

Apéndice D. Evaluación experimental de los modelos de inteligencia artificial

En este apéndice se presentan los resultados obtenidos durante la evaluación de los modelos de inteligencia artificial desarrollados para el sistema automatizado de clasificación de frutos cítricos. Se incluyen los procedimientos de validación empleados, las curvas de entrenamiento, las métricas de desempeño y los análisis realizados para los modelos de clasificación y detección implementados.

D.1 Evaluación del modelo RGB

D.1.1 Validación cruzada estratificada

Con el propósito de evaluar la capacidad de generalización del modelo RGB y reducir la dependencia de una única partición del conjunto de datos, se implementó un procedimiento de validación cruzada estratificada (Stratified K-Fold) con cinco particiones.

Durante cada iteración, una de las particiones fue utilizada como conjunto de validación, mientras que los cuatro restantes conformaron el conjunto de entrenamiento. Este procedimiento se repitió cinco veces, de manera que cada muestra del conjunto de datos participó exactamente una vez como dato de validación y cuatro veces como dato de entrenamiento.

Tabla D.1. Configuración empleada durante la validación cruzada del modelo RGB.

Parámetro	Valor
Estrategia de validación	Stratified K-Fold
Número de <i>folds</i>	5
Entrenamientos independientes	5
Callback 1	EarlyStopping
Callback 2	ReduceLROnPlateau
Criterio de selección	Mejor pérdida de validación

Durante el entrenamiento se emplearon los callbacks mencionados anteriormente (Apéndice C) detener automáticamente el entrenamiento cuando la pérdida de validación dejaba de mejorar y ajustar dinámicamente la tasa de aprendizaje cuando el proceso alcanzaba una meseta. Para cada fold se almacenó el modelo correspondiente a la época con mejor desempeño sobre el conjunto de validación, el cual fue utilizado posteriormente para calcular las métricas de evaluación.

D.1.2 Resultados de la validación cruzada

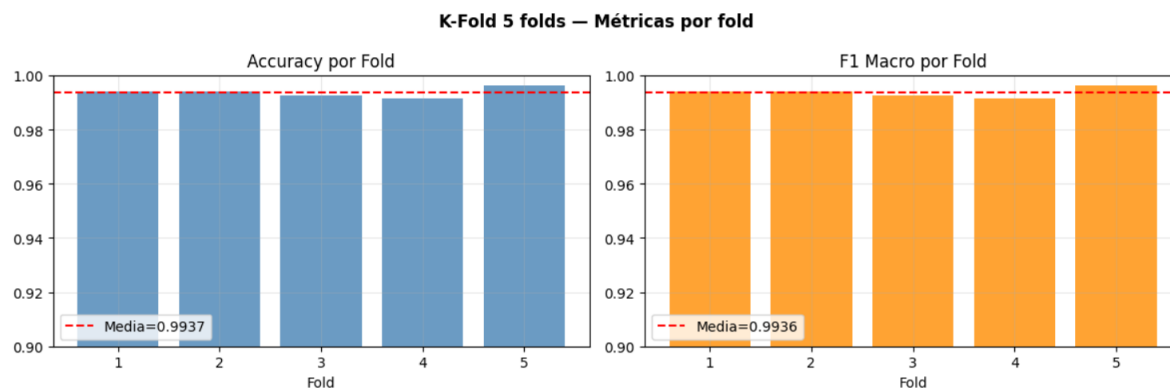
La Tabla D.2 resume las métricas obtenidas durante cada una de las cinco particiones de la validación cruzada estratificada. En todos los casos se alcanzaron valores de accuracy y F1 superiores al 99 %, evidenciando un comportamiento consistente del modelo frente a las diferentes particiones del conjunto de datos.

Tabla D.2. Metricas de particiones de la validación cruzada estratificada

Fold	Accuracy	F1 Macro
1	0,9939	0,9939
2	0,9939	0,9939
3	0,9927	0,9926
4	0,9915	0,9914
5	0,9963	0,9963
Promedio $\pm$ DE	0,9937 $\pm$ 0,0016	0,9936 $\pm$ 0,0016

En los resultados obtenidos se evidencia una baja variabilidad entre las diferentes particiones, lo que indica que el modelo presentó un comportamiento estable durante todo el procedimiento de validación cruzada y una adecuada capacidad de generalización sobre no utilizados durante el entrenamiento.

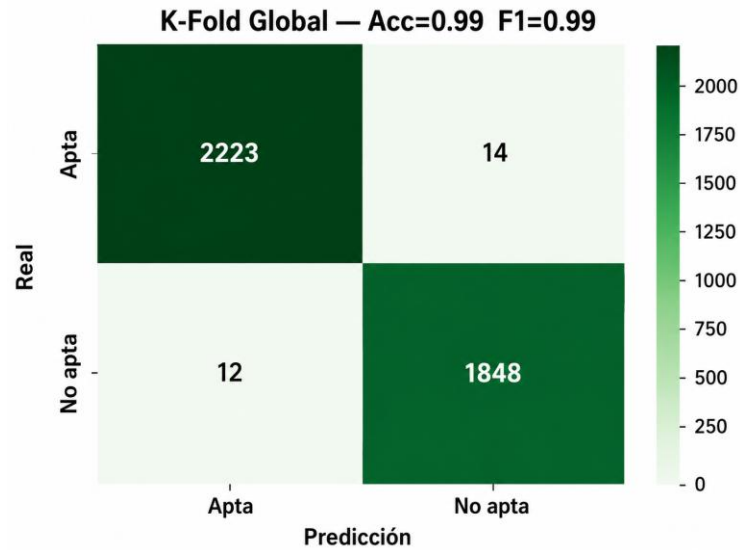
Figura D.2. Accuracy y F1 obtenidos en cada una de las cinco particiones de la validación cruzada.



Posteriormente, se realizó un análisis conjunto de todas las predicciones generadas durante la validación cruzada con el fin de obtener una estimación global del desempeño del modelo. El

reporte de clasificación mostró valores de precision, recall y F1 cercanos al 99% para ambas categorías, indicando un comportamiento equilibrado entre las clases apta y no apta.

Figura D.3. *Matriz de confusión global obtenida a partir de las predicciones realizadas durante la validación cruzada.*



La matriz de confusión evidencia que la mayoría de las muestras fueron clasificadas correctamente, registrándose únicamente 14 mandarinas aptas clasificadas como no aptas (FP) y 12 mandarinas no aptas clasificadas como aptas (FN), para un total de 26 errores sobre las 4097 muestras evaluadas. Estos resultados confirman la capacidad del modelo para discriminar ambas categorías con un elevado nivel de exactitud.

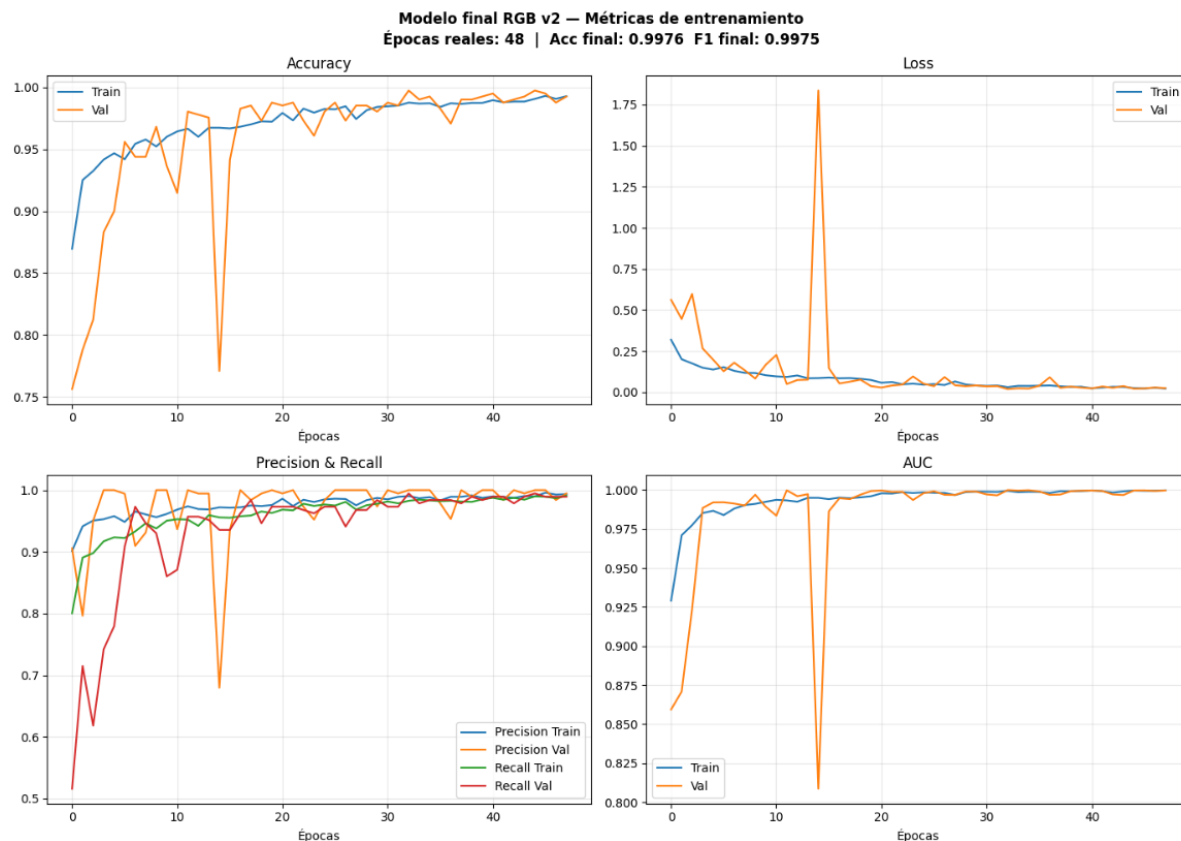
D.1.3 Entrenamiento del modelo final

Una vez verificada la estabilidad de la arquitectura mediante la validación cruzada estratificada, se realizó el entrenamiento del modelo definitivo empleando la totalidad del conjunto de datos RGB, conformado por 4097 imágenes correspondientes a las categorías apta y no apta.

Con el propósito de definir una duración adecuada del entrenamiento y evitar depender del comportamiento particular de un único fold, el número de épocas de referencia se estimó como el promedio de la mejor época obtenida durante las cinco particiones de la validación cruzada. A partir de este valor se estableció un límite máximo de 77 épocas, permitiendo que el criterio EarlyStopping determinara automáticamente el punto óptimo de finalización del entrenamiento.

Durante esta etapa se mantuvo la arquitectura, el aumento de datos y la configuración de hiperparámetros descritos en el ¡Error! No se encuentra el origen de la referencia.. Para supervisar el proceso se empleó una partición interna del 10 % del conjunto de datos, utilizada exclusivamente como referencia para los callbacks EarlyStopping, ReduceLROnPlateau y ModelCheckpoint, sin intervenir en la selección de la arquitectura ni en la validación cruzada previamente realizada.

Figura D.4. *Curvas de entrenamiento del modelo RGB definitivo.*



El entrenamiento finalizó automáticamente en la época 48, restaurando los pesos correspondientes a la época 33, en la cual se obtuvo el menor valor de pérdida sobre el conjunto de validación interno. Finalmente, el modelo entrenado fue almacenado en formato HDF5 (.h5) para su posterior utilización durante la etapa de inferencia y como punto de partida para la comparación con los modelos espectrales.

D.1.4 Evaluación del modelo final

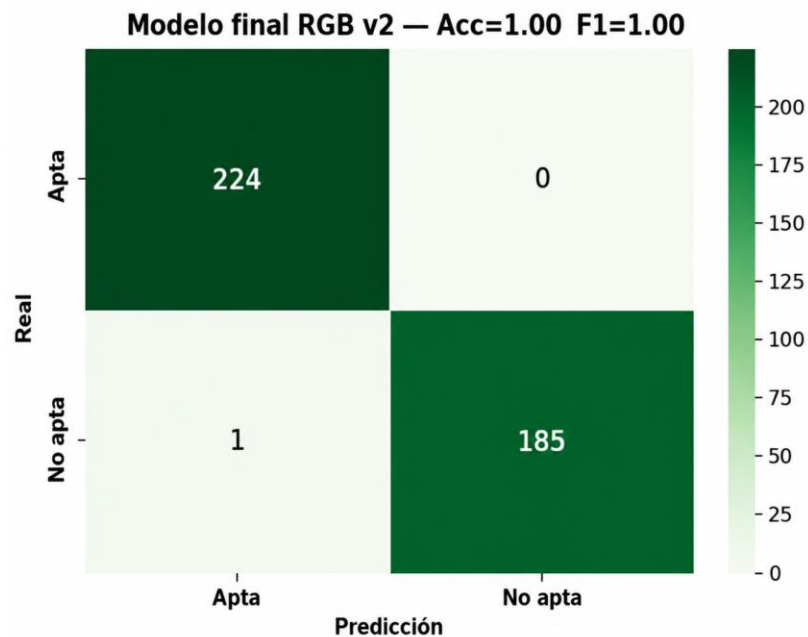
Para verificar el desempeño del modelo definitivo obtenido durante el entrenamiento, se realizó una evaluación sobre el conjunto de validación interno empleado durante esta etapa.

Tabla D.3 *resumen las principales métricas obtenidas por el modelo.*

Métrica	Valor
Accuracy	99,76 %
Precision	100,00 %
Recall	99,46 %
F1-score	99,75 %

Los resultados obtenidos evidencian un desempeño sobresaliente del modelo, alcanzando una exactitud cercana al 100 % y un valor de F1 superior al 99 %, lo que indica un adecuado equilibrio entre precisión y sensibilidad para ambas categorías evaluadas.

Figura D.5. *Matriz de confusión del modelo RGB definitivo.*



La matriz de confusión muestra que el modelo clasificó correctamente las 224 muestras correspondientes a la categoría apta. Para la categoría no apta, únicamente una muestra fue clasificada incorrectamente como apta, mientras que los 185 restantes fueron identificados correctamente. En conjunto, el modelo realizó una única clasificación errónea sobre las 410 muestras evaluadas, confirmando su elevada capacidad para discriminar entre ambas categorías bajo las condiciones del conjunto de validación.

Estos resultados permitieron establecer el modelo RGB como línea base para las comparaciones posteriores con los modelos desarrollados a partir de información espectral, las cuales se presentan en las siguientes secciones del presente apéndice.

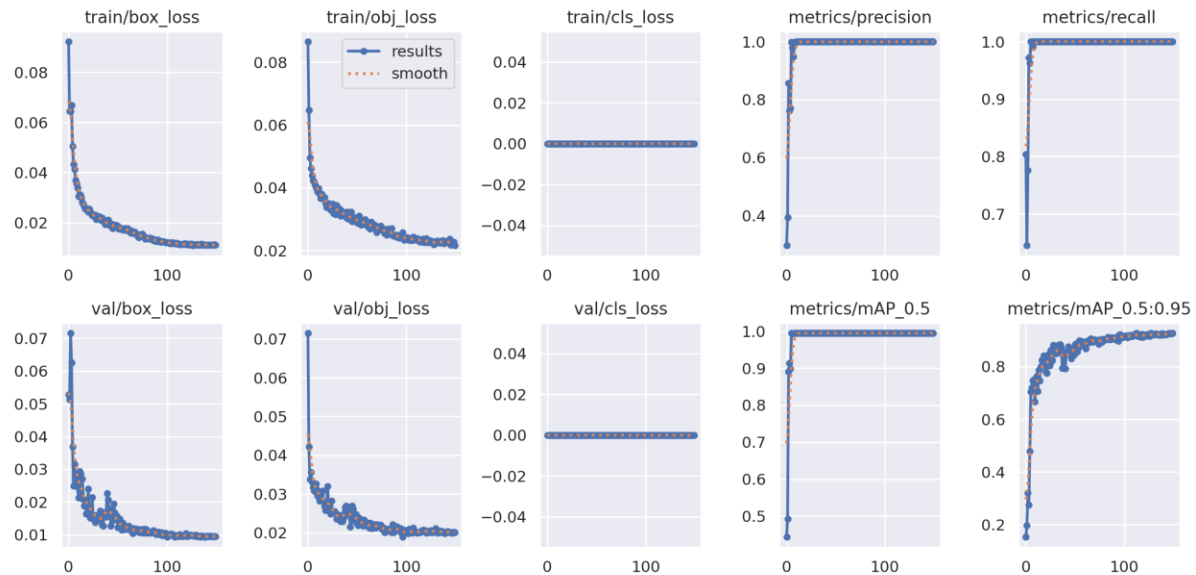
D.2 Evaluación del detector YOLO

D.2.1 Métricas obtenidas durante el entrenamiento

La Figura D.6 se presentas las curvas de pérdida y métricas registradas a lo largo de las 150 épocas del entrenamiento final. Las pérdidas de localización (box loss) y de objetividad (obj loss), tanto en entrenamiento como en validación, muestran una reducción sostenida desde las primeras épocas, con una caída en los primeros 20 ciclos seguida de una convergencia gradual hacia valores estables cercanos a 0,011 y 0,020 respectivamente al finalizar el entrenamiento. La pérdida de clasificación (cls loss) permaneció en cero durante todo el proceso, lo cual es el comportamiento esperado dado que el modelo fue configurado para detectar una única clase.

En cuanto a las métricas de desempeño sobre el conjunto de validación, la precisión y el recall alcanzaron valores superiores a 0,999 y 1,000 respectivamente desde la época 5, manteniéndose estables durante el resto del entrenamiento. El mAP@0.5 se estabilizó igualmente en 0,995 desde etapas tempranas. El mAP@0.5:0.95, que evalúa la precisión geométrica de los bounding boxes bajo múltiples umbrales de IoU, mostró una mejora progresiva y más lenta a lo largo de todo el entrenamiento, alcanzando su valor máximo de@ 0,927 en la época 127. Al finalizar las 150 épocas, el valor registrado fue de 0,925, un poco menor al pico alcanzado en la mejor época. El criterio de early stopping con paciencia de 25 épocas no llegó a activarse, lo que indica que el modelo continuó presentando mejoras, aunque marginales, hasta completar las 150 épocas definidas.

Figura D.6. *Curvas de pérdida y métricas registradas durante el entrenamiento.*



D.2.2 Curvas Precision–Recall, F1–Confidence y métricas de validación

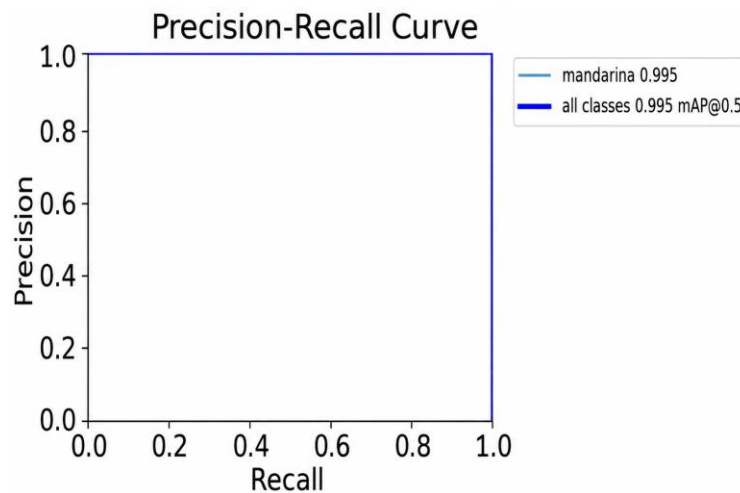
La evaluación del modelo sobre el conjunto de validación que estuvo compuesto por 22 imágenes y 107 instancias de mandarinas mostró los siguientes resultados presentados en la Tabla D.4.

Tabla D.4. *Resultados del proceso de entrenamiento final del modelo YOLOv5.*

Métrica	Valor
Precisión	0,979
Recall	0.999
mAP0.5	0.993
mAP0.5:0.95	0.917
Imágenes evaluadas	22
Instancias evaluadas	107

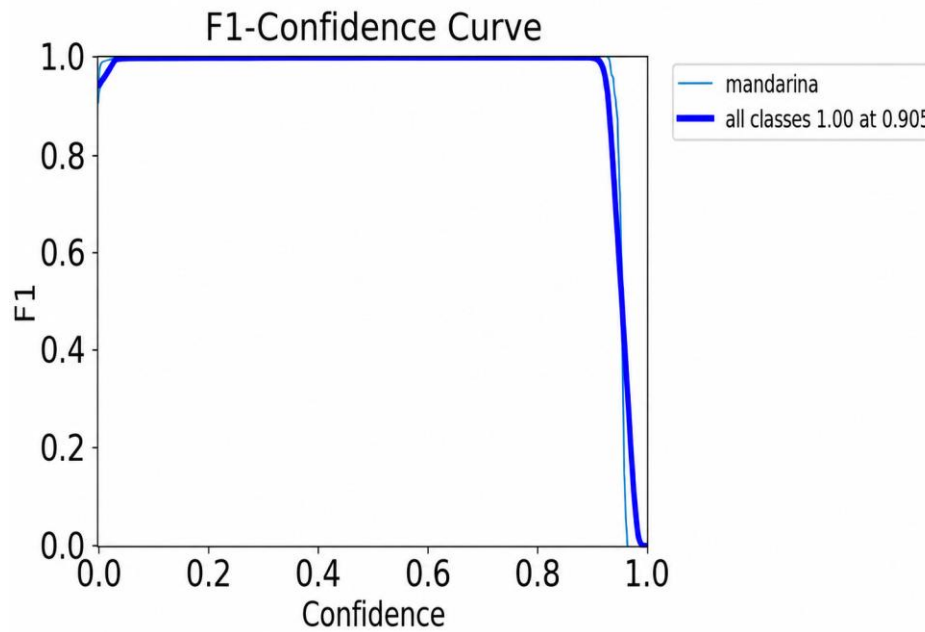
La curva de Precision-Recall(Figura D.7.) deja ver un comportamiento casi ideal, con un área bajo la curva de 0,993. La curva mantiene una precisión cercana a 1.0 a lo largo de todo el rango de recall, esto indica que el detector no requiere sacrificar precisión para detectar la totalidad de las instancias presentes en las imágenes.

Figura D.7. Curva Precision–Recall del detector YOLOv5m sobre el conjunto de validación.



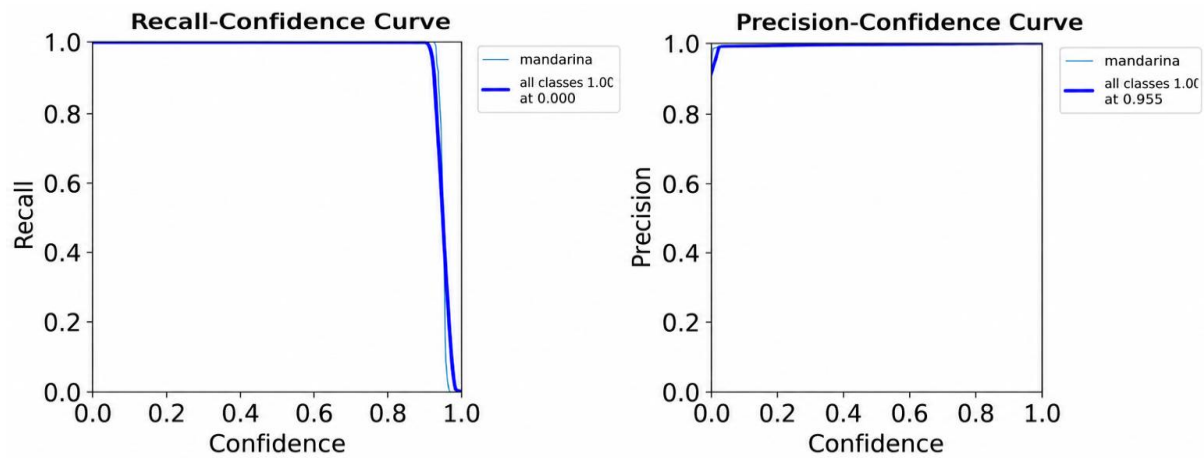
La curva F1-Confidence (Figura D.8.) muestra que el modelo alcanza un F1 igual a 1,00 con un umbral de confianza de 0,905, además mantiene valores de F1 superiores a 0.97 en un rango amplio de umbrales entre 0,0 y 0,90. Esto indica que el detector es robusto ante variaciones del umbral operativo, lo cual es benéfico para la configuración en el sistema final, ya que para este se estableció un umbral de confianza de 0,60 que es un valor conservador dentro del rango máximo de desempeño identificado en la curva.

Figura D.8. Curva F1–Confianza del detector YOLOv5m.



La curva Recall-Confidence confirma que el modelo detecta la totalidad de las mandarinas presentes para cualquier umbral de confianza por debajo de 0,90, sin presentar detecciones perdidas en condiciones operativas normales. La curva Precision-Confidence muestra que la precision se sostiene en 1,0 para umbrales superior a aproximadamente 0,05, indicando una tasa de falsos positivos prácticamente nula en el conjunto de validación.

Figura D.9. *Curvas de Recall–Confianza (izquierda) y Precisión–Confianza (derecha) del detector YOLOv5m sobre el conjunto de validación.*



D.2.3 Ejemplos de detección

La Figura D.10. presenta ejemplos de las detecciones realizadas por el modelo sobre imágenes del conjunto de validación. En cada imagen se observan las mandarinas presentes en la escena sobre el fondo negro de la banda transportadora, junto con los bounding boxes predichos y sus respectivos valores de confianza.

En la totalidad de los casos mostrados, el detector identifica correctamente cada mandarina presente en la escena, con bounding boxes ajustados adecuadamente a la región ocupada por cada fruto, sin presentar detecciones duplicadas ni falsas sobre el fondo, Este comportamiento es consistente en imágenes provenientes de los distintos lotes de captura incluidos en el dataset, lo que sugiere que el modelo generaliza de forma satisfactoria.

Figura D.10 Ejemplos de detección sobre el conjunto de validación. Izquierda: etiquetas. Derecha: predicciones del modelo YOLOv5m con sus respectivos puntajes de confianza.



D.3 Evaluación del modelo espectral

Con el propósito de seleccionar la arquitectura más adecuada para la clasificación espectral, se evaluaron diferentes modelos de aprendizaje profundo utilizando un mismo protocolo experimental. Todas las arquitecturas fueron entrenadas y validadas empleando las mismas particiones de entrenamiento y validación mediante validación cruzada estratificada de cinco particiones, además de un conjunto Holdout completamente independiente conformado por cubos espectrales provenientes de sesiones de adquisición diferentes. Esta estrategia permitió realizar una comparación objetiva entre los modelos evaluados y seleccionar la arquitectura con mayor capacidad de generalización.

D.3.1 Comparación entre arquitecturas

Durante el desarrollo del sistema se implementaron tres arquitecturas de clasificación espectral: un modelo basado en fine-tuning progresivo, una arquitectura CNN Dual-Branch y

finalmente una arquitectura híbrida CNN-Transformer. Todas fueron entrenadas bajo las mismas condiciones experimentales y posteriormente evaluadas utilizando el mismo conjunto Holdout.

Tabla D.5. Comparación del desempeño de las arquitecturas evaluadas.

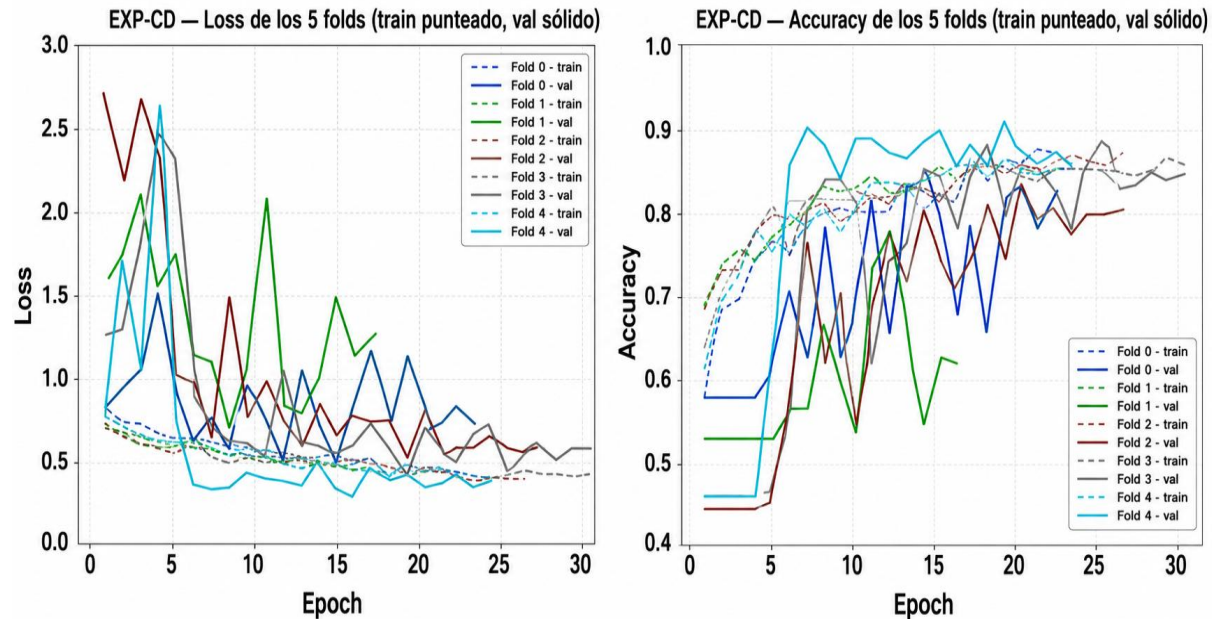
Arquitectura	CV (F1)	Holdout (F1)	Selección
Fine-tuning	0.7847	0.7494	—
progresivo			
CNN Dual-Branch	0.8129	0.6000	—
CNN-Transformer	0.8075	0.7749	✓

Aunque la arquitectura CNN Dual-Branch obtuvo el mayor valor de F1 durante la validación cruzada, su desempeño disminuyó considerablemente al evaluarse sobre el conjunto Holdout independiente, evidenciando dificultades para generalizar frente a nuevas sesiones de adquisición. Por el contrario, la arquitectura CNN-Transformer presentó un comportamiento más consistente entre la validación cruzada y la evaluación final, obteniendo el mejor desempeño sobre datos completamente independientes. Por esta razón, esta arquitectura fue seleccionada para su integración en el sistema automatizado de clasificación.

D.3.2 Validación cruzada del modelo seleccionado

Una vez seleccionada la arquitectura CNN-Transformer, se analizó su comportamiento durante el proceso de validación cruzada. La Figura D.11 presenta la evolución de la función de pérdida (Loss) y la evolución de la exactitud (Accuracy) para los cinco folds utilizados en el entrenamiento.

Figura D.11. Evolución de pérdida y exactitud durante la validación cruzada del modelo CNN-Transformer.



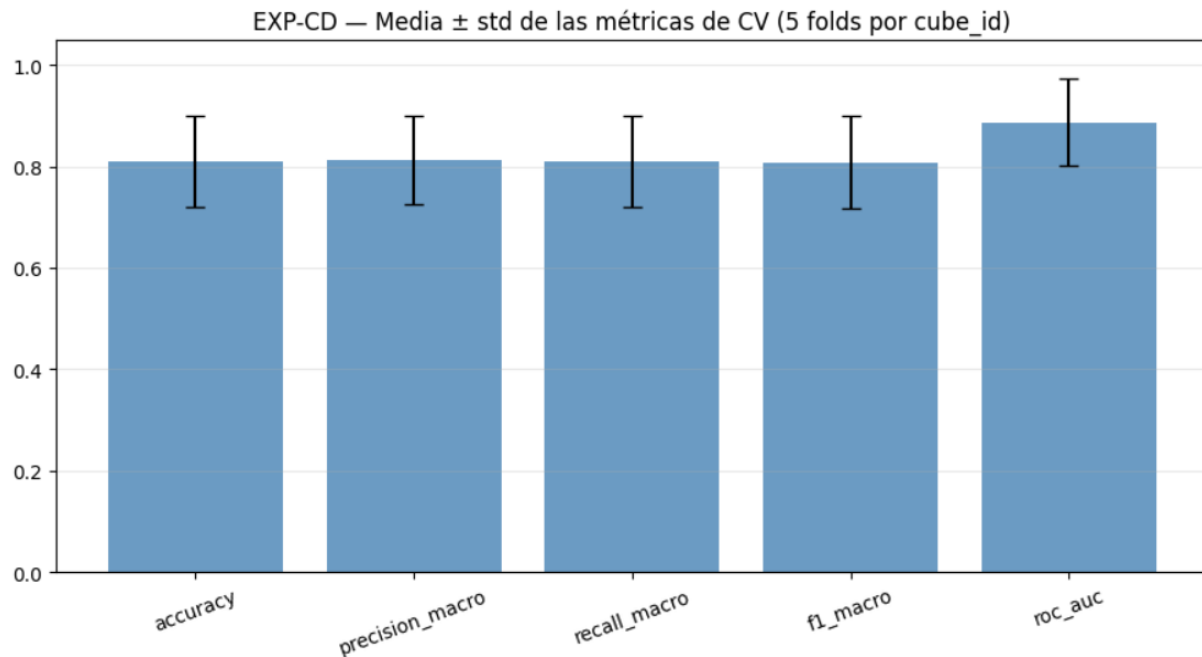
En términos generales, las curvas muestran un comportamiento estable durante el entrenamiento, observándose una reducción progresiva de la función de pérdida y un incremento de la exactitud en la mayoría de las particiones. Aunque se presentan fluctuaciones entre algunos folds, atribuibles a la variabilidad inherente de los datos espectrales, el modelo mantuvo un comportamiento consistente durante las diferentes iteraciones de validación.

Tabla D.6. Desempeño promedio obtenido durante la validación cruzada (5 folds).

Métrica	Media $\pm$ Desviación estándar
Accuracy	0.8102 $\pm$ 0.0896
Precision Macro	0.8126 $\pm$ 0.0867
Recall Macro	0.8096 $\pm$ 0.0908
F1 Macro	0.8075 $\pm$ 0.0916
ROC-AUC	0.8870 $\pm$ 0.0857

La figura D.12 resume gráficamente estas métricas, evidenciando una variabilidad moderada entre las diferentes particiones y un desempeño consistente durante el proceso de validación.

Figura D.12. Resumen de las métricas promedio obtenidas durante la validación cruzada.



D.3.3 Entrenamiento del modelo final

Una vez seleccionada la arquitectura definitiva, se realizó un entrenamiento final utilizando el conjunto de desarrollo completo, manteniendo una partición interna de validación destinada exclusivamente al monitoreo del entrenamiento mediante los callbacks implementados. Antes del entrenamiento se verificó que no existieran cubos espectrales compartidos entre los subconjuntos de entrenamiento, validación y Holdout, garantizando una separación estricta basada en el identificador de cada cubo (cube_id) y evitando cualquier fuga de información entre las etapas de desarrollo y evaluación.

Figura D.13. Verificación de la separación entre los conjuntos de entrenamiento, validación y Holdout mediante *cube_id*.

```
=====
CD-FINAL-0 - VERIFICACIÓN DEL SPLIT FINAL
=====
cube_id Train final : 117
cube_id Val final   : 30
Train n Val        = 0
Train n Holdout    = 0
Val n Holdout      = 0

✅ Las tres intersecciones son cero.

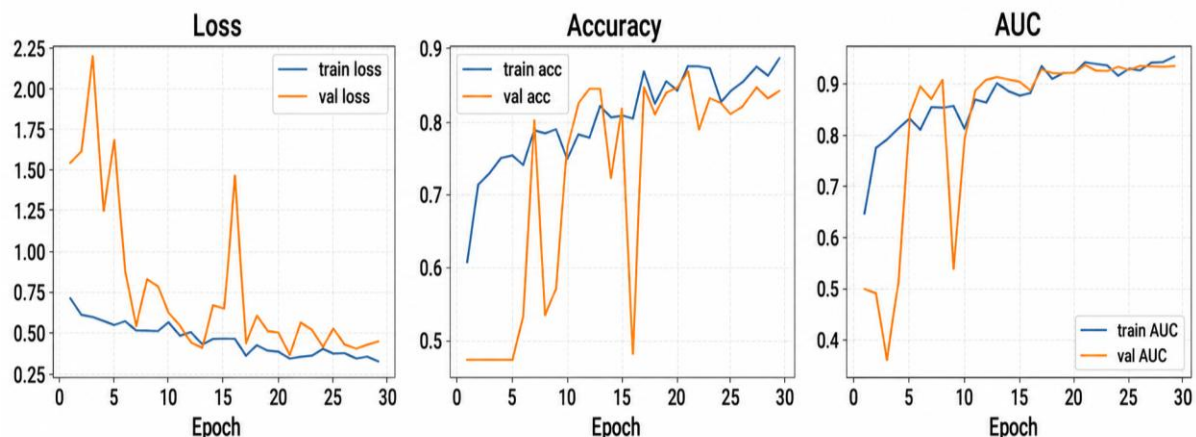
Muestras Train final : 520
Muestras Val final   : 134

Good/Bad Train final:
clase
bad    267
good   253

Good/Bad Val final:
clase
bad     70
good    64
```

La figura D.14 presenta la evolución de las principales métricas registradas durante el entrenamiento final del modelo, incluyendo la función de pérdida, la exactitud y el área bajo la curva ROC (AUC). En conjunto, estas curvas muestran una convergencia estable del proceso de aprendizaje y un comportamiento consistente entre los conjuntos de entrenamiento y validación.

Figura D.14. Curvas de entrenamiento del modelo CNN-Transformer definitivo.



D.4 Comparación entre modelos

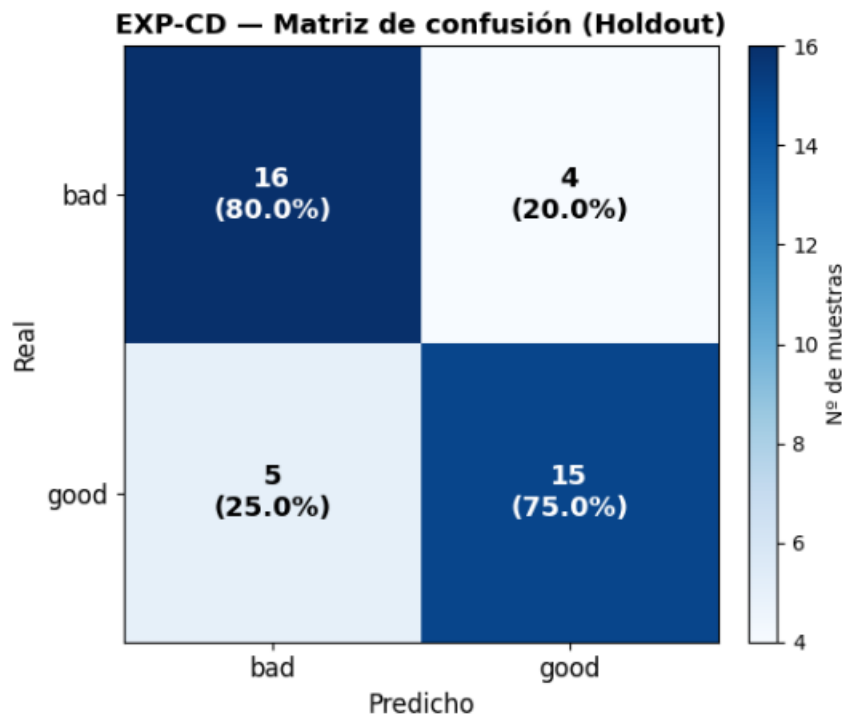
Finalmente, el modelo fue evaluado sobre un conjunto Holdout completamente independiente, conformado por muestras provenientes de sesiones de adquisición no utilizadas durante el entrenamiento ni durante la validación cruzada. Esta evaluación permitió estimar la capacidad de generalización del modelo frente a datos completamente nuevos.

Tabla D.7. *resultados obtenidos sobre el conjunto Holdout.*

Métrica	Valor
Accuracy	0.7750
Precision Macro	0.7757
Recall Macro	0.7750
F1 Macro	0.7749
ROC-AUC	0.8125

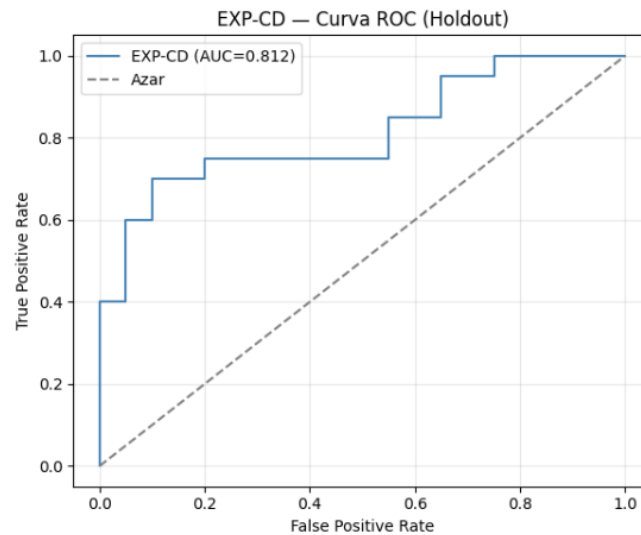
La figura D.15 presenta la matriz de confusión obtenida durante la evaluación Holdout. Se observa que el modelo clasificó correctamente la mayoría de las muestras pertenecientes a ambas categorías, alcanzando un equilibrio entre la identificación de aptas y no aptas.

Figura D.15. Matriz de confusión del modelo CNN-Transformer sobre el conjunto Holdout.



Por otra parte, la figura D.16 muestra la curva ROC obtenida durante esta evaluación, alcanzando un área bajo la curva (AUC) de 0.81 lo que indica una adecuada capacidad discriminativa para distinguir entre ambas clases utilizando únicamente información espectral.

Figura D.16 Curva ROC obtenida durante la evaluación del modelo sobre el conjunto Holdout.



En conjunto, los resultados evidencian que la arquitectura CNN-Transformer presentó el comportamiento más consistente entre las diferentes etapas de evaluación, manteniendo un desempeño estable tanto durante la validación cruzada como sobre el conjunto Holdout independiente. Esta capacidad fue el principal criterio para su selección como modelo de clasificación espectral dentro del sistema automatizado.