Iraqi Journal for Computers and Informatics
University of Information Technology and Communications
Vol. 51, No. 1, 2025, pp. 75-82
DOI: https://doi.org/10.25195/ijci.v51i1.551
Print ISSN: 2313-190X, Online ISSN: 2520-4912

Research Article

Prediction of Hypertension Patients with Machine Learning Algorithm

[1]Eko Priyono
Nusa Mandiri University
And BMKG
Jakarta, Indonesia
14220040@nusamandiri.ac

ARTICLE INFO
Article History
Received: 07/02/2025
Accepted: 02/04/2025
Published:05/06/2025

ABSTRACT
Hypertension, known as the "silent killer," is one of the leading causes of global mortality, with a steadily increasing prevalence. Worldwide, the prevalence of hypertension reaches approximately 30%, with only 50% of cases being diagnosed and a low level of treatment adherence. Hypertension symptoms are often invisible, making early detection crucial to preventing serious complications. This paper aims to develop a hypertension prediction system using the Decision Tree and Random Forest algorithms, which are machine learning techniques used to solve classification and regression problems. These algorithms can predict hypertension risk based on clinical data, such as age, medical history, and lifestyle. The findings of this paper indicate that the Decision Tree and Random Forest algorithms are effective in predicting hypertension risk, achieving accuracies of 99.6% and 99.5%, respectively. This prediction system can provide fast and accurate information, assisting healthcare professionals in designing appropriate intervention strategies while also supporting better medical decision-making.

Keywords: *Decision Tree, Deteksi Dini, Hypertension, Machine Learning, Random Forest ensemble, Decision Tree classifier*

1. INTRODUCTION

Hypertension, or high blood pressure, is among the main causes of death worldwide, often referred to as the "silent killer" due to its asymptomatic but deadly nature [1]. Globally, hypertension has caused approximately 9.4 million deaths, with the highest prevalence found in the 31-44 age group (20%), age groups of 45–54 (35%), and 55–64 (45%). Total hypertension prevalence 30%, only 50% have been diagnosed, and 40% of those diagnosed do not regularly take medication [2]. Additionally, 30% of patients fail to adhere to treatment schedules, reflecting a lack of awareness and proper management among individuals with hypertension [3].

Common symptoms of hypertension include headaches, dizziness, nausea, vomiting, neck pain, fatigue, anxiety, shortness of breath, nosebleeds, and loss of consciousness. Although previously more common in older adults, hypertension is now increasingly observed among younger generations. Risk factors include gender, age, genetics, stress, obesity, alcohol consumption, smoking, high salt intake, absence of exercise, as well as a history of diabetes or kidney issues. The rising prevalence of hypertension highlights the need for early detection to reduce the risk of severe complications [4], [5].

Machine learning algorithms offer a solution for detecting and predicting hypertension risk more effectively and accurately. Algorithms such as Random Forest, Decision Tree, and Logistic Regression can assist healthcare professionals in predicting hypertension risk based on clinical data. These algorithms process information from various risk factors, including age, medical history, and lifestyle, to provide accurate predictions about an individual's hypertension status [6], [7], [8].

Iraqi Journal for Computers and Informatics
University of Information Technology and Communications
Vol. 51, No. 1, 2025, pp. 75-82
DOI: https://doi.org/10.25195/ijci.v51i1.551
Print ISSN: 2313-190X, Online ISSN: 2520-4912

Studies have shown that Random Forest and Decision Tree algorithms are highly effective in predicting diseases, including hypertension. With high accuracy rates of 99.6% for Decision Tree and 99.5% for Random Forest, these algorithms can enhance understanding and provide valuable insights for hypertension risk prediction [9], [10]. The prediction results based on machine learning algorithms can provide fast and accurate information, assisting in the formulation of effective prevention strategies. Thus, the application of machine learning algorithms in hypertension analysis and prediction not only accelerates diagnosis but also contributes to better medical intervention planning.

2. TECHNIQUES METHOD

Figure 1 illustrates the process by which the predictive analysis results for the classification of hypertension disease were obtained.

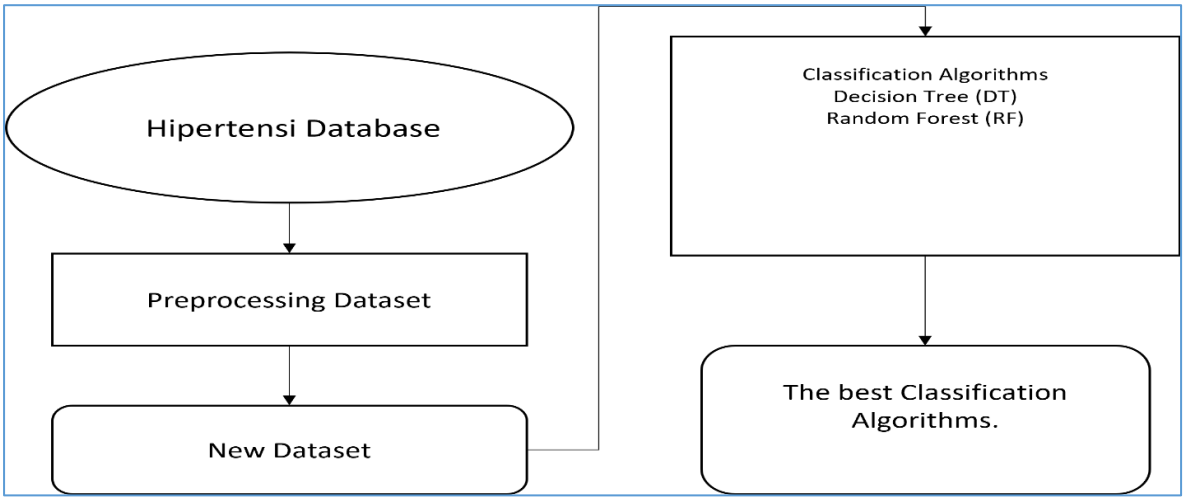

Fig. 1 Proposed work

This work employed a thorough analysis using Python version 3.9.12, enhanced by the integration of key modules such as NumPy, Matplotlib, and Scikit-Learn. Python 3.9.12 leverages NumPy, a powerful library that facilitates the handling of multidimensional arrays and matrices. NumPy not only ensures efficient storage and manipulation of numerical data but also offers a wide range of advanced mathematical operations, making it indispensable for tasks involving complex computations and array processing. Complementing NumPy, Matplotlib serves as a dedicated visualization tool, enabling the creation of graphs and plots that provide clear and insightful representations of analytical outcomes. This enhances the interpretability and communication of data-driven findings.

For the machine learning components of this work, Scikit-Learn (also referred to as sklearn) was utilized. Scikit-Learn is an open-source machine-learning library tailored for Python, offering a comprehensive suite of algorithms and tools for model training, evaluation, and deployment. Its integration with Python 3.9.12 allowed for the seamless implementation of various machine-learning techniques, ensuring accurate and reliable results. By combining the capabilities of NumPy for numerical operations, Matplotlib for data visualization, and Scikit-Learn for machine learning, this work achieved a robust and efficient analytical framework. This integrated approach facilitated in-depth data processing, analysis, and interpretation, underscoring the effectiveness of these tools in advancing the work objectives [16]–[18].

2.1 Dataset

The dataset was obtained from Kaggle ML, a Machine Learning Repository, and is used for work related to hypertension issues. The dataset is in CSV format (https://www.kaggle.com). The work process involved the following steps: data collection, where data was sourced from Kaggle ML, focusing on hypertension-related problems. Class imbalance handling was addressed using oversampling techniques, particularly SMOTE, to overcome class imbalance within the dataset [19]. Attributes in the hypertension dataset include different characteristics or factors that are assessed or observed for every sample in the dataset. Attributes provide information about patient characteristics, hypertension symptoms, or other relevant factors, as shown in Table 1.

Iraqi Journal for Computers and Informatics
University of Information Technology and Communications
Vol. 51, No. 1, 2025, pp. 75-82
DOI: https://doi.org/10.25195/ijci.v51i1.551
Print ISSN: 2313-190X, Online ISSN: 2520-4912

Table I. Attributes and descriptions of Hypertension data

Attribute Name	Information
actividad_total	Total physical activity performed by an individual over a certain period, usually measured by minutes or hours.
circunferencia_de_la_pantorrilla	Calf circumference, usually measured by centimetres, which can indicate muscle mass or nutritional status.
concentracion_hemoglobina	The blood's haemoglobin content, expressed in grams per decilitre (g/dL), used to assess anemia or blood health.
distancia_rodilla_talon	Knee-to-heel distance, which can be used to assess leg length or body growth.
edad	Age of the individual, usually measured by years.
estatura	The person's height usually measured by centimetres or meters.
medida_cintura	The person's height, usually measured by kilograms.
peso	Waist circumference, used to assess health risks related to obesity or fat distribution.
resultado_glucosa	Glucose level result in the blood, used to assess the risk or presence of diabetes.
resultado_glucosa_promedio	Average glucose level result over a certain period, providing an overview of long-term glucose control.
riesgo_hipertension	Hypertension risk, which can be measured through various indicators like blood pressure or other risk factors.
segundamedicion_cintura	Second waist circumference measurement, usually for data verification or consistency.
segundamedicion_estatura	Second height measurement, to verify or ensure data accuracy.
segundamedicion_peso	segundamedicion_peso: Second weight measurement, for data verification or consistency data.
sexo	Gender of the individual, usually classified as male or female.
sueno_horas	Number of hours of sleep per day, which can affect overall health and well-being.
temperatura_ambiente	Ambient temperature where the measurement is taken, usually in degrees Celsius.
tension_arterial	Blood pressure, measured by millimetres of mercury (mmHg) and used to assess cardiovascular health.
valor_acido_urico	Uric acid level in the blood, used to assess the risk of gout or metabolic disorders.
valor_albumina	Albumin level in the blood, reflecting nutritional status and liver function
valor_colesterol_HDL	level of "good" cholesterol, or HDL (High-Density Lipoprotein) cholesterol.
valor_colesterol_LDL	level of "bad" cholesterol, or LDL (Low-Density Lipoprotein) cholesterol.
valor_colesterol_total	Total cholesterol level in the blood, including HDL, LDL, and other components.
valor_creatina	Creatinine level in the blood, used to assess kidney function.
valor_ferritina	Ferritin level in the blood, reflecting the body's iron stores.
valor_folato	Folate level in the blood, important for red blood cell formation and DNA function.

Iraqi Journal for Computers and Informatics
University of Information Technology and Communications
Vol. 51, No. 1, 2025, pp. 75-82
DOI: https://doi.org/10.25195/ijci.v51i1.551
Print ISSN: 2313-190X, Online ISSN: 2520-4912

valor_hemoglobina_glucosilada	Glycated haemoglobin (HbA1c) level, used to assess long-term blood glucose control.
valor_homocisteina	Homocysteine level in the blood, which may be associated with cardiovascular disease risk.
valor_insulina	Insulin level in the blood, important for blood glucose control and metabolism.
valor_proteinac_reactiva	C-Reactive Protein (CRP) level, reflecting inflammation in the body.
valor_transferrina	Transferrin level in the blood, a protein that transports iron.
valor_trigliceridos	Triglyceride level in the blood, a type of fat used as an energy source.
valor_vitamina_bdoce	Vitamin B12 level, important for nerve function and red blood cell formation.
valor_vitamina_d	Vitamin D level, important for bone health and immune system function.

2.2 Preprocessing

The data preparation stage involves several key steps. Categorical data is converted into a numeric format to ensure that the entire dataset can be processed by models implemented in Python. Data standardization is performed to prevent certain attributes from dominating the analysis, using the Min-Max Normalization method [20]. Once processed, the data is ready for use in the machine learning modeling process. A plot illustrates data points in which the X-axis stands for The Y-axis and Data_Value indicate *High_Confidence_Limit*. A linear relationship is indicated by the points' tendency to form a straight line the two variables' connection. The scatter plot illustrates the connection between *Data_Value* and *High_Confidence_Limit*. From the plot, a positive linear pattern is observed, showing that as *Data_Value* increases, *High_Confidence_Limit* also tends to increase. Overall, this analysis indicates a strong relationship between the two variables.

2.3. Data Test

The data is recorded in CSV (Comma-Separated Values) format and utilized as input for the classification stage following the completion of the last preparation step. Entering the classification phase, the data is split into training and testing datasets for use with two models: Random Forest (RF) and Decision Tree (DT). Training and testing sets of data are separated out utilizing Python 3.9.12's sklearn (sci-kit-learn) module. For this work, the data is split into 80% for training and 20% for testing. This random split proportion is chosen because it is an easy and effective method, especially handling big datasets. 1,500 samples make up the whole dataset used in this work. Using this splitting proportion, a validation dataset is obtained, ensuring a reliable process for model evaluation.

2.4. Techniques for Classifying Data Using Machine Learning

As detailed in the following subsection, we employed a number of popular machine-learning techniques.

2.4.1 Random Forest RF

Ensembles of numerous individual decision trees, or “random forests,” are formed by using random data choices, often known as "bagging," in the RF machine learning technique. Apart than bagging, RF uses random feature selection and random data subsets to build trees. Each tree in the RF anticipates a category, and the model predicts the most popular category [21].

$$Entropy\ (S) = -\sum -P_1\ log_2 P_1 \qquad (1)$$

i=1 = Number of partitions
S = Set of cases n, fraction of S to S = Pi

2.4.2 Decision Tree Classifier DT

DT, known as a tree diagram, is a graphical depiction of a series of choices or occurrences. It is used to assist discovering the optimum course of action based on particular conditions or criteria by visualizing the processes in a

Iraqi Journal for Computers and Informatics
University of Information Technology and Communications
Vol. 51, No. 1, 2025, pp. 75-82
DOI: https://doi.org/10.25195/ijci.v51i1.551
Print ISSN: 2313-190X, Online ISSN: 2520-4912

decision-making process. DT often has practical significance that can be used to aid treatment decisions by drawing reasonable medical conclusion [22].

The entropy of a dataset can be calculated by using Equation (1):

$$E(S) = \sum_{i=1}^{c} \text{Pi} - P_1 log_2 P_1 \qquad (2)$$

S = stands for starting condition,
i = arrange a class on S,
Pi = likelihood or a node's share of class I

3. RESULTS AND DISCUSSION

This work tended to use machine learning to create a model for the diagnosis and management of hypertension. The work techniques include data processing and models development using six ML algorithms, and evaluation of performance using metrics like F1-score, recall, accuracy, and precision. Based on the findings, 99.5% accuracy was attained by Random Forest (RF), whereas the Decision Tree (DT) demonstrated excellence in diagnosing hypertension, achieving the highest accuracy of 99.6%. To create an optimal model, an automated algorithm selection procedure was also applied. With its effective hypertension detection methods, this work has significant practical implications and could have a meaningful impact on public health practices, particularly in high-income regions with hedonistic lifestyles. The results demonstrate progress and superiority of the suggested paradigm over earlier papers. While additional validation is required, this work offers a solid basis for upcoming advancements in hypertension identification and management. The evaluation of machine learning methods' performance in this work is presented Table 2 and Table 3.

Table II . Values of various methods.

Algorithm	Accuracy	Precision	Recall	F1 Score
Decision Tree	**0.996**	0.991	0.987	0.989
Random Forest	0.995	0.913	0.984	0.997

Table III. Compare these methods with recent works.

AUTHOR(S)	PUBLICATION YEAR	CLASSIFIER	PERFORMANCE (%ACCURACY)
Idongesit Umoh et al. [11]	2024	SVM	90%
Niharika Patil et al. [12]	2024	RF	94.85%
Aref Andishgar et al. [13]	2024	LGBM, RF	0.67%, 65%
Lailil Muflikhah et al. [14]	2024	LSTM	89%
Oliver Danjuma [15]	2024	LSTM, RF, DT	-

The ROC Curve (Receiver Operating Characteristic Curve) is a plot used to assess the performance of a classification model by showing how the True Positive Rate TPR and False Positive Rate FPR vary with different decision thresholds.

True Positive Rate TPR: TPR, often called Sensitivity or Recall, assesses how effectively a model detects positive cases. **False Positive Rate FPR:** The frequency with which the model misclassifies negative cases as positive is measured by the FPR.

Iraqi Journal for Computers and Informatics
University of Information Technology and Communications
Vol. 51, No. 1, 2025, pp. 75-82
DOI: https://doi.org/10.25195/ijci.v51i1.551
Print ISSN: 2313-190X, Online ISSN: 2520-4912

In the context of this explanation, the model's AUC (Area Under the Curve) of 0.93 suggests that it can separate positive cases (hypertension patients) from negative cases (non-hypertension patients). With an AUC value of 0.93, it is a strong indication that your classification model performs exceptionally well in distinguishing between individuals with hypertension and those without it. Therefore, an ROC Curve with such a high AUC value reflects the good performance of the model in detecting hypertension. See Figure 2 for illustration.

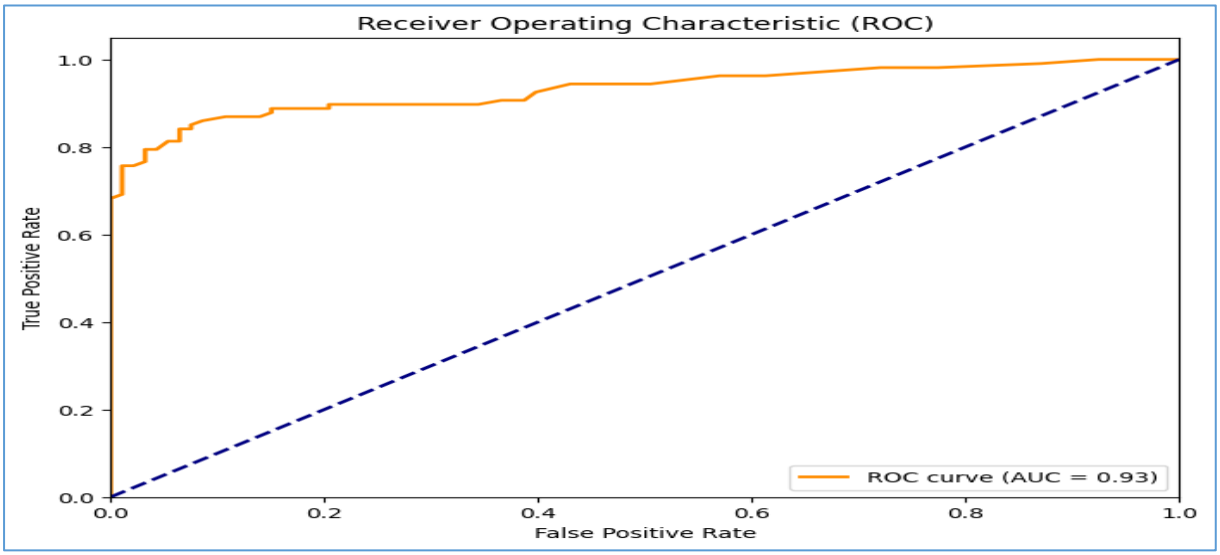

Fig 2 ROC Curve

3.1. Evaluation Metrics

It is evident from table 2 and table 3 that the Decision Tree prototype demonstrates exceptional performance compared to most models in other studies. Random Forest had an accuracy of 94.85%, whereas the Decision Tree had an accuracy of 99.6%, in the work Niharika Patil et al. The accuracy difference is: 99.6% - 94.85% = 4.75%. Therefore, the Decision Tree's accuracy is 4.75% higher than the Random Forest accuracy reported in Niharika Patil et al.'s work The model significantly outperforms those using algorithms like SVM, LSTM, and various other algorithm combinations.

These findings highlight the outstanding potential of the Decision Tree algorithm in addressing classification problems. However, to draw stronger conclusions, further analysis is needed, considering factors such as problem complexity, dataset size, and model generalization. The work findings have important ramifications for advancing our knowledge of hypertension, providing crucial contributions to better clinical judgment, and potentially improving results for patients. Integration of machine learning with transformers in the door to a deeper understanding of the complexities of hypertension is opened through our work, which can guide more precise treatment and intervention strategies.

Thus, added value in the clinical context is offered through our work, which not only focuses on scientific advancement but also drives significant improvements in health practices related to hypertension.

4. CONCLUSION

The purpose of this work is to assess the efficacy of classification algorithms in detecting hypertension and emphasize the importance of early detection. Two classification algorithms, DT and RF, or decision trees and random forests, were evaluated for detection accuracy using a dataset comprising various clinical and biological features related to hypertension. The results revealed that with an accuracy of 99.6%, the Decision Tree outperformed Random Forest, which came in second with 99.5%. These results offer insightful information about the performance of classification algorithms for early hypertension detection, forming a basis for developing more efficient detection methods. In the context of classification, we adopted metrics for performance evaluation, including F1-score, recall, accuracy, and precision. While the Decision Tree excelled in accuracy, the integration of DT and RF also highlighted key risk variables. The significance of these findings lies in the clinical relevance of integrating machine learning, particularly with RF and DT. Clinical decision-making is improved, and patient outcomes are potentially enhanced through our work's findings. By utilizing a model that integrates statistical features and two classification algorithms, our results

Iraqi Journal for Computers and Informatics
University of Information Technology and Communications
Vol. 51, No. 1, 2025, pp. 75-82
DOI: https://doi.org/10.25195/ijci.v51i1.551
Print ISSN: 2313-190X, Online ISSN: 2520-4912

significantly outperform previous studies, underscoring substantial progress. The integration of machine learning not only reflects scientific advancements but also adds value in a clinical context, opening avenues for a deeper understanding of the complexities of hypertension. These implications can guide more precise treatment and intervention strategies, ultimately strengthening healthcare practices related to hypertension.

REFERENCES

[1] A. Ostrowska *et al.*, "The impact of the COVID-19 Pandemic on hypertension phenotypes (ESH ABPM COVID-19 study)," *Eur. J. Intern. Med.*, vol. 131, no. August 2024, pp. 58–64, 2024, doi: 10.1016/j.ejim.2024.08.027.

[2] X. Wang *et al.*, "Prevalence, awareness, treatment, and control rates of hypertension in the general population of Australia: a systematic review and meta-analysis," *J. Hypertens.*, no. June, pp. 1–6, 2024, doi: 10.1097/hjh.0000000000003854.

[3] D. A. Nisdayanti, D. T. Lestari, and A. M. Rahmawati, "Dominant Factors Barriers to Hypertension Diet Management in Hypertension Sufferers," vol. 8, no. 1, pp. 48–61, 2025.

[4] E. Priyono and S. Ma, "Effects of Diet and Physical Activity on Coronary Heart Disease Risk Among Badminton Players," pp. 55–59, doi: 10.37034/medinftech.v2i2.36.

[5] D. Machnik *et al.*, "Risk factors associated with complications of palliative drainage of ascites with tunneled peritoneal catheters," *Therap. Adv. Gastroenterol.*, vol. 18, pp. 1–10, 2025, doi: 10.1177/17562848241310183.

[6] E. Priyono, "Prediction of tuberculosis patients with machine learning algorithms," vol. 9, no. 4, pp. 2334–2341, 2024.

[7] E. Priyono, T. Al Fatah, S. Ma'mun, and F. Aziz, "Tubercolusis Segmentation Based on X-ray Images," *J. Med. Informatics Technol.*, pp. 101–104, 2023, doi: 10.37034/medinftech.v1i4.22.

[8] U. N. Emeruwa *et al.*, "Lasix for the prevention of de novo postpartum hypertension: a randomized placebo-controlled trial (LAPP Trial)," *Am. J. Obstet. Gynecol.*, no. January, 2024, doi: 10.1016/j.ajog.2024.04.016.

[9] L. Xu *et al.*, "Abdominal perfusion pressure is critical for survival analysis in patients with intra-abdominal hypertension: mortality prediction using incomplete data," *Int. J. Surg.*, no. June 2024, pp. 371–381, 2024, doi: 10.1097/js9.0000000000002026.

[10] M. Magaz *et al.*, "Porto-sinusoidal vascular liver disorder with portal hypertension: Natural History and Long-Term Outcome.," *J. Hepatol.*, pp. 72–83, 2024, doi: 10.1016/j.jhep.2024.07.035.

[11] I. Umoh and V. Essien, "Optimizing Hypertension Risk Classification through Machine Learning," vol. 186, no. 14, pp. 21–29, 2024.

[12] R. Aarthi, P. Vanitha, P. Rajalakshmi, S. J. Thomas, and V. Maadhesh, "Brain Stroke Prediction Using Machine Learning," *Lect. Notes Networks Syst.*, vol. 1046 LNNS, no. 1, pp. 296–304, 2024, doi: 10.1007/978-3-031-64813-7_31.

[13] A. Andishgar *et al.*, "Machine learning-based models to predict the conversion of normal blood pressure to hypertension within 5-year follow-up," *PLoS One*, vol. 19, no. 3 March, pp. 1–17, 2024, doi: 10.1371/journal.pone.0300201.

[14] L. Muflikhah *et al.*, "Single nucleotide polymorphism based on hypertension potential risk prediction using LSTM with Adam optimizer," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 33, no. 2, pp. 1126–1139, 2024, doi: 10.11591/ijeecs.v33.i2.pp1126-1139.

[15] O. Danjuma, "MACHINE LEARNING-BASED PREDICTION OF STROKE RISK FACTORS," vol. 14, pp. 1–6, 2024.

[16] R. K. Hamad, "Integrating Machine Learning and Genetic Algorithms to Enhance Gene-Disease Classification : An XBNet-Based Framework," vol. 2025, pp. 1–12, 2025.

[17] C. Gudiato, L. Frigia, and M. Horhoruw, "G-Tech : Journal of Applied Technology Learning Approach to Big Data Using the Naive Bayes Method," vol. 9, no. 1, pp. 381–389, 2025.

[18] R. Samuelsson, "A Mixed Methods Approach to Analyzing Embodied Interaction: The Potentials of Integrated

Iraqi Journal for Computers and Informatics
University of Information Technology and Communications
Vol. 51, No. 1, 2025, pp. 75-82
DOI: https://doi.org/10.25195/ijci.v51i1.551
Print ISSN: 2313-190X, Online ISSN: 2520-4912

Mixed Methods Analysis of Video Interaction Data," *J. Mix. Methods Res.*, vol. 19, no. 1, pp. 41–57, 2023, doi: 10.1177/15586898231225496.

[19] T. Sugihartono, B. Wijaya, A. F. Alkayes, and H. A. Anugrah, "Optimizing Stunting Detection through SMOTE and Machine Learning : a Comparative Study of XGBoost , Random Forest , SVM , and k-NN," vol. 6, no. 1, pp. 667–682, 2025.

[20] P. Kaleeswari, R. Ramalakshmi, T. A. Prasath, A. Muthukumar, R. Kottaimalai, and M. T. Raj, "DABiG : Breath pattern classification using the hybrid deep learning with," 2025, doi: 10.1177/09287329241303368.

[21] G. James, "Sciences Analysis of support vector machine and random forest models for predicting the scalability of a broadband network," vol. 6, pp. 1–10, 2024.

[22] A. Gaballah, A. E. B. Abu-Elanien, and A. I. Megahed, "A Decision Tree Based Ultra-high-speed Protection Scheme for Meshed MMC-MTDC Grids with Hybrid Lines," *J. Electr. Eng. Technol.*, vol. 19, no. 2, pp. 887–900, 2024, doi: 10.1007/s42835-024-01808-9.