MethodsX 14 (2025) 103198

Contents lists available at ScienceDirect

MethodsX

journal homepage: www.elsevier.com/locate/methodsx

Towards combining self-organizing maps (SOM) and convolutional neural network (CNN) for improving model accuracy: Application to malaria vectors phenotypic resistance ☆

Komi Mensah Agboka [a,*], Elfatih M. Abdel-Rahman [a,b], Daisy Salifu [a], Brian Kanji [a], Frank T. Ndjomatchoua [c], Ritter A.Y. Guimapi [d], Sunday Ekesi [a], Landmann Tobias [a]

[a] International Centre of Insect Physiology and Ecology (ICIPE), P.O. Box 30772 00100, Kenya
[b] School of Agricultural, Earth, and Environmental Sciences, University of KwaZulu-Natal, Pietermaritzburg 3209, South Africa
[c] Department of Plant Sciences, School of the Biological Sciences, University of Cambridge, Cambridge CB2 3EA, United Kingdom
[d] Biotechnology and Plant Health Division, Norwegian Institute of Bioeconomy Research (NIBIO), P.O. Box 115, Ås NO-1431, Norway

ARTICLE INFO

Method name:
Combining Self-Organizing Maps (SOM) and Convolutional Neural Network (CNN) for improving model accuracy

Keywords:
Unsupervised and supervised learning
Data science
Decision making
Insecticide resistance

ABSTRACT

This study introduces a hybrid approach that combines unsupervised self-organizing maps (SOM) with a supervised convolutional neural network (CNN) to enhance model accuracy in vector-borne disease modeling. We applied this method to predict insecticide resistance (IR) status in key malaria vectors across Africa. Our results show that the combined SOM/CNN approach is more robust than a standalone CNN model, achieving higher overall accuracy and Kappa scores among others. This confirms the potential of the SOM/CNN hybrid as an effective and reliable tool for improving model accuracy in public health applications.

- The hybrid model, combining SOM and CNN, was implemented to predict IR status in malaria vectors, providing enhanced accuracy across various validation metrics.
- Results indicate a notable improvement in robustness and predictive accuracy over traditional CNN models.
- The combined SOM/CNN approach demonstrated higher Kappa scores and overall model accuracy.

Specifications table

Subject area:	Environmental Science
More specific subject area:	*Vector borne disease modelling*
Name of your method:	*Combining Self-Organizing Maps (SOM) and Convolutional Neural Network (CNN) for improving model accuracy*
Name and reference of original method:	Ramirez-Quintana JA, Chacon-Murguia MI. Self-adaptive SOM-CNN neural system for dynamic object detection in normal and complex scenarios. Pattern Recognition. 2015 Apr 1;48(4):1137-49.
Resource availability:	https://github.com/komimensah/Data-to-reproduce-SOM-CNN

☆ **Related research article:** Ramirez-Quintana JA, Chacon-Murguia MI. Self-adaptive SOM-CNN neural system for dynamic object detection in normal and complex scenarios. Pattern Recognition. 2015 Apr 1;48(4):1137-49.

* Corresponding author.
E-mail address: kagboka@icipe.org (K.M. Agboka).

https://doi.org/10.1016/j.mex.2025.103198
Received 12 November 2024; Accepted 29 January 2025
Available online 30 January 2025

Background

Recent advancements in data science have led to the utilization of both unsupervised and supervised methods for data analytics, enabling researchers to extract insights and patterns from complex data sets [1–4]. Indeed, studies have shown that unsupervised clustering, such as self-organizing maps (SOM), has been effective in identifying patterns and trends in large and complex datasets [5,6]. On the other hand, supervised learning techniques, such as convolutional neural network (CNN), have also been utilized to showcase various citizen science applications [7–9]. Hence, the combination of these algorithms is gaining significant attention for enhancing model accuracy and improving overall understanding [10,11].

In the context of public health data analysis, SOM and or CNN have been overlooked or rarely used. Recent applications include those by [12,13] among others. This could be attributed to the lack of full understanding of those algorithms [10]. In combination with CNN, SOMs can be used to cluster and classify mosquito resistance data into distinct groups based on similar resistance profiles, aiding in the identification of different resistance mechanisms and their spatial distribution. This approach has the potential to enhance our understanding of real-life data such as health cases related to vector-borne disease (VBD) and inform the development of targeted control strategies. One of the deadliest VBDs especially in Africa is malaria which is caused by *Plasmodium falciprum*.

The spread of insecticide resistance (IR) among mosquito populations poses a severe threat to current vector control strategies, such as insecticide-treated nets and indoor residual spraying. Climate variability further complicates the spread of resistance by impacting mosquito population dynamics, making accurate modeling of insecticide resistance necessary for developing effective and sustainable control measures [14]. Therefore, there is an urgent need for comprehensive approaches that can accurately understand landscape suitability for the development of IR in mosquito vectors.

Hence, in this study, we propose a methodology for enhancing mosquito resistance models accuracy using hybrid modelling approach. This approach combines the SOM algorithm to cluster and reduce the dimensionality of mosquito resistance data with the CNN for prediction to enhance predictive accuracy. We apply this methodology on *Anopheles gambiae* and *Anopheles funestus* IR data in Africa, using defined environmental factors such as land use/land cover, climatic variables, and human population density.

Method details

Data description and preprocessing

The primary data consists of IR phenotypic data obtained through bioassay from the world health organization (WHO) standard test on malaria vector complex and subgroups across Africa with georeferences retrieved from IR MAPPER. The data set comprised a total of 13,619 cases spanning between (1955 and 2018). In addition to the IR data, we selected and extracted the corresponding major environmental variables such as precipitation, temperature [15,16], humidity [17], human presence [18], and land cover [19], which indirectly and directly contribute to mosquito population dynamics. It's important to note that the selection of these variables does not imply they are the only factors influencing insecticide resistance. They were chosen based on their well-documented effects and to ensure computational feasibility. An exhaustive data-driven approach for modeling IR, as described by Hancock et al. [31], offers insights into a broader range of variables that could be considered for comprehensive IR modeling.

Data sources include, 'Envidat' (https://www.envidat.ch/#/metadata/chelsa_cmip5_ts) for precipitation, temperature [20], relative humidity data (https://www.envidat.ch/#/metadata/bioclim_plus); 'Google Earth Engine' for population count (GPWv411: Population Count (Gridded Population of the World Version 4.11) | Earth Engine Data Catalog | Google Developers) and land use and land cover data (https://developers.google.com/earthengine/datasets/catalog/MODIS_061_MCD12Q1). Python programming language [21] was used for data extraction and pre-processing, including conversion, projection, and transformation.

According to the WHO [22], a mosquito population is considered susceptible if the percentage of mortality is 98 % or greater, while the population is considered resistant if the percentage of mortality is less than 90 %. Populations with mortality rates between 90 % and 98 % are considered to have possible resistance that needs further investigation. Therefore, we created the resistance status column using if-then logic in Matlab [23].

We further subset the data to match the time span of the least distributed environmental variable (i.e., land cover 2001–2021), resulting in a total of 5295 observations available for analysis. Before developing the model, we removed dimensionality in the data using the principal components analysis (PCA). Fig. 1 illustrates the correlation circle plot of predictor variables, including LULC (Land Use and Land Cover), rainfall, temperature, population density, and humidity.

The dataset was split into training and testing subsets using stratified sampling to ensure balanced representation across resistance categories. This method was chosen for its efficiency in handling imbalanced data, maintaining proportional representation of all classes in both subsets [24]. While algorithms like Kennard-Stone are effective in regression tasks for evenly distributing data in feature space, stratified sampling is better suited for classification tasks [25]. It is computationally simpler and ensures the model performs well on minority classes, reducing bias and enhancing generalizability [26]. This choice strikes a balance between computational efficiency and accuracy, making it ideal for this study. The SOM algorithm was first applied to the training dataset, creating topological maps that efficiently captured the underlying structure and relationships within the mosquito resistance data. Next, the CNN model was trained on the clustered data produced by the SOM, to classify mosquito populations based on their resistance categories using environmental variables. Upon completion of the training process, we applied our method to the testing dataset to assess its predictive capabilities. The entire method was implemented using Matlab [23]. The workflow of our methodology is shown in Fig. 2.

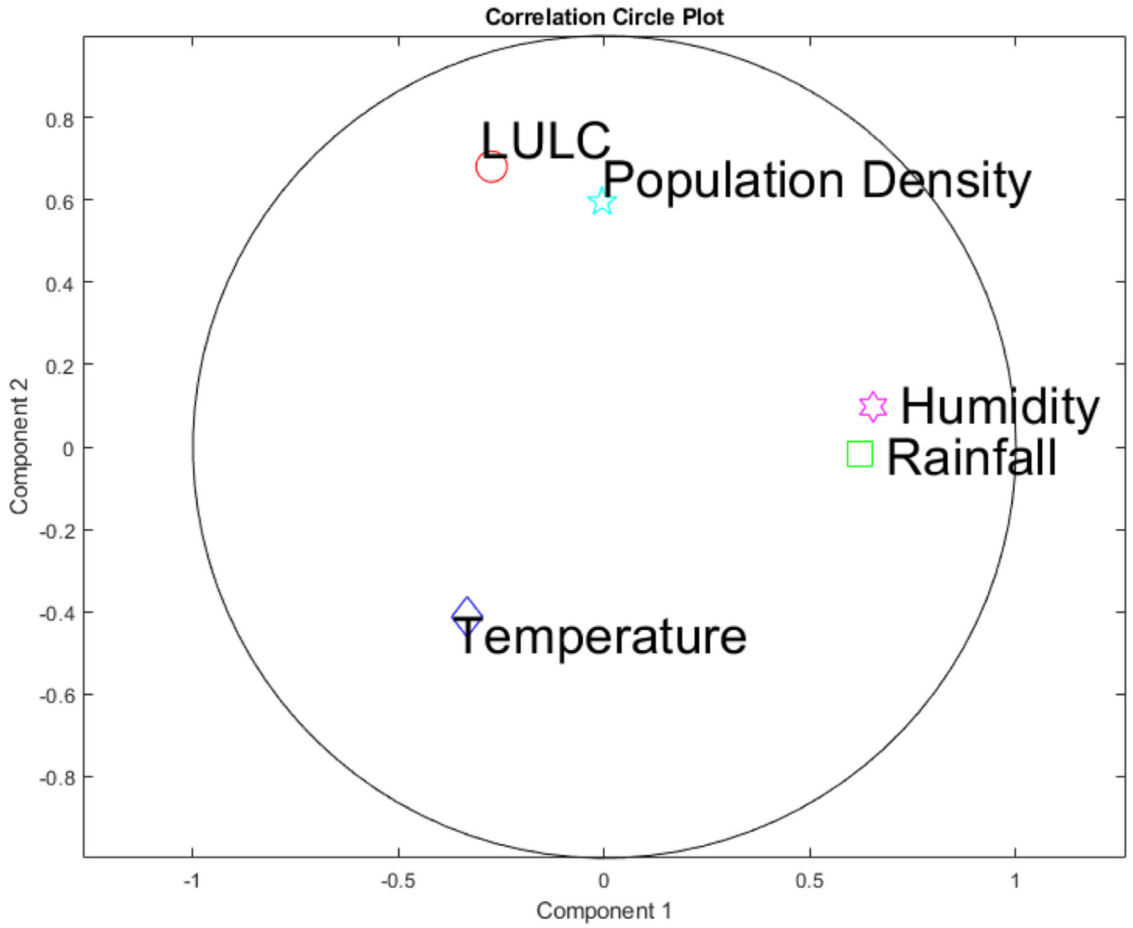

Fig. 1. Correlation circle plot of predictor variables. Each variable is represented with a distinct color and marker shape, as shown in the legend. The distance between two variables in the plot indicates their degree of correlation, with variables closer to each other having a stronger correlation. The position of the variables relative to the circle's circumference represents the contribution of the variables to the principal components.

Self-organizing map (SOM)

Self-organizing map is an unsupervised neural network algorithm that is used for clustering and visualization purposes. It projects high-dimensional data onto a low-dimensional grid, preserving the topological structure and similarity relationships among the data points [27]. In this study, SOM was designed using a 10 × 10 grid with 20 epochs of training. The euclidean distance was used to measure the similarity between neurons in the grid. The best matching unit (BMU) was found for each data point, and the transformed data was obtained by setting the BMU to 1 and all other neurons to 0. The transformed data were then split into training and testing sets using a holdout method with a 0.3 test fraction. The 3-dimensional visualization of the U-Matrix and resistance class distribution on the SOM for the malaria data is presented in Fig. 3.

Convolutional neural network (CNN)

Convolutional neural network (CNN) is a type of neural network that automatically and adaptively learns spatial hierarchies of features from input images, with each layer of the network learning increasingly complex features at a higher level of abstraction [28,29]. The CNN model was trained using the MATLAB Deep Learning Toolbox. The model architecture consists of 10 layers, including an input layer, convolutional layers, batch normalization layers, ReLU activation layers, max-pooling layers, and fully connected layers (Fig. 4). The input to the model is a 1 × 40 × 1 image, where 40 is the product of the dimensions of the SOM, which was used for feature extraction.

The training data consisting of the least collinear variables (rainfall, land use/land cover data, population density, temperature) were normalized using feature scaling using the Min-Max normalization and then passed through the SOM to obtain the transformed data. The transformed data were then split into training and testing sets using a 70/30 ratio. The training data was further pre-

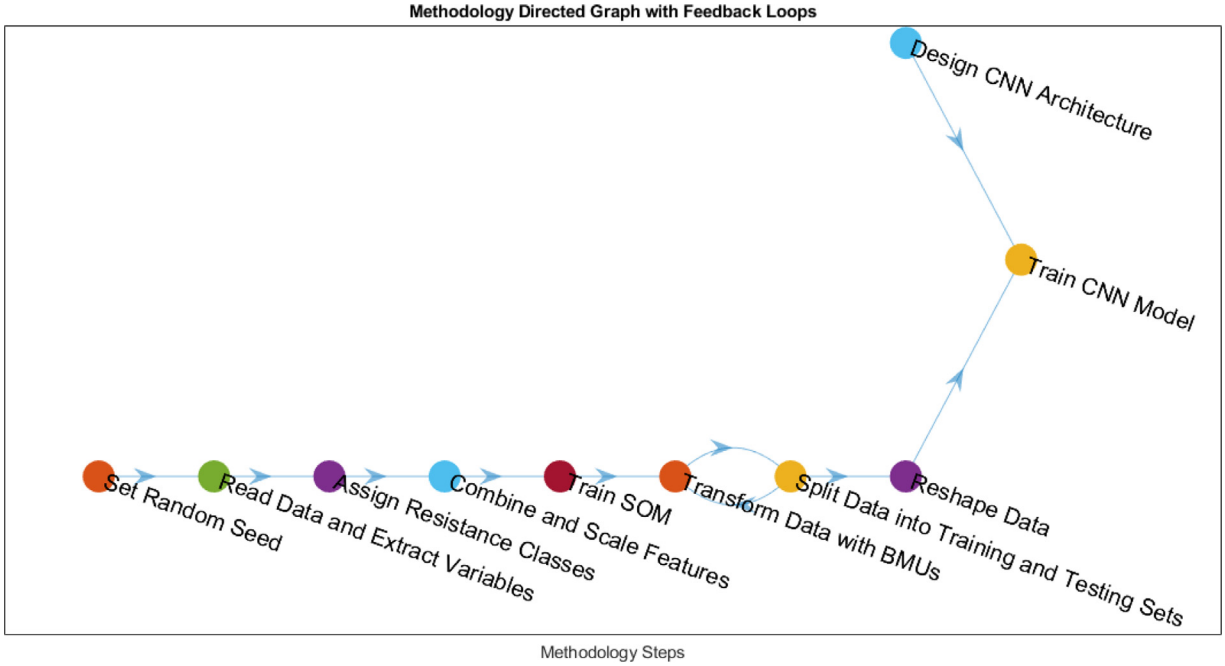

Fig. 2. A diagrammatic workflow illustrating the proposed methodology, which employs a self-organizing map (SOM) combined with a convolutional neural network (CNN) to analyze mosquito resistance data.

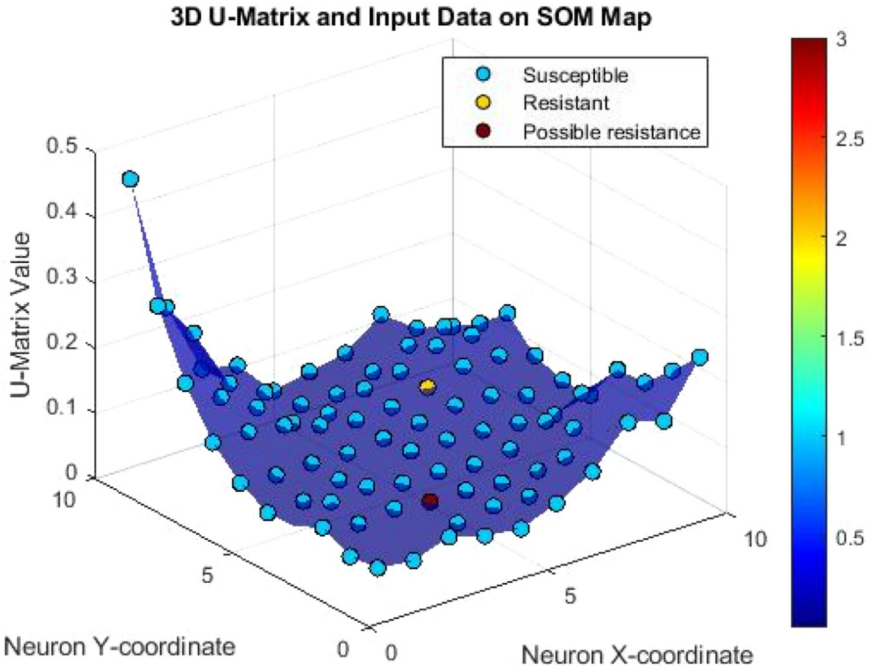

Fig. 3. The 3-dimensional visualization of the U-Matrix and resistance class distribution on the self-organizing map (SOM) for the malaria data. The U-Matrix values represent the average distance between neurons and their neighbors, with higher values indicating larger dissimilarities. The input data points are plotted on the SOM and are colored based on their resistance class: susceptible (class 1), resistant (class 2), and possible resistance (class 3). The distribution of resistance classes on the SOM allows for the identification of patterns and relationships between the input features and resistance levels.

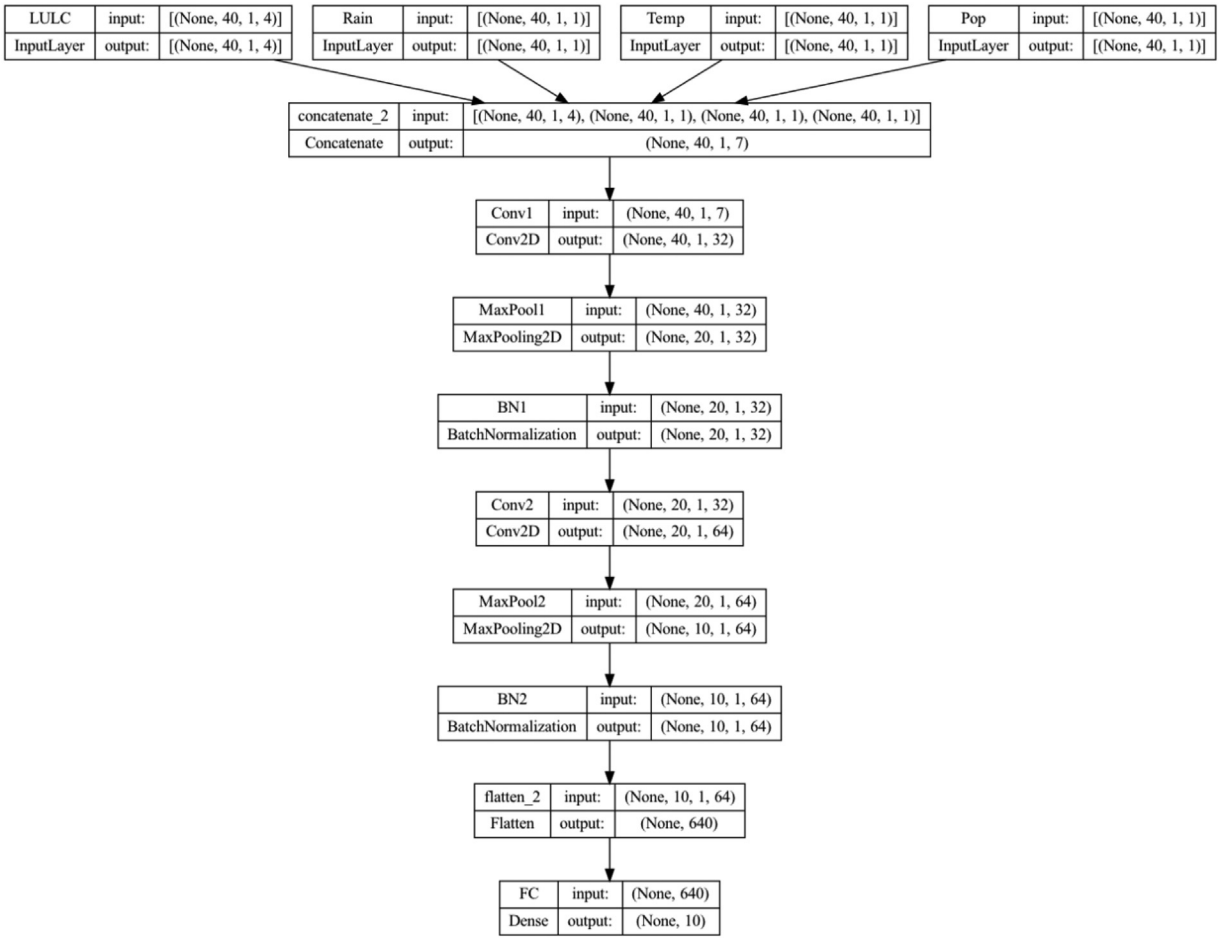

Fig. 4. The architecture of the developed convolutional neural network (CNN). The architecture is organized into a directed graph with nodes representing different layers and edges denoting connections between the layers. The layered layout provides a clear visual representation of the network's structure. The input predictor variables include LULC (land use and land cover), rain (rainfall), temperature (temperature), pop (population density), and output IR (insecticide resistance) class.

processed by reshaping it to a $1 \times 40 \times 1$xN array, where N is the number of training samples. The testing data was also pre-processed by reshaping it to a $1 \times 40 \times 1$xM array, where M is the number of testing samples.

The model was trained using the 'adam' optimizer with a learning rate of 0.001, a maximum of 40,000 epochs, and a mini-batch size of 128. The training progress was monitored using the 'training-progress' option, which displays the training progress in real time. The validation data was also specified to monitor the model's performance on unseen data during training. After training the model, the testing data was used to evaluate the model's performance. The predicted labels were compared with the true labels to calculate the classification accuracy and Cohen's Kappa coefficient.

Evaluation of the performance and robustness of our approach

To demonstrate the robustness of our proposed SOM/CNN methodology, we compared its performance with the standalone CNN model. Our objective was to investigate whether the combination of SOM/CNN could enhance the classification accuracy of the CNN model when applied to mosquito resistance data. Additionally, we aimed to ensure a fair comparison by using the same hyperparameters for both models, preventing one model from overfitting. To compare the performance of the SOM/CNN model with a standalone CNN model, the CNN model was also trained without the SOM feature extraction step. The standalone CNN model had the same architecture as the SOM/CNN model but with an input layer of size $1 \times 160 \times 1$ because the was no prior pre-processing using SOM. The training data for the standalone CNN model was pre-processed by reshaping it to a $1 \times 160 \times 1$xN array, where N is the number of training samples. The testing data was also pre-processed by reshaping it to a $1 \times 160 \times 1$xM array, where M is the number of testing samples. The performance of the SOM/CNN model and the standalone CNN model was compared using bar plots for overall accuracy and Cohen's Kappa coefficient to ensure a fair and consistent comparison between the models.

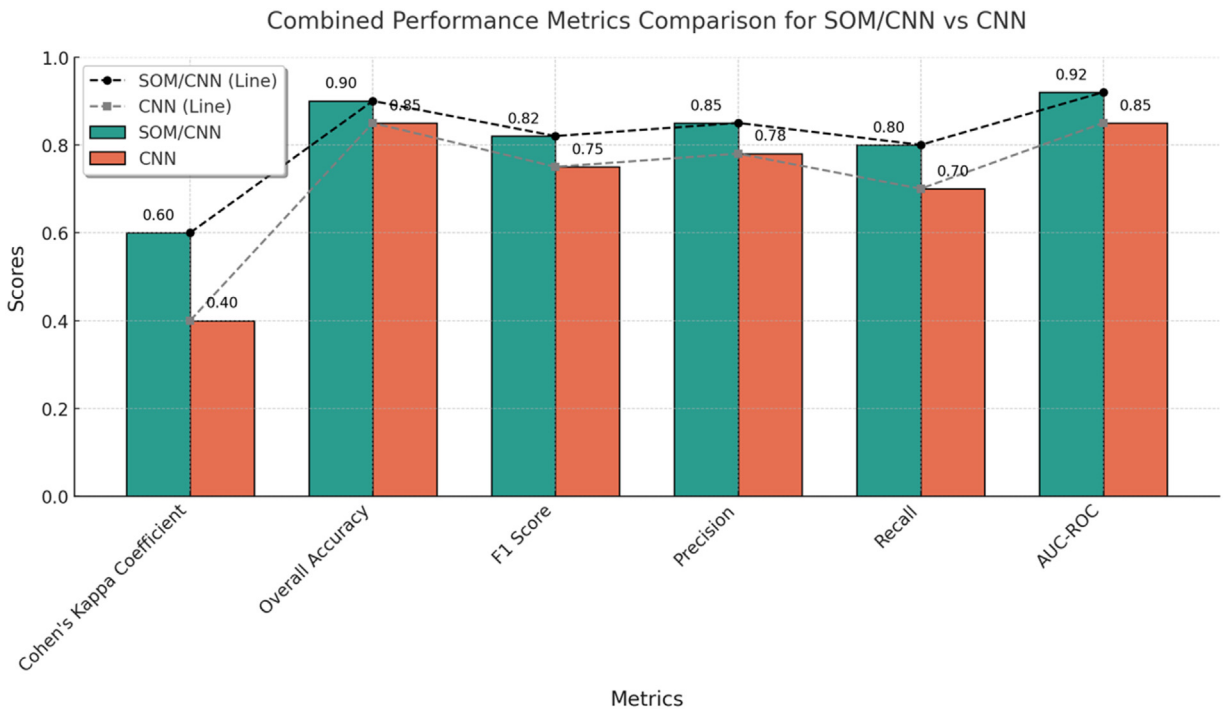

Fig. 5. Comparison between self-organizing maps (SOM) combined with a convolutional neural network (CNN) and a standalone convolutional neural network (CNN).

Method validation

After comparing the performance of our SOM/CNN methodology with the standalone CNN model, the results suggest that our approach demonstrated superior robustness and accuracy Fig. 5. By using the same hyperparameters for both models, we ensured that the comparison was fair and not biased towards one model due to overfitting.

The enhanced performance of the SOM/CNN methodology can be attributed to the integration of SOM for unsupervised clustering and dimensionality reduction. This integration enables the CNN model to focus on the most relevant features and relationships within the data, effectively capturing the complex patterns and interactions underlying mosquito resistance and its associated environmental factors. This comparative analysis with the standalone CNN model highlights the ability of our SOM/CNN methodology to improve classification accuracy when combined with CNN. Moreover, the method's performance highlights the benefits of combining unsupervised clustering and supervised classification techniques to enhance predictive accuracy [1–4]. This finding suggests that our approach could be an effective tool for analyzing mosquito resistance data and informing targeted interventions in malaria control efforts.

Overall the combination of SOM with CNN in vector-borne disease modeling brings several notable advantages. SOM excel at unsupervised clustering and dimensionality reduction, making them particularly effective for managing the complex and high-dimensional datasets often encountered in this field [30]. By grouping similar data points and minimizing noise, SOM allow CNN to focus on the most relevant features and patterns within the data, significantly improving classification and prediction accuracy. This hybrid approach capitalizes on the strengths of SOM for exploratory analysis and CNN for supervised learning, creating a more robust and interpretable model [11]. Additionally, this method will enhances our understanding of the intricate relationships between environmental, biological, and resistance factors, offering critical insights to inform targeted disease control strategies.

Limitations

Nevertheless, a potential limitation of converting tabular data into images for use in CNNs is that it may introduce biases. The transformation process might distort relationships between features, especially when spatially mapping high-dimensional data into a two-dimensional space. This can affect the model's performance and interpretation. To mitigate this, exploring tabular-specific deep learning algorithms, which directly operate on structured data, could provide a more accurate representation and may enhance model effectiveness without the need for image. Despite its limitations, our study represents an important step toward developing more accurate models for combating mosquito-borne diseases. It lays the groundwork for creating improved models that can better inform vector control strategies. Future research could apply our approach to model malaria insecticide resistance (IR) more effectively, incorporating the comprehensive environmental variables documented by Hancock et al. [31] conversion.

Ethics statements

In this manuscript no, human participants or animals their data or biological material, are not involved

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Komi Mensah Agboka: Conceptualization, Methodology, Software, Visualization, Investigation, Data curation, Writing – original draft, Writing – review & editing. **Elfatih M. Abdel-Rahman:** Supervision. **Daisy Salifu:** Visualization. **Brian Kanji:** Data curation. **Frank T. Ndjomatchoua:** Validation. **Ritter A.Y. Guimapi:** Validation. **Sunday Ekesi:** Supervision. **Landmann Tobias:** Supervision.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by: the Swedish International Development Cooperation Agency (Sida); the Swiss Agency for Development and Cooperation (SDC); the Australian Centre for International Agricultural Research (ACIAR); the Government of Norway; the German Federal Ministry for Economic Cooperation and Development (BMZ); and the Government of the Republic of Kenya. The views expressed herein do not necessarily reflect the official opinion of the donors.

References

[1] O.E.L. Aissaoui, Y.E.L.A. El Madani, L. Oughdir, Y.E.L Allioui, Combining supervised and unsupervised machine learning algorithms to predict the learners' learning styles, Procedia Comput. Sci. 148 (2019) 87–96, doi:10.1016/j.procs.2019.01.012.
[2] S. Gupta, B. Parekh, A. Jivani, A hybrid model of clustering and classification to enhance the performance of a classifier BT, in: A.K. Luhach, D.S. Jat, K.B.G. Hawari, X.Z. Gao, P. Lingras (Eds.), Advanced Informatics For Computing Research, Springer, Singapore, 2019, pp. 383–396.
[3] F. Maturo, R. Verde, Combining unsupervised and supervised learning techniques for enhancing the performance of functional data classifiers, Comput. Stat. (2022), doi:10.1007/s00180-022-01259-8.
[4] Y. Zhao, Z. Shao, W. Zhao, J. Han, Q. Zheng, R. Jing, Combining unsupervised and supervised classification for customer value discovery in the telecom industry: a deep learning approach, Computing (2023), doi:10.1007/s00607-023-01150-4.
[5] T. Kohonen, Learning vector quantization BT, in: T. Kohonen (Ed.), Self-Organizing Maps, Springer, Berlin Heidelberg, 2001, pp. 245–261, doi:10.1007/978-3-642-56927-2_6.
[6] J. Vesanto, J. Himberg, E. Alhoniemi, J. Parhankangas, Self-organizing map in MATLAB: the SOM Toolbox, in: Proceedings of the MATLAB DSP Conference, 99, 1999, pp. 16–17.
[7] A. Ajit, K. Acharya, A. Samanta, A review of convolutional neural networks, in: Proceedings of the 2020 International Conference on Emerging Trends in Information Technology and Engineering (Ic-ETITE), 2020, pp. 1–5, doi:10.1109/ic-ETITE47903.2020.049.
[8] L. Alzubaidi, J. Zhang, A.J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M.A. Fadhel, M. Al-Amidie, L. Farhan, Review of deep learning: concepts, CNN architectures, challenges, applications, future directions, J. Big Data 8 (1) (2021) 53, doi:10.1186/s40537-021-00444-8.
[9] A. Dhillon, G.K. Verma, Convolutional neural network: a review of models, methodologies and applications to object detection, Prog. Artif. Intell. 9 (2) (2020) 85–112, doi:10.1007/s13748-019-00203-0.
[10] R. Kamimura, Partially black-boxed collective interpretation and its application to SOM-based convolutional neural networks, Neurocomputing 450 (2021) 336–353, doi:10.1016/j.neucom.2021.04.019.
[11] J.A. Ramirez-Quintana, M.I. Chacon-Murguia, Self-adaptive SOM-CNN neural system for dynamic object detection in normal and complex scenarios, Pattern Recognit. 48 (4) (2015) 1137–1149.
[12] Z. Yu, H. Gong, M. Li, D. Tang, Hollow prussian blue nanozyme-richened liposome for artificial neural network-assisted multimodal colorimetric-photothermal immunoassay on smartphone, Biosens. Bioelectron. 218 (2022) 114751.
[13] Z. Yu, D. Tang, Artificial neural network-assisted wearable flexible sweat patch for drug management in Parkinson's patients based on vacancy-engineered processing of g-C3N4, Anal. Chem. 94 (51) (2022) 18000–18008.
[14] K.M. Agboka, M. Wamalwa, J.M. Mutunga, H.E.Z. Tonnang, A mathematical model for mapping the insecticide resistance trend in the Anopheles gambiae mosquito population under climate variability in Africa, Sci. Rep. 14 (1) (2024) 9850, doi:10.1038/s41598-024-60555-z.
[15] C. Christiansen-Jucht, P.E. Parham, A. Saddler, J.C. Koella, M.G. Basáñez, Temperature during larval development and adult maintenance influences the survival of Anopheles gambiae s.s, Parasites Vectors 7 (2014) 489, doi:10.1186/s13071-014-0489-3.
[16] P.E. Parham, & E. Michael, (2010). Modelling climate change and malaria transmission. Modelling Parasite Transmission and Control, 184–199.
[17] L.F. Chaves, C.J.M. Koenraadt, Climate change and highland malaria: fresh air for a hot debate, Q. Rev. Biol. 85 (1) (2010) 27–55.
[18] J.E. Gimnig, M. Ombok, L. Kamau, W.A. Hawley, Characteristics of larval anopheline (Diptera: culicidae) habitats in Western Kenya, J. Med. Entomol. 38 (2) (2001) 282–288.
[19] K.P. Paaijmans, M.B. Thomas, The influence of mosquito resting behaviour and associated microclimate for malaria risk, Malar. J. 10 (1) (2011) 1–7.
[20] D.N. Karger, & N.E. Zimmermann, (2019). Climatologies at high resolution for the earth land surface areas CHELSA V1. 2: technical specification. Swiss Federal Research Institute WSL, Switzerland.
[21] Guido Van Rossum, Python Tutorial: Centrum voor wiskunde en informatica Amsterdam, The Netherlands, 1995.
[22] W.H. Organization, (2016). Test procedures for insecticide resistance monitoring in malaria vector mosquitoes.
[23] Version, M.A.T. L.A.B. "9.10. 0.1613233 (R2021a); The Mathworks." Inc.: Natick, MA, USA (2021).
[24] F. Vilariño, P. Spyridonos, J. Vitrià, P. Radeva, Experiments with SVM and stratified sampling with an imbalanced problem: detection of intestinal contractions, in: Proceedings of the International Conference on Pattern Recognition and Image Analysis, 2005, pp. 783–791.
[25] N. Baughan, H.M. Whitney, K. Drukker, B. Sahiner, T. Hu, G.H. Kim, M. McNitt-Gray, K.J. Myers, M.L. Giger, Sequestration of imaging studies in MIDRC: stratified sampling to balance demographic characteristics of patients in a multi-institutional data commons, J. Med. Imaging 10 (6) (2023) 064501.

[26] L. Cao, H. Shen, CSS: handling imbalanced data by improved clustering with stratified sampling, Concurr. Comput. Pract. Exp. 34 (2) (2022) e6071.
[27] T. Kohonen, The self-organizing map, Proc. IEEE 78 (9) (1990) 1464–1480.
[28] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Commun. ACM 60 (6) (2017) 84–90.
[29] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.
[30] I. Kaur, A.K. Sandhu, Y. Kumar, Artificial intelligence techniques for predictive modeling of vector-borne diseases and its pathogens: a systematic review, Arch. Comput. Methods Eng. 29 (6) (2022) 3741–3771.
[31] P.A. Hancock, A. Lynd, A. Wiebe, M. Devine, J. Essandoh, F. Wat'senga, E.Z. Manzambi, F. Agossa, M.J. Donnelly, D. Weetman, C.L. Moyes, Modelling spatiotemporal trends in the frequency of genetic mutations conferring insecticide target-site resistance in African mosquito malaria vector species, BMC Biol. 20 (1) (2022) 46, doi:10.1186/s12915-022-01242-1.