Machine Learning como herramienta para monitoreo e identificación rápida de cianobacterias

Silvina M. Rosa[1], Lilen Yema[3], Patricia L. M. Torres[1,2], María E. Buemi[4], Rocío Balderrama[5], María Sofía Plastani[6], Agustín Sanguinetti[1,2] y Cristian Martínez [7]

Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Argentina: [1]DBBE; [2]IBBEA-CONICET; [3]IEGEBA-CONICET; [4]DC, ICC-CONICET; [5]DM, IMAS-CONICET, [6]IDEAN-CONICET.

[7]Departamento de Informática, Universidad Nacional de Salta, Argentina.

cmartinez@di.unsa.edu.ar

Abstract. Las floraciones de cianobacterias son crecimientos masivos de estos microorganismos acuáticos, que tienen alta relevancia ecológica, sanitaria y económica. Las instituciones públicas y privadas responsables de la gestión de cuerpos de agua no siempre cuentan con personal especializado, por lo que disponer de herramientas automáticas, precisas y sencillas que sirvan de soporte para su detección y caracterización es de suma importancia. Proponemos una solución basada en Machine Learning para el reconocimiento de cianobacterias formadoras de floraciones mediante dos fases: detección y clasificación. Para medir la calidad de nuestra propuesta, se utilizó un dataset propio de imágenes de cianobacterias de diversos cuerpos de agua de Argentina, siendo los géneros seleccionados los principales responsables de floraciones en Argentina en base a relevamientos previos. A través de una red SOM para la detección de cianobacterias y una red CNN pre-entrenada para la clasificación de géneros, se alcanzaron resultados muy prometedores, considerando la poca disponibilidad de muestras, la complejidad de los microorganismos y su poco tratamiento en la literatura.

Keywords: aprendizaje automático - cianobacterias - identificación taxonómica - redes neuronales.

1 Introducción

Las cianobacterias (phylum Cyanobacteria) son un grupo de microorganismos fotosintéticos oxigénicos que se encuentran principalmente en cuerpos de agua continentales (lagos, lagunas, embalses, ríos). Con un total de 5844 especies agrupados en más de 200 géneros [6], estos microorganismos pueden generar sobrecrecimientos masivos denominados floraciones (*blooms*) cuando las condiciones ambientales son propicias. Estas floraciones, que suelen estar dominadas por una o pocas especies, representan un importante problema ambiental, sanitario y económico debido a la producción de toxinas y olores, con un impacto negativo en la salud humana, la fauna y los costos asociados a la potabilización del agua. En las últimas décadas, las floraciones han aumentado en intensidad y frecuencia debido a la eutrofización antrópica y al cambio climático, y la tendencia indica que seguirán aumentando globalmente [2,4]. Esta situación plantea un desafío para las instituciones encargadas de la gestión ambiental de los cuerpos de agua dulce, quienes deben contar

con sistemas de monitoreo eficientes que incluyan alertas tempranas para detectar el inicio de las floraciones de cianobacterias. La implementación de este sistema requiere personal capacitado en la identificación taxonómica de estos microorganismos y en la estimación de su abundancia, lo cual conlleva tiempo y recursos especializados cada vez más escasos.

En la última década, el uso de Machine Learning (Aprendizaje Automático) para la identificación taxonómica de algas y cianobacterias, dio lugar a publicaciones con resultados que destacan la importancia de ahondar en esta línea. Mosleh *et al.* [9] utiliza procesamiento de imágenes y redes neuronales sobre un dataset de imágenes de muestras de agua de un lago de Malasia donde se incluyen 3 géneros de cianobacteria y 2 de algas, totalizando 100 imágenes por género. Para el reconocimiento y clasificación utilizaron una Multilayer Perceptron (MLP) de 3 capas, con una capa de entrada de 21 neuronas, una capa oculta de 8 y una capa de salida de 5, alcanzando una precisión de 93%. Schulze *et al.* [11] proponen un enfoque similar para la identificación de fitoplancton en Alemania. En este caso el dataset está conformado por 10 clases (10 especies, incluyendo 1 cianobacteria), con 600 imágenes por clase. Presentan una MLP con 2 capas ocultas (la primera con 50 neuronas y la segunda con 30), donde la capa de entrada recibe características de formas, texturas, dimensión, entre otras, obteniendo una precisión de 94.7%. Gandola *et al.* [5] proponen la clasificación de 5 géneros de cianobacterias filamentosas y dan un reporte detallado por imagen: cantidad, género y dimensiones de las células detectadas en el dataset que consta de imágenes de cianobacterias recolectadas en lagos de Italia, conteniendo 120 por clase. Para clasificar los géneros, usaron 17 indicadores numéricos y 15 parámetros de superficie. La clasificación se realizó a través de un modelo Random Forest (RF) de 200 árboles. Informan una matriz de confusión (de test) con valores de clasificación correcta que varían del 89.1% al 100%. Más recientemente, Baek *et al.* [3] presentan una solución basada en Deep Learning para la detección, identificación y conteo de células de cuatro géneros de cianobacterias. El dataset está constituido por 250 imágenes por clase provenientes de muestras de ríos de Corea del Sur. La identificación de cianobacterias es realizada a través de una Fast R-CNN mientras que el conteo de células de *Microcystis* es a través de una CNN.

A diferencia de los trabajos previos, nuestro estudio se enfoca en utilizar un dataset propio de cianobacterias de Argentina. Según una revisión reciente sobre la distribución de cianobacterias formadoras de floraciones en nuestro país [10], se ha observado que los géneros más abundantes en los cuerpos de agua afectados por este fenómeno son *Anabaenopsis* (8%), *Cylindrospermopsis* (7%), *Dolichospermum* (33%), *Microcystis* (21%), *Planktothrix* (3%), y *Raphidiopsis* (11%). Aguilera *et al.* [1] también presentan a los géneros *Cylindrospermopsis*, *Dolichospermum* y *Microcystis* como los más abundantes de las floraciones en Argentina. Nuestro estudio propone el uso de Machine Learning para generar una herramienta que detecte la presencia de cianobacterias y clasifique los géneros más frecuentes que causan floraciones en Argentina. Este enfoque tiene como objetivo facilitar una respuesta temprana ante el riesgo de floraciones nocivas para la salud y el ambiente mediante el procesamiento de imágenes de muestras de agua en laboratorio.

2 Materiales y Métodos

El dataset se compone de imágenes obtenidas de muestras tomadas de diversos cuerpos de agua de Argentina: lagos urbanos de la Ciudad Autónomas de Buenos Aires (Lugano, Rosedal), lagunas pampeanas de la provincia de Buenos Aires (Cuero de Zorro, Chascomús, Los Horcones, Lobos, Hinojo, Gómez, Los Coipos), lagos y lagunas patagónicas (Cervecero, La Soñada), embalse Ramos Mexía de El Chocón, ríos (Río Uruguay, Río de la Plata); y de cultivos de laboratorio.. En base a los trabajos de Aguilera *et al.* [1] y O'Farrell *et al.* [10], se seleccionaron los géneros de cianobacterias más frecuentes en floraciones tóxicas en la Argentina. Las muestras de los cuerpos de agua se tomaron con red de fitoplancton, para el análisis cualitativo en un microscopio óptico (Olympus CX41), y/o se obtuvieron sin concentrar para el análisis cuantitativo mediante el método de Utermöhl (1958) [15] con un microscopio óptico invertido (Olympus CKX31). En el caso de los cultivos, las imágenes se obtuvieron con el mismo microscopio óptico, durante el análisis cuantitativo con cámara de Palmer. El dataset es heterogéneo, con imágenes de dimensiones variadas (desde 259*195 hasta 2048*1536 píxeles) y resoluciones desde 96 a 300 ppi; consta de 20 imágenes del género *Anabaenopsis*, 18 de *Cylindrospermopsis*, 20 de *Dolichospermum*, 21 de *Microcystis*, 21 de *Planktothrix* y 23 de *Raphidiopsis*. En Fig. 1 se presentan algunos ejemplos de imágenes de las muestras de cada género que son tratadas.

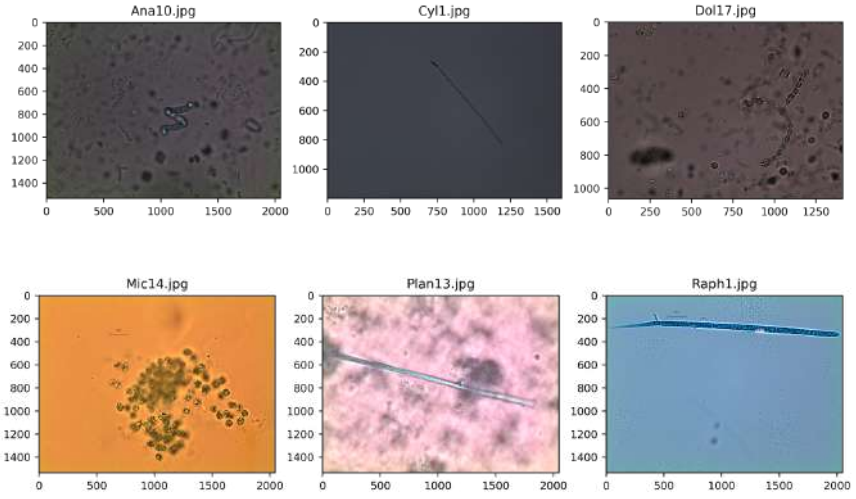

Fig. 1. Imágenes del dataset.

Este trabajo comprende 2 etapas; en primer lugar la detección de cianobacteria a partir de una imagen y si corresponde, la clasificación según su género. Para la detección, se usó un mapa auto-organizado (SOM, conocido como mapas de Kohonen) [8] por su rapidez, calidad de los resultados y posibilidad de evaluación numérica de la red. El

modelo SOM desarrollado toma como datos de entrada el dataset de imágenes, las cuales son pre-procesadas y luego reducidas en su dimensión aplicando Componentes Principales [7]. Una vez entrenado, el modelo es evaluado usando un conjunto de imágenes de cianobacterias que no pertenecen a los géneros elegidos.

Para las pruebas se consideraron los siguientes parámetros (basados en experimentos previos): redimensión de la imagen de entrada → [(168,200), (205,250), (246, 300)]; iteraciones → [1000, 5000, 10000]; sigma → [0.5, 1, 2] ; learning rate → [0.001, 0.5]; grilla → [(20, 20), (25, 25), (30,30)]; vecindario → [triangle, gaussian, mexican hat].

Respecto a la clasificación de géneros, y considerando la características del dataset, se propone el uso de CNN. La elección está motivada por los excelentes resultados que han tenido en clasificación de imágenes. Por otra parte, una forma de trabajar con ellas es usando redes previamente pre-entrenadas con grandes bases de datos y adaptarlas al problema de estudio. En nuestro caso se usaron: la red profunda residual ResNet-18 [12], la red profunda separable MobileNet-v2 [13] y la red basada en el aprendizaje de capas de arquitecturas Inception conocida como GoogleNet[14]; previamente se aplicaron tareas de preprocesamiento sobre las imágenes.

A los efectos de una comparación más objetiva de los modelos para clasificación de género de cianobacterias, fueron considerados los siguientes parámetros: redimensión de la imagen de entrada → [(134,160), (168,200), (205,250), (246, 300)]; epochs → [50, 100]; batch_size → [16, 32, 64]; learning rate → 0.001; y optimizador → ADAM.

Para el entrenamiento y prueba de los 3 modelos basados en CNN pre-entrenadas, se dividió el dataset en 80% para entrenamiento y 20% para prueba. Del total de imágenes para entrenamiento, se usó un 20% para validación.

3 Resultados y análisis

Respecto a los experimentos realizados sobre el modelo SOM, los mejores valores de los parámetros son: redimensión de imagen → (205,250), iteraciones → 10000, sigma → 1, learning rate → 0.5, grilla → (25,25), vecindario → gaussian. El error de cuantización del modelo entrenado fue de 0.94. Para la prueba de detección, se usó un umbral de 0.90 sobre el mayor error de cuantización de los datos de entrada del modelo. Se usaron 25 imágenes externas al dataset y se calculó el error de cuantización correspondiente. En todos los casos, los mismos estuvieron por encima del umbral usado. Los resultados alcanzados indican que el modelo SOM ha sido correctamente entrenado para distinguir cianobacterias incluidas en el dataset respecto de otros géneros. En Fig. 2 se muestra la convergencia del modelo y la detección correspondiente.

En cuanto a las pruebas sobre clasificación de géneros, de acuerdo a los experimentos realizados, los mejores resultados en etapa de prueba se alcanzaron con los siguientes valores de parámetros/hiper-parámetros: epochs → 100, redimensión de imagen → (205,250), batch_size → 16. El mejor valor de accuracy (o precisión) fue 84% alcanzado por ResNet-18 y MobileNet-v2, y 76% para GoogleNet; si bien lo obtenido por GoogleNet es menor, no deja de ser una alternativa ya que utiliza menos

recursos. Sólo a modo de ejemplo, en Fig. 3 se presenta la convergencia del modelo GoogleNet durante la fase de entrenamiento.

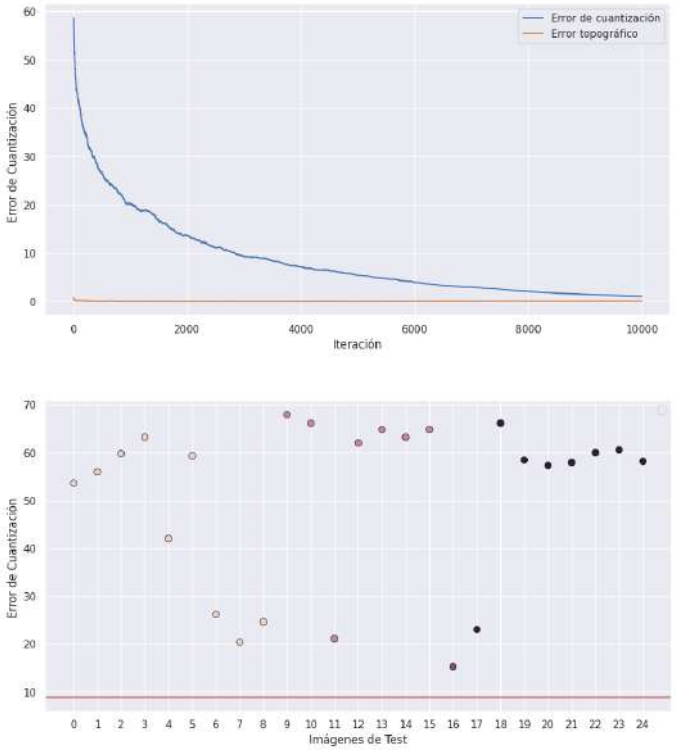

Fig. 2. a) arriba, se muestra la convergencia del modelo SOM estabilizándose luego de 6000 iteraciones aproximadamente. **b) abajo,** se muestra la clasificación obtenida sobre las 25 imágenes externas usadas para test, donde cada punto representa el error de cuantización de la imagen externa i-ésima; se observa que en todos los casos, los errores asociados están por encima del umbral (línea horizontal de color rojo).

4 Conclusiones y Trabajo Futuro

Este es el primer trabajo de ML sobre un dataset local de imágenes por microscopía de cianobacterias en Argentina que tengamos conocimiento. El uso de la metodología propuesta arrojó resultados alentadores a pesar del bajo número de imágenes que contiene el dataset, de resolución heterogénea y de morfología similar de algunos géneros. Los resultados obtenidos en fase de detección y clasificación indican una elección correcta de técnicas basadas en procesamiento de imaǵenes, reducción de dimensión y modelos (SOM y CNN pre-entrenadas) e hiper-parámetros asociados. El próximo paso será analizar nuevas muestras de diversos cuerpos de agua que presentan los géneros seleccionados para obtener imágenes con una metodología estandarizada de manera de incrementar el tamaño del dataset. Además del

reconocimiento taxonómico, se espera incorporar el recuento automático y estimar el biovolumen de los organismos. Esta información es necesaria para establecer el nivel de riesgo para diversos usos de los cuerpos de agua por parte de la sociedad, y suele ser difícil de obtener, debido a que se necesita personal especializado, es un trabajo tedioso y lento, lo que resulta crítico en una situación de emergencia sanitaria generada por floraciones de cianobacterias.

Agradecimientos. Este estudio fue financiado por la Fundación Ciencias Exactas y Naturales (FUNDACEN) bajo el proyecto +4i "*Uso del aprendizaje automático para la identificación taxonómica de muestras microscópicas de interés económico y ambiental*". Agradecemos a los integrantes del laboratorio de Limnología (IEGEBA, UBA-CONICET) por facilitar las imágenes que forman parte del Dataset. Los autores no tienen intereses en competencia que declarar que sean relevantes para el contenido de este artículo.

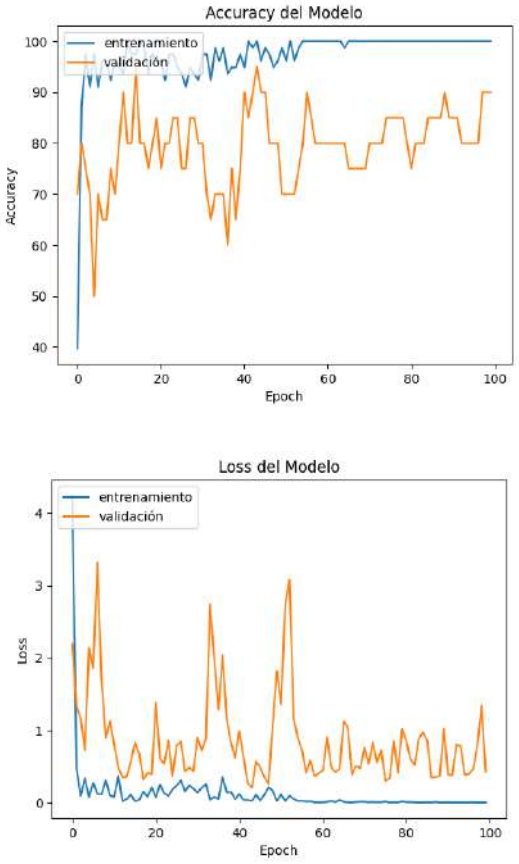

Fig. 3. a) arriba, se muestra la convergencia del accuracy (precisión) del modelo GoogleNet en fase de entrenamiento (incluida la validación). **b) abajo,** el valor de loss (función de pérdida) cercanos a 0.

Referencias

1. Aguilera, A., Haakonsson, S., Martin, M.V., Salerno, G.L., Echenique, R.O.: Bloom-forming cyanobacteria and cyanotoxins in Argentina: a growing health and environmental concern. Limnologica 69, 103–114 (2018)
2. Aguilera, A., Almanza, V., Haakonsson, S., Palacio, H., Benítez Rodas, G., Capelo-Neto, J., Urrutia, R. Aubriot, L., Bonilla, S.: Cyanobacterial bloom monitoring and assessment in Latin America. Harmful Algae **125**, 102429 (2023)
3. Baek, S., Pyo, J., Pachepsky, Y., Park, Y., Ligaray, M., Ahn, C., Kim, Y., Chun, J., Cho, K.: Identification and enumeration of cyanobacteria species using a deep neural network. Ecological Indicators **115**, 106395 (2020)
4. Chorus, I., Welker, M.: Toxic Cyanobacteria in Water, 2nd edition. CRC Press, Boca Raton (FL), on behalf of the World Health Organization, Geneva, CH. (2021)
5. Gandola, E., Antonioli, M., Traficante, A., Franceschini, S., Scardi, M., Congestri, R.: ACQUA: Automated cyanobacterial quantification algorithm for toxic filamentous genera using spline curves, pattern recognition and machine learning. Journal of Microbiological Methods **124**, 48-56 (2016)
6. Guiry, M.D. & Guiry, G.M. AlgaeBase. World-wide electronic publication, National University of Galway. https://www.algaebase.org. Último acceso el 16/04/2024
7. Jolliffe, I., Cadima, J.: Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **374** (2065), 20150202 (2016)
8. Kohonen, T.: Self-Organizing Maps. In: Huang, T., Kohonen, T., Schroeder, M. (eds.) Springer Series in Information Sciences, vol. 30, Springer, Heidelberg (1995)
9. Mosleh, M., Manssor, H., Malek, S., Milow, P., Salleh, A.: A preliminary study on automated freshwater algae recognition and classification system. BMC Bioinformatics **13** (17), S25 (2012)
10. O'Farrell, I., Motta, C., Forastier, M., Polla W., Otaño, S., Meichtry, N., Devercelli, M., Lombardo, R.: Ecological meta-analysis of bloom-forming planktonic Cyanobacteria in Argentina, Harmful Algae **83**, 1–13 (2019)
11. Schulze, K., Tillich, U., Dandekar, T., Frohme, M.: Planktovision - an automated analysis system for the identification of phytoplankton. BMC Bioinformatics **14**, 115 (2013)
12. K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016, pp. 770–778.
13. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “Mobilenetv2: Inverted residuals and linear bottlenecks,” in Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2018, pp.4510–4520.
14. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2015.
15. H. Utermöhl, Zur vervollkommnung der quantitativen phytoplankton-methodik: Mit 1 Tabelle und 15 abbildungen im Text und auf 1 Tafel. Internationale Vereinigung für theoretische und angewandte Limnologie: Mitteilungen **9** (1), 1-38 (1958).