Análisis centrado en protocolos de validación aleatoria y cronológica en clasificación de imágenes agrícolas

A Protocol-Centric Analysis of Random and Chronological Validation in Agricultural Image Classification

Misael Hernández Sandoval[1], Sebastián Salazar Colores[1*]

[1]Centro de Investigaciones en Óptica A.C., 37150, León, Guanajuato, México
*Autor de correspondencia: sebastian.salazar@cio.mx

Resumen

La evaluación de modelos de visión artificial en agricultura depende no solo de la arquitectura empleada, sino también del criterio utilizado para separar los datos. Este trabajo presenta una versión de difusión centrada en el efecto del protocolo de validación, utilizando como caso de estudio la clasificación de estados nutrimentales en lechuga a partir de imágenes RGB. Se emplea un subconjunto de LettuceDeF con 4,405 imágenes y cuatro clases: nutrición completa de la variedad Blanes (FN-Blanes) y deficiencias de nitrógeno (-N), fósforo (-P) y potasio (-K). Se contrastan dos esquemas con proporciones 70/10/20: Protocolo A, selección aleatoria estratificada, y Protocolo B, selección cronológica. Nueve redes neuronales convoluciones (CNN) preentrenadas fueron evaluadas con 12 semillas por modelo y protocolo. El F1-macro medio global disminuyó de 0.978 bajo el Protocolo A a 0.828 bajo el Protocolo B. ConvNeXt Base alcanzó 0.995 bajo validación aleatoria, mientras DenseNet201 obtuvo el mayor valor cronológico, 0.862. La degradación media fue de 0.15 puntos de F1-macro, con un rango intermodelo de 0.13 a 0.20. Los resultados muestran que una partición aleatoria puede producir una estimación sustancialmente más favorable que una evaluación que reserva observaciones tardías del ciclo para prueba. El objetivo de esta investigación no es establecer un nuevo ranking exhaustivo de CNN, sino destacar la importancia de alinear el protocolo de validación con el escenario temporal de uso del modelo.

Palabras clave: validación temporal; partición cronológica; clasificación agrícola; deficiencias nutrimentales; redes neuronales convolucionales; F1-macro.

1. Introducción

La lechuga (Lactuca sativa L.) es una hortaliza de interés alimentario por su contenido de compuestos bioactivos y polifenoles (Shi *et al*., 2022; Mitra *et al*., 2022), y diferentes trabajos han descrito propiedades y beneficios asociados a su consumo y composición (Sepehri *et al*., 2022; Lafarga *et al*., 2020). En México, su producción tiene relevancia agrícola y económica, incluida la región de Guanajuato (SIAP, 2023). En este contexto, la clasificación automática de deficiencias nutrimentales mediante imágenes RGB ofrece una alternativa no destructiva para apoyar el monitoreo de cultivos.

Las deficiencias de macronutrientes producen alteraciones cromáticas y morfológicas que pueden aprovecharse mediante visión por computadora (Xu *et al*., 2015; de Bang *et al*., 2021). Sin embargo, el diagnóstico visual puede ser subjetivo y los métodos analíticos convencionales implican costos y muestreo destructivo; por ello, el aprendizaje automático y profundo ha recibido atención creciente para el diagnóstico foliar (Nikitha *et al*., 2025; Soares *et al*., 2026; Barbedo, 2019). Aun así, la transferencia de estas herramientas a la agricultura aplicada requiere evaluar su comportamiento fuera de condiciones idealizadas, un problema también señalado para herramientas digitales de apoyo agronómico (Sida *et al*., 2023).

En este contexto, la forma de construir los conjuntos de entrenamiento, validación y prueba condiciona la interpretación del desempeño. Una partición aleatoria estratificada distribuye las unidades experimentales entre los subconjuntos manteniendo la proporción de clases y representa el escenario convencional en el que

las muestras futuras se asumen provenientes de la misma distribución. En cambio, una partición cronológica ordena las observaciones por fecha y reserva las etapas finales del ciclo para prueba.

Ambos protocolos responden a preguntas diferentes. El Protocolo A permite estimar el rendimiento bajo una distribución aproximadamente estacionaria, mientras que el Protocolo B introduce un desplazamiento temporal asociado con la evolución fenológica y las condiciones de adquisición a lo largo del ciclo. Por ello, una diferencia entre ambos resultados constituye evidencia de sensibilidad del sistema al criterio temporal de evaluación, sin aislar causalmente un único factor biológico o ambiental.

El objetivo de esta investigación es presentar, a partir de los experimentos, una lectura centrada en el protocolo de evaluación. Se utiliza el mismo experimento de nueve CNN como caso de estudio, pero el énfasis se desplaza del ranking exhaustivo de arquitecturas hacia tres preguntas: cómo se construyen las dos particiones, cuánto cambia la estimación de F1-macro y qué patrones de error aparecen cuando la prueba se desplaza hacia observaciones posteriores. La secuencia metodológica completa, desde la consolidación y procesamiento de los datos hasta la evaluación estadística de los Protocolos A y B, se resume en la Figura 1.

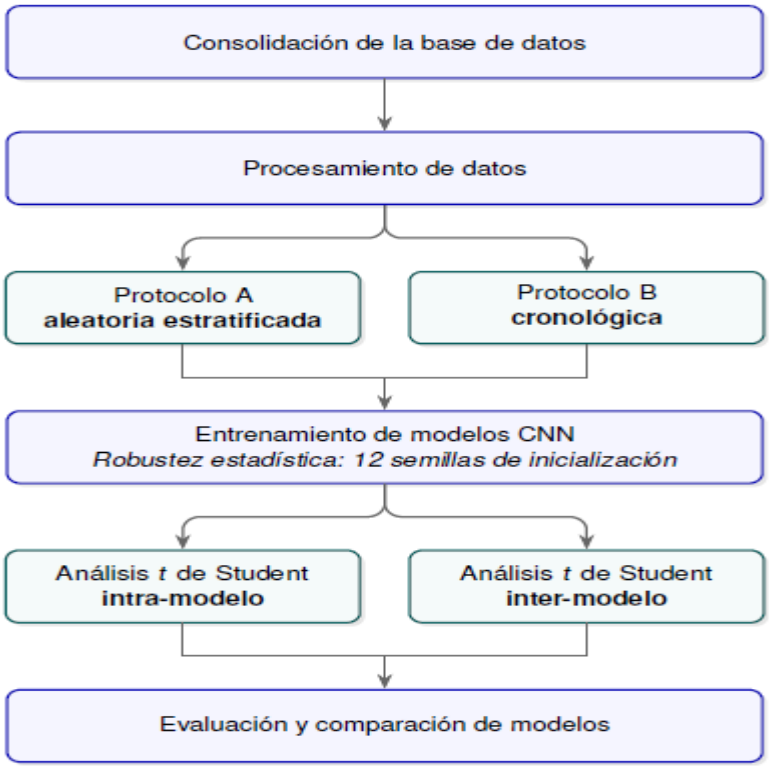


Figura 1. Diagrama de flujo de la metodología general, desde la consolidación de la base de datos hasta la evaluación, incluyendo los Protocolos A y B, las 12 semillas y los análisis estadísticos.
Fuente: Autoría propia.

2. Diseño Experimental y Protocolos de Evaluación

2.1. Caso de estudio y subconjunto de datos

El corpus LettuceDeF fue adquirido en invernadero mediante imágenes JPEG capturadas con un dispositivo Xiaomi 23100RN82L en posición cenital, durante el intervalo 09:00-17:00 h y bajo iluminación natural. Para el experimento de clasificación se seleccionaron cuatro clases: FN-Blanes como control y las clases -N, -P y -K. La clasificación automatizada de deficiencias en lechuga mediante visión artificial cuenta con

antecedentes recientes que incluyen modelos ligeros, métodos de visión por computadora y marcos de clasificación para sistemas hidropónicos (Sikati *et al*., 2023; Lu *et al*., 2023; Abidi *et al*., 2024). El subconjunto utilizado aquí contiene 4,405 imágenes y presenta un índice de desequilibrio ρ = 1.15; por ello se adoptó F1-macro como métrica principal, en concordancia con la necesidad de considerar el desbalance de clases en problemas de aprendizaje supervisado (He & Garcia, 2009). La distribución de imágenes por clase se presenta en la Tabla 1. La configuración de adquisición y ejemplos representativos de la variabilidad visual del conjunto se muestran en la Figura 2, incluyendo pigmentación verde, clorosis y patrones de necrosis observados a lo largo del experimento.

Tabla 1. Distribución del subconjunto LettuceDeF utilizado en el experimento.

Clase	Imágenes
Sana (FN-Blanes)	1,000
Deficiencia de nitrógeno (-N)	1,150
Deficiencia de fósforo (-P)	1,153
Deficiencia de potasio (-K)	1,102
Total	4,405

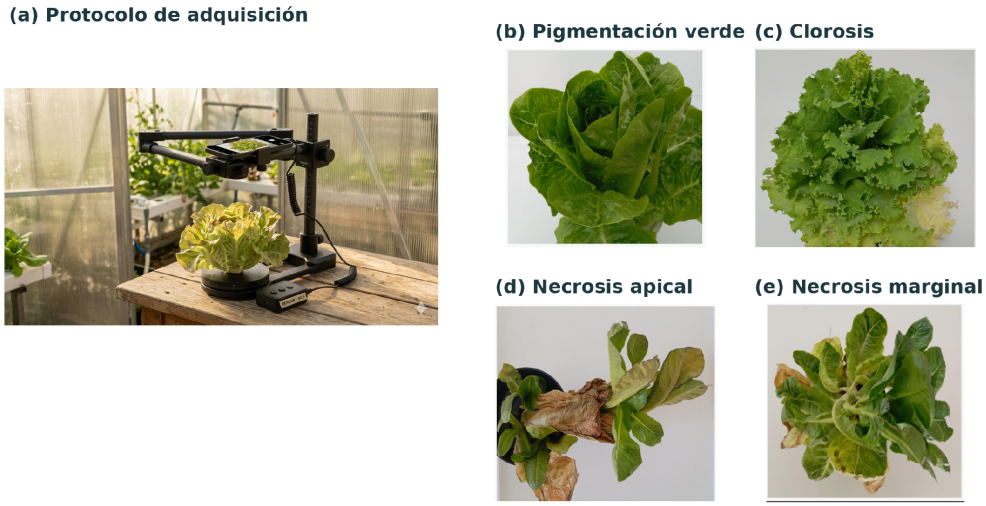


Figura 2. Contexto visual del conjunto LettuceDeF utilizado en el experimento: (a) configuración de adquisición cenital en invernadero; (b) pigmentación verde; (c) clorosis; (d) necrosis apical; y (e) necrosis marginal. Los paneles ilustran la variabilidad visual presente durante la clasificación de estados nutrimentales.
Fuente: Autoría propia.

2.2. Protocolos A y B

El subconjunto se dividió en entrenamiento (70%, 3,083 imágenes), validación (10%, 440) y prueba (20%, 881). En el Protocolo A, las unidades experimentales se asignaron aleatoriamente a los tres subconjuntos preservando la proporción relativa de las cuatro clases. En el Protocolo B, las observaciones se ordenaron temporalmente según la fecha de adquisición durante el ciclo de ocho semanas; entrenamiento y validación utilizaron registros de las primeras semanas y la prueba se reservó para imágenes de las etapas finales. En ambos protocolos se conservaron las mismas proporciones de partición. La Figura 3 ilustra de forma comparativa ambas estrategias de partición y su ejecución con nueve CNN y 12 semillas por protocolo.

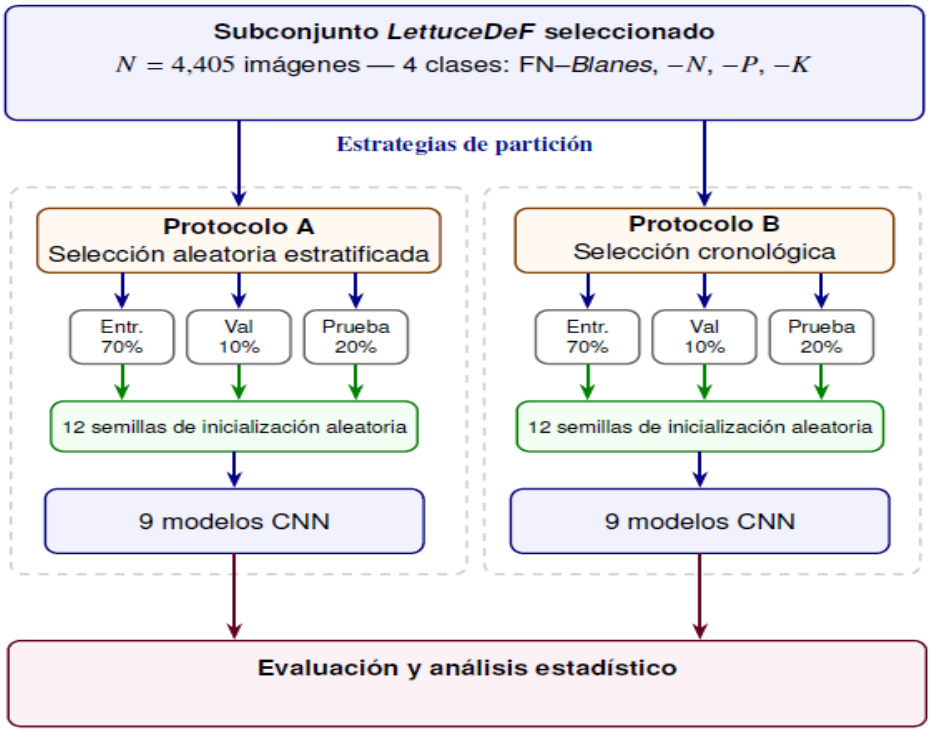


Figura 3. Esquema de partición del subconjunto de 4,405 imágenes, mostrando Protocolo A (selección aleatoria estratificada), Protocolo B (selección cronológica), proporciones 70/10/20, nueve CNN y 12 semillas por protocolo.
Fuente: Autoría propia.

3. Modelos y Evaluación

3.1. Arquitecturas y entrenamiento

Se evaluaron nueve arquitecturas CNN preentrenadas: ConvNeXt Base, DenseNet201, EfficientNet-B3, InceptionV3, MobileNetV2, NASNetMobile, ResNet50, VGG16 y Xception. Los modelos cubren un rango amplio de complejidad, desde 3.4 M parámetros en MobileNetV2 hasta 138 M en VGG16. La adaptación al dominio agrícola se realizó mediante aprendizaje por transferencia con ajuste fino completo y una capa de clasificación para las cuatro clases. El entrenamiento empleó AdamW, tasa de aprendizaje inicial de 5e-5, weight decay de 0.01, CosineAnnealingLR, un máximo de 30 épocas, early stopping con paciencia de 7 y pérdida de entropía cruzada. Las principales características de las nueve arquitecturas se resumen en la Tabla 2.

Tabla 2. Arquitecturas CNN consideradas en el caso de estudio.

Modelo	Parámetros [M]	Tamaño [px]	Lote
ConvNeXt Base	89	224	64
DenseNet201	20	224	128
EfficientNet-B3	12	300	64
InceptionV3	24	299	96
MobileNetV2	3.4	224	256
NASNetMobile	5.3	224	256

ResNet50	25.6	224	128
VGG16	138	224	64
Xception	22.9	299	96

3.2. Robustez estadística

Cada combinación modelo-protocolo se ejecutó con 12 semillas de inicialización aleatoria. El desempeño se resumió mediante F1-macro medio, desviación estándar e intervalo de confianza del 95%. Para cada arquitectura se compararon los resultados de los Protocolos A y B mediante una prueba t de Student pareada usando las mismas semillas como pares experimentales, con corrección de Bonferroni para controlar comparaciones múltiples. La comparación gráfica de los valores medios y sus intervalos de confianza se presenta en la Figura 4.

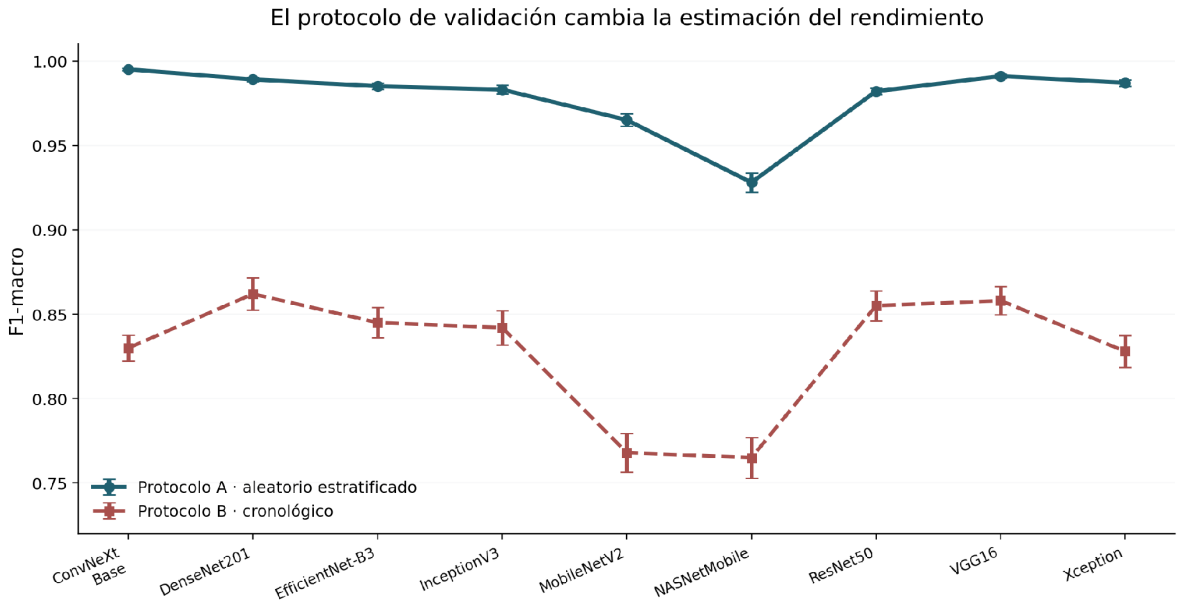


Figura 4. Comparación del F1-macro medio de las nueve CNN bajo los Protocolos A y B. Los puntos representan la media de 12 semillas y las barras de error el intervalo de confianza del 95%.
Fuente: Autoría propia.

3.3. Lectura centrada en el protocolo

La comparación no se interpreta únicamente como una competencia entre arquitecturas. El interés principal es observar si la conclusión sobre el rendimiento cambia cuando se modifica el criterio de partición. Como se observa en la Figura 4, el desempeño bajo el Protocolo A se mantiene próximo al techo para varias arquitecturas, mientras que bajo el Protocolo B aparece una reducción sistemática. Para ilustrar este efecto se consideran el mejor resultado del Protocolo A (ConvNeXt Base), el mejor del Protocolo B (DenseNet201) y la magnitud global de la degradación entre protocolos.

4. Resultados: Efecto del Protocolo de Validación

Tabla 3. F1-macro de las nueve CNN bajo los Protocolos A y B (media de 12 semillas).

Modelo	Protocolo A	Protocolo B	Diferencia A-B
ConvNeXt Base	0.995	0.830	0.165
DenseNet201	0.989	0.862	0.127
EfficientNet-B3	0.985	0.845	0.140
InceptionV3	0.983	0.842	0.141
MobileNetV2	0.965	0.768	0.197
NASNetMobile	0.928	0.765	0.163
ResNet50	0.982	0.855	0.127
VGG16	0.991	0.858	0.133
Xception	0.987	0.828	0.159
Media global	0.978	0.828	0.150

Los valores resumidos en la Tabla 3 muestran que, bajo el Protocolo A, seis de las nueve arquitecturas superaron 0.98 de F1-macro y ConvNeXt Base alcanzó 0.995. Bajo el Protocolo B, el rendimiento se redujo para todas las arquitecturas; DenseNet201 obtuvo el mayor F1-macro cronológico (0.862), mientras MobileNetV2 y NASNetMobile quedaron por debajo de 0.77. La media global pasó de 0.978 a 0.828.

La prueba t pareada intra-modelo mostró diferencias significativas entre protocolos para las nueve arquitecturas ($p < 0.001$ en todos los casos). La degradación media fue de 0.15 puntos de F1-macro (DE = 0.035), con un rango intermodelo de aproximadamente 0.13 para DenseNet201 a 0.20 para MobileNetV2. Estas diferencias, visibles en la Figura 4 y cuantificadas en la Tabla 3, muestran que la estimación de desempeño depende de manera importante del protocolo de evaluación y que la magnitud del efecto no es uniforme entre arquitecturas.

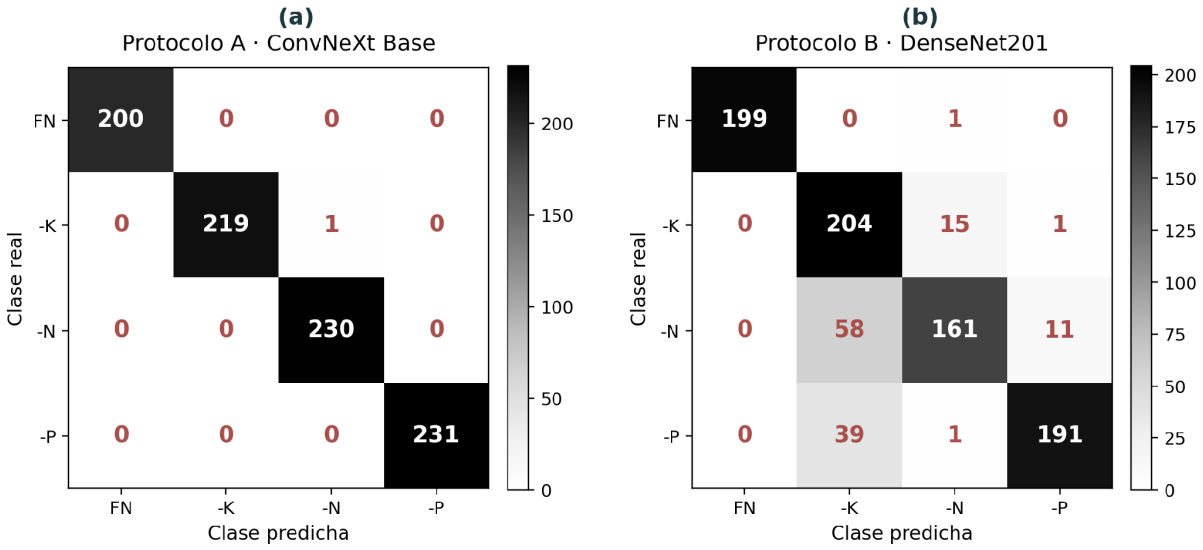


Figura 5. Matrices de confusión de los modelos de mayor rendimiento en cada protocolo: (a) ConvNeXt Base bajo Protocolo A y (b) DenseNet201 bajo Protocolo B. Las etiquetas corresponden a FN, -K, -N y -P.
Fuente: Autoría propia.

4.1. Contraste de las matrices de confusión

Como se muestra en la Figura 5(a), la matriz de ConvNeXt Base bajo el Protocolo A presenta 880 clasificaciones correctas de 881 instancias de prueba: FN 200/200, -K 219/220, -N 230/230 y -P 231/231. El único error corresponde a una instancia -K clasificada como -N. Este resultado representa el comportamiento casi saturado que puede observarse cuando entrenamiento y prueba se construyen mediante selección aleatoria estratificada.

En contraste, la Figura 5(b) muestra que bajo el Protocolo B DenseNet201 conserva un desempeño alto en FN (199/200) y -K (204/220), pero aparecen errores estructurados en las clases de deficiencia. De 230 instancias -N, 161 se clasificaron correctamente, 58 como -K y 11 como -P. Para -P, 191 de 231 fueron correctas y 39 se confundieron con -K. El contraste entre ambas matrices evidencia que el desplazamiento temporal modifica no solo el F1-macro agregado, sino también la estructura de los errores entre clases.

4.2. Implicación para la evaluación de modelos agrícolas

Tabla 4. Síntesis del contraste entre los dos protocolos.

Aspecto	Síntesis
Datos	Mismo subconjunto: 4,405 imágenes, cuatro clases
Partición	70% entrenamiento / 10% validación / 20% prueba
Protocolo A	Selección aleatoria estratificada por clase
Protocolo B	Selección cronológica; prueba reservada para etapas finales
Repeticiones	12 semillas por modelo y protocolo
Resultado global	F1-macro: 0.978 (A) vs. 0.828 (B)
Interpretación	La partición cronológica revela una brecha de rendimiento oculta por la evaluación aleatoria

La síntesis comparativa de ambos esquemas se presenta en la Tabla 4. Los resultados apoyan una recomendación metodológica concreta: cuando el objetivo de una aplicación agrícola incluye predecir observaciones adquiridas después del periodo utilizado para desarrollar el modelo, una evaluación aleatoria estratificada debe complementarse con una partición que respete el orden temporal. En este caso de estudio, el Protocolo A produce resultados próximos al techo para varias CNN, mientras que el Protocolo B revela una reducción sistemática y una mayor dificultad para distinguir deficiencias con sintomatología superpuesta. Este comportamiento es coherente con la literatura que reconoce tanto el potencial de la visión proximal y el aprendizaje profundo para detectar deficiencias nutricionales (Barbedo, 2019; Nikitha *et al*., 2025; Soares *et al*., 2026) como la necesidad de considerar limitaciones de transferencia y escalamiento en aplicaciones agronómicas digitales (Sida *et al*., 2023).

5. Conclusiones y Trabajo Futuro

El caso de estudio demuestra que el criterio de partición puede cambiar sustancialmente la estimación del rendimiento de clasificadores agrícolas. Como sintetizan la Tabla 3 y la Figura 4, con el mismo subconjunto de 4,405 imágenes las nueve CNN alcanzaron una media global de F1-macro de 0.978 bajo selección aleatoria estratificada y de 0.828 bajo selección cronológica. La caída se observó en todas las arquitecturas y fue estadísticamente significativa.

La contribución de esta versión de difusión es enfatizar la lectura metodológica del experimento: un valor elevado obtenido mediante partición aleatoria no implica que el mismo rendimiento se conserve cuando el

conjunto de prueba se desplaza hacia observaciones posteriores del ciclo. Por ello, la selección del protocolo debe corresponder con la pregunta de generalización que se desea responder.

Como trabajo futuro, se plantea ampliar la evaluación a múltiples ciclos de cultivo, variedades y condiciones ambientales, así como incorporar información temporal explícita y arquitecturas capaces de modelar secuencias. Estas extensiones permitirían estudiar la robustez temporal en escenarios más amplios que el ciclo de invernadero considerado en el experimento actual.

Referencias

Shi, M., Gu, J., Wu, H., & Rauf, A. (2022). Phytochemicals, nutrition, metabolism, bioavailability, and health benefits in lettuce—A comprehensive review. Antioxidants, 11(7), 1158. https://doi.org/10.3390/antiox11061158

Mitra, S., Tareq, A. M., Das, R., Emran, T. B., Nainu, F., Chakraborty, A. J., Ahmad, I., Tallei, T. E., Idris, A. M., & Simal-Gandara, J. (2022). Polyphenols: A first evidence in the synergism and bioactivities. Food Reviews International, 1–23. https://doi.org/10.1080/87559129.2022.2026376

Sepehri, N. Z., Parvizi, M. M., Habibzadeh, S., & Handjani, F. (2022). Lettuce as an effective remedy in uremic pruritus: Review of the literature supplemented by an in silico study. Evidence-Based Complementary and Alternative Medicine, 2022, Article 4231854. https://doi.org/10.1155/2022/4231854

Lafarga, T., Villaró, S., Rivera, A., Bobo, G., & Aguiló-Aguayo, I. (2020). Bioaccessibility of polyphenols and antioxidant capacity of fresh or minimally processed modern or traditional lettuce (Lactuca sativa L.) varieties. Journal of Food Science and Technology, 57, 754–763. https://doi.org/10.1007/s13197-019-04108-7

Servicio de Información Agroalimentaria y Pesquera (SIAP). (2023). Estadísticas de producción de hortalizas.

Xu, S., Lin, W., Wu, W., & Zhao, H. (2015). Image diagnosis of rapeseed deficiency based on color features. Chinese Journal of Oil Crop Sciences, 37(4), 576–582.

de Bang, T. C., Husted, S., Laursen, K. H., Persson, D. P., & Schjoerring, J. K. (2021). The molecular–physiological functions of mineral macronutrients and their consequences for deficiency symptoms in plants. New Phytologist, 229(5), 2446–2469. https://doi.org/10.1111/nph.17074

Nikitha, S., Prabhanjan, S., & Sathyanarayan, A. (2025). Plant nutritional deficiency detection: A survey of predictive analytics approaches. Iran Journal of Computer Science, 8(1), 83–101. https://doi.org/10.1007/s42044-024-00209-5

Soares, C. C., Regazzo, J. R., da Silva, T. L., Tavares, M. S., Devechio, F. F. S., Silva, R. M., Tech, A. R. B., & Baesso, M. M. (2026). Applications of machine learning and deep learning for foliar nutritional deficiency: A systematic review. AgriEngineering, 8(3), 101. https://doi.org/10.3390/agriengineering8030101

Sida, T. S., Gameda, S., Chamberlin, J., Andersson, J. A., Getnet, M., Woltering, L., & Craufurd, P. (2023). Failure to scale in digital agronomy: An analysis of site-specific nutrient management decision-support tools in developing countries. Computers and Electronics in Agriculture, 108060. https://doi.org/10.1016/j.compag.2023.108060

Barbedo, J. G. A. (2019). Detection of nutrition deficiencies in plants using proximal images and machine learning: A review. Computers and Electronics in Agriculture, 162, 482–492.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Sikati, J., *et al*. (2023). YOLO-NPK: A lightweight deep network for lettuce nutrient deficiency classification based on improved YOLOv8 nano. In 10th International Electronic Conference on Sensors and Applications (ECSA-10), 58, 31. https://doi.org/10.3390/ecsa-10-16256

Lu, J., *et al*. (2023). Lettuce plant trace-element-deficiency symptom identification via machine vision methods. Agriculture, 13(8), 1614. https://doi.org/10.3390/agriculture13081614

Abidi, M. H., *et al*. (2024). Elucidation of intelligent classification framework for hydroponic lettuce deficiency using enhanced optimization strategy and ensemble multi-dilated adaptive networks. IEEE Access. https://doi.org/10.1109/ACCESS.2024.3392482