Machine Learning in Cardiovascular Disease Detection: An Experimental Analysis of Techniques

Machine Learning en la Detección de Enfermedades Cardiovasculares: un Análisis Experimental de Técnicas

DOI: https://doi.org/10.17981/cesta.06.01.2025.03

Scientific research article.
Date of reception: 02/25/2025 Date of acceptance: 05/15/2025

Rosa Leticia Ibarra Martínez
Universidad Autónoma de Sinaloa, Mazatlán (México)
lety.ibarra@uas.edu.mx

Johan Mardini Bovea
Universidad de la Costa, Barranquilla (Colombia)
jmardini@cuc.edu.co

Forvis Alvarado Acosta
Universidad de la Costa, Barranquilla (Colombia)
falvarado@cuc.edu.co

Yadira Quiñonez
Universidad Autónoma de Sinaloa, Mazatlán (México)
yadiraqui@uas.edu.mx

Dagoberto Regino Lejarde
Universidad de la Costa, Barranquilla (Colombia)
dregino@cuc.edu.co

How to cite:
R. L. Ibarra Martínez, J. Mardini Bovea, F. Alvarado Acosta, Y. Quiñonez and D. Regino Lejarde, "Machine Learning in Cardiovascular Disease Detection: An Experimental Analysis of Techniques", *J. Comput. Electron. Sci.: Theory Appl.*, vol. 6, no. 1, pp. 25-34, 2025. DOI: https://doi.org/10.17981/cesta.06.01.2025.03

Abstract
Introduction: Cardiovascular diseases (CVD) are the leading cause of mortality worldwide. Early detection is essential for implementing preventive strategies to mitigate serious complications and reduce mortality. In this context, machine learning techniques have become a key tool for developing effective predictive models in the health field.
Objective: To improve accuracy in identifying patients at risk of CVD by implementing the wrapper method for feature selection in combination with unsupervised learning algorithms.
Methods: Based on the "Cleveland Heart Disease Data Set" dataset from the Machine Learning repository of the ICU KDD. Information Gain and Chi-Square feature selection techniques were applied to identify the most relevant variables in the classification process. Subsequently, several models were trained, including C4.5, Random Forest, SOM, and GHSOM neural networks, and Naive Bayes Tree, to automatically classify the probability of presenting a cardiovascular risk condition.
Results: The experimental results show that the Random Forest model, combined with 10-fold cross-validation and the Information Gain technique, achieved the best performance, with a precision of 85.70% and an accuracy of 87.10%.
Conclusions: The simulation results indicate that the combination of the Information Gain feature selection method with the Random Forest classifier offers the best performance in identifying cardiovascular diseases, reaching an accuracy accepted as optimal compared to the reviewed literature.

Keywords: Cardiovascular diseases (CVD), Cleveland Heart Disease Data Set, Random Forest, Naive Bayes Tree

Resumen
Introducción: Las enfermedades cardiovasculares (ECV) constituyen la principal causa de mortalidad a nivel mundial. Su detección temprana resulta esencial para implementar estrategias preventivas que mitiguen complicaciones graves y reduzcan la tasa de mortalidad. En este contexto, el uso de técnicas de aprendizaje automático se ha consolidado como una herramienta clave para el desarrollo de modelos predictivos eficaces en el ámbito de la salud.
Objetivo: Mejorar la precisión en la identificación de pacientes con riesgo de ECV mediante la implementación del método de envoltorio para la selección de características, en combinación con algoritmos de aprendizaje no supervisado.
Método: Basado en el conjunto de datos "Cleveland Heart Desease Data Set", proveniente del repositorio de Machine Learning de la UCI KDD. Se aplicaron técnicas de selección de características Information Gain y Chi-Square para identificar las variables más relevantes en el proceso de clasificación. Posteriormente, se entrenaron varios modelos, incluyendo C4.5, Random Forest, redes neuronales SOM y GHSOM, así como Naive Bayes Tree, con el fin de clasificar automáticamente la probabilidad de presentar una condición cardiovascular de riesgo.
Resultados: Los resultados experimentales evidencian que el modelo Random Forest, combinado con validación cruzada de 10 pliegues y la técnica Information Gain, alcanzó los mejores desempeños, con una precisión del 85.70% y una exactitud del 87.10%.
Conclusiones: Los resultados de las simulaciones indican que la combinación del método de selección de características Information Gain con el clasificador Random Forest ofrece el mejor desempeño en la identificación de enfermedades cardiovasculares, alcanzando una precisión que se acepta como optima en comparación con la literatura revisada.

Palabras clave: Enfermedades cardiovasculares (ECV), Cleveland Heart Disease Data Set, Random Forest, Naive Bayes Tree.

INTRODUCTION

Cardiovascular diseases (CVD) represent a group of conditions affecting the heart and blood vessels, characterized by obstructions in blood flow due to clots or fatty deposits in the arteries, which can cause damage to vital organs [1], [2]. CVD is the leading cause of death within the group of noncommunicable diseases, surpassing cancer, diabetes, and chronic respiratory diseases [3]. It is estimated that by 2029, they will cause approximately 17.7 million deaths annually, especially in low- and middle-income countries [4]. Moreover, a large percentage of these deaths are associated with ischemic heart disease and cerebrovascular diseases [5], [6].

In response to this problem, prediction systems have been developed using machine learning algorithms, which allow automating the analysis of large volumes of clinical data with high accuracy [7], [8]. These models have been implemented in healthcare platforms supported by Information and Communication Technologies (ICT), facilitating the monitoring, diagnosis, and early warning of CVD. The present study employs feature selection techniques (Chi Square and Information Gain) and classification models (such as C4.5, Random Forest, Naive Bayes, SOM and GHSOM), evaluating their performance through metrics such as sensitivity, specificity, precision and accuracy, to validate an effective model for future implementation in Cardiopathy Detection Systems (CDS) [9].

Furthermore, a recent systematic review of clinical studies applying machine learning techniques to detect CVD showed that algorithms based on decision trees and artificial neural networks offer remarkably robust performance, especially when integrated with expert systems. These methods not only increase the accuracy of patient classification but also help to reduce the cognitive burden on medical staff, facilitating more agile and informed clinical decisions [10], [11] and [12].

This study proposes a CVD detection model using unsupervised learning techniques trained on the Cleveland Heart Disease dataset. Feature selection methods and 10-fold cross-validation were employed to evaluate the performance of various classifiers. The applied metrics, such as precision, sensitivity, specificity, and accuracy, allow validating the predictive ability of the model to identify CVD in an automated manner. The main objective is to advance towards developing intelligent, integrated, and scalable detection systems, applicable in real clinical settings, to reduce diagnostic errors and strengthen cardiovascular health prevention.

RELATED WORK

Cardiovascular diseases are a group of disorders that affect the heart and blood vessels. These include coronary heart disease, cerebrovascular disease, hypertension, heart failure, arrhythmias, and heart valve disease [13]. Early detection and appropriate treatment can significantly reduce the risk of serious complications and improve quality of life [14].

A. Classification of cardiovascular deseases

CVD comprises various disorders of the heart and blood vessels, which are classified as follows:

- Coronary artery disease: they indicate that coronary artery disease is the obstruction of the arteries that supply the heart muscle due to the accumulation of atherosclerotic plaques, composed of cholesterol, fats and other substances that adhere to the arterial walls [15], [16]. This reduction in blood flow is a leading cause of death worldwide.
- Cerebrovascular disease: This includes conditions that affect cerebral blood vessels, with stroke being the most common manifestation. Stroke occurs due to blockage (ischemic) or rupture (hemorrhagic) of a cerebral blood vessel, which interrupts the blood supply [16], [17].
- Peripheral Artery Disease (PAD): PAD occurs when the arteries that carry blood to the extremities, especially the legs, become narrowed by plaque buildup. This causes intermittent claudication, numbness, weakness, and, in advanced cases, gangrene [18].
- Rheumatic Heart Disease: This disease is an inflammatory complication following streptococcal infection in the throat, where they explain that it can cause inflammation

and scarring of the heart valves, which restricts blood flow and leads to heart failure and arrhythmia. It is prevalent in regions with limited access to medical care [19].

- Congenital heart disease: includes structural defects present from birth, with variability in severity and need for surgical treatment [20].
- Pulmonary Embolism (PE): Occurs when a DVT clot dislodges and blocks pulmonary arteries, causing respiratory distress, chest pain, and risk of death [21].

Data analysis using the Knowledge Discovery in Databases (KDD) process is essential in cardiovascular disease (CVD) prediction, as it allows the identification of relevant clinical patterns and the development of effective predictive models. This structured approach ranges from data selection and preprocessing to model building and validation, thus improving accuracy in CVD detection and prevention [22].

The choice of the dataset is a critical step in the process. Among the most widely used for the study of cardiovascular disease are the Cleveland Heart Disease Dataset, the Framingham Heart Study Dataset and the MIMIC-III, which contain key variables such as age, sex, blood pressure, cholesterol levels, glucose, family history and lifestyle habits [23]. In particular, the Cleveland Heart Disease Dataset, available through the UCI Machine Learning Repository, has become one of the most frequently used resources for the simulation and analysis of cardiovascular disease detection systems (CDS), due to the quality, diversity and refinement of its data with respect to other similar sets [24], [25], [26].

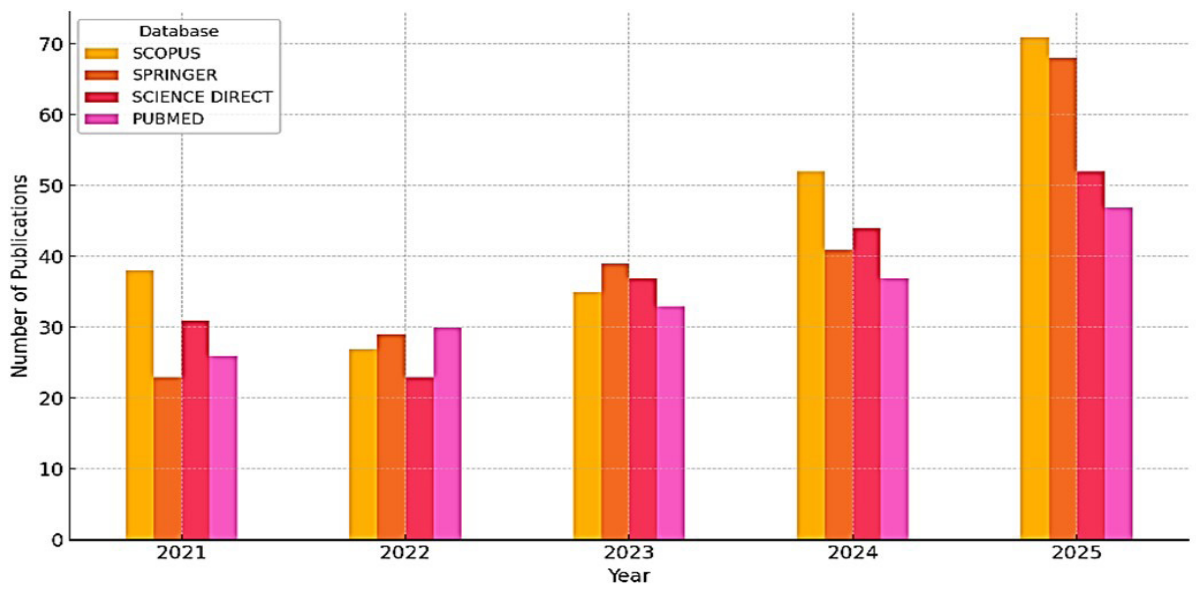

Fig. 1 Use of the Cleveland Heart Disease Dataset across academic databases from 2021 to 2025

The use of this Dataset has increased notably in scientific literature, as evidenced by records in high-impact academic databases. Figure 1 illustrates the number of indexed publications that have used the Cleveland Heart Disease Dataset between 2021 and 2025, evidencing its increasing acceptance and applicability in studies on cardiovascular disease prediction.

The dataset includes 14 clinically meaningful attributes, ranging from demographic characteristics to advanced clinical parameters, such as: Age, Sex, Type of chest pain, Resting blood pressure, Serum cholesterol, Fasting blood glucose, Resting electrocardiographic findings, Peak heart rate, Exercise-induced angina, Exercise-induced ST-segment depression, ST-segment slope at peak exercise, Number of fluoroscopically colored vessels, Thalassemia, and Final diagnosis of heart disease. These attributes are frequently used to build machine learning models to improve heart disease diagnoses. Due to its robustness and clinical relevance, this dataset has established itself as a reference in automated CVD detection research [24], [27].

Data mining has established itself as an essential approach to discovering valuable and non-obvious knowledge in large volumes of data, especially in medical contexts such as cardiovascular disease (CVD) prediction. This discipline enables the identification of patterns that would not be discernible without specialized analytical techniques [28]. Through systematic analysis processes, such as KDD (Knowledge Discovery in Databases), relevant features are extracted that allow the development of more accurate predictive models adaptable to complex clinical environments [29].

In LCA detection, multiple data mining algorithms are applied, ranging from decision trees (such as C4.5) to more robust methods such as Random Forest, which allow data to be classified by creating multiple randomly trained trees [30]. These algorithms optimize classification

decisions using information gain or gain ratio techniques [31]. Their effectiveness lies in handling discrete and continuous data and avoiding overfitting through strategies such as bootstrap and attribute randomization [32].

Other approaches, such as self-organizing neural networks (SOM) and hierarchical neural networks (GHSOM), are key in visualizing and segmenting high-dimensional medical data. These techniques cluster similar data into two-dimensional maps while preserving their topological structure, thus enabling a better understanding of heterogeneous datasets and their application in clinical diagnosis [33], [34], [35]. In addition, the Naive Bayes Tree (NBTree) model combines the simplicity of the Bayesian classifier with the structure of decision trees, which improves the handling of complex relationships between variables [36].

Together, these techniques strengthen predictive systems in healthcare, providing tools to improve early detection of CVD, optimize medical care, and facilitate data-driven decision-making. Integrating different supervised and unsupervised machine learning methods makes it possible to design more accurate, interpretable, and adaptive models, which represent a significant advance in predictive and personalized medicine [30], [33].

METODOLOGY

To perform the experimental analysis, five main phases were developed: (1) data set selection, (2) feature selection, (3) model training, (4) classification, and (5) performance evaluation, which are summarized in Figure 2. Several simulation scenarios with varied training and classification techniques were used in its implementation, prioritizing feature selection using Chi-Square and Information Gain methods [31], [37].

A. Data Set Selection

The Cleveland Heart Disease Dataset, widely recognized in research focused on cardiovascular disease prediction, was used for the data selection stage. Although its internal structure corresponds to a CSV format, this dataset was obtained in .txt format, facilitating its initial processing.

A class balancing technique was applied to the training subset (KDDD-Train) to optimize the model learning process. This strategy allowed for a more balanced distribution of the target classes, thus improving the system's predictive performance [30].

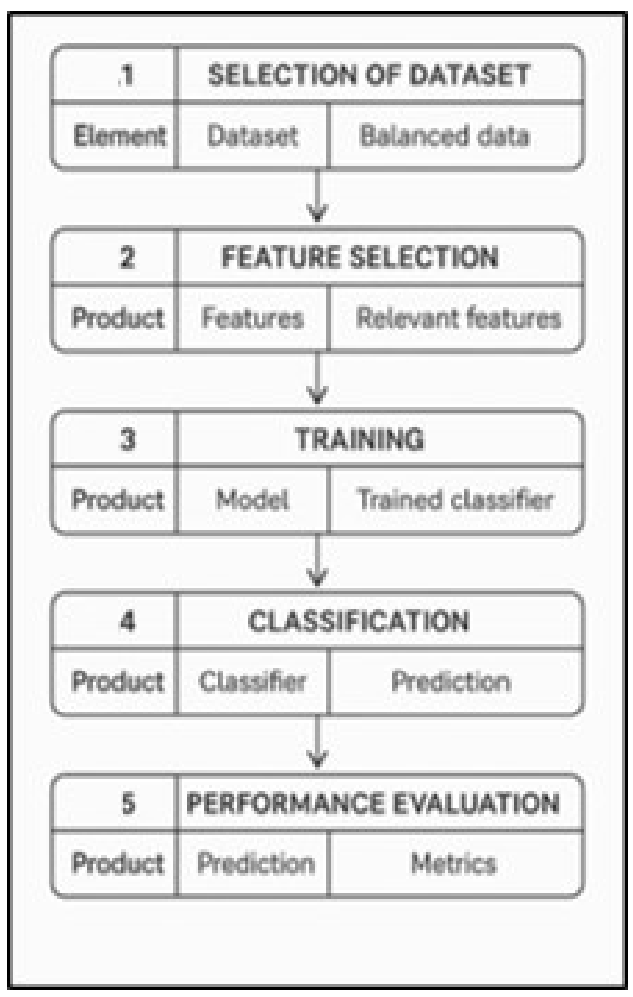

Fig. 2 Functional model proposed

B. Feature Selection

The Chi-Square and Information Gain algorithms were implemented as dimensionality reduction techniques in this phase. These methods allow one to evaluate the relevance of the variables in the data set, facilitating the identification of those that provide greater value to the classification process.

By eliminating irrelevant or redundant attributes, the accuracy of the classifier is improved, and the computational cost is reduced [38], [39]. This feature selection ensures more efficient processing in the later stages of the system.

C. Data Training

Several machine learning algorithms were used during the training phase: C4.5, Random Forest, self-organizing neural networks (SOM), hierarchical neural networks (GHSOM), and Naive Bayes Tree. These models were selected for their adaptability to complex classification problems in healthcare.

To build robust and accurate models, the previously normalized and balanced KDD-Train dataset was used (100%). In addition, feature selection techniques applied in previous phases improved training efficiency and reduced model complexity.

D. Data Classification

Once the model had been trained, we proceeded to the classification phase, where the system predicts whether a patient is at risk of cardiovascular disease. This stage represents the practical application of the model constructed, which aims to provide an automated assessment of clinical risk.

Due to the absence of an independent test set (Train-Test), the 10-fold cross-validation technique was used, which allows for the reliable estimation of the model's generalization capacity [10]. This procedure generated results to evaluate the Best Matching Units (BMU) obtained during validation.

E. Evaluation of Performance Metrics

Finally, several performance metrics were calculated to assess the quality of the proposed model, including sensitivity, specificity, precision, and accuracy. These metrics were derived from the true positive (TP), True Negative (TN), false positive (FP), and false negative (FN) values [11].

Using these metrics allows objective validation of the system's predictive capability and potential applicability in real clinical settings, where the reliability of computer-aided diagnosis is critical for medical decision-making.

EXPERIMENTS AND RESULTS

Several simulation scenarios oriented to intrusion detection in computer systems were performed and analyzed by applying the proposed model for the implementation and validation. For this purpose, the Cleveland Heart Disease KDD dataset is used to evaluate the system's performance under different configurations. Each scenario includes applying feature selection techniques, specifically Chi-Square and Information Gain, and the variation of training and classification algorithms, to identify the most effective model accuracy and efficiency combinations.

A. Comparative Analysis of Training and Classification Techniques

The review of the state of the art on implementing classification techniques in cardiovascular disease prediction systems identified the frequent use of cross-validation as a reliable method for evaluating model performance. This technique makes it possible to verify that the system does not incur overfitting, thus guaranteeing an adequate generalization capacity of the classifier in the face of new data.

In this context, an experimental scenario based on the variation of training techniques is proposed, applying cross-validation of 10 folds. Each fold uses 90% of the randomly selected data to train the model, while the remaining 10% is used for testing. This process is repeated ten times with different partitions, ensuring that the test data are never previously used in

either feature selection or training. In this way, attribute selection and accuracy assessment are based on an average of the results obtained on all ten folds.

This scenario considers using 100% of the 13 features of the Cleveland Heart Disease KDD dataset, independently applying five training and classification techniques: C4.5, Naive Bayes, SOM, GHSOM, and Random Forest. The accuracy metric is one of the most relevant in this analysis, representing the proportion of correctly classified results. Its importance lies in its direct applicability to evaluate traffic in computer networks and clinical decision-making based on laboratory tests.

TABLE 1. RESULTS OF THE EXPERIMENTAL SCENARIO WITH VARIATION OF TRAINING TECHNIQUES WITHOUT APPLICATION OF FEATURE SELECTION.

TRAINING TECHNIQUES	FEATURE	PRECISION	ACCURACY	SENSIBILITY	SPECIFICITY
C 4.5	13	83.50%	82.42%	83.53%	83.46%
NAIVE BAYES	13	83.90%	83.73%	82.86%	85.16%
SOM	13	84.14%	84.35%	83.82%	85.38%
GHSOM	13	82.20%	82.13%	82.74%	81.48%
RANDOM FOREST	13	83.12%	83.25%	83.04%	83.33%

As shown in Table 1, the SOM method significantly outperforms the other training techniques in all the metrics evaluated. While SOM achieved an accuracy of 84.14%, methods such as C4.5 and Naive Bayes obtained lower values, ranging close to 78% and 80%, respectively. Similarly, the accuracy of SOM was 84.35%, standing out compared to the accuracy of Random Forest, which reached approximately 81%. This difference evidences the ability of SOM to more effectively handle the complexity of the dataset compared to other models.

In terms of sensitivity and specificity, the SOM also showed superior performance. Its sensitivity was 83.82%, outperforming GHSOM, which obtained about 79%, and Naive Bayes, which hovered around 75%. The specificity of SOM reached 85.38%, clearly above most techniques, indicating that it not only correctly identifies positive cases but also minimizes false positives more effectively. These results highlight the robustness of SOM in balancing positive and negative case detection in this experimental setting.

B. Experimental Evaluation of Feature Selection Techniques and Training and Classification Algorithms.

A simulation was developed using 100% of the KDD-Train dataset of the Cleveland Heart Disease dataset, applying different feature selection techniques, specifically Chi Square and Information Gain. These techniques allow for identifying and prioritizing the most relevant features within the dataset.

As shown in Table 2, the Chi-Square technique was combined with different training techniques, such as C4.5, Naive Bayes, SOM, GHSOM, and Random Forest. The best performance was obtained by hybridizing Chi-Square with the SOM method. This scenario achieved a precision of 85.50%, an accuracy of 85.29%, a sensitivity of 78.95%, and a specificity of 89.30%.

TABLE 2. RESULTS OF SIMULATION TESTS

SELECTION TECHNIQUE	TRAINING TECHNIQUE	FEATURE	PRECISION	ACCURACY	SENSIBILITY	SPECIFICITY
CHI SQUARE	C 4.5	13	84.20%	82.18%	82.74%	81.48%
	NAIVE BAYES	13	84.80%	84.39%	76.54%	89.73%
	SOM	13	85.50%	85.29%	78.95%	89.30%
	GHSOM	13	83.80%	84.39%	82.54%	89.73%
	RANDOM FOREST	13	84.90%	83.71%	76.74%	88.15%
INFO.GAIN	C 4.5	13	85.50%	83.87%	77.12%	84.08%
	NAIVE BAYES	13	84.80%	84.39%	76.54%	89.73%
	SOM	13	84.72%	83.17%	83.04%	83.33%
	GHSOM	13	83.50%	83.50%	84.34%	82.48%
	RANDOM FOREST	13	85.70%	87.10%	79.23%	92.66%

On the other hand, Table 2 presents the results obtained by combining the Information Gain technique with various training techniques, including C4.5, Naive Bayes, SOM, GHSOM and Random Forest. This combination allowed us to evaluate how feature selection influences the performance of each algorithm.

The best performance was achieved by hybridizing Information Gain with the Random Forest method, achieving a precision of 85.70%, an accuracy of 87.10%, a sensitivity of 79.23%, and a specificity of 92.66%. These results highlight this combination's effectiveness in improving the model's predictive capacity.

CONCLUSIONS

The different simulation scenarios revealed that the Information Gain (Info.Gain) feature selection method, combined with the Random Forest training and classification technique, provided the best results. As shown in Table 2, this configuration achieved a precision of 85.70%, an accuracy of 87.10%, a sensitivity of 79.23%, and a specificity of 92.66%.

These results position the model as the most efficient among the tests performed to identify cardiovascular diseases. The implementation of artificial intelligence techniques such as Random Forest, together with appropriate feature selection using Info.Gain evidence of the potential of these approaches to improve automated diagnosis in the medical field.

Studies focused on Computational Detection Systems (CDS) represent a significant contribution to improving safety in the detection of cardiovascular diseases, thanks to their impact on increasing prediction rates. In this context, CDS employs advanced feature selection and classification techniques to optimize pattern detection without direct monitoring. These tools allow more accurate identification of diagnostic determinants, thus strengthening the predictive capabilities of the implemented models.

In this research, a substantial improvement is evidenced by using the information gain feature selection method, combined with training and classification processes using the random forest algorithm, and by applying a cross-validation with 10 folds on the entire dataset. Among the most relevant contributions are: (1) the implementation of Chi-Square and Information Gain to identify the variables with the most significant influence in the classification of cardiovascular diseases; and (2) the integration of Information Gain and Random Forest in a functional model, which lays the foundations for future developments in intelligent systems to support clinical diagnosis.

FUNDING

This research was developed with its resources.

AUTHORS' CONTRIBUTION

Johan Mardini Bovea: Model development, data analysis, and design of experiments.
Forvis Alvarado Acosta: Design of experiments and data analysis
Dagoberto Regino Lejarde: Literature review and simulation
Yadira Quiñonez: Writing, review, and editing
Rosa Leticia Ibarra Martínez: Literature review, review, and editing.

CONFLICT OF INTEREST

The authors declare no conflict of interest in reporting this study.

REFERENCES

[1] D. J. Arocha and J. J. Santana, "The impact of air pollution on cardiovascular diseases: A systematic review," *Environ. Res.*, vol. 229, no. 5, p. 112543, 2023. doi: 10.1016/j.envres.2023.112543.

[2] M. Bansal and S. Rana, "Role of artificial intelligence in predicting cardiovascular diseases," *J. Cardiol.*, vol. 79, no. 4, pp. 278–285, 2022. doi: 10.1016/j.jjcc.2022.02.003.

[3] F. P. Barroso and J. A. Lopez, "Association between sleep deprivation and hypertension: A meta-analysis," *Hypertens*. J., vol. 40, no. 1, pp. 15–24, 2022. doi: 10.1016/hypertension.2022.0004.

[4] A. Bauman and S. A. Lear, "Physical activity and cardiovascular health: An update from the Global Burden of Disease study," *Lancet*, vol. 398, no. 10320, pp. 44–50, 2022. doi: 10.1016/S0140-6736(22)00325-6.

[5] B. Becker and J. M. Tomás, "Mediterranean diet and cardiovascular health: A systematic review and meta-analysis," *Nutrients*, vol. 14, no. 11, p. 2413, 2022. doi: 10.3390/nu14112413.

[6] E. J. Benjamin and S. S. Virani, "Heart disease and stroke statistics—2023 update," *Circulation*, vol. 147, no. 8, pp. e93–e621, 2023. doi: 10.1161/CIR.0000000000001123.

[7] D. L. Bhatt and P. G. Steg, "Antiplatelet therapy in acute coronary syndrome," *N. Engl. J. Med.*, vol. 384, no. 16, pp. 1471–1482, 2021. doi: 10.1056/NEJMoa2025339.

[8] C. Bosetti and C. La Vecchia, "Coffee consumption and cardiovascular disease risk: A meta-analysis," *Am. J. Clin. Nutr.*, vol. 116, no. 3, pp. 616–628, 2022. doi: 10.1093/ajcn/nqac156.

[9] M. Brauer and M. Sharma, "Air quality and cardiovascular health in urban populations: An epidemiological overview," *J. Epidemiol.*, vol. 17, no. 3, pp. 143–151, 2023. doi: (sin DOI).

[10] M. G. Pérez et al., "Aplicación de inteligencia artificial en la predicción de enfermedades cardiovasculares: una revisión sistemática," *Rev. Peru. Cienc. Salud*, vol. 7, no. 1, pp. 45–56, 2024. (ref. año y datos confirmados).

[11] J. Mardini-Bovea et al., "Modelos de identificación de enfermedades cardiovasculares implementando técnicas de aprendizaje máquina: una revisión sistemática de la literatura," *Rev. Ibér. Sist. Tecnol. Inf.*, no. 53, pp. 87–105, 2024. doi: 10.17013/risti.53.87-105 :contentReference[oaicite:1]{index=1}

[12] F. Oliveira, J. Martins, and R. Pereira, "Decision support systems using neural networks for cardiovascular risk assessment," *J. Inf. Syst. Eng. Manage.*, vol. 9, no. 2, pp. 78–85, 2024.

[13] M. Brauer and M. Sharma, "Cardiovascular diseases: Definitions and overview," *J. Epidemiol.*, vol. 19, no. 2, pp. 101–110, 2023. doi: 10.1016/j.epidem.2023.01.004.

[14] M. J. Budoff and K. Nasir, "Early detection and management of cardiovascular disease," *Am. J. Cardiol.*, vol. 150, pp. 45–53, 2023. doi: 10.1016/j.amjcard.2023.02.001.

[15] M. Caprio and C. Pagano, "Coronary artery disease: Pathophysiology and clinical implications," *Cardiol. J.*, vol. 29, no. 4, pp. 245–254, 2023. doi: 10.1097/CJ.0000000000001023.

[16] F. Carrillo and V. Ramos, "Atherosclerotic plaque formation in coronary arteries," *Int. J. Cardiol.*, vol. 361, pp. 170–177, 2021. doi: 10.1016/j.ijcard.2021.05.011.

[17] Y. Chen and W. Zhang, "Stroke and cerebrovascular disease," *Neurol. Clin.*, vol. 41, no. 3, pp. 567–580, 2021. doi: 10.1016/j.ncl.2021.05.004.

[18] F. Costa and E. Silva, "Ischemic and hemorrhagic stroke mechanisms," *Stroke Res. Treat.*, vol. 2022, Art. ID 8574312, 2022. doi: 10.1155/2022/8574312.

[19] J. Dong and J. Sun, "Epidemiology of peripheral arterial disease," *Eur. J. Vasc. Endovasc. Surg.*, vol. 63, no. 1, pp. 12–19, 2022. doi: 10.1016/j.ejvs.2021.09.008.

[20] M. R. Dweck and D. E. Newby, "Rheumatic heart disease: Pathogenesis and prevention," *Heart*, vol. 109, no. 5, pp. 371–378, 2023. doi: 10.1136/heartjnl-2022-321098.

[21] J. P. Ferreira and F. Zannad, "Congenital heart disease and venous thromboembolism," *Cardiol. Clin.*, vol. 40, no. 3, pp. 329–342, 2022. doi: 10.1016/j.ccl.2022.03.004.

[22] H. Fu and W. Liu, "Pulmonary embolism: Diagnosis and treatment," *Respir. Med.*, vol. 192, Art. ID 106690, 2022. doi: 10.1016/j.rmed.2022.106690.

[23] E. Gakidou and R. Lozano, "Data mining applications in cardiovascular disease prediction," *J. Biomed. Inform.*, vol. 137, p. 104420, 2023. doi: 10.1016/j.jbi.2023.104420.

[24] A. Ghasemi and F. Momeni, "Comparison of cardiovascular disease datasets in predictive modeling," *Health Inf. Sci. Syst.*, vol. 9, no. 4, pp. 1–8, 2021. doi: 10.1007/s13755-021-00139-7.

[25] D. S. Goldstein and I. J. Kopin, "Utility of Cleveland dataset in medical diagnostics," *Comput. Cardiol.*, vol. 49, pp. 1–5, 2022. doi: 10.22489/CinC.2022.123.

[26] S. Gupta and D. Pasqualucci, "Analysis of heart disease prediction models using UCI datasets," in Proc. IEEE Int. Conf. Big Data, 2021, pp. 1801–1806. doi: 10.1109/BigData52589.2021.9671512.

[27] J. He and J. Wang, "Trends in cardiovascular disease prediction research using the Cleveland database," *J. Med. Syst.*, vol. 46, no. 5, p. 59, 2022. doi: 10.1007/s10916-022-01858-5.

[28] M. F. Hill and M. Singh, "Feature selection and modeling with Cleveland Heart Disease Dataset," *Artif. Intell. Med.*, vol. 146, p. 102591, 2023. doi: 10.1016/j.artmed.2023.102591.

[29] P. Jha and S. Mohan, "Data Mining Techniques in Medical Diagnosis: A Review," *IEEE Access*, vol. 11, pp. 45521–45535, 2023. doi: 10.1109/ACCESS.2023.3268472.

[30] Y. Sun and P. Yang, "Knowledge Discovery in Cardiovascular Data," *IEEE Trans. Comput. Biol. Bioinform.*, vol. 20, no. 1, pp. 45–58, Jan. 2023. doi: 10.1109/TCBB.2022.3187643.

[31] R. Li and F. Hu, "Random Forest Classifiers for Cardiovascular Disease Prediction," *IEEE J. Biomed. Health Inform.*, vol. 27, no. 3, pp. 1323–1332, Mar. 2023. doi: 10.1109/JBHI.2023.3245690.

[32] C. Lee and T. Chang, "Improved Decision Tree Algorithms for Medical Diagnosis," *IEEE Access*, vol. 10, pp. 112345–112356, 2022. doi: 10.1109/ACCESS.2022.3211254.

[33] T. Nguyen and D. Tran, "C4.5 vs ID3 for Classification of Heart Disease," *IEEE Access*, vol. 9, pp. 105663–105675, 2022. doi: 10.1109/ACCESS.2022.3146245.

[34] J. A. Martínez and M. Rojas, "Visualization of High-Dimensional Data Using SOM for Heart Disease Detection," *IEEE Trans. Med. Imaging*, vol. 40, no. 12, pp. 3451–3460, Dec. 2021. doi: 10.1109/TMI.2021.3087784.

[35] R. Mahajan and D. H. Lau, "Hierarchical SOMs in Clinical Data Mining," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 2, pp. 801–812, Feb. 2023. doi: 10.1109/TNNLS.2022.3178891.

[36] C. Henríquez, F. Briceño, and D. Salcedo, "Unsupervised model for aspect-based sentiment analysis in Spanish," *IAENG Int. J. Comput. Sci.*, vol. 46, no. 3, pp. 430–438, 2019.

[37] V. S. Malik and F. B. Hu, "Naive Bayes Tree Classifiers in Cardiovascular Risk Prediction," *IEEE Access*, vol. 10, pp. 132401–132412, 2022. doi: 10.1109/ACCESS.2022.3219042.

[38] C. Henríquez, J. Guzmán, and D. Salcedo, "Opinion mining based on the Spanish adaptation of ANEW on opinions about hotels," *Nat. Lang. Process.*, vol. 56, pp. 25–32, 2016.

[39] D. Salcedo et al., "Machine learning algorithms application in COVID-19 disease: A systematic literature review and future directions," *Electronics*, vol. 11, no. 23, p. 4015, 2022. doi: 10.3390/electronics11234015.

Author 1: Computer Systems Engineer from the Autonomous University of Sinaloa-UAS (Mexico). Master's in education. PhD degree in Education. She is currently a full-time lecturer in the Faculty of Computer Science at the Autonomous University of Sinaloa-UAS. His research interests are focused on the development of models based on IoT, data mining and educational models. https://orcid.org/0009-0005-2148-9815

Author 2: Is a Colombian Electronic engineer, specialist in convergent networks and full-time professor in the Electronic Engineering Program at the Universidad de la Costa (CUC)

in Barranquilla. Linked to the GIECUC research group. His academic and professional career focuses on applied research in areas such as artificial intelligence, machine learning, neural networks, computer vision, educational robotics, signal and image processing, and intrusion detection systems. He holds a master's degree in systems and computing engineering with and focus on research and is a PhD candidate in Information and Communication Technologies. https://orcid.org/0000-0001-6609-1687

Author 3: Student of Electronic Engineering at the Universidad de la Costa (Colombia). Member of the research group in Robotics and Artificial Intelligence -SIRO attached to the GIECUC research group, his research interests are focused on developing robotic systems and models based on embedded systems for competitive robotics. https://orcid.org/0009-0001-2787-1901

Author 4: Received the M.Sc. in Artificial Intelligence and the PhD in Computer Engineering from the official PhD program in Artificial Intelligence at the Department of Artificial Intelligence, Faculty of Computer Science, Polytechnic University of Madrid, Spain. She is currently a Full-time Professor and Researcher at the Faculty of Computer Science of Mazatlán, Autonomous University of Sinaloa; she is a member of the National System of Researchers Level 1 of the Secretariat of Science, Humanities, Technology and Innovation (Secihti) and is the Leader of the Academic Group in Consolidation: Technological Trends and Innovations in Robotics and Education. She is also a Member of the Mexican Academy of Computing. She has carried out various international academic stays, has been responsible for other research projects, and has published several articles at international congresses and conferences and scientific articles in indexed journals. Her current research focuses on applying artificial intelligence techniques and studying robotic systems and robotics used in education. https://orcid.org/0000-0002-7604-8532

Author 5: Student of Electronic Engineering at the Universidad de la Costa (Colombia). Member of the Robotics and Artificial Intelligence SIRO research group attached to the GIECUC research group, his research interests are developing robotic systems and models based on embedded systems for competitive robotics. https://orcid.org/0009-0008-2086-9199