Comparación empírica de clasificadores neuronales One-vs-All y multi-salida: rendimiento, coste y robustez arquitectónica

Isabel de la Higuera[1 [0009−0004−6668−3403]], Nicolás Segura-Kunsagi[1 [0009−0002−1413−2277]], Francisco Herrera[1,2 [0000−0002−7283−312X]] y María Victoria Velasco[3 [0000−0003−4957−3275]]

[1] Instituto Andaluz Interuniversitario en Ciencia de Datos e Inteligencia Computacional (DaSCI)
[2] Departamento de Ciencias de la Computación e Inteligencia Artificial (DECSAI)
[3] Departamento de Análisis Matemático, Facultad de Ciencias
Universidad de Granada, Granada, España
isadelahiguera@correo.ugr.es {nkunsagi,vvelasco}@ugr.es
herrera@decsai.ugr.es

Resumen Este trabajo estudia la equivalencia práctica entre clasificadores neuronales multi-salida y clasificadores One-vs-All (OVA). La motivación parte de resultados teóricos recientes que muestran que, bajo una formulación funcional de interpolación de norma mínima, el problema de aprendizaje asociado a una red neuronal multi-salida de arquitectura fija puede descomponerse en problemas escalares, uno por cada componente de salida. Sin embargo, esta descomposición teórica no determina por sí sola el comportamiento empírico de modelos entrenados con datos finitos, funciones de pérdida distintas, restricciones arquitectónicas y costes computacionales reales. Se presenta una evaluación experimental sobre conjuntos clásicos de aprendizaje automático, tareas de clasificación de imágenes y escenarios de imagen biomédica. La comparación considera rendimiento predictivo, tiempo de entrenamiento, paralelización y robustez frente a reducciones arquitectónicas, utilizando F1 macro y tiempo efectivo de entrenamiento como criterios principales. Los resultados muestran que OVA puede ser prácticamente equivalente a los modelos multi-salida en varios conjuntos, mientras que ofrece ventajas claras en casos como Digits y BRISC. En el conjunto BRISC de tumores cerebrales, OVA alcanza un F1 macro de 0.9306 frente a 0.9054 del modelo multi-salida, con un tiempo efectivo comparable al paralelizar el entrenamiento. Estos resultados indican que la descomposición teórica no caracteriza por completo el comportamiento práctico, y que OVA puede ser una alternativa competitiva cuando se consideran conjuntamente rendimiento, coste y errores por clase.

Keywords: Clasificación multiclase · Clasificadores One-vs-All · Redes neuronales multi-salida · Clasificación de imágenes biomédicas

1. Introducción

La clasificación multiclase es un problema fundamental en aprendizaje automático y en tareas modernas de clasificación de imágenes [6]. En aprendizaje profundo, el enfoque dominante consiste en emplear una única red neuronal multi-salida entrenada de forma conjunta, donde se aprende una representación compartida y la capa final produce una puntuación por clase [6,16].

Una estrategia alternativa consiste en descomponer el problema multiclase en una colección de tareas de clasificación binaria. El enfoque One-vs-All (OVA) realiza esta descomposición entrenando un clasificador binario para cada clase frente al resto. Sin embargo, en el entrenamiento práctico de redes neuronales profundas, OVA y multi-salida no se implementan de forma idéntica: OVA optimiza tareas binarias independientes mediante pérdidas logísticas, mientras que el modelo multi-salida se entrena conjuntamente mediante una pérdida softmax acoplada. Además, OVA requiere entrenar un modelo por clase, lo que incrementa el coste computacional respecto a un único modelo multi-salida.

Esta diferencia abre una cuestión práctica relevante. Por un lado, OVA puede favorecer la especialización por clase y permite paralelizar el entrenamiento de los clasificadores binarios. Por otro lado, el modelo multi-salida comparte representaciones entre clases y resuelve el problema mediante una única optimización acoplada. Por tanto, la comparación entre ambas formulaciones no debería limitarse al rendimiento predictivo, sino incluir también el coste efectivo de entrenamiento, el comportamiento bajo restricciones arquitectónicas y la estabilidad en distintos tipos de datos.

Este trabajo parte de la motivación teórica proporcionada por de la Higuera et al. [7], donde redes neuronales multi-salida de arquitectura fija se estudian mediante espacios RKBS vectoriales y el problema de aprendizaje asociado admite una descomposición por componentes. A partir de esta relación, la pregunta que abordamos no es demostrar una nueva equivalencia teórica, sino evaluar si dicha descomposición se refleja en una equivalencia práctica cuando los modelos se entrenan con datos finitos, funciones de pérdida habituales y restricciones computacionales reales. Esta cuestión es especialmente importante en clasificación de imágenes biomédicas, donde la robustez del modelo, el comportamiento por clase y el coste de entrenamiento pueden ser tan importantes como la exactitud global.

En este trabajo se presenta una comparación empírica entre clasificadores neuronales multi-salida entrenados de forma conjunta y clasificadores OVA entrenados de forma independiente. El estudio incluye conjuntos de datos clásicos de aprendizaje automático y tareas de clasificación de imágenes, incluyendo escenarios de imagen biomédica. Se comparan ambas formulaciones en términos de rendimiento predictivo, tiempo de entrenamiento, paralelización y robustez frente a reducciones arquitectónicas.

Las principales contribuciones de este trabajo son:

- una comparación empírica de clasificadores neuronales multi-salida y OVA en conjuntos de datos multiclase clásicos y basados en imágenes;

- un análisis del compromiso entre rendimiento predictivo y coste computacional, incluyendo entrenamiento OVA secuencial y paralelo;
- un estudio del efecto de reducir la arquitectura sobre ambas formulaciones;
- un análisis de escenarios de clasificación de imágenes biomédicas como casos de estudio para evaluar el comportamiento práctico de ambas formulaciones.

2. Trabajo relacionado

Reducciones multiclase y pérdidas. Las reducciones de problemas multiclase a problemas binarios han sido estudiadas ampliamente. Allwein et al. [2] propusieron una visión unificada de estas reducciones para clasificadores de margen, incluyendo esquemas como OVA, One-vs-One y códigos de corrección de errores. En esta línea, Rifkin y Klautau [14] argumentaron que OVA no debe considerarse simplemente un método de referencia básico, mostrando que clasificadores OVA bien ajustados pueden alcanzar un rendimiento comparable al de métodos multiclase más sofisticados.

En redes neuronales profundas, la formulación multi-salida con softmax se ha consolidado como el enfoque estándar para clasificación multiclase [6]. No obstante, las estrategias basadas en descomponer el problema multiclase en tareas binarias han sido revisitadas recientemente en este contexto. Pawara et al. [12] estudiaron clasificadores One-vs-One en redes neuronales profundas, mientras que Padhy et al. [11] y Wójcik et al. [18] analizaron formulaciones OVA desde la perspectiva de la incertidumbre predictiva, la calibración y la detección de muestras fuera de distribución. Estos últimos trabajos implementan OVA mediante una parametrización de salidas sigmoidales dentro de una única red compartida, sin entrenar K modelos independientes; nuestro trabajo se centra en cambio en el esquema OVA clásico de K clasificadores entrenados de forma independiente, que es el que admite paralelización real y al que se refiere la descomposición por componentes de [7]. Comparar ambas parametrizaciones OVA entre sí queda como línea de trabajo futuro.

La comparación entre modelos multi-salida y OVA también está relacionada con la función de pérdida utilizada durante el entrenamiento: entropía cruzada softmax en el caso multi-salida y pérdidas logísticas binarias en OVA. Li et al. [10] mostraron que estas pérdidas pueden inducir dinámicas diferentes, especialmente en escenarios multiclase de etiqueta única y con desbalance de clases.

Motivación teórica. Desde una perspectiva funcional, trabajos recientes han mostrado que redes neuronales multi-salida de arquitectura fija pueden estudiarse mediante espacios de Banach con núcleo reproductor vectoriales y características destacadas, y que, bajo una formulación de interpolación de norma mínima, el problema multi-salida puede descomponerse en problemas escalares, uno por cada componente de salida [7]. Nuestro trabajo se sitúa en la intersección entre esta motivación teórica y la evaluación experimental. Aunque dicha formulación funcional no coincide directamente con los protocolos habituales de entrenamiento de redes neuronales, proporciona una base para estudiar si el aprendizaje por componentes, implementado en la práctica mediante clasificadores OVA independientes, se comporta de forma comparable a una red multi-salida

entrenada conjuntamente bajo restricciones reales de datos, coste, paralelización y capacidad arquitectónica.

3. Experimentación

El objetivo de la experimentación es comparar el comportamiento práctico de clasificadores neuronales multi-salida y OVA bajo condiciones controladas, atendiendo a rendimiento predictivo, coste computacional y robustez frente a reducciones de arquitectura. No se busca maximizar el rendimiento absoluto en cada conjunto, sino analizar el comportamiento relativo de ambas formulaciones manteniendo constantes la familia arquitectónica, el protocolo de entrenamiento y las particiones de datos.

En este contexto, la equivalencia práctica se entiende como una situación en la que ambas formulaciones presentan diferencias predictivas pequeñas, compatibles con el margen de equivalencia definido, y no introducen una penalización temporal desproporcionada.

En la formulación multi-salida se entrena una única red con tantas salidas como clases mediante entropía cruzada softmax. En OVA se entrena un clasificador binario independiente por clase mediante pérdida logística, y durante la inferencia se selecciona la clase con mayor puntuación sigmoidal.

La Tabla 1 resume los conjuntos utilizados: problemas clásicos, referencias estándar de clasificación de imágenes y dos escenarios biomédicos. En los conjuntos clásicos se reserva un 20 % para el conjunto de prueba y un 15 % del resto para el de validación. En MNIST, CIFAR-10 y BRISC se mantiene la partición de prueba original y se obtiene el conjunto de validación del conjunto de entrenamiento. En TB X-ray se emplea una partición estratificada aproximada 70/15/15.

Tabla 1. Resumen de los conjuntos de datos utilizados. La columna Tamaño indica la resolución de entrada empleada en los modelos de imagen.

Conjunto	Clases	Muestras	Tipo	Tamaño	Train/Val/Test
Iris [5]	3	150	tabular	–	102/18/30
Wine [1]	3	178	tabular	–	120/22/36
Breast Cancer [19]	2	569	tabular	–	386/69/114
Digits [3]	10	1797	imagen	8×8	1221/216/360
MNIST [9]	10	70000	imagen	28×28	51000/9000/10000
CIFAR-10 [8]	10	60000	imagen RGB	32×32	42500/7500/10000
BRISC [4]	4	6000	imagen médica	128×128	4250/750/1000
TB X-ray [13]	2	4200	imagen médica	128×128	2940/630/630

3.1. Arquitecturas neuronales

Para que la comparación entre formulaciones sea directa, en cada conjunto de datos se utiliza la misma arquitectura base para el modelo multi-salida y para

los clasificadores OVA. La única diferencia se encuentra en la capa de salida: el modelo multi-salida dispone de tantas unidades como clases, mientras que cada clasificador OVA tiene una única salida binaria asociada a una clase frente al resto.

En los conjuntos tabulares se emplean perceptrones multicapa (MLP). La arquitectura de referencia consta de dos capas ocultas con tamaños $[32, 16]$. Para analizar la sensibilidad de ambas formulaciones a la capacidad del modelo, también se consideran dos arquitecturas reducidas, con tamaños $[24, 12]$ y $[16, 8]$.

En los conjuntos de imágenes se utilizan redes convolucionales compactas de tipo VGG [16], formadas por tres bloques convolucionales con canales configurables. La arquitectura de referencia emplea canales $[32, 64, 128]$, mientras que las arquitecturas reducidas utilizan $[24, 48, 96]$ y $[16, 32, 64]$. Se opta por arquitecturas secuenciales sin conexiones residuales para mantener una estructura fija y directamente comparable entre las formulaciones multi-salida y OVA. De este modo, las diferencias observadas se atribuyen principalmente a la formulación de salida y al esquema de entrenamiento, y no a cambios en la familia arquitectónica.

Las arquitecturas empleadas son redes *feed-forward* de arquitectura fija, sin conexiones residuales ni mecanismos dinámicos. Esta elección mantiene los modelos dentro del tipo de redes consideradas en la motivación funcional de [7] y evita introducir factores arquitectónicos adicionales en la comparación entre multi-salida y OVA.

3.2. Entrenamiento y paralelización

Todos los modelos se entrenan utilizando el mismo protocolo experimental dentro de cada conjunto de datos. Para reducir la dependencia de una única inicialización, los experimentos se repiten con diez semillas aleatorias. En cada semilla se emplean las mismas particiones de entrenamiento, validación y prueba para el modelo multi-salida y para los clasificadores OVA correspondientes.

Todos los modelos de referencia se optimizan mediante Adam, con una tasa de aprendizaje de 10^{-3}, un máximo de 50 épocas y *early stopping* basado en la pérdida de validación con paciencia de 10 épocas. En las configuraciones de referencia se utiliza *batch size* 32 en los conjuntos clásicos y biomédicos, y *batch size* 64 en MNIST y CIFAR-10. En cada conjunto, el modelo multi-salida y su contraparte OVA de referencia se entrenan con el mismo *batch size*.

Los modelos multi-salida se entrenan como una única red neuronal con tantas salidas como clases. En cambio, la formulación OVA entrena un clasificador binario independiente por cada clase. Esta independencia aumenta el coste si los clasificadores se entrenan secuencialmente, pero también permite paralelizarlos. Por ello, para OVA se reporta tanto el tiempo secuencial, calculado como la suma de los tiempos por clase, como el tiempo paralelo efectivo, calculado como el máximo de dichos tiempos. Esta segunda medida aproxima el coste cuando se dispone de recursos para entrenar los clasificadores simultáneamente.

El rendimiento predictivo se evalúa mediante exactitud, precisión macro, recall macro y F1 macro. La métrica principal es F1 macro, ya que pondera por

igual el rendimiento de cada clase y resulta más informativa en escenarios con posible desbalance. Además, se reporta el tiempo de entrenamiento para analizar el compromiso entre rendimiento y coste computacional.

Para el análisis de robustez arquitectónica, se compara el rendimiento de las arquitecturas de referencia con el de las arquitecturas reducidas. Esta comparación permite evaluar si una formulación mantiene mejor su rendimiento cuando se disminuye la capacidad del modelo.

3.3. Análisis estadístico

El análisis estadístico se realiza sobre las diferencias pareadas de F1 macro entre OVA y multi-salida, calculadas utilizando las mismas semillas experimentales. En primer lugar, se ejecuta una evaluación inicial con diez semillas por conjunto de datos. A partir de estas diferencias preliminares se realizan dos análisis de potencia mediante simulación Monte Carlo basada en remuestreo bootstrap de las diferencias observadas.

El primero está asociado a la prueba de rangos con signo de Wilcoxon y permite estimar el número de semillas necesario para detectar una diferencia relevante de 0,02 en F1 macro con un nivel de significación $\alpha = 0{,}05$. El segundo está asociado al análisis de equivalencia y permite estimar el número de semillas necesario para concluir equivalencia práctica cuando la diferencia real entre modelos se encuentra dentro del margen de equivalencia especificado.

Se considera una potencia objetivo de 0,8 en los conjuntos generales y de 0,9 en los escenarios biomédicos. Cuando las diez semillas iniciales no son suficientes para alcanzar dicha potencia, se amplía el número de semillas antes de aplicar los contrastes finales.

Sobre los resultados finales disponibles para cada conjunto de datos se aplica la prueba de rangos con signo de Wilcoxon [17] para contrastar si existen diferencias sistemáticas entre OVA y multi-salida. La prueba se aplica de forma pareada sobre las diferencias OVA menos multi-salida, utilizando como unidad experimental la semilla.

Además del contraste de diferencias, se evalúa la equivalencia práctica mediante un procedimiento inspirado en *Two One-Sided Tests* (TOST) [15]. Se fija un margen de equivalencia de 0,02 en F1 macro, de modo que dos modelos se consideran prácticamente equivalentes si la diferencia de rendimiento queda dentro del intervalo $[-0{,}02, 0{,}02]$. Para evaluar esta condición se construyen intervalos de confianza bootstrap bilaterales del 90 % para la mediana de las diferencias pareadas. Este intervalo se estima mediante bootstrap: se generan 1000 remuestreos con reemplazo de la muestra simulada, se calcula la mediana en cada remuestreo y se toman los percentiles 5 y 95 de esas medianas para obtener un intervalo de confianza del 90 %. Se concluye equivalencia práctica cuando el intervalo completo queda contenido dentro del margen predefinido.

Este procedimiento permite distinguir entre diferencias estadísticamente significativas, equivalencia práctica respaldada por la evidencia estadística y situaciones en las que la ausencia de significación o de equivalencia puede atribuirse a una potencia insuficiente.

4. Resultados y discusión

En esta sección se analiza si la relación teórica entre modelos multi-salida y esquemas OVA se traduce en un comportamiento práctico equivalente al entrenar redes neuronales bajo condiciones finitas de datos, arquitectura y tiempo de cómputo. En primer lugar se comparan ambas formulaciones con las arquitecturas de referencia: MLP [32, 16] para los conjuntos tabulares y VGG [32, 64, 128] para los conjuntos de imágenes. Posteriormente se estudia si configuraciones OVA reducidas o ajustadas pueden mantener la capacidad predictiva con menor coste efectivo de entrenamiento.

La interpretación de los resultados predictivos toma como referencia el número final de semillas determinado por el análisis de potencia. Con este criterio, la mayoría de conjuntos se evaluaron con las diez semillas iniciales. Iris se amplió a 18 semillas para reforzar el contraste de diferencias, mientras que BRISC se amplió a 44 semillas por el análisis de potencia asociado a la equivalencia en el escenario biomédico. Los tiempos de entrenamiento se resumen sobre las diez semillas iniciales, usando el mismo protocolo para todos los conjuntos.

4.1. Comparación con arquitecturas de referencia

La Figura 1 muestra el F1 macro obtenido por los modelos multi-salida y OVA con las arquitecturas de referencia. Los resultados muestran que la equivalencia teórica entre ambas formulaciones no implica necesariamente equivalencia práctica en todos los escenarios. OVA mejora el rendimiento en varios conjuntos, especialmente en Digits y BRISC, mientras que en MNIST las diferencias son prácticamente despreciables. En CIFAR-10, por el contrario, el modelo multi-salida mantiene una ligera ventaja.

El caso de BRISC es el más claro: el modelo OVA de referencia alcanza un F1 macro de 0.9306 frente a 0.9054 del modelo multi-salida. En Digits también se observa una mejora relevante, con 0.9796 frente a 0.9635. En MNIST, ambos modelos presentan un rendimiento prácticamente indistinguible, mientras que en CIFAR-10 la formulación multi-salida obtiene una ligera ventaja.

El análisis estadístico confirma esta interpretación. En Digits se rechaza la hipótesis nula de igualdad mediante la prueba de Wilcoxon ($p = 0{,}00195$), y el análisis de equivalencia no permite considerar ambas formulaciones equivalentes dentro del margen de 0,02 en F1 macro. En BRISC, tras ampliar el número de semillas según el análisis de potencia, también se rechaza la hipótesis nula y la diferencia observada supera el margen de equivalencia. En ambos casos, la no equivalencia práctica favorece a OVA.

En el resto de conjuntos, el análisis no muestra diferencias estadísticamente significativas y el procedimiento TOST respalda la equivalencia práctica bajo el margen considerado. Por tanto, los resultados sugieren que la relación teórica entre ambas formulaciones se traduce con frecuencia en equivalencia práctica, aunque en determinados conjuntos la descomposición OVA puede aportar mejoras predictivas relevantes.

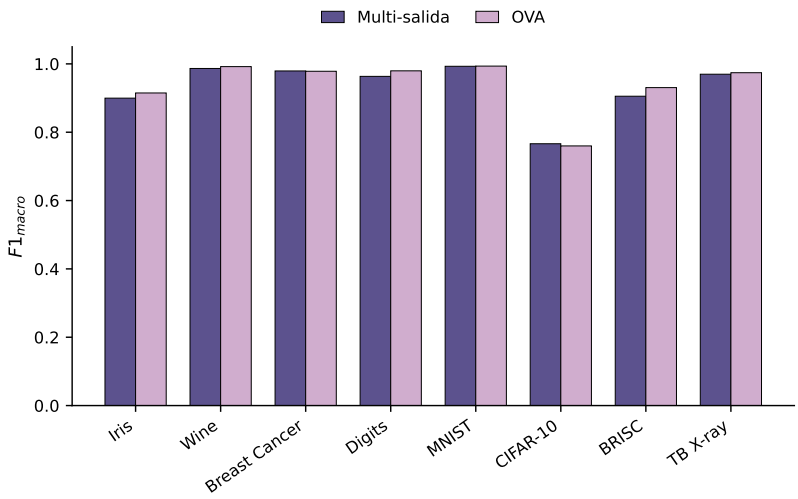


Figura 1. Comparación de F1 macro entre modelos multi-salida y OVA con arquitecturas de referencia.

En los escenarios biomédicos, además del F1 macro, resulta relevante analizar los errores por clase. La Tabla 2 muestra la tasa de falsos negativos en BRISC, agregando las predicciones sobre el conjunto de prueba en diez semillas experimentales. OVA reduce el número total de falsos negativos por clase de 934 a 697 respecto al modelo multi-salida. La reducción es especialmente apreciable en `meningioma` y `pituitary`.

Tabla 2. Falsos negativos por clase en BRISC, agregados sobre las predicciones del conjunto de prueba en diez semillas experimentales. FN indica el número de falsos negativos y FNR la tasa de falsos negativos.

Clase	FN multi	FN OVA	FNR multi	FNR OVA
`glioma`	315	275	0.124	**0.108**
`meningioma`	533	380	0.174	**0.124**
`no_tumor`	6	5	0.004	0.004
`pituitary`	80	37	0.027	**0.012**
Total	934	**697**	–	–

En tuberculosis también se observa una reducción de la tasa macro de falsos negativos, aunque de menor magnitud, pasando de 0.0377 en el modelo multi-salida a 0.0332 en OVA. Este comportamiento es relevante en entornos médicos, donde minimizar los falsos negativos puede ser prioritario cuando el coste de no detectar una clase es elevado.

4.2. Configuraciones OVA prácticas

La selección de configuraciones OVA prácticas no persigue maximizar exclusivamente el rendimiento predictivo, sino identificar soluciones con un equilibrio razonable entre F1 macro y coste efectivo de entrenamiento (véase la Sección 4.3). Para cada conjunto se evaluaron combinaciones de arquitectura e hiperparámetros de entrenamiento, considerando como candidatas aquellas configuraciones cuyo rendimiento en el conjunto de validación permanecía próximo al OVA de referencia.

Entre ellas se seleccionó la configuración con menor tiempo paralelo efectivo, siempre que la pérdida de F1 macro en validación respecto al mejor modelo OVA quedara dentro del margen práctico de 0,02 considerado en el análisis de equivalencia. El conjunto de prueba no interviene en esta selección y se reserva únicamente para reportar los resultados finales. Este criterio busca aproximar un escenario práctico donde pequeñas diferencias de rendimiento pueden aceptarse si reducen significativamente el coste computacional.

Tabla 3. F1 macro medio y desviación estándar para el modelo multi-salida de referencia, el OVA de referencia y la configuración OVA seleccionada. Los valores se calculan sobre las diez semillas iniciales.

Conjunto	Multi ref.	OVA ref.	OVA sel.	Config. OVA sel.
Iris	$0{,}900 \pm 0{,}081$	$\mathbf{0{,}915 \pm 0{,}071}$	$\mathbf{0{,}915 \pm 0{,}071}$	$[32, 16]$, $b = 32$
Wine	$0{,}987 \pm 0{,}019$	$0{,}992 \pm 0{,}013$	$\mathbf{0{,}992 \pm 0{,}013}$	$[16, 8]$, $b = 32$
Breast Cancer	$0{,}979 \pm 0{,}011$	$0{,}978 \pm 0{,}010$	$\mathbf{0{,}982 \pm 0{,}009}$	$[16, 8]$, $b = 64$
Digits	$0{,}963 \pm 0{,}008$	$0{,}980 \pm 0{,}003$	$\mathbf{0{,}980 \pm 0{,}003}$	$[32, 16]$, $b = 64$
MNIST	$0{,}993 \pm 0{,}001$	$0{,}994 \pm 0{,}001$	$\mathbf{0{,}994 \pm 0{,}001}$	$[32, 64, 128]$, $b = 128$
CIFAR-10	$\mathbf{0{,}766 \pm 0{,}010}$	$0{,}760 \pm 0{,}007$	$0{,}766 \pm 0{,}007$	$[32, 64, 128]$, $b = 128$
BRISC	$0{,}905 \pm 0{,}020$	$\mathbf{0{,}931 \pm 0{,}011}$	$0{,}921 \pm 0{,}014$	$[32, 64, 128]$, $b = 32$
TB X-ray	$0{,}970 \pm 0{,}009$	$\mathbf{0{,}974 \pm 0{,}010}$	$0{,}974 \pm 0{,}009$	$[16, 32, 64]$, $b = 32$

Los resultados muestran que las configuraciones OVA seleccionadas mantienen un rendimiento competitivo en la mayoría de los conjuntos. En Wine, Breast Cancer y Digits, el modelo OVA seleccionado mejora el F1 macro del modelo multi-salida de referencia. En MNIST y CIFAR-10 las diferencias son muy pequeñas, lo que sugiere equivalencia práctica desde el punto de vista predictivo. En BRISC, la configuración seleccionada conserva una mejora frente al modelo multi-salida, además de reducir el coste respecto al OVA de referencia. En tuberculosis, la configuración seleccionada mantiene un F1 macro muy próximo al OVA de referencia y mejora sustancialmente el tiempo paralelo efectivo.

Para analizar la robustez arquitectónica, las configuraciones OVA seleccionadas se comparan también con modelos multi-salida equivalentes, es decir, modelos multi-salida entrenados con la misma arquitectura y el mismo *batch size* que el OVA seleccionado. Esta comparación permite distinguir si la mejora procede únicamente de cambiar la capacidad del modelo o si la formulación OVA conserva mejor el rendimiento bajo restricciones similares. En esta comparación, OVA

mejora al multi-salida equivalente en Wine (+0.030), Breast Cancer (+0.003), Digits (+0.017), MNIST (+0.0002), BRISC (+0.015) y tuberculosis (+0.010). En CIFAR-10, en cambio, el multi-salida equivalente mantiene ventaja (-0.008 para OVA). Estos resultados sugieren que OVA puede ser más robusto frente a reducciones o cambios de capacidad arquitectónica en ciertos conjuntos, aunque este comportamiento no es universal.

4.3. Coste temporal

La Tabla 4 resume los tiempos medios de entrenamiento para el modelo multi-salida de referencia, los modelos OVA de referencia secuencial y paralelo, y el modelo OVA seleccionado paralelo.

Tabla 4. Tiempos medios de entrenamiento para las arquitecturas de referencia y para la configuración OVA seleccionada. Los tiempos se expresan en segundos.

Conjunto	Multi ref.	OVA sec.	OVA par.	OVA sel. par.
Iris	**0.47**	1.20	0.98	0.98
Wine	**0.66**	2.03	0.87	0.77
Breast Cancer	**1.04**	2.31	1.99	1.41
Digits	**3.02**	27.28	5.66	3.06
MNIST	112.84	1204.15	161.21	**109.76**
CIFAR-10	128.00	1095.24	124.97	**109.61**
BRISC	135.96	517.88	137.91	**134.31**
TB X-ray	106.38	216.67	117.10	**73.04**

Los resultados muestran que el coste secuencial de OVA puede ser elevado, aunque su impacto depende del tamaño y complejidad del conjunto de datos. En los conjuntos clásicos, la diferencia temporal existe pero es menos crítica debido al bajo coste absoluto de entrenamiento. El caso más claro dentro de este grupo es Digits, donde OVA secuencial requiere 27.28 segundos frente a 3.02 segundos del modelo multi-salida. Sin embargo, al usar la configuración OVA seleccionada y considerar el tiempo paralelo efectivo, el coste se reduce a 3.06 segundos, prácticamente igualando al modelo multi-salida.

El efecto de la paralelización es mucho más relevante en los conjuntos de imágenes. En MNIST y CIFAR-10, el OVA secuencial supera ampliamente los 1000 segundos, mientras que el tiempo paralelo efectivo se reduce a valores comparables al modelo multi-salida. El caso de BRISC combina ambas ventajas: el OVA seleccionado mantiene la mejora predictiva frente al modelo multi-salida y reduce marginalmente el tiempo efectivo, con 134.31 frente a 135.96 segundos. En tuberculosis, la configuración OVA seleccionada reduce el tiempo efectivo de 106.38 a 73.04 segundos.

Estos resultados matizan la principal desventaja práctica atribuida a OVA: entrenar un modelo binario por clase puede parecer mucho más costoso que

entrenar un único modelo multi-salida. Sin embargo, la independencia entre clasificadores permite paralelizar el entrenamiento de forma natural. En los conjuntos de imágenes estudiados, el tiempo efectivo de OVA paralelo alcanza valores comparables al modelo multi-salida e incluso inferiores en las configuraciones seleccionadas de MNIST, CIFAR-10, BRISC y TB X-ray. Por tanto, el coste de OVA debe interpretarse considerando no solo el tiempo secuencial, sino también su capacidad de paralelización.

5. Conclusiones y trabajo futuro

Este trabajo ha estudiado la equivalencia práctica entre clasificadores neuronales multi-salida y clasificadores OVA, tomando como punto de partida una relación teórica previa basada en la descomposición por componentes. Los resultados muestran que ambas formulaciones presentan con frecuencia un comportamiento predictivo compatible con equivalencia práctica bajo el margen considerado. Sin embargo, esta equivalencia no es universal: en Digits y BRISC la diferencia es estadísticamente significativa y favorable a OVA.

En conjunto, los resultados refuerzan la idea de que la elección entre una formulación multi-salida y una estrategia OVA no debe basarse únicamente en consideraciones teóricas de representabilidad, sino también en criterios prácticos asociados al coste, la paralelización y el comportamiento por clase. Aunque los modelos multi-salida siguen siendo una opción natural y eficiente en muchos escenarios, OVA ofrece una alternativa competitiva cuando la especialización de los clasificadores binarios permite mejorar clases concretas o reducir errores críticos. Este aspecto es especialmente relevante en dominios sensibles, como la imagen biomédica, donde pequeñas diferencias agregadas pueden traducirse en cambios importantes en falsos negativos o en la fiabilidad de la decisión por clase.

El estudio presenta varias limitaciones: se restringe a MLP y redes VGG compactas de arquitectura fija, y la evaluación se centra en tareas de clasificación. Como trabajo futuro, se plantea ampliar el análisis a arquitecturas más complejas, incluyendo modelos con conexiones residuales o mecanismos de atención, así como a tareas de regresión y a más escenarios tabulares, visuales y biomédicos. En estos últimos será especialmente relevante estudiar el comportamiento por clase mediante métricas sensibles al coste del error y determinar, junto con especialistas, cuándo la especialización por clase de OVA compensa frente al aprendizaje conjunto en modelos multi-salida.

Agradecimientos Esta investigación es parte del Proyecto TSI-100927-2023-1, financiado por Plan de Recuperación, Transformación y Resiliencia Europeo, de la Unión Europea-Next Generation EU a través del Ministerio para la Transformación Digital y de la Función Pública y está parcialmente financiada por el Ministerio de Ciencia y Tecnología de España en el marco del proyecto PID2023-150070NB-I00, financiado por MICIU/AEI/10.13039/501100011033 y por FEDER, UE.

Declaración de intereses No existe ningún conflicto de intereses.

Referencias

1. Aeberhard, S., Forina, M.: Wine. UCI Machine Learning Repository (1992)
2. Allwein, E.L., Schapire, R.E., Singer, Y.: Reducing multiclass to binary: A unifying approach for margin classifiers. Journal of Machine Learning Research **1**, 113–141 (2000)
3. Alpaydin, E., Kaynak, C.: Optical Recognition of Handwritten Digits. UCI Machine Learning Repository (1998)
4. Fateh, A., Rezvani, Y., Moayedi, S., Rezvani, S., Fateh, F., Fateh, M., Abolghasemi, V.: Brisc: Annotated dataset for brain tumor segmentation and classification. Scientific Data **13**(1), 361 (2026)
5. Fisher, R.A.: Iris. UCI Machine Learning Repository (1936)
6. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016)
7. de la Higuera, I., Herrera, F., Velasco, M.V.: Featured reproducing kernel banach spaces for learning and neural networks (2026), `https://arxiv.org/abs/2602.07141`
8. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Tech. Rep. 0, University of Toronto, Toronto, Ontario (2009)
9. LeCun, Y., Cortes, C., Burges, C.: Mnist handwritten digit database. ATT Labs [Online]. **2** (2010)
10. Li, X., Li, X., Pan, D., Zhu, D.: On the learning property of logistic and softmax losses for deep neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 4739–4746 (2020)
11. Padhy, S., Nado, Z., Ren, J., Liu, J., Snoek, J., Lakshminarayanan, B.: Revisiting one-vs-all classifiers for predictive uncertainty and out-of-distribution detection in neural networks. In: ICML UDL (2020)
12. Pawara, P., Okafor, E., Groefsema, M., He, S., Schomaker, L.R.B., Wiering, M.A.: One-vs-one classification for deep neural networks. Pattern Recognition **108**, 107528 (2020)
13. Rahman, T., Khandakar, A., Kadir, M.A., Islam, K.R., Islam, K.F., Mahbub, Z.B., Ayari, M.A., Chowdhury, M.E.H.: Reliable tuberculosis detection using chest x-ray with deep learning, segmentation and visualization. IEEE Access **8**, 191586–191601 (2020)
14. Rifkin, R.M., Klautau, A.: In defense of one-vs-all classification. Journal of Machine Learning Research **5**, 101–141 (2004)
15. Schuirmann, D.J.: A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. Journal of Pharmacokinetics and Biopharmaceutics **15**, 657–680 (1987)
16. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015)
17. Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics Bulletin **1**(6), 80–83 (1945)
18. Wójcik, B., Grela, J., Śmieja, M., Misztal, K., Tabor, J.: Uncertainty estimation using single label one-vs-all classifier. Applied Soft Computing **126**, 109219 (2022)
19. Zwitter, M., Soklic, M.: Breast Cancer. UCI Machine Learning Repository (1988)