizaje se estableció en .001, y la pérdida se estableció entre .05 y .005. El tiempo total requerido para llevar a cabo el entrenamiento fue de 17 a 31 minutos, y se realizó en una laptop con GPU NVIDIA Geforce GTX 1660 Ti y un procesador Intel Core i7 10th GEN.

De igual importancia, se mencionan las métricas utilizadas por los diferentes modelos, las cuales se detallan a continuación:

F1-Score: La puntuación F1, es la media armónica de precisión y recuperación. Y está dada por:

$$F_1 = \frac{2}{P^{-1} + R^{-1}} = 2 \cdot \frac{P \cdot R}{P + R} = \frac{\mathrm{TP}}{\mathrm{TP} + \frac{1}{2}(\mathrm{FP} + \mathrm{FN})}$$

Figura 33. Fórmula para obtener la métrica F1-Score.

Mean Precision: Es la proporción de todos los positivos predichos correctamente. Esto refleja cuán confiable es el modelo con respecto a la detección de muestras positivas. Y está dada por:

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

Figura 34. Fórmula para obtener la métrica Mean Precision.

Mean Recall: Es la proporción de todos los positivos predichos correctamente a todos los positivos reales.

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

Figura 35. Fórmula para obtener la métrica Mean Recall.

Por otra parte, la visión artificial se utiliza en todas las áreas exigentes de imágenes y permiten nuevas soluciones de automatización para el Internet industrial de las cosas al proporcionar tecnologías de alta gama como visión 3D, aprendizaje profundo y visión integrada. La visión artificial se desarrolla y diseña en la sede de Múnich, Alemania. Una herramienta de aprendizaje

profundo puede etiquetar fácilmente los datos gracias a la interfaz de usuario intuitiva.

Estos datos se pueden integrar perfectamente en otros *softwares* para realizar detección de objetos basada en aprendizaje profundo, clasificación, segmentación semántica, detección de anomalías y *OCR* profundo.

Cómo ya se mencionó anteriormente entre los modelos que podemos entrenar con estas herramientas se encuentran: *Compact, Enhanced, ResNet-50, MobileNetV2 y AlexNet.* El entrenamiento se realizó utilizando estos 5 modelos, los modelos son redes pre-entrenadas que tienen algunas diferencias entre ellas.

CAPÍTULO 5. RESULTADOS Y ANÁLISIS DEL TRABAJO DESARROLLADO

En este capítulo se discute acerca de las imágenes utilizadas para desarrollar el trabajo de tesis presentado, describiendo el resultado de los análisis obtenidos por los distintos modelos. También se menciona la generación de imágenes sintéticas y los métodos utilizados. Finalmente se discuten los resultados obtenidos, y se comparan por medio de los distintos métodos ejecutados, con la finalidad de dar respuesta a las preguntas de investigación planteadas al inicio de este trabajo.

5.1 Imágenes sintéticas tradicionales

Para la generación de imágenes sintéticas tradicionales se creó un procedimiento para poder obtener más imágenes. Se aplicaron 4 transformaciones geométricas a la imagen original, las cuales fueron: rotación a 90°, rotación a 180°, rotación a 270° y efecto espejo, adicionalmente se modificó el contenido de las imágenes sumando el ruido sal y pimienta, contraste, brillo e iluminación, obteniendo así un total de 12,000 imágenes sintéticas.

5.2 Imágenes sintéticas poligonales

Para obtener las imágenes poligonales se creó un procedimiento que consiste precisamente en obtener formas poligonales, es decir convierte un polígono de una región de interés (ROI) en una máscara binaria.

Primeramente, se genera un polígono en coordenadas polares, posteriormente el polígono se extrapola al plano cartesiano generando con ello la máscara binaria.

Empleando esta máscara se recorta primeramente el interior de una imagen, este procedimiento se aplica a 2 imágenes distintas y ambos resultados se combinan, el polígono puede variar en cada imagen. Lo mismo sucede con la parte exterior de la máscara. En total se obtuvieron 12,000 imágenes con este nuevo conjunto de datos y se entrenaron los 5 diferentes modelos explicados anteriormente.

5.3 Análisis de los modelos de aprendizaje

Como ya se mencionó anteriormente entre los modelos que se entrenaron, se encuentran: *Compact, Enhanced, ResNet-50, MobileNetV2 y AlexNet.*
El entrenamiento se realizó utilizando los 5 modelos, los cuales son redes pre-entrenadas que tienen algunas diferencias entre ellas.

5.3.1 Fase de entrenamiento:

Para entrenar los diferentes modelos de redes neuronales se realizó lo siguiente:

Primero se definen las imágenes que se van a utilizar para el entrenamiento, posteriormente se etiquetan las imágenes, para este caso; se colocaron dos etiquetas una para células sanas y otra para células enfermas.

Posteriormente se definen los parámetros de entrenamiento de la red. Se deben ajustar los valores predeterminados comunes utilizados para entrenar a los clasificadores. Ejecutar el entrenamiento.

El último paso es evaluar los resultados, esto se puede hacer mediante las medidas de calidad y precisión. Se obtendrá una matriz de confusión y un mapa de calor mostrando los detalles del rendimiento de los modelos.

Se entrenaron 5 modelos y se muestra un ejemplo del entrenamiento realizado con 5000 imágenes sintéticas tradicionales utilizando el modelo **MobileNetV2.**

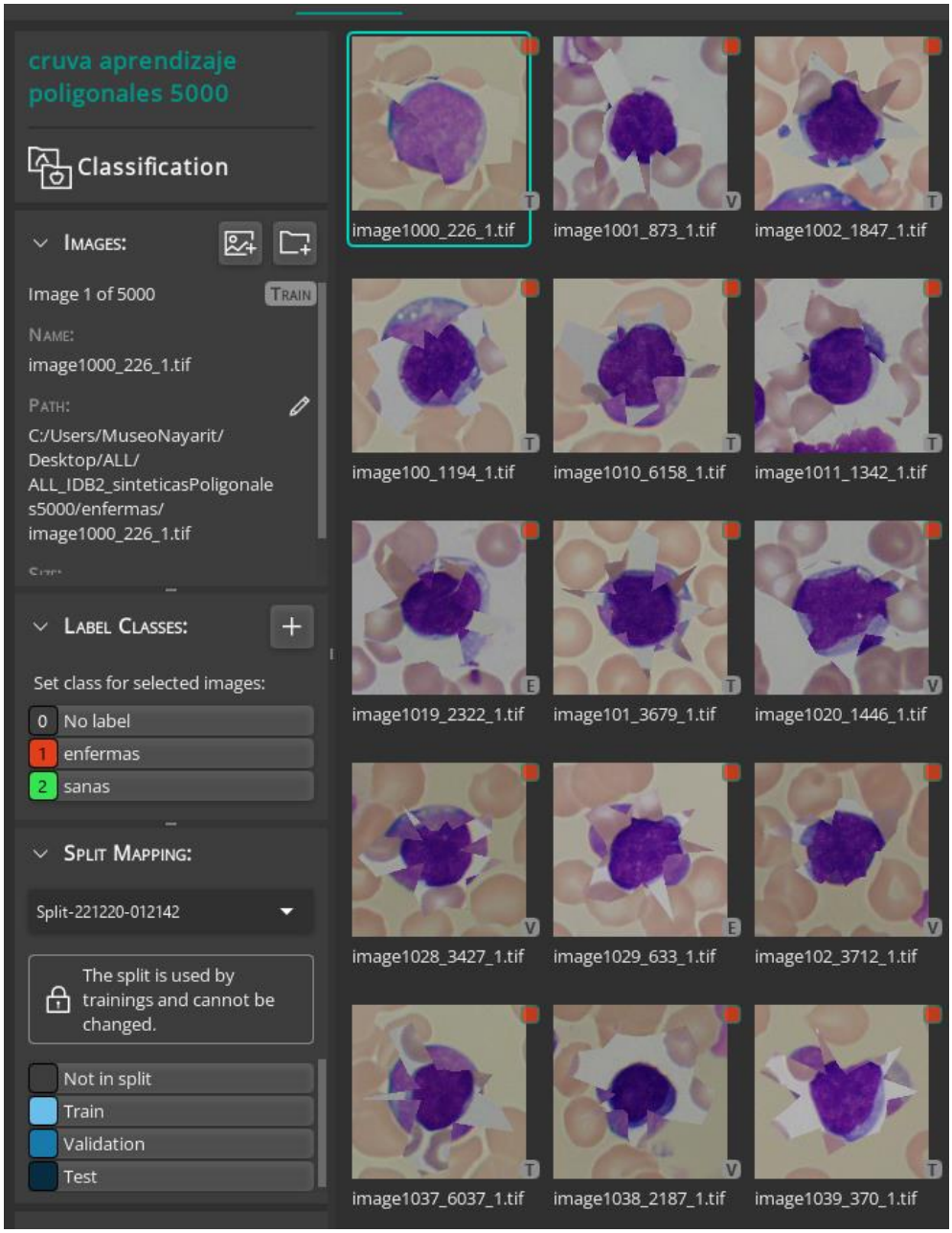


Figura 36. Ejemplos de imágenes etiquetadas como células enfermas. Aquí se definen las imágenes que se van a utilizar para el entrenamiento del modelo.

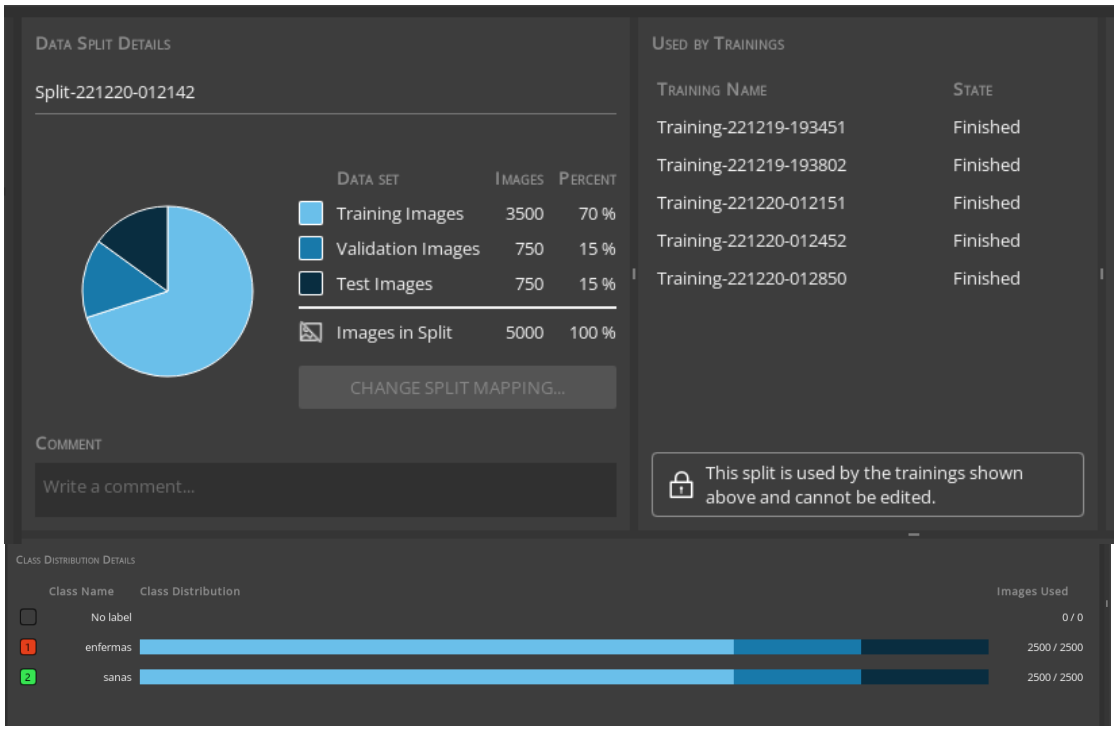


Figura 37. Creación de los parámetros del Split. Aquí se determinan los detalles de los valores que se utilizaran para entrenamiento, validación y prueba, y puede apreciar la distribución de etiquetas asignadas para células sanas y enfermas.

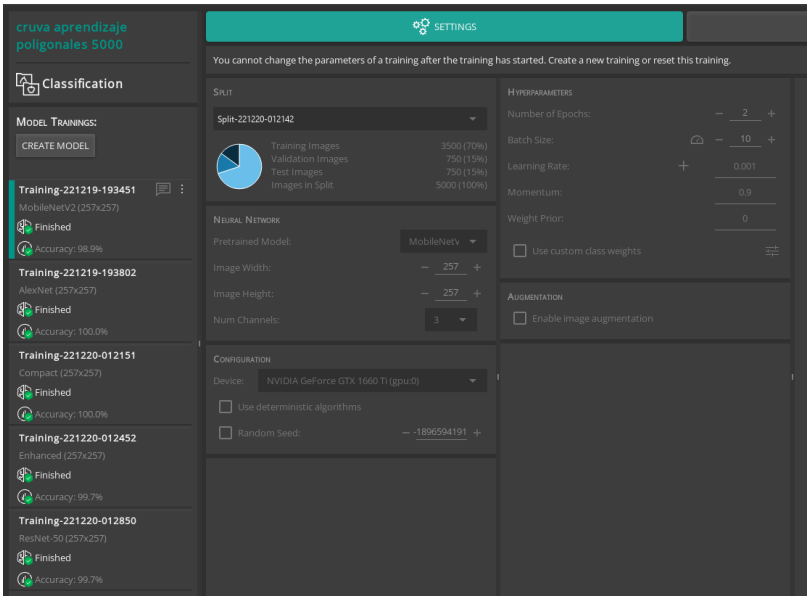


Figura 38. Parámetros e hiperparámetros que se asignaron para realizar el entrenamiento.

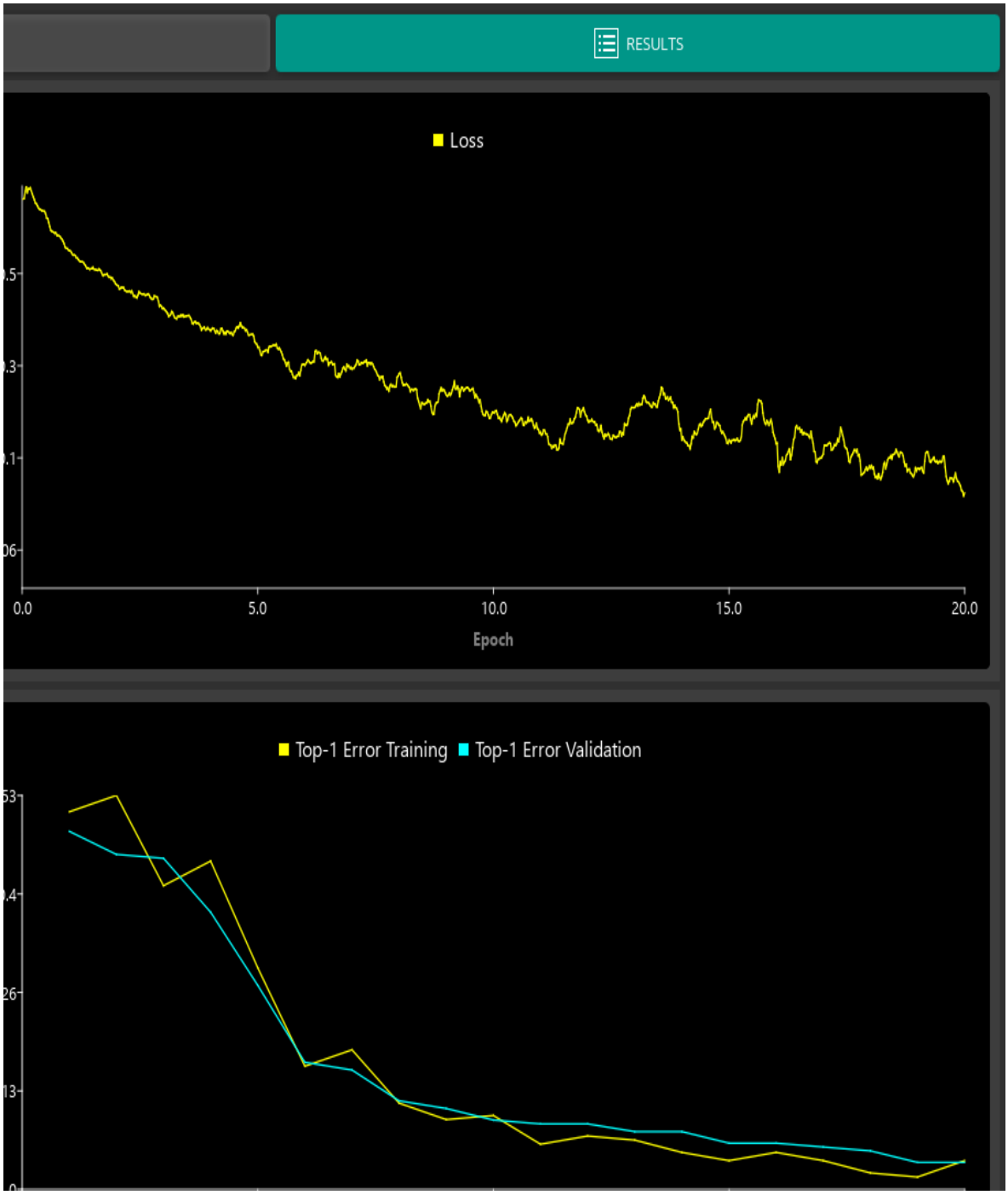


Figura 39. Función de pérdida, donde ambas gráficas tienden a cero, lo que indica que se realizó un buen entrenamiento.

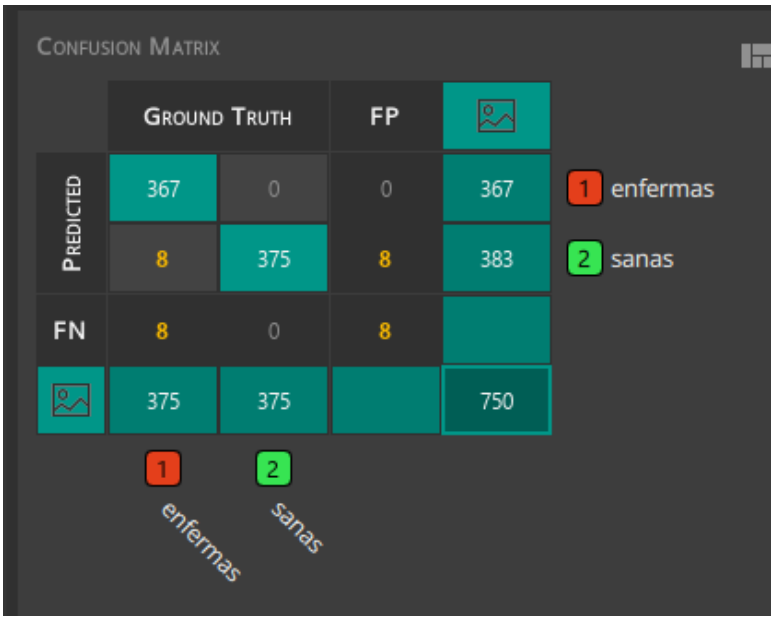


Figura 40. Matriz de confusión donde se puede observar la precisión que alcanzó el modelo entrenado.

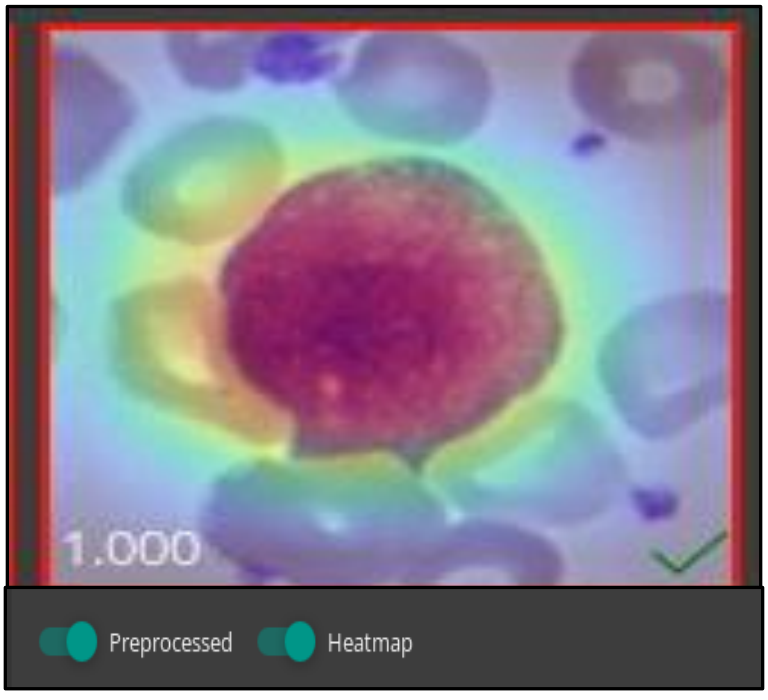


Figura 41. Mapa de calor donde se observa el área de interés que utilizó el modelo para llegar a esa clasificación.

Tipo de división	Porcentaje	Número de imágenes
Train	70.08 %	911
Validación	14.92 %	194
Prueba	15 %	195

Figura 42. Número de imágenes que utilizó el método para realizar el entrenamiento, la validación y la prueba.

Nombre del parámetro	Valor
Nombre	Formación-221120-185958
Modelo preentrenado	pretrained_dl_classifier_mobilenet_v2.hdl
Ancho	257
Altura	257
Número de canales	3
Número de épocas	20
Tamaño del lote	20
Tasa de aprendizaje	1E-04
Momento	0.9
Peso anterior	0
Dispositivo	NVIDIA GeForce GTX 1660 Ti (gpu:0)
Usar algoritmos deterministas	Deshabilitado
Semilla aleatoria	897851917

Figura 43. Parámetros utilizados para realizar el entrenamiento, como son: el tamaño de la imagen, el número de canales de la imagen, el número de épocas, entre otros.

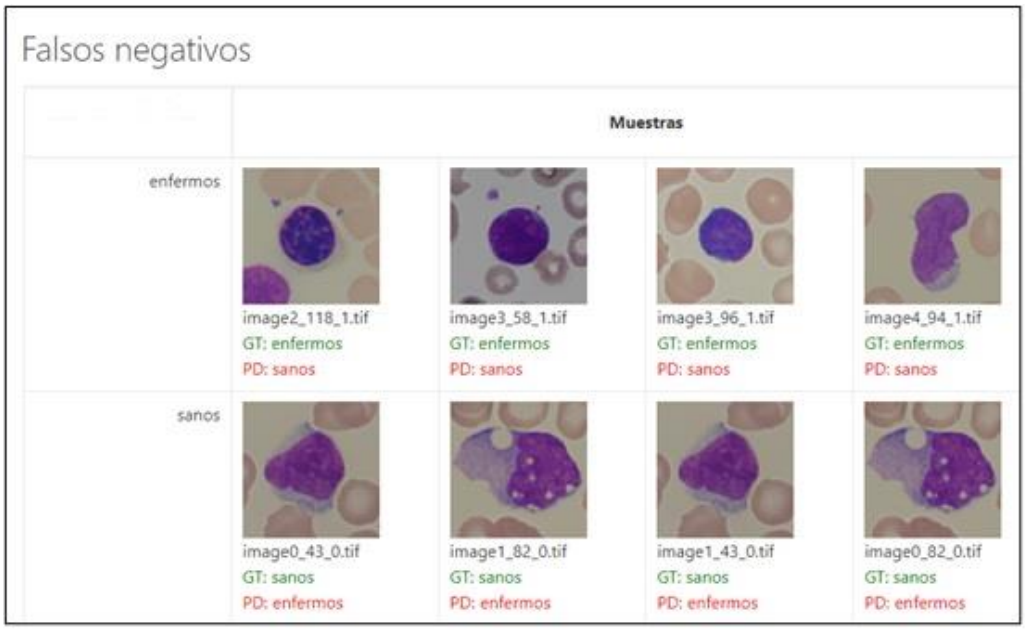


Figura 44. Aquí se observan los falsos negativos que presentó el entrenamiento.

Nombre del resultado	Valor
Exactitud	96.62 %
Tiempo de inferencia*	3,91 ms
Tiempo de preprocesamiento*	0,41 ms
Tiempo total*	4,32 ms
Top1-Error	3.38 %
Puntuación F1	96.62 %
Precisión	96.62 %
Recordar	96.62 %

Figura 45. Finalmente, se observan los resultados de las métricas de evaluación como son: exactitud, tiempo de inferencia, puntuación F1, y precisión.

5.4 Interpretación de los resultados del entrenamiento

El conjunto de datos utilizado para realizar el entrenamiento fue el siguiente: primero se entrenaron 260 imágenes originales, 12000 sintéticas tradicionales y 12000 sintéticas poligonales, las pruebas se realizaron con los 5 modelos, obteniendo los siguientes resultados.

En la tabla 5 se muestra la precisión obtenida por el entrenamiento realizado a las imágenes sintéticas tradicionales. Se realizaron entrenamientos con 1000, 2500, 5000, 8000 y 12000 imágenes sintéticas tradicionales.

Precisión obtenida con imágenes Sintéticas Tradicionales					
Modelo Entrenado	**1000 imágenes**	**2500 imágenes**	**5000 imágenes**	**8000 imágenes**	**12000 imágenes**
Compact	72.67%	88.50%	90.80%	92.25%	95.22%
Enhanced	84.67%	84.50%	95.07%	95.50%	97.39%
ResNet-50	70.67%	68.50%	81.74%	87.25%	96.72%
MobileNetV2	78.00%	75.67%	66.67%	73.67%	93.78%
AlexNet	57.33%	64.44%	58.90%	70.75%	74.44%

Tabla 5. Precisión obtenida por el entrenamiento de las imágenes sintéticas tradicionales.

En la siguiente figura se muestra la gráfica de los resultados de la precisión alcanzada durante los entrenamientos realizados a las imágenes sintéticas tradicionales con los 5 modelos. En donde se aprecia que la precisión más alta la obtuvo el modelo Enhanced, y la precisión más baja fue para el modelo AlexNet.

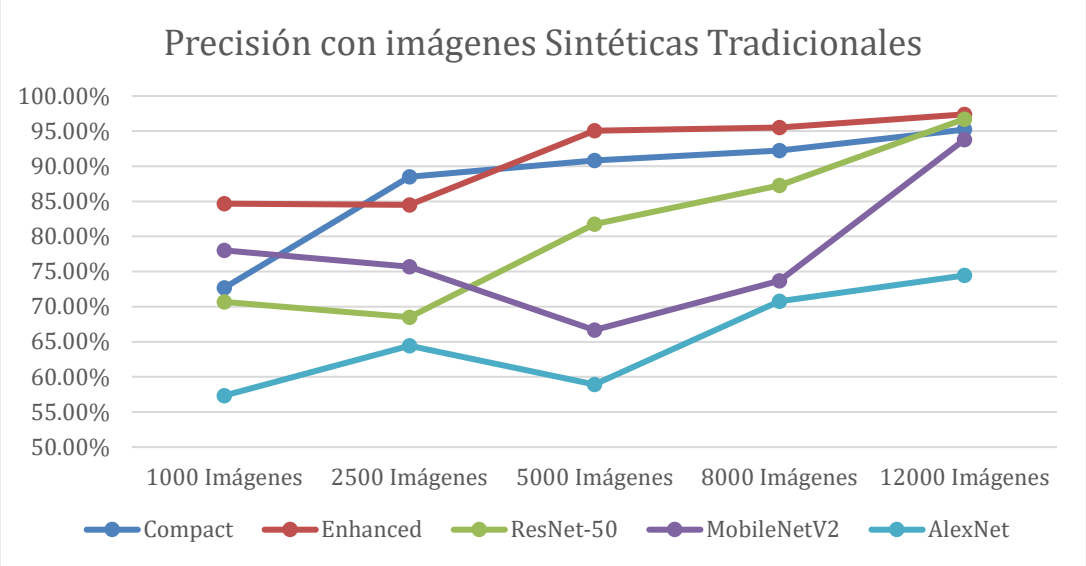


Figura 46. Precisión de imágenes sintéticas tradicionales

En la tabla 6 se muestra la precisión obtenida por el entrenamiento realizado a las imágenes sintéticas poligonales.

Precisión obtenida con imágenes Sintéticas Poligonales					
Modelo Entrenado	**1000 imágenes**	**2500 imágenes**	**5000 imágenes**	**8000 imágenes**	**12000 imágenes**
Compact	98.67%	98.40%	100.00%	99.92%	99.94%
Enhanced	98.00%	98.93%	99.73%	99.75%	100.00%
ResNet-50	96.00%	98.93%	99.73%	99.92%	99.94%
MobileNetV2	96.00%	95.45%	98.90%	99.58%	100.00%
AlexNet	96.00%	96.26%	100.00%	100.00%	100.00%

Tabla 6. Precisión de las imágenes sintéticas poligonales.

En la siguiente figura se muestra la gráfica de los resultados de la precisión alcanzada durante los entrenamientos realizados a las imágenes sintéticas poligonales, en donde se aprecia que la precisión más alta la obtuvo el modelo AlexNet, y la precisión más baja fue para el modelo MobileNetV2.

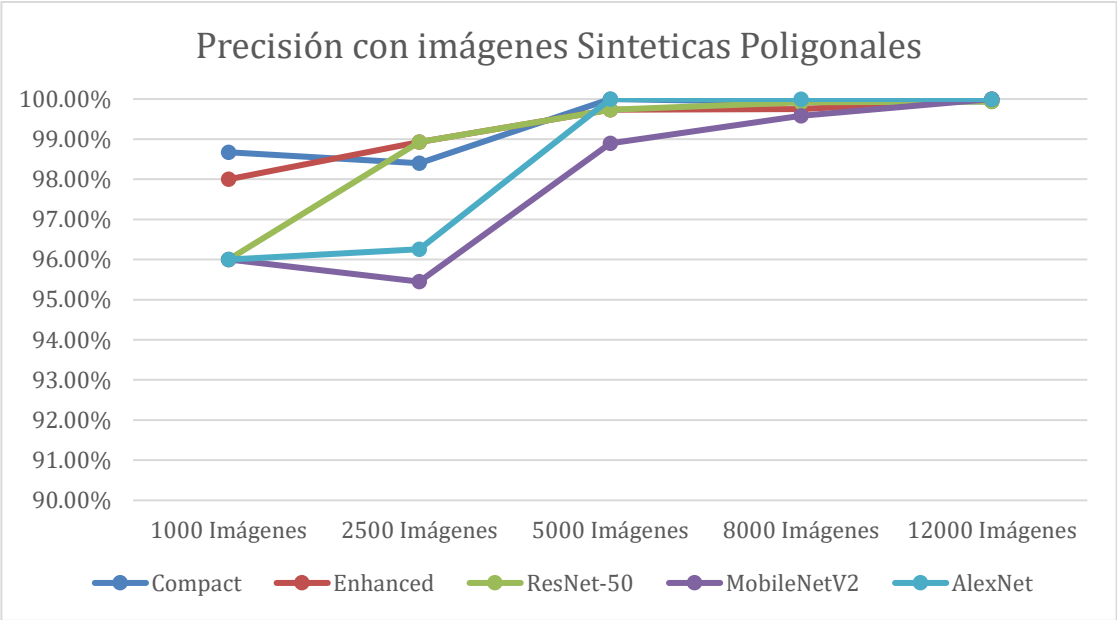


Figura 47. Gráfica del resultado de las imágenes sintéticas poligonales.

5.5 Comparación de los resultados

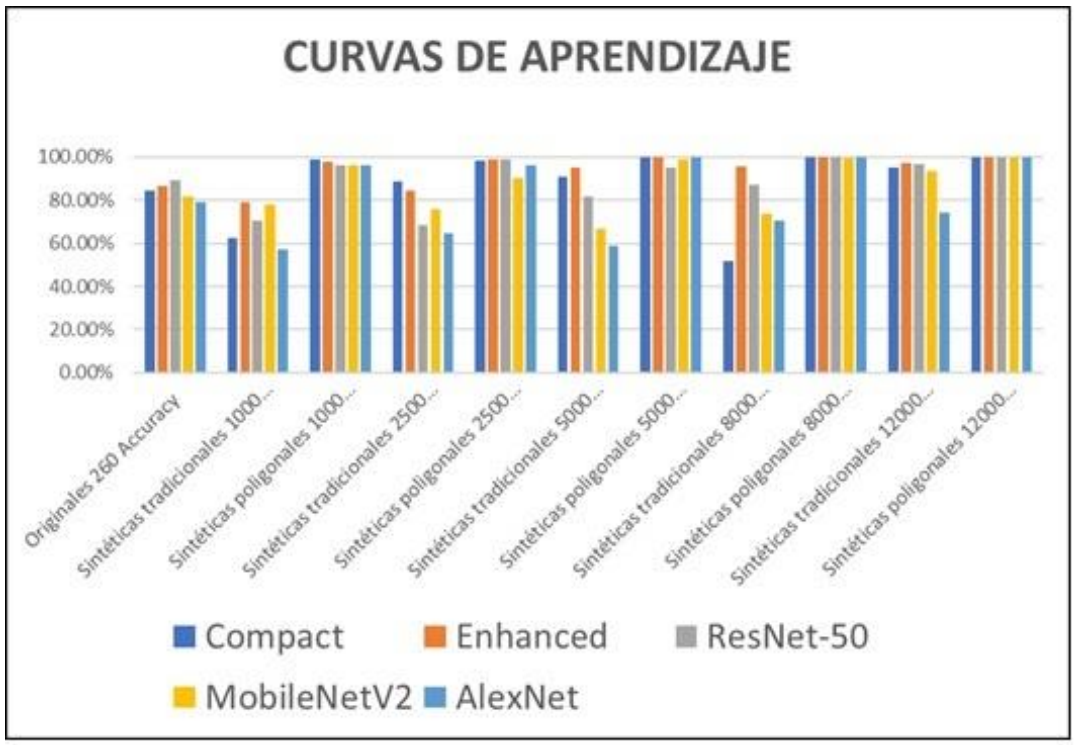


Figura 48. Gráfica comparativa de la precisión de los 5 modelos.

5.6 Análisis

Una vez obtenidos los resultados de los entrenamientos realizados con los distintos modelos, se puede responder a las preguntas de investigación.

¿Es posible aumentar el desempeño de los modelos de Deep Learning utilizando imágenes sintéticas en el área hematológica? Se lograron resultados muy positivos y se demostró que con la utilización de imágenes sintéticas se aumenta la precisión para el reconocimiento de células.

¿En qué porcentaje se mejora el desempeño de los modelos con el uso de imágenes sintéticas? A partir de los porcentajes obtenidos podemos concluir que los 5 modelos resultan eficaces para realizar el reconocimiento de células sanguíneas, sin embargo, con las imágenes poligonales es con las que se logra una mayor precisión. Por ejemplo, para el conjunto de 5,000 imágenes, en las imágenes tradicionales los porcentajes van de 58.90% al 95.07%, mientras que con las imágenes poligonales los porcentajes se sitúan entre 98.90% y 100%.

¿Cuál es el número máximo de imágenes sintéticas para lograr un desempeño óptimo? Con base en los resultados obtenidos se deduce que, el valor óptimo para entrenar las redes con imágenes sintéticas poligonales se sitúa entre 2500 y 5000 imágenes.

CAPÍTULO 6. CONCLUSIONES

6.1 Conclusiones

En este trabajo de tesis, se experimentó con técnicas de aumento de datos para la generación de imágenes sintéticas, logrando realizar el reconocimiento de células sanguíneas mediante aprendizaje profundo.

A partir de los resultados obtenidos se puede concluir que, de los 5 modelos probados, los modelos Compact y Enhanced obtuvieron la precisión más alta, superando significativamente a los otros 3 modelos.

Se observa que con las imágenes sintéticas tradicionales el modelo AlexNet obtuvo la precisión más baja, mientras que el modelo Enhanced alcanzó la precisión más alta. Y en las imágenes sintéticas poligonales el modelo MobileNet V2 fue el más bajo y el modelo AlexNet el más alto.

Se deduce que el valor óptimo de imágenes generadas con el método poligonal para entrenar las redes se sitúa entre 2500 y 5000 imágenes para los 5 modelos. Por ello, el aumento de datos es una excelente opción para hacer viables proyectos que con datos reducidos no podrían desarrollarse.

Se determinó que, al aumentar el número de imágenes de la base de datos mediante distintos algoritmos se incrementa el porcentaje de precisión de la red. Y cuando la función de pérdida tiende a cero el entrenamiento mostrará un resultado óptimo.

También se comprobó que cuando se utilizan muchas imágenes sintéticas, el error tiende a no bajar más, y eso provoca pérdida de precisión, por lo que se determina que cuando en la gráfica las líneas de errores se crucen, será el momento perfecto para no seguir aumentando imágenes al entrenamiento.

El objetivo general y los objetivos específicos se cumplieron satisfactoriamente.

6.2 Trabajo futuro

Como trabajo futuro se propone realizar las siguientes actividades:

- Probar otros métodos para generar imágenes sintéticas y compararlos con los resultados obtenidos en esta tesis, con el fin de conocer la cantidad de imágenes necesarias para cada nuevo modelo entrenado.

- Investigar y aplicar la técnica de adaptación de dominio para mejorar aún más el rendimiento de los modelos.

Referencias

[1] Jordi Torres. Data Augmentation y Transfer Learning. https://torres.ai/data-augmentation-y-transfer-learning-en-keras-tensorflow/. Septiembre 2019.

[2] Pol Bertran Prieto. Las 10 enfermedades sanguíneas más comunes. https: //medicoplus.com/medicina-general/enfermedades-sanguineas-comunes. 2023.

[3] Gerard Andrews. ¿Qué son los datos sintéticos? https://la.blogs.nvidia.com/2021/07/20/que-son-los-datos-sinteticos. Julio 2021.

[4] José de Jesús Velázquez Arreola. Identificación de peatones en imágenes aéreas con redes neuronales explicativas y fusión de sensores. Tesis de maestría, febrero 2019.

[5] A. Hernández Merino. Anemias en la infancia y adolescencia. Clasificación y diagnóstico. Volumen XX, Número 5. https://www.pediatriaintegral.es/wp-content/uploads/2016/07/Pediatria-Integral-XX-05_WEB.pdf#page=7. Julio 2016.

[6] E. Clemente Lirola. Anemias. https://www.elsevier.es/es-revista-medicina-familia-semergen-40-pdf-S1138359303742543. Semergen 2003.

[7] Ekin Tiu. Understanding Latent Space in Machine Learning. https://towardsdatascience.com/ understanding-latent-space-in-machine-learning - de5a7c687d8d. Feb 4, 2020.

[8] El mundo salud, leucemias agudas, tipos de leucemias. http://www.elmundo.es/elmundosalud/especiales/cancer/leuc agudas2.html. fecha de consulta: 27 de mayo, 2021.

[9] Hematología. Leucemia y trastornos mieloproliferativos. http://www.altillo.com/medicina/monografias/leucemia.asp. fecha de consulta: 15 de abril, 2021.

[10] Melanie Re. Anemia microcítica: causas, síntomas y tratamiento. https: //www.onsalus.com/anemia-microcitica-causas-sintomas-y-tratamiento-19470.html. mayo 2021.

[11] Evan M. Braunstein, MD, PhD, Johns Hopkins. Anemias macrocíticas megalo-blásticas. University School of Medicine. Revisado médicamente, https://www.msd- manuals.com/es-mx/professional/hematolog %C3 %ADa-y-onco log % C3 % Ada /anemias-causadas-por-deficiencia-de-la-

eritropoyesis/anemias-macroc% C3% AD ticas -megalobl% C3%A1sticas#:~:text = Cuando % 20 est % C 3% A1% 20 com-mpletamente % 20 desarrollada % 2C% 20la, la % 20forma% 20 de% 20los %20- eritrocitos. septiembre 2021.

[12] A. Batlle, J. Núñez, C. Montes Gaisán, A. Insunza. Protocolo diagnóstico de las anemias normocíticas. Volumen 11, Número 20, páginas 1238-1241. noviembre 2012.

[13] Gary H. Gibbons, M.D. Trastornos plaquetarios. Trombocitopenia. National Heart, Lung, and Blood Institute. https://www.nhlbi.nih.gov/es/salud/trombocitopenia. Julio 2022.

[14] Samuel Antonio Sánchez Amador. Trombocitopenia: síntomas, causas y tratamiento. https://psicologiaymente.com/salud/trombocitopenia. Junio 2021.

[15] Generative Adversarial Networks for Text Generation — Part 1 | by Karthik Chintapalli | Becoming Human: Artificial Intelligence Magazine. 2021.

[16] Siddharth Das. CNN Architectures: LeNet, AlexNet, VGG, GoogLeNet, ResNet and more… https://medium.com/analytics-vidhya/cnns-architectures-lenet-alexnet-vgg-googlenet-resnet-and-more-666091488df5. Noviembre 2017.

[17] Gregor Guncar, Matjaz Kukar, et al. An application of machine learning to haematological diagnosis. Scientific Reports. https://medium.com/analytics-vidhya/cnns-architectures-lenet-alexnet-vgg-googlenet-resnet-and-more-6660919-1488df5. January 2018.

[18] H. Molina “Clasificadores para el reconocimiento automático de células blásticas en leucemias agudas linfoides y mieloides. Universitat Oberta de Catalunya. España 2018.

[19] R. Tambe, S. Mahajan, U. Shah, M. Agrawal, and B. Garware, “Towards Designing an Automated Classification of Lymphoma subtypes using Deep Neural Networks,” in *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, Kolkata, India, 2019, pp. 143–149.

[20] Fahad Kamal Alsheref, Wael Hassan Gomaa. Detección de enfermedades de la sangre mediante algoritmos clásicos de aprendizaje automático. International Journal of Advanced Computer Science and Applications (IJACSA), volumen 10, número 7, 2019.

[21] M. Font, “Clasificación automática de linfocitos anormales procedentes de Linfomas de baja prevalencia utilizando few-shot learning,” 2020.

[22] Meifang Wang, Chunxia Dong, et al. Un modelo de aprendizaje profundo para el reconocimiento automático de anemia aplásica, síndromes mielodisplásicos y leucemia mieloide aguda basada en frotis de médula ósea. Volumen 12, 2022.

[23] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, Youngjoon Yoo Cut Mix: Regularization Strategy to Train Strong Classifiers with Localizable Features. https://arxiv.org/search/cs?searchtype=author&query=Yun%2C+S. 2020.

[24] Jie Qin 1,2,1, Jiemin Fang 3,4,1, et al. ResizeMix: Mixing Data with Preserved Object Information and True Labels. 2020.

[25] Ethan Harris, Antonia Marcu, Matthew Painter, Mahesan Niranjan, Adam Prügel-Bennett, Jonathon Hare. FMix: Enhancing Mixed Sample Data Augmentation. 2020.

[26] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez Paz. Mixup: Beyond empirical risk minimization. . arXiv preprint arXiv:1710.09412, 2017.

[27]Fabio Scotti. ALL-IDB2. Acute Lymphoblastic Leukemia Image Database for Image Processing. Department of Computer Science – Universit’a degli Studi di Milano. 2010.

Anexos

A. Ejemplos de imágenes sintéticas tradicionales

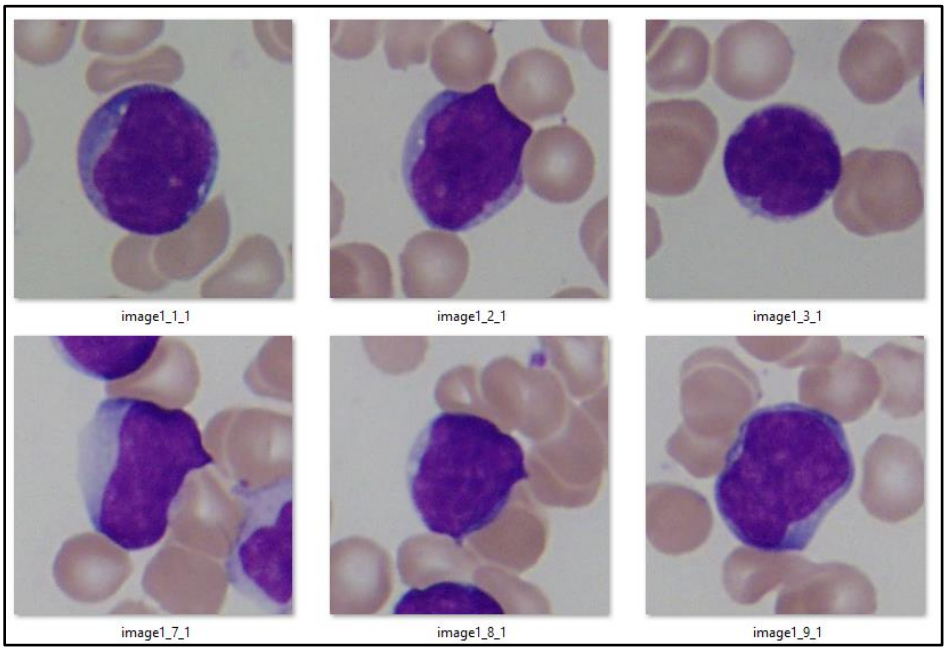


Figura 49. Ejemplos de imágenes sintéticas tradicionales de células enfermas

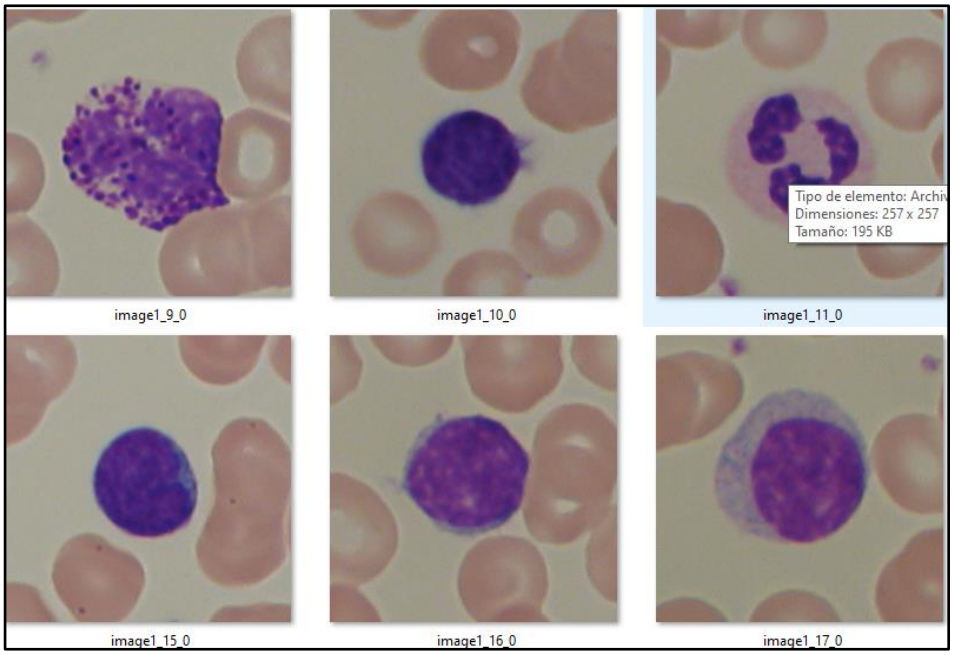


Figura 50. Ejemplos de imágenes sintéticas tradicionales de células sanas

B. Ejemplos de imágenes sintéticas poligonales

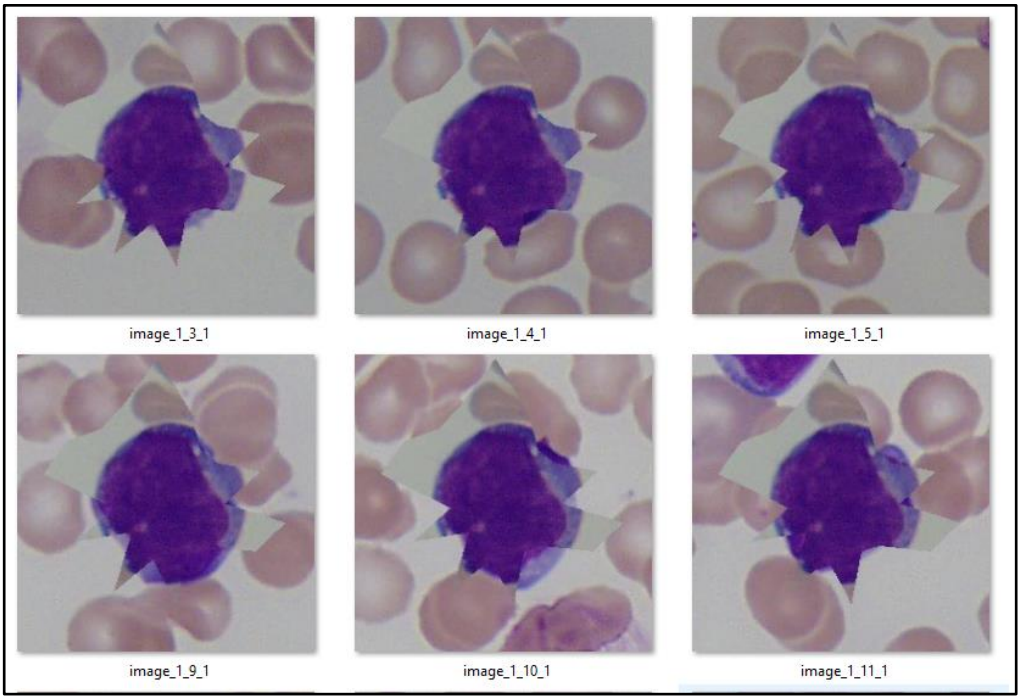


***Figura 51**. Ejemplos de imágenes sintéticas poligonales de células enfermas*

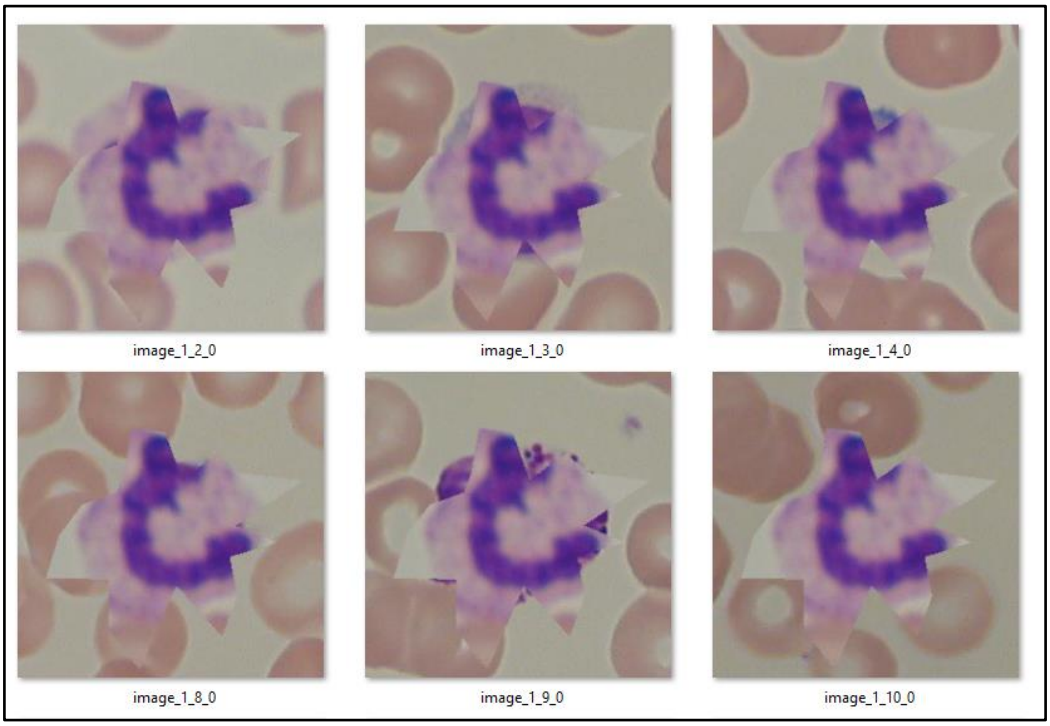


***Figura52**. Ejemplos de imágenes sintéticas poligonales de células sanas*

C. Glosario

Ajuste de hiperparámetros: Ejecutar varios entrenamientos con diferentes conjuntos de hiperparámetros para encontrar el conjunto que conduce al mejor error de validación.

Ajuste inferior: Se produce cuando el modelo generaliza demasiado. En otras palabras, no es capaz de describir la complejidad de la tarea. Esto se refleja directamente en el error en el conjunto de entrenamiento, que no disminuye significativamente.

Época: Hiperparámetro; Una iteración completa sobre todos los datos de entrenamiento Como el número total de muestras que un modelo "ve" durante el entrenamiento depende del número de iteraciones y del tamaño del lote, este valor permite comparar dos modelos con diferentes tamaños de lote. El número requerido de épocas depende del problema individual. Cuanto más complejo sea el problema, más tiempo necesitará el entrenamiento para obtener buenos resultados. Por lo tanto, es beneficioso aumentar el valor.

Hiperparámetro: Variable establecida manualmente con un valor predeterminado que no se optimiza durante el entrenamiento, por ejemplo, tasa de aprendizaje o tamaño de lote.

Inferencia: A grandes rasgos, la inferencia consiste en poner en práctica lo que el modelo ha aprendido en el entrenamiento. Una vez que el modelo aprende se utilizará para resolver y / o clasificar datos con respecto al problema.

Iteraciones: Número de muestras dividido por el tamaño del lote (redondeado). Como solo se procesa simultáneamente un lote de muestras, se necesitan varias iteraciones para procesar todo el conjunto de datos. El número de iteraciones necesarias para procesar todos los datos una vez está determinado por el número total de muestras y cuántas muestras se procesan simultáneamente (por lo tanto, el tamaño del lote).

Mapa de calor: Medida de evaluación; Muestra qué partes de una imagen tienen una fuerte influencia en la inferencia en una determinada clase.

Matriz de confusión: Medida de evaluación; Tabla que compara las afiliaciones de clase previstas y la verdad básica.

Modelo de aprendizaje profundo: En HALCON, este término comprende métodos que utilizan una red neuronal con múltiples capas ocultas. Los modelos se pueden almacenar como archivos "*HALCON Deep Learning" (HDL)* y *HDevelop* los puede leer.

Momento: Hiperparámetro, abreviado μ; utilizado dentro de la optimización de la función de pérdida. Cuando se actualizan los pesos (después de haber calculado el gradiente), se añade una fracción del vector de actualización anterior (del paso de iteración anterior). Esto tiene el efecto de amortiguar las oscilaciones. Cuando se establece en 0, el método de momento no tiene influencia. En palabras simples, cuando actualizamos los argumentos de la función de pérdida, todavía recordamos el paso que hicimos para la última actualización. Ahora vamos un paso en la dirección del gradiente con una longitud acorde a la tasa de aprendizaje y adicionalmente repetimos el paso que hicimos la última vez, pero esta vez solo dos veces más largo.

Pérdida: Función optimizada durante el proceso de entrenamiento para adaptar la red a una tarea específica. Compara la predicción de la red con la información dada de lo que debe encontrar en la imagen y penaliza las desviaciones.

Puntuación F1: Medida de evaluación; medio armónico de precisión y recuerdo; Representa la precisión y el recuerdo con un solo número.

Precisión: Medida de evaluación; Proporción de todos los positivos predichos correctamente ("verdaderos positivos") a todos los positivos previstos (verdaderos positivos y falsos positivos).

Recall: Medida de evaluación; proporción de todos los positivos predichos correctamente (verdaderos positivos) a todos los positivos reales (verdaderos positivos y falsos negativos); También se llama "tasa positiva verdadera" o "sensibilidad".

Semilla aleatoria: Número utilizado para inicializar el generador de números (pseudo) aleatorios.

Sobreajuste: Ocurre cuando la red comienza a "memorizar" datos de entrenamiento en lugar de aprender a encontrar reglas generales para la clasificación. Esto se hace visible cuando el modelo continúa minimizando el error en el conjunto de entrenamiento, pero el error en el conjunto de validación aumenta.

Tamaño del lote: Hiperparámetros; Cada conjunto de datos se divide en subconjuntos más pequeños de datos, denominados lotes. Durante el entrenamiento, las muestras dentro de un lote se seleccionan al azar. El tamaño del lote determina el número de imágenes tomadas en un lote y, por lo tanto, procesadas simultáneamente. Es recomendable establecer el tamaño del lote lo más alto posible para la memoria de GPU disponible.