REVISTA AMBIENTAL
AGUA, AIRE Y SUELO

Extracción De Características Visuales Para La Clasificación Y Análisis De Plántulas En El Dataset Plant Seedlings Empleando Ciencia De Datos

Visual Feature Extraction For Seedling Classification And Analysis In The Plant Seedlings Dataset Using Data Science.

Niño-Rondón, C. V.[1]; López-Bustamante, O. A.[2]

*[1]**Magister en Ciencia de Datos Carlos Vicente Niño Rondón.** Programa de Ingeniería Electrónica. Facultad de Ingeniería. Universidad Francisco de Paula Santander. e-mail: carlosvicentenr@ufps.edu.co. Orcid: https://orcid.org/0000-0002-3781-4564*
*[2] **Magister en Educación Matemática Oriana Alexandra López Bustamante.** Programa de Ingeniería Electrónica. Facultad de Ingeniería. Universidad Francisco de Paula Santander. e-mail: orianaalexandralb@ufps.edu.co. Orcid: https://orcid.org/0000-0003-4601-1111*

Universidad Francisco de Paula Santander
Avenida Gran Colombia No. 12E-96, Cúcuta, Norte de Santander

Recibido: 03/02/2025 / Aceptado: 04/07/2025

Resumen

Este estudio detalla un enfoque para obtener, examinar y ponderar atributos visuales tomados de fotos del conjunto de datos Plant Seedlings, buscando perfeccionar la segmentación automática de brotes jóvenes. Se utilizaron métodos de edición de imágenes para conseguir indicadores de trama (LBP y rasgos de Haralick), cromaticidad (histogramas en el ambiente HSV), silueta (bordes y medidas geométricas) y datos básicos (promedio, diferencia estándar, curtosis y sesgo). Después, se realizó un análisis estadístico explicativo de los atributos recopilados, que reveló una gran diversidad entre variables y poca multicolinealidad, lo que sugiere que los indicadores recogen facetas variadas y que se complementan en las fotos. Los resultados sugieren que la combinación de múltiples tipos de descriptores mejora la capacidad discriminativa para la clasificación de especies vegetales. Esta metodología sienta las bases para el desarrollo de sistemas inteligentes aplicados a la agricultura de precisión, integrando ingeniería electrónica y ciencia de datos para lograr decisiones agronómicas más eficientes y sostenibles.

Palabras clave: Ciencia de datos, agricultura, procesamiento de imágenes, extracción de características, histogramas.

ISSN 1900-9178 Volumen 16 N°2 - 2025
UNIVERSIDAD DE PAMPLONA

Abstract

This study details an approach to obtain, examine and weight visual attributes taken from photos of the Plant Seedlings dataset, seeking to refine the automatic segmentation of young shoots. Image editing methods were used to obtain indicators of raster (LBP and Haralick features), chromaticity (histograms in the HSV environment), silhouette (edges and geometric measures), and basic data (average, standard difference, kurtosis, and skewness). An explanatory statistical analysis of the attributes collected was then performed, revealing a high diversity among variables and little multicollinearity, suggesting that the indicators capture varied facets and that they complement each other in the photos. The results suggest that combining multiple types of descriptors improves the discriminative power for plant species classification. This methodology lays the foundation for the development of intelligent systems in precision agriculture, integrating electronic engineering and data science to support more efficient and sustainable agronomic decision-making.

Keywords: Data science, agriculture, image processing, feature extraction, histograms.

1. INTRODUCCIÓN

La agricultura de precisión ha experimentado un crecimiento exponencial en las últimas décadas gracias a la integración de tecnologías avanzadas como la ingeniería electrónica, la visión por computadora y la ciencia de datos (Altalak *et al.,* 2022). En este contexto, el monitoreo temprano y preciso del desarrollo vegetal se ha convertido en un componente crucial para optimizar la productividad, reducir el uso de insumos y minimizar impactos ambientales negativos. Particularmente, la identificación y clasificación de plántulas en etapas iniciales del crecimiento son fundamentales para la toma de decisiones en manejo fitosanitario, fertilización y riego, lo cual impacta directamente en la eficiencia agrícola (Juwono *et al.,* 2023) .

El dataset Plant Seedlings, ampliamente utilizado en investigaciones de clasificación automática de especies vegetales, contiene imágenes digitales de plántulas pertenecientes a diversas especies (Sharma *et al.,* 2020).

La obtención de características visuales es una fase fundamental en este procedimiento (Wang *et al.,* 2020), que implica la recolección de descriptores que reflejan detalles importantes sobre la textura, color, forma y cualidades estadísticas de las imágenes (Zebari *et al.,* 2020). Entre los descriptores relacionados con la textura, métodos como el patrón binario local (Local Binary Pattern, LBP) y las características de Haralick derivadas de matrices de co-ocurrencia ayudan a identificar patrones espaciales y la rugosidad de las hojas, los cuales pueden diferir entre distintas especies y fases de crecimiento (Zhou *et al.,* 2020). Por otra parte, los histogramas en el espacio de color (Hue, Saturation, Value, HSV) proporcionan una representación sólida ante cambios en la iluminación, lo que permite distinguir los matices y niveles de saturación característicos de cada planta. Respecto a la forma, la obtención de contornos y medidas geométricas (como área, perímetro y excentricidad) ayuda a cuantificar la morfología, que es un factor clave para diferenciar especies con estructuras foliares distintivas (Castro

Casadiego *et al.,* 2021). Asimismo, se añaden descriptores estadísticos básicos como la media, la desviación estándar y medidas de asimetría, que enriquecen la descripción general de la imagen y pueden ser útiles para reconocer variaciones dentro de una especie o condiciones específicas del cultivo. La integración de estos descriptores produce un vector de características multidimensional que sirve como cimiento para análisis exploratorios posteriores, clasificación supervisada y modelado predictivo (Ali *et al.,* 2017).

En este texto se ofrece un enfoque detallado para obtener y analizar características visuales del conjunto de datos Plant Seedlings, utilizando métodos sólidos de procesamiento de imágenes y análisis de datos. Se aborda la capacidad diferenciadora de cada tipo de descriptor, su papel en la clasificación correcta de especies y su posible uso en sistemas electrónicos para la agricultura de precisión. Así, se ayuda a fortalecer la colaboración entre la ingeniería electrónica y el análisis de datos en el campo agroambiental, creando oportunidades para soluciones tecnológicas sostenibles y efectivas.

2. METODOLOGÍA

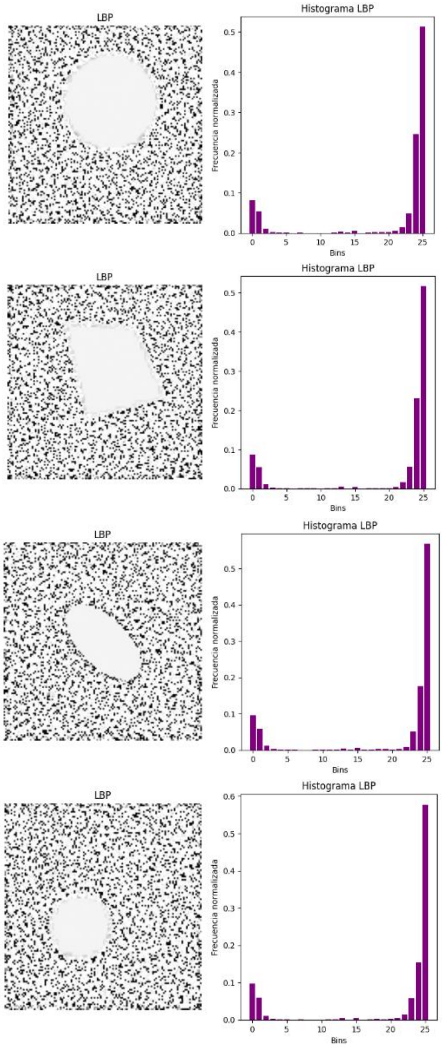
Inicialmente, se procede a reunir y adecuar las fotos del conjunto de datos de Plántulas. En esta etapa, se bajan las imágenes, se clasifican según el tipo de plántula y se ajustan su tamaño y resolución para garantizar uniformidad en el análisis. Adicionalmente, se aplican técnicas de optimización de imagen, como la normalización del histograma y la corrección de la luz, para disminuir las diferencias provocadas por el entorno al tomar las fotos. Luego, se identifican atributos visuales concretos, que se transforman en la base cuantitativa para el siguiente análisis de datos. Para definir la textura de las plántulas, se emplean descriptores como el Patrón Binario Local (LBP), que cifra la textura local en patrones binarios, y las propiedades de Haralick extraídas de la matriz de co-ocurrencia, permitiendo captar atributos como contraste, homogeneidad y entropía. Estos descriptores brindan información sobre la estructura superficial y los patrones reiterativos en la imagen, cruciales para identificar diferentes especies. Simultáneamente, se consiguen propiedades de color del espacio HSV (Han & Lee, 2020), que separa la información en matiz, saturación y brillo (*Hue, Saturation, Value*, HSV). Se crean histogramas de color para cada parte, aportando una representación robusta ante cambios de luz y facilitando la identificación de tonalidades singulares de cada especie. Este estudio es muy importante dado que el color es una característica innata y fácilmente reconocible en las plantas jóvenes. Para registrar la forma y morfología, se efectúa una segmentación sencilla de la planta en relación al fondo usando umbralización adaptativa y métodos para descubrir contornos (Yadav & Yadav, 2020). Partiendo de los contornos detectados, se calculan métricas geométricas como área, perímetro, circularidad, excentricidad y relación de aspecto. Estas mediciones describen la forma general y complejidad de las plántulas, ofreciendo información que se añade al análisis de textura y color para una caracterización más completa. Para terminar, se recogen estadísticas básicas de cada imagen, incluyendo media, desviación estándar, curtosis y asimetría, las cuales aportan datos adicionales sobre la distribución de intensidades y características generales. La combinación

ISSN 1900-9178 Volumen 16 N°2 - 2025
UNIVERSIDAD DE PAMPLONA

de todos estos descriptores conforma un vector de características multidimensional para cada muestra, que sirve como insumo para los análisis posteriores de clasificación y exploración de datos.

3. RESULTADOS

En la figura 1 se muestran las extracciones LBP y los histogramas para 5 imágenes de muestras aleatorias.



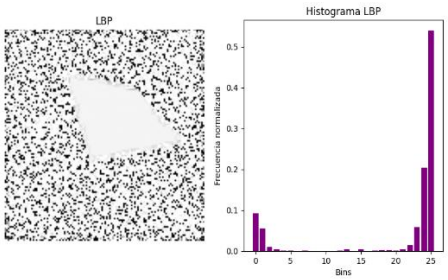
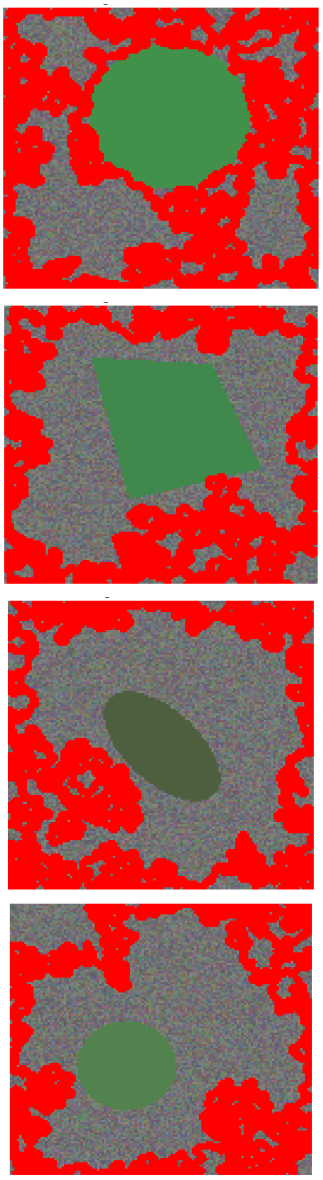


Figura 1. *Histogramas y extracciones LBP.*

De igual forma, en la figura 2 se presenta la extracción y análisis de contornos en las imágenes de prueba aleatorias.



ISSN 1900-9178 Volumen 16 N°2 - 2025
UNIVERSIDAD DE PAMPLONA

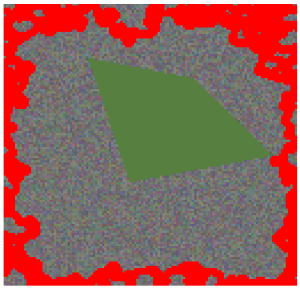
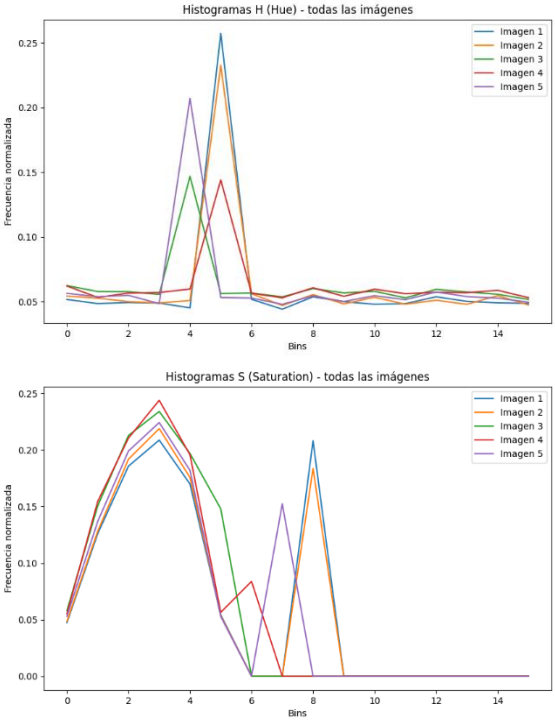


Figura 2. *Extracción de contornos.*

Asimismo, en la figura 3 se muestran los histogramas HSV, para H, S y V respectivamente.



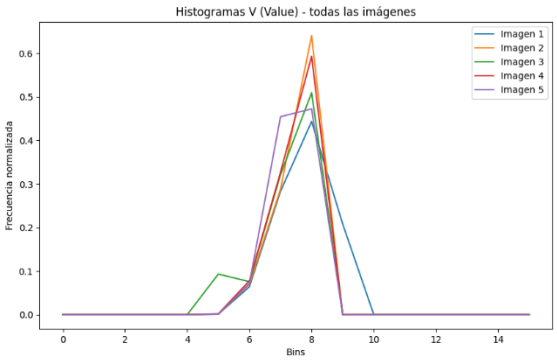


Figura 3. *Histogramas HSV*

Asimismo, se presenta el análisis estadístico de las características extraídas. La distribución estadística se presenta en la figura 4.

ISSN 1900-9178 Volumen 16 N°2 - 2025
UNIVERSIDAD DE PAMPLONA

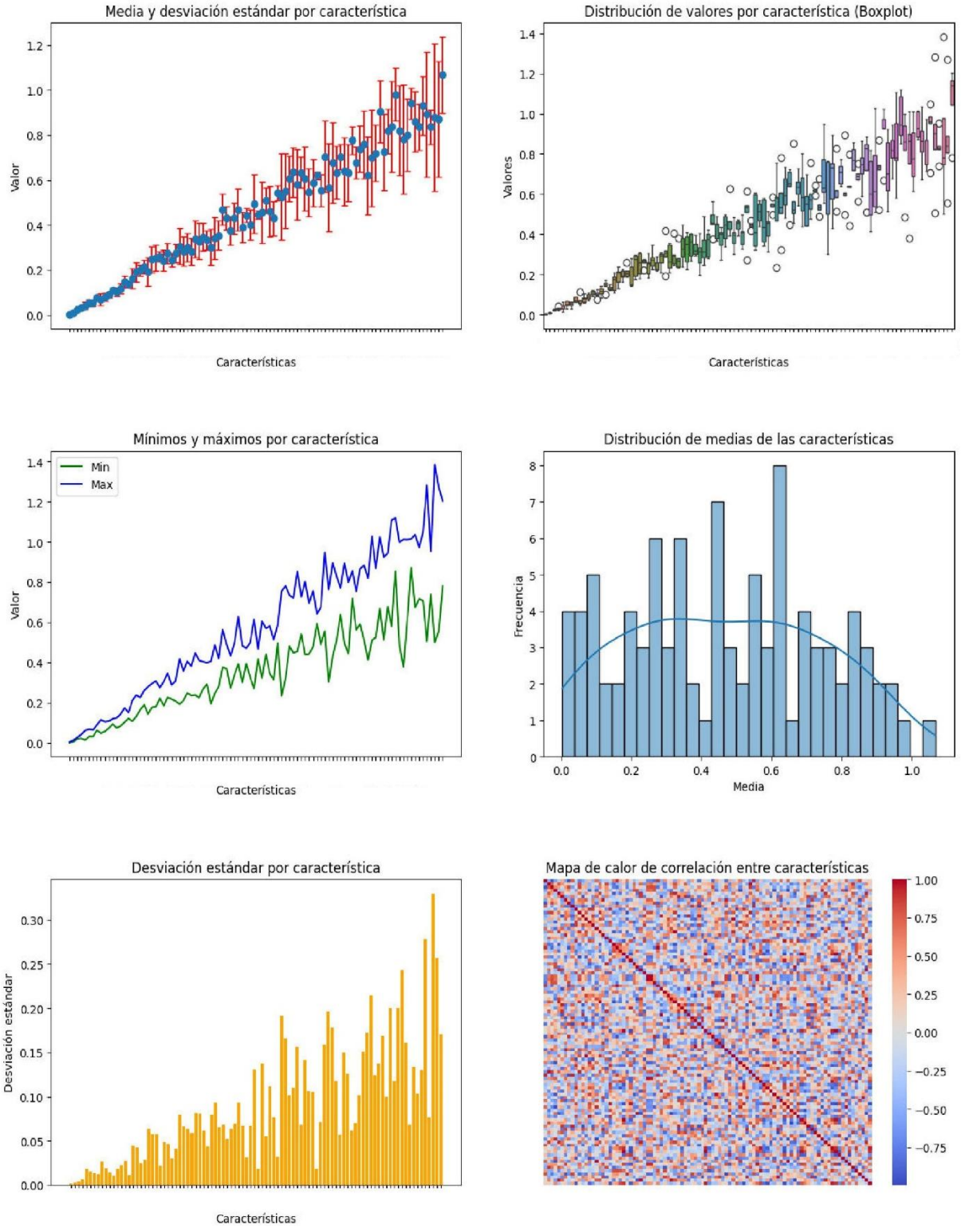


Figura 4. *Análisis estadístico de características del conjunto de datos*

ISSN 1900-9178 Volumen 16 N°2 - 2025
UNIVERSIDAD DE PAMPLONA

4. ANÁLISIS Y DISCUSIÓN DE RESULTADOS

Tras una inspección inicial de la información, se observa que los promedios de los atributos se mueven, más o menos, entre 0. 0 y 1. 2, viéndose un aumento gradual a medida que avanzamos en el eje de atributos. En el primer gráfico, las líneas de error nos dicen que la dispersión típica anda entre 0. 01 y 0. 3, notándose un incremento importante hacia los últimos atributos. Los diagramas de caja apoyan esta variabilidad, demostrando que casi todos los atributos tienen datos atípicos y una dispersión importante en los valores, llegando algunos hasta 1. 4. Los valores más bajos están generalmente entre 0. 0 y 0. 6, y los más altos llegan hasta 1. 4, lo que indica una amplitud mayor por atributo. La forma en que se distribuyen los promedios muestra una concentración entre 0. 3 y 0.6, con algunos atributos llegando a un promedio cercano a 1. 0, lo que insinúa una ligera inclinación hacia valores medio-altos. La gráfica de dispersión típica apoya esta idea: la mayoría de los atributos tienen dispersiones menores a 0. 2, pero algunos pasan de 0. 3, señalando mayor dispersión. Finalmente, el mapa de calor de correlación muestra una matriz en su mayoría con correlaciones bajas entre atributos, con coeficientes que van de -0. 75 a 0. 75, lo que indica poca multicolinealidad e independencia relativa entre las variables.

5. CONCLUSIONES

La obtención de atributos visuales diversos (como la textura, el color, la configuración geométrica y datos numéricos) posibilita una descripción extensa y sólida de las imágenes de las plantas jóvenes, lo cual ayuda a distinguir las diferentes especies desde sus primeros momentos de vida. El análisis estadístico exploratorio reveló una alta variabilidad entre características, con medias entre 0.0 y 1.2 y desviaciones estándar de hasta 0.3, lo cual demuestra la riqueza informativa del vector de características y su potencial discriminativo. La baja correlación observada entre características (coeficientes entre -0.75 y 0.75) sugiere que los descriptores capturan propiedades distintas de las imágenes, lo que favorece el desempeño de modelos clasificadores al evitar redundancia en los datos. Las técnicas utilizadas, como LBP, Haralick e histogramas HSV, resultaron eficaces para representar patrones específicos de cada especie, mostrando su utilidad en aplicaciones reales de clasificación vegetal. Este enfoque metodológico puede ser integrado en sistemas electrónicos para agricultura de precisión, proporcionando una herramienta valiosa para monitoreo automatizado, manejo agronómico y toma de decisiones sustentables en entornos agrícolas.

6. REFERENCIAS BIBLIOGRÁFICAS

Ali, H., Lali, M. I., Nawaz, M. Z., Sharif, M., & Saleem, B. A. (2017). Symptom based automated detection of citrus diseases using color histogram and textural descriptors. *Computers and Electronics in Agriculture*, *138*, 92–104. https://doi.org/10.1016/j.compag.2017.04.008

Altalak, M., Uddin, M. A., Alajmi, A., & Rizg, A. (2022). Smart Agriculture Applications Using Deep Learning Technologies: A Survey. *Applied Sciences (Switzerland)*, *12*(12). https://doi.org/10.3390/app12125919

Castro Casadiego, S. A., Medina Delgado, B., Guevara Ibarra, D., Puerto López,

K., Sánchez Mojica, K., & Niño Rondón, C. V. (2021). Efecto de los filtros morfológicos en los procesos de detección de objetos en movimiento. *Mundo Fesc*, *11*(21), 87–95. https://www.fesc.edu.co/Revistas/OJS/index.php/mundofesc/article/view/676

Han, J., & Lee, C. (2020). Color Lane Line Detection Using the Bhattacharyya Distance. *11th International Conference on Information, Intelligence, Systems and Applications, IISA 2020*, 6–9. https://doi.org/10.1109/IISA50023.2020.9284147

Juwono, F. H., Wong, W. K., Verma, S., Shekhawat, N., Lease, B. A., & Apriono, C. (2023). Machine learning for weed–plant discrimination in agriculture 5.0: An in-depth review. In *Artificial Intelligence in Agriculture* (Vol. 10, pp. 13–25). KeAi Communications Co. https://doi.org/10.1016/j.aiia.2023.09.002

Sharma, P., Berwal, Y. P. S., & Ghai, W. (2020). Performance analysis of deep learning CNN models for disease detection in plants using image segmentation. *Information Processing in Agriculture*, *7*(4), 566–574. https://doi.org/10.1016/j.inpa.2019.11.001

Wang, C., Li, G., Xue, P., & Wu, Q. (2020). A Comparative Study of Face Recognition Classification Algorithms. *International Journal of Advanced Network, Monitoring and Controls*, *5*(3), 23–29. https://doi.org/10.21307/ijanmc-2020-024

Yadav, S. P., & Yadav, S. (2020). Image fusion using hybrid methods in multimodality medical images. *Medical and Biological Engineering and Computing*, *58*(4), 669–687. https://doi.org/10.1007/s11517-020-02136-6

Zebari, R., Abdulazeez, A., Zeebaree, D., Zebari, D., & Saeed, J. (2020). A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction. *Journal of Applied Science and Technology Trends*, *1*(2), 56–70. https://doi.org/10.38094/jastt1224

Zhou, W., Gao, S., Zhang, L., & Lou, X. (2020). Histogram of Oriented Gradients Feature Extraction from Raw Bayer Pattern Images. *IEEE Transactions on Circuits and Systems II: Express Briefs*, *67*(5), 946–950. https://doi.org/10.1109/TCSII.2020.2980557