**Clasificación de lesiones cutáneas con *deep learning* y late fusion: comparación de ResNet-18, MobileNetV3-Small y EfficientNet-B0**

**Introducción**

La detección temprana de cáncer de piel es un objetivo fundamental en la dermatología moderna, pues una identificación a tiempo puede marcar la diferencia entre un tratamiento sencillo y una enfermedad mortal. En los últimos años, el **aprendizaje profundo** (*deep learning*) ha demostrado una gran capacidad para extraer patrones complejos de imágenes médicas. En este trabajo describo, desde cero y de forma detallada, cómo entrené, validé y comparé tres arquitecturas de redes neuronales convolucionales (CNN): **ResNet-18**, **MobileNetV3-Small** y **EfficientNet-B0**, con el fin de clasificar once tipos de lesiones cutáneas a partir de imágenes clínicas y dermatoscópicas.

Un elemento central de este proyecto es la estrategia de **late fusion**, que permite tomar una decisión diagnóstica **a nivel de lesión** cuando existen varias imágenes por caso. Esto refleja mejor la práctica clínica y aumenta la estabilidad de las predicciones. A lo largo de este documento explico los fundamentos teóricos de cada modelo, cómo se implementó la fusión, los pasos de entrenamiento y las métricas obtenidas, para finalmente justificar la elección del mejor modelo para participar en la competencia **MILK10k Challenge**.

**Datos utilizados y preprocesamiento**

El conjunto de datos **MILK10k** incluye imágenes clínicas y dermatoscópicas de distintas lesiones cutáneas. Cada lesión se representa con, al menos, **dos imágenes** diferentes, lo que proporciona información complementaria. Las clases abarcan once categorías, entre ellas carcinoma basocelular (BCC), melanoma (MEL) y queratosis seborreica (BKL), entre otras.

Las imágenes originales tienen una resolución de aproximadamente **600×450 píxeles**. Para entrenar modelos preentrenados en **ImageNet**, redimensioné cada imagen a **224×224 píxeles**. Apliqué normalización con las medias y desviaciones estándar de ImageNet. En el conjunto de entrenamiento usé aumentos de datos como giros horizontales, rotaciones de ±10 grados y ligeros cambios de brillo y contraste, con el objetivo de que los modelos aprendan a reconocer patrones robustos y no memoricen únicamente las imágenes originales.

Dividí el conjunto de entrenamiento en dos: **entrenamiento** y **validación**, asegurándome de que ninguna lesión apareciera en ambos subconjuntos. De esta forma, la evaluación en validación mide la capacidad real de generalización y no solo la memoria del modelo.

**Fundamentos teóricos de los modelos utilizados**

**ResNet-18**

ResNet-18 es una red de 18 capas que introdujo un avance clave: los **bloques residuales**. Estos bloques incluyen **conexiones de salto** (*skip connections*) que permiten que la señal de entrada se sume directamente a la salida del bloque. Matemáticamente, cada bloque calcula:

donde es la transformación aprendida por el bloque (dos convoluciones de 3×3 con normalización y ReLU). Esta estructura evita el problema del **desvanecimiento del gradiente**, permitiendo redes más profundas sin degradar el aprendizaje.  
En mi implementación, ResNet-18 sirvió como **modelo base**: es estable y rápido de entrenar, pero su menor capacidad de representación lo hizo menos competitivo frente a arquitecturas más modernas.

**MobileNetV3-Small**

MobileNetV3-Small fue diseñada para ser **eficiente y ligera**, ideal para dispositivos con recursos limitados. Su arquitectura combina varios elementos:

* **Convoluciones separables en profundidad** (*depthwise separable convolutions*): reducen drásticamente el número de operaciones al separar la convolución espacial de la combinación de canales.
* **Bloques Inverted Residual (IR)**: primero expanden el número de canales, luego aplican la convolución depthwise y finalmente proyectan a un espacio más pequeño, manteniendo un camino de identidad cuando las dimensiones coinciden.
* **Atención Squeeze-and-Excitation (SE)**: recalibra cada canal según su importancia, realzando las características más relevantes.
* **Activación h-swish**: una variante eficiente de la función swish que mejora la no linealidad sin aumentar el costo computacional.

Esta combinación permite a MobileNetV3-Small lograr una gran relación entre precisión y consumo de recursos. En este proyecto logró un **F1-macro** mayor que ResNet-18 y una exactitud muy cercana a EfficientNet-B0, mostrando que la eficiencia no implica sacrificar demasiado la calidad.

**EfficientNet-B0**

EfficientNet-B0 es una arquitectura de nueva generación que parte de una idea fundamental: el **escalamiento compuesto** (*compound scaling*). En lugar de aumentar de forma independiente la profundidad, el ancho o la resolución, EfficientNet los escala de forma coordinada usando un solo coeficiente, buscando el mejor equilibrio entre capacidad de representación y costo de cómputo.

Su bloque básico, **MBConv**, es una evolución del Inverted Residual, con atención SE y relaciones de expansión bien calibradas. Este diseño permite capturar patrones de textura y color con gran detalle y de manera eficiente. En mi experiencia, EfficientNet-B0 consiguió el **mejor balance global**, con una alta capacidad de generalización incluso frente al desbalance de clases.

**Entrenamiento y optimización**

Los tres modelos se inicializaron con pesos preentrenados en **ImageNet**, reemplazando su capa final por una capa totalmente conectada con once salidas. Para la optimización utilicé **Adam** con una tasa de aprendizaje de 0.001 y la función de pérdida **CrossEntropyLoss**, que compara las salidas del modelo con las clases reales:

donde es el logit de la clase verdadera. Para acelerar el cálculo y reducir el consumo de memoria implementé **precisión mixta** con autocast y GradScaler. Además, empleé un planificador ReduceLROnPlateau que reduce la tasa de aprendizaje cuando la métrica de validación deja de mejorar.

**Evaluación con late fusion**

Un elemento central de este trabajo es la estrategia de **late fusion**. Cada lesión tiene al menos dos imágenes, por lo que evaluar cada imagen por separado no representa la realidad clínica. En lugar de combinar las imágenes al inicio (*early fusion*), permito que cada una pase individualmente por la red y genere su propio vector de logits. Después, en la fase de decisión, **promedio los logits de todas las imágenes de la misma lesión**:

y la predicción final de la lesión es la clase con mayor valor en . Esta estrategia reduce el impacto de imágenes borrosas o mal iluminadas y garantiza una decisión única y robusta por lesión.

**Resultados experimentales**

Tras el entrenamiento, evalué los modelos en el conjunto de validación con late fusion. Los resultados globales fueron los siguientes:

* **ResNet-18**: F1-macro **0.301**, F1-weighted **0.635**, accuracy **0.675**.
* **MobileNetV3-Small**: F1-macro **0.441**, F1-weighted **0.702**, accuracy **0.721**.
* **EfficientNet-B0**: F1-macro **≈0.47**, F1-weighted **≈0.72**, accuracy **≈0.73**.

Las matrices de confusión mostraron que todas las redes reconocen bien las clases abundantes, como BCC, con recall superior al 90 %. Sin embargo, las clases poco representadas (BEN\_OTH, DF, INF, MAL\_OTH, VASC) siguen siendo difíciles. En ResNet-18 algunas de ellas ni siquiera obtuvieron predicciones positivas, mientras que MobileNetV3-Small y EfficientNet-B0 ofrecieron una ligera mejora.

EfficientNet-B0 también logró mejores resultados en clases visualmente complejas, como BKL y MEL, donde la textura y los bordes finos son determinantes. Su arquitectura MBConv con atención SE parece capturar mejor estas sutilezas.

**Discusión**

Los resultados reflejan la teoría detrás de cada modelo. **ResNet-18**, con su enfoque residual, es excelente para entrenar redes profundas sin perder gradiente, pero su profundidad y ancho limitados lo hacen menos potente para problemas muy desbalanceados. **MobileNetV3-Small**, gracias a sus convoluciones separables y a la atención SE, demuestra que la eficiencia puede ir de la mano de una buena precisión, siendo ideal para entornos con recursos limitados. **EfficientNet-B0**, al equilibrar de manera óptima profundidad, ancho y resolución, obtiene la mejor capacidad de generalización, lo que se traduce en el mayor F1-macro.

Además, la **late fusion** demostró ser un componente crucial: al combinar las salidas de todas las imágenes de cada lesión, el modelo se vuelve más robusto frente a variaciones en iluminación, ángulo o calidad de las fotos, reflejando mejor cómo un dermatólogo integra varias vistas antes de dar un diagnóstico.

**Conclusiones y decisión final**

El análisis conjunto de las métricas muestra que **EfficientNet-B0** es el modelo más equilibrado y confiable para este problema. Su desempeño superior en F1-macro y en *accuracy* a nivel de lesión indica que no solo acierta más, sino que lo hace de manera más homogénea entre las distintas clases, algo fundamental en medicina.

Por estas razones, las predicciones que enviaré al **MILK10k Challenge** se basarán en **EfficientNet-B0**. Su arquitectura moderna, su capacidad de extraer patrones finos y la solidez de la estrategia de **late fusion** ofrecen la mejor oportunidad para obtener un resultado competitivo en el conjunto de prueba oficial.