A Comparative Analysis of Hypertension Risk Detection Using Machine Learning and Deep Learning Techniques

Areen Arabiat [1,*], Hamza Abu Owida [2], and Muneera Altayeb [1]

[1] Department of Communication and Computer Engineering, Al-Ahliyya Amman University, Amman, Jordan
[2] Department of Medical Engineering, Al-Ahliyya Amman University, Amman, Jordan
Email: a.arabiat@ammanu.edu.jo (A.A.); h.abuowida@ammanu.edu.jo (H.A.O.); m.altayeb@ammanu.edu.jo (M.A.)
*Corresponding author

***Abstract*—A major global health concern that raises health risks considerably is hypertension. Improving early diagnosis and management of hypertension and improving patient outcomes represent the 2 benefits of utilizing Machine Learning (ML) algorithms for early detection. A lot of attention has been paid to them. This work tackles the important issue of hypertension identification by utilizing a sizable dataset from Kaggle that contains a range of physiological and demographic features. To accurately evaluate this dataset, the suggested model incorporates a number of ML classifiers, including Adaptive Boosting (AdaBoost), Gradient Boosting (GB), Random Forest (RF), Decision Tree Classifier (DT), Artificial Neural Network (ANN) and Covering Number 2 (CN2) rule inducer, to investigate their prediction potential. Performance criteria such as accuracy, precision, sensitivity, specificity, and F1-score were used to assess each classifier's effectiveness. According to the results, the GB approach had the highest accuracy (99.2%), followed by AdaBoost (96.4%). The 2 algorithms that performed exceptionally well were RF and CN2 rule inducers, which had respective accuracy rates of 93.8% and 93.5%. These results demonstrate the possibility of accurately predicting hypertension with state-of-the-art ML techniques, offering useful information to support healthcare providers in making knowledgeable decisions about patient care.**

***Keywords*—artificial intelligence, classification, gradient boosting, hypertension, Kaggle, machine learning**

I. INTRODUCTION

Hypertension, commonly referred to as high blood pressure, is a major non-communicable disease risk factor and a central contributor to the global burden of Cardiovascular Disease (CVD). It is estimated that over 1.2 billion people worldwide are currently living with hypertension, with projections suggesting continued growth due to aging populations, urbanization, and adverse lifestyles [1]. The condition is frequently asymptomatic in its early stages, earning the description "silent killer", yet it significantly increases the risk of stroke, ischemic heart disease, chronic kidney disease, and premature mortality. Because of this, early identification of individuals at high risk of developing hypertension is a critical public health and clinical priority [2].

Hypertension is inherently multifactorial, arising from the interplay of genetic, biological, behavioral, and environmental determinants. Non-modifiable risk factors such as advancing age, family history, and certain genetic polymorphisms establish a baseline predisposition [3]. Modifiable contributors, however, account for a substantial share of attributable risk, including obesity, high dietary sodium intake, low potassium intake, physical inactivity, excessive alcohol consumption, smoking, and comorbidities such as diabetes and dyslipidemia [4]. Additionally, new research emphasizes the significance of environmental exposures (such as air pollution), socioeconomic status, and psychosocial stress as relevant variables [5]. The complexity of the etiology of hypertension and the demand for prediction frameworks that can account for nonlinear, interacting, and population-specific effects are highlighted by the variety of these risk factors.

Regression-based risk scores and statistical screening models have been the backbone of traditional risk prediction techniques. The Framingham hypertension risk score and comparable models obtained from extensive cohort studies are 2 examples [6]. Based on easily measured clinical and demographic features such age, sex, baseline blood pressure, Body Mass Index (BMI), and family history, these methods have been used extensively to stratify patients. They have the benefits of interpretability, clinical familiarity, and transparency. However, their limited flexibility across varied populations, incapacity to properly utilize high-dimensional or longitudinal datasets, and assumptions of linearity and additivity frequently restrict their predictive efficacy. As a result, standard models may not always be as accurate in identifying high-risk individuals in diverse, real-world groups [7].

Manuscript received October 27, 2025; revised December 19, 2025; accepted January 5, 2026; published May 13, 2026.

The significance of early prediction extends beyond individual-level care. At the clinical level, timely identification of individuals with elevated risk allows for targeted interventions, ranging from lifestyle counseling and structured physical-activity programs to pharmacological prevention in selected high-risk cases [8]. From a health-system perspective, accurate risk stratification can optimize resource allocation, guide community screening programs, and inform the design of population-level prevention policies. Importantly, delaying or preventing hypertension onset has been shown to substantially reduce long-term healthcare costs and morbidity, strengthening the argument for robust predictive tools [9]. However, ML has emerged as a powerful approach for enhancing risk prediction in complex diseases such as hypertension. Unlike traditional regression-based models, ML algorithms, including Random Forest (RF)'s, support vector machines, Gradient Boosting (GB) methods, and deep learning architectures, do not require strict a priori assumptions about the nature of predictor–outcome relationships [10]. Instead, they can flexibly model nonlinearities, interactions, and high-dimensional data structures. ML approaches are particularly well suited to integrate diverse data sources, including Electronic Health Records (EHRs), laboratory findings, wearable-device data, imaging biomarkers, and even genetic or epigenetic information [11].

Several empirical studies have demonstrated the utility of ML techniques in predicting incident hypertension, highlighting their potential to improve early risk stratification, enhance the accuracy of predictive models compared to traditional statistical methods, and ultimately support timely clinical interventions aimed at prevention and management. Wang *et al.* [12] developed and validated a ML ensemble model using 6 algorithms, including Adaptive Boosting and logistic regression to predict 5-year incident hypertension. They trained the model on the South Korea National Health Insurance Service National Sample Cohort (n = 244,814) and externally validated it using the Japanese Medical Data Center cohort (n = 1,296,649). The ensemble achieved a balanced accuracy of 0.812 (sensitivity = 0.806, specificity = 0.818) and an AUC of 0.901 in the Korean data, with external validation showing a balanced accuracy of 0.741 and an AUC of 0.824. Key predictor features were age, diastolic and systolic blood pressure, BMI, and fasting blood glucose. Schjerven *et al.* [13] used data from the Trøndelag Health (HUNT) Study in Norway (n = 17,852, follow-up about 11 years), Schjerven *et al.* [13] compared several ML models Extreme Gradient Boosting (XGBoost), Support Vector Machine (SVM), RF, elastic regression, and k-nearest neighbor to logistic regression and recalibrated Framingham risk scores. ML models (especially XGBoost, SVM, and elastic regression) slightly outperformed references in discrimination, though improvements were modest. Important predictors identified included age, both systolic and diastolic blood pressure, BMI, height, and family history of hypertension, with ML models well-calibrated and offering clinical net benefit.

Du *et al.* [14] built a hypertension risk prediction system based on 10 ML algorithms, including RF, Light Gradient Boosting Machine (LightGBM) gradient boosting, SVM, and neural networks, using health check-up data from 1617 anonymized records. LightGBM performed best in terms of accuracy, F1-score, and Receiver Operating Characteristic (ROC). Shapley Additive explanations (SHAP) analysis revealed that age, alkaline phosphatase, and triglycerides were key contributory features. The study also introduced a web-based visualization tool to convey individualized risk and key contributing factors. Mroz *et al.* [15] investigated the prediction of hypertension control within 12 months using ML in a retrospective cohort of 350,008 hypertensive patients from the Cleveland Clinic EHRs. Using a sliding-window framework, ML models (details unspecified) achieved a pooled Area Under the Curve (AUC) of 0.76 (95% CI 0.75–0.76), sensitivity of approximately 61.5%, specificity of approximately 75.7%, PPV of approximately 67.8%, and NPV of approximately 70.5%. This demonstrates moderate predictive ability for blood pressure control using routine clinical data.

Islam *et al.* [16] trained 4 ML algorithms (artificial neural network, SVM, RF, XGBoost) using 16 risk factors selected via Boruta. The XGBoost model performed best, and SHAP analysis identified age, weight, fat mass, income, BMI, diabetes, dietary salt, hypertension in the household Hierarchical Hidden Temporal Networks (HHTN), drinking, and smoking as the most influential risk factors for hypertension prediction. Li *et al.* [17] implemented a longitudinal China Health and Retirement Longitudinal Study (CHARLS) cohort study (n = 4948, baseline normotensive Chinese adults aged ≥ 45 years, follow-up about 9 years) applied Logistic Regression (LR), RF, XGBoost, and linear SVM. XGBoost showed best performance on test data (AUC = 0.710; accuracy = 0.664; sensitivity = 0.652; specificity = 0.671; F1 = 0.592). SHAP analysis highlighted baseline systolic blood pressure, age, Triglyceride-Glucose (TyG) index, and BMI as top predictors. Time-varying effects of TyG index and depressive symptoms Center for Epidemiologic Studies Depression Scale 10 (CES-D10) were also identified via survival analysis. Mohammadi *et al.* [18] developed ML models logistic regression and recurrent neural networks to predict patients at risk of uncontrolled hypertension within 3 months, using EHR data from 14,407 (training) and 3009 (test) patients. The best model achieved an Area Under the Receiver Operating Characteristic Curve (AUROC) of 0.719, outperforming a simple last-BP baseline predictor (0.634), although the Recurrent Neural Network (RNN) did not outperform logistic regression. Ballinger *et al.* [19] used semi-supervised Long Short-Term Memory (LSTM) models on wearable heart rate sensor data (57,675 person-weeks) to detect multiple conditions. The model achieved an AUC of 0.8086 for high blood pressure detection, demonstrating the utility of sensor-based digital signals in cardiovascular risk stratification.

Nevertheless, their potential, ML techniques also present difficulties that need to be resolved in order to accomplish clinical translation. The “black-box” problem of interpretability, the risk of overfitting, and restricted generalizability when models are trained on small populations are important concerns. Furthermore, since predictive models may erroneously replicate health inequities if trained on unbalanced datasets, consideration of fairness and bias is necessary when integrating ML technologies into clinical practice. Recent developments in interpretable ML, explainable AI, and calibration methods are therefore very significant in the study of hypertension [20–22].

While advanced models can uncover important risk indicators and assist with specific risk assessment, they often have trouble integrating real-world inputs from wearable technology or electronic health records, capturing short-term risk dynamics, and incorporating longitudinal health data. To overcome these obstacles, a thorough multi-model strategy is needed, one that guarantees better calibration, interpretability, and transferability in addition to utilizing the advantages of several techniques. More dependable instruments for the early identification and successful prevention of hypertension may be provided by such an approach. This study aims to predict the risk of hypertension using multiple ML models, including ensemble methods, gradient boosting, support vector machines, and neural networks. By analyzing demographic, clinical, and lifestyle data, the study seeks to identify key risk factors, compare model performance, and provide accurate, individualized risk predictions to support early detection and prevention of hypertension

The novelty of this study is summarized by offering a novel approach that makes use of a synthetic, clinically driven Kaggle dataset designed exclusively for educational and benchmarking purposes. This novel approach not only makes testing more transparent and reproducible, but it also emphasizes essential features of model evaluation in ML. While this dataset enables systematic testing and development, it also highlights key drawbacks, most notably the lack of complete real-world clinical variability. By addressing these issues, our study contributes to the continuing debate over the balance between reproducibility and practical applicability of ML models in various healthcare contexts.

In conclusion, this work presents 3 major contributions: (1) a systematic comparison of 6 established ML classifiers for predicting hypertension risk using a publicly accessible Kaggle dataset; (2) a fully reproducible methodology that encompasses detailed preprocessing, feature management, and hyperparameter tuning; (3) a critical evaluation of the effects of utilizing a synthetic, balanced dataset on model performance, emphasizing both the benchmarking opportunities and the limitations of direct.

II. Methodology

This work was utilized for developing a ML model for detecting hypertension, as it is depicted in Fig. 1. However, the Hypertension detection dataset consists of 10 features, it was acquired from Kaggle [23].

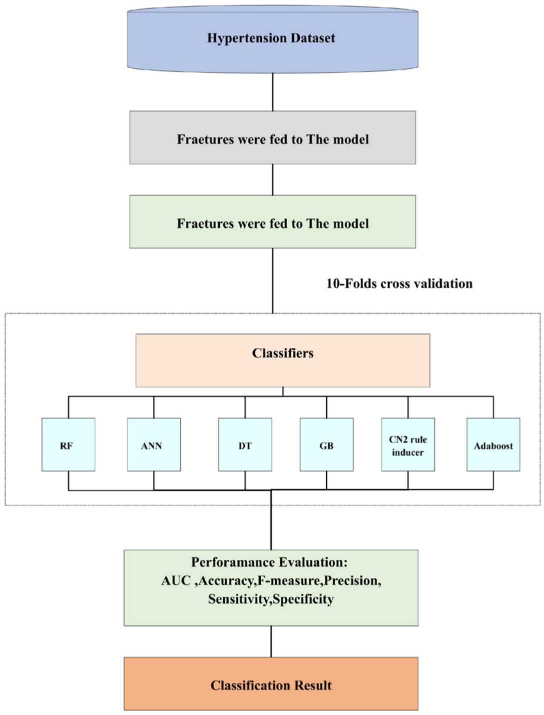

Fig. 1. Hypertension detection model structure.

This dataset used comprises 1985 records, featuring a nearly balanced distribution of the target variable, with approximately 52% indicating hypertension and 48% non-hypertensive cases. Importantly, there are no missing values. Despite the dataset’s balance, we emphasize the significance of overall accuracy and also report precision, sensitivity, specificity, F1-score, and AUC using a stratified 10-fold cross-validation strategy, which maintains class distribution in each fold. To enhance generalizability and reliability, we employed a 10-fold cross-validation method to create varying subsets of data for training and validation. This approach minimizes the risk of overfitting and provides a more accurate assessment of the model’s performance. Following the training phase, several fundamental algorithms were utilized to detect hypertension, including GB, Covering Number 2 (CN2) rule inducer, AdaBoost, DT, RF, and Artificial Neural Network (ANN).

However, grid search on training folds was used to pick tree-based ensemble hyperparameters (number of estimators, maximum depth, minimum samples per leaf, and learning rate for GB/AdaBoost) using mean F1-score as the criterion. The final RF and GB models used 200 trees with short depths (3–5 levels), whereas AdaBoost used 200 decision stumps with a learning rate of 0.05. We tested numerous shallow topologies for the ANN and chose 1 with 2 hidden layers (32 and 16 neurons, Rectified Linear Unit (ReLU) activation) and a sigmoid output neuron, trained with the Adam optimizer (learning rate 0.001), batch size 32, and early halting on validation loss.

Additionally, improving interpretability in hypertension prediction is essential to fostering confidence in the model's findings. This involves finding important predictors using feature importance analysis and explaining how features affect predictions using SHAP values or other Explainable AI techniques. These risk factors will be interpreted clinically in order to link them to patient outcomes and health. The objective is to increase the model's interpretability so that medical professionals can use it as a useful tool for controlling hypertension.

Although the relevance and interpretability of features are acknowledged, complete explain ability analyses such as SHAP were not fully utilized in the current study. The future work will focus on incorporating explainable AI methods to promote clinician trust and transparency.

The primary aim of this research is to evaluate the performance metrics of the developed model, focusing particularly on F-measure, accuracy, sensitivity, specificity, and precision post-training. Additionally, a confusion matrix will be employed to determine the percentage of incorrectly classified cases. The workflow of the proposed model, illustrated in Fig. 1, outlines the systematic process followed in developing and evaluating the machine learning approach for hypertension detection.

A. Dataset

The study's dataset incorporates both categorical and continuous features for hypertension prediction. Categorical features such as BP History, Medication, Family History, Exercise Level, and Smoking Status were encoded differently for various ML algorithms. One-hot encoding was applied to these features for the ANN, converting them into binary columns for effective interpretation. Conversely, for tree-based models like DT and RF, label encoding was utilized, assigning unique integers to each category. This preprocessing method ensures proper formatting of features for accurate analysis and predictions in the study.

TABLE I. FEATURES OF THE HYPERTENSION DATASET

Feature	Description
Age	Patient's age (in years)
Salt Intake	Daily salt intake (in grams)
Stress Score	Scale of 0–10 measuring psychological stress level
BP History	Previous blood pressure Normal, Prehypertension, Hypertension
Sleep Duration	Average sleep hours per day
BMI	Body Mass Index (weight/height-based obesity measure)
Medication	None, Beta Blocker, Diuretic, ACE Inhibitor, Other
Family History	Family history of hypertension: Yes/No
Exercise Level	Physical activity level: Low, Moderate, High
Smoking Status	Whether the patient is a Smoker or Non-Smoker

By modeling real-world scenarios, this dataset provides a helpful tool for analyzing patterns, risk factor identification, and predictive model creation to help better understand and treat hypertension. Its systematic methodology allows for comprehensive analysis for the objective of advancing research and improving health outcomes related to this prevalent condition. This allocation technique allows users to train the model on large amounts of data, increasing prediction accuracy and enabling results to be extrapolated to new, untested data. In addition to improving the model's dependability, this method facilitates more accurate performance evaluation. The features of the hypertension dataset utilized in the proposed model are detailed in Table I [23]. This table provides a comprehensive overview of the various attributes included in the dataset, which are critical for predicting hypertension risk.

B. Preprocessing

However, Preprocessing is an important step in ML that improves model performance by preprocessing raw data. This study contains crucial preprocessing procedures such as checking the dataset's completeness, applying label encoding to categorical variables like BP History and Smoking Status for tree-based models, and one-hot encoding for ANN to convert them to binary forms. To boost the ANN's convergence, continuous variables such as age and BMI were normalized for a zero mean and unit variance. To prevent data leaking, changes were applied within each training fold during cross-validation.

C. Classification Methods

ML is the use of Artificial Intelligence (AI), which improves the fact by offering a programming-free environment. ML aims to develop systems that can gather information and utilize it to learn how to make judgments in the future [24]. The methods used in this work to classify data into different readings in order to assess the strength of concrete are examined in this part. These algorithms include RF, DT, ANN, GB, AdaBoost, and CN2 rule inducer techniques.

1) Random Forest (RF)

During training, the RF classifier, an ensemble ML approach, mixes many trees [25]. Each split is informed by a randomly chosen subset of characteristics, and each tree within the forest is trained using a randomly chosen fraction of the training data. The forecast of a Random Forest (RF) classifier is determined by aggregating the majority votes from all individual decision trees in the ensemble, as illustrated in Fig. 2. The majority vote from each tree is combined to determine the prediction of a RF Classifier. $Ti(x)$ represents the *ith* tree forecast, while $\hat{Y}(x)$ denotes an estimator or a prediction of a value based on input. Eq. (1) determines the RF ensemble's final prediction, where the input data point is denoted by x [26].

$$\hat{Y}(x) = \text{mode}\,\{T_1(x), T_2(x), \cdots, T_n(x)\} \tag{1}$$

2) Gradient Boosting (GB)

GB is an ensemble learning technique that iteratively adds weak learners. The GB approach trains each tree using the residual errors of the previous trees, as opposed to the RF approach [28]. The model's subsequent trees give the remaining flaws priority. The GB Classifier calculates its prediction by adding up all of the weak learners' predictions. Eq. (2) provides the final prediction of the GB ensemble:

$$\hat{Y}(x) = \sum_{i=1}^{n} f_i(x) \tag{2}$$

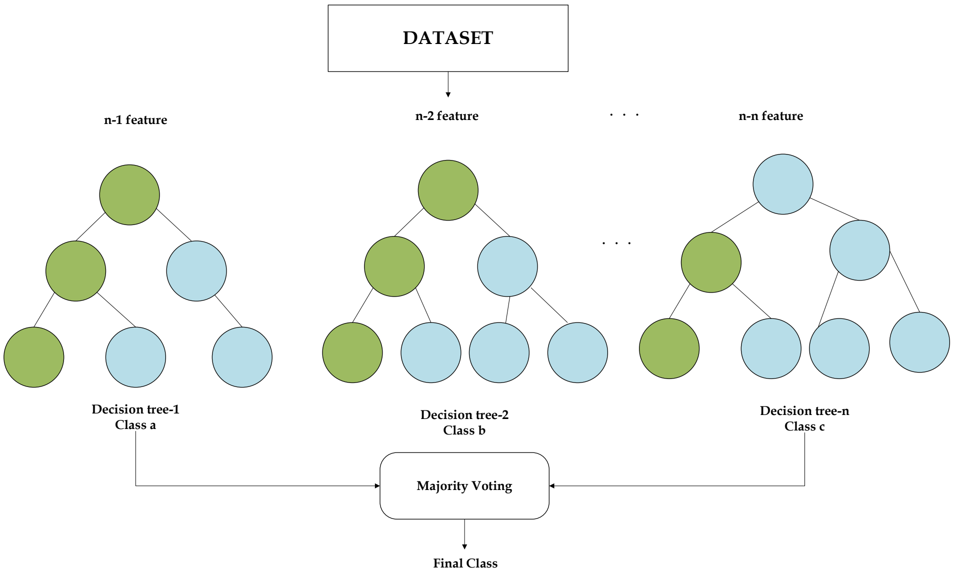

Fig. 2. RF architecture [27].

where $f_i(x)$ is the prediction of the *ith* weak learner, where x is the input data point, and n is the total number of weak learners and $Y(x)$ in this context represents the total estimate or prediction derived from aggregating the outputs of all the individual functions [29].

3) *AdaBoost*

A technique called AdaBoost combines weak base classifiers one after the other to create a powerful classifier, as depicted in Fig. 3, operates through a sequential ensemble method that focuses on improving classification performance by combining the predictions of multiple weak classifiers.3. It is among ML's most important advancements [30]. Shallow decision trees serve as its foundational classifiers. The training sample is reweighted at each iteration so that the subsequent DT [31]. Eq. (3) demonstrates that when a new tree model is presented, the general tree is deleted and only the strongest tree is incorporated into the system. With this approach, the model's overall performance steadily becomes better as more computations are made. The newly added tree is represented by $y_{i,}$, the *ith* tree, and the entire model is represented by $F_n(x)$ when the inserted tree is indicated by $h(x_i)$. In the other hand, m denotes the iteration number in a sequential modeling process, while L represents a loss function that quantifies the difference between predicted values and observed outcomes. This formulation is commonly used in ensemble methods to iteratively improve predictions [32].

$$F_n(x) = F_{m-1}(x) + \text{argminh}_h \sum_{i=1}^{n} L(y_{i,} F_{m-1}(x_i) + h(x_i) \quad (3)$$

4) *Decision Tree (DT)*

One popular ML algorithm is the decision tree. Layer by layer, it outputs the judgment findings after evaluating the data attributes in a tree structure. The last leaf node indicates the outcome of a classification or prediction as shown in Fig. 4 which illustrates the architecture of a DT. Additionally DT is a classification model that is frequently used in data mining. Each tree is made up of nodes and branches. Every subset specifies a value that the node can accept, and each node represents features in a category that has to be classified [34, 35].

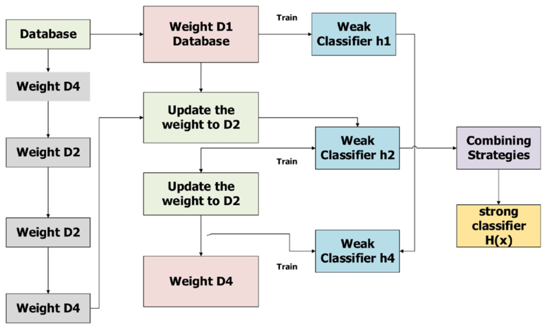

Fig. 3. AdaBoost classifier structure [33].

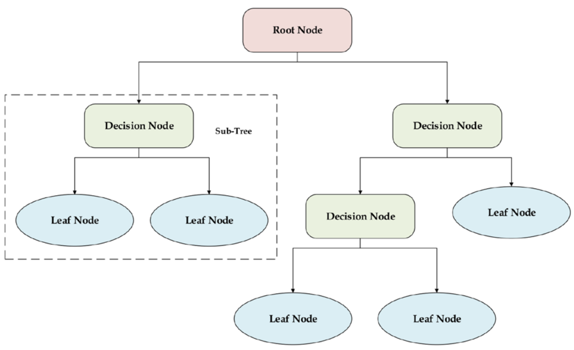

Fig. 4. DT architecture [35].

5) *Artificial Neural Network (ANN)*

Artificial Neural Network (ANN) is a computational model made up of individual units known as artificial neurons that are connected by weights. ANNs are information-driven tools that can capture complex and nonlinear relationships between input and output datasets without any prior understanding of the underlying processes [36]. Typically, these networks are made up of an input layer, one or more hidden layers, and an output layer. The input layer sends data to the first hidden layer, which serves as a feature detector, and the output layer is responsible for providing outcomes that correlate to specific inputs. Choosing the suitable network architecture is a major difficulty in system modeling [37, 38]. A schematic illustration of a general 3-layer Artificial Neural Network (ANN) model is presented in Fig. 5. This diagram provides a clear visualization of the architecture and components of the network. In this sudy the following ANN architectural options were set to improve reproducibility. The model uses the Rectified Linear Unit (ReLU) activation function and has 100 neurons in the hidden layers. The Adaptive Moment Estimation (ADAM) optimizer is used in the optimization process, and regularization is implemented using an Alpha parameter of 0.05 to reduce overfitting. Furthermore, during the 200 iterations of training, an error threshold of 0.0001 guarantees sufficient convergence of the model. By offering these particular criteria, we hope to increase the transparency of our methodology and make it easier for future research to duplicate or expand upon our findings in the prediction of hypertension.

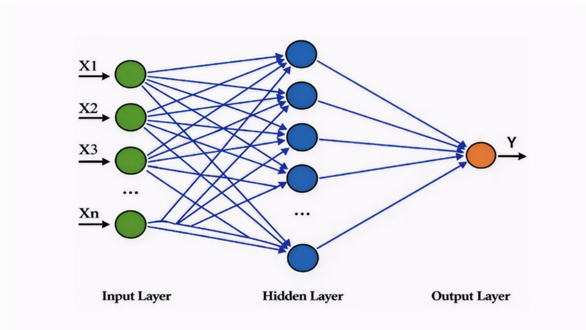

Fig. 5. ANN classifier structure [39].

6) *Rule Induction (CN2)*

To produce a more accurate model, a rule-based categorization algorithm iteratively develops and refines rules. By focusing on patterns that were overlooked by earlier regulations included in the final conclusion, each additional rule in the ensemble increases the accuracy of the model [40]. Through an iterative procedure, the CN2 algorithm, which we use in this study, produces a ranked rule list. In the process, the algorithm iteratively looks for good rules that encompass as many occurrences of the data set as possible. An evaluation function (such as the Laplace estimate or entropy) is used to estimate the goodness of a given rule [41]. While the remaining observations are progressively covered by other rules, the occurrences covered by this rule are then eliminated. If every instance is covered and no additional rules are discovered, the procedure finally comes to an end [42].

III. Performance Evaluation

A. *Confusion Matrix*

The confusion matrix generates measures F1-score, recall/sensitivity, accuracy, specificity and precision by contrasting predicted labels with real labels in the test dataset, so the confusion matrix used to evaluate the model's performance [43, 44]. True Positives (TPs), True Negatives (TNs), False Positives (FPs), and False Negatives (FNs) are displayed in the confusion matrix, a 2 × 2 table that assesses classification results. Higher TP and TN values indicate improved model performance [45–47]. The confusion matrix analysis in Table II illustrates the discrepancy between actual and anticipated classifications.

TABLE II. Confusion Matrix

Category	Predicted	
Actual	True Positive (TP)	False Negative (FN)
	False Positive (FP)	True Negative (TN)

The confusion matrix is employed to evaluate the effectiveness of classifiers in this study, as illustrated in Fig. 6. This figure presents the confusion matrices for 6 different models: (a) the confusion matrix of ANN, (b) confusion matrix of DT, (c) the confusion the matrix of RF, (d) the confusion matrix of CN2 rule inducer, (e) the confusion matrix of GB, and (f) the confusion matrix of AdaBoost.

Actual \ Predicted	No	Yes	∑
No	855	98	953
Yes	104	928	1,032
∑	959	1,026	1,985

(a)

Actual \ Predicted	No	Yes	∑
No	910	43	953
Yes	44	988	1,032
∑	954	1,031	1,985

(b)

Actual \ Predicted	No	Yes	∑
No	903	50	953
Yes	73	959	1,032
∑	976	1,009	1,985

(c)

Actual \ Predicted	No	Yes	∑
No	899	54	953
Yes	76	956	1,032
∑	975	1,010	1,985

(d)

Actual \ Predicted	No	Yes	∑
No	946	7	953
Yes	9	1023	1,032
∑	955	1,030	1,985

(e)

Actual \ Predicted	No	Yes	∑
No	919	34	953
Yes	37	995	1,032
∑	956	1,029	1,985

(f)

Fig. 6. Confusion matrix of all classifiers. (a) Confusion matrix of ANN classifier. (b) Confusion matrix of DT classifier. (c) Confusion matrix of RF classifier. (d) Confusion matrix of CN2 rule inducer classifier. (e) Confusion matrix of GB classifier. (f) Confusion matrix of AdaBoost classifier.

B. Performance Metrics

Validity and efficiency can be confirmed using the performance evaluation approach. The 10-folds Cross-validation approach is used to eliminate the over-fitting issue and evaluate the data's predicted success after training on the training set. This work used 6 ML classification classifiers: ANN, RF, DT, GB, CN2 rule inducer, and AdaBoost. Every classifier's performance was compared. A confusion matrix indicates which predictions a classification model made correctly and incorrectly using a sample of test data.

Accuracy: defined as the ratio of correctly identified samples to all samples. Although accuracy is a useful indicator, it may not be enough for datasets that are imbalanced, since it disproportionately favors the dominant class, as shown in Eq. (4) [48, 49].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

Sensitivity: is the model's ability to predict positive cases. It is calculated by dividing the total number of positive forecasts by the number of positive predictions as described in Eq. (5) [50]:

$$Sensitivity = \frac{TP}{TP+FN} \tag{5}$$

Precision: shows the percentage of correctly identified positive data that is really anticipated; high precision reduces false positives, as shown in Eq. (6) [51, 52]:

$$Precision = \frac{TP}{TP+FP} \tag{6}$$

F-measure: harmonic mean of the precision and sensitivity as shown in Eq. (7) [53]:

$$F\text{-}measure = \frac{2 \times Precision \times Sensitivity}{Precision+Sensitivity} \tag{7}$$

Specificity: evaluated the model's capacity to accurately detect true negatives, assisting in determining how well it prevented false positives on cases that were actually negative, as shown in Eq. (8) [54]:

$$Specificity = \frac{TN}{TN+FP} \tag{8}$$

Fig. 7 illustrates various performance metrics used to evaluate the effectiveness of classifiers in this study. These metrics provide a comprehensive assessment of the models' predictive capabilities: (a) the AUC, (b) the Accuracy, (c) the F-measure, (d) precision, (e) the sensitivity, and (f) the specificity.

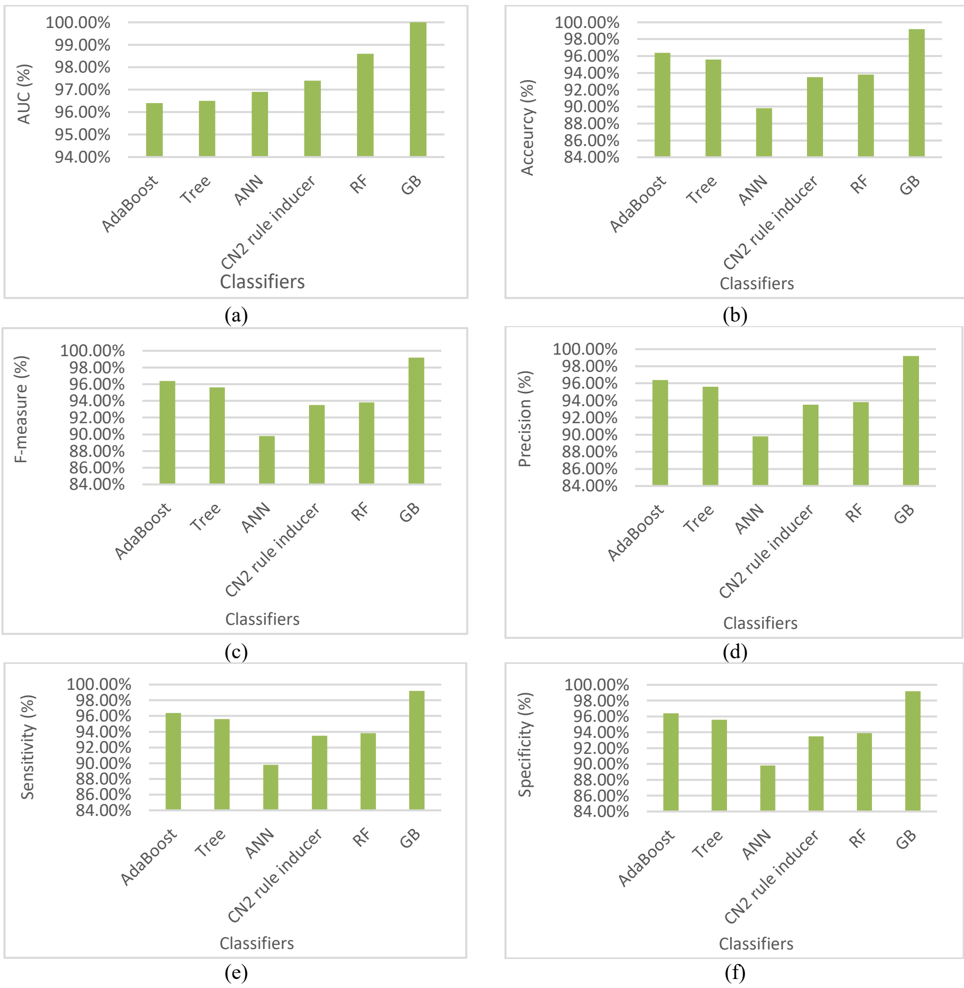

Fig. 7. Performance metrics of all classifiers. (a) AUC of all classifiers. (b) Accuracy of all classifiers. (c) F-measure of all classifiers. (d) Precision of all classifiers. (e) Sensitivity of all classifiers. (f) Specificity of all classifiers.

Beyond the mathematical definition of evaluation metrics, each metric has specific clinical and methodological implications. Accuracy measures overall correctness but can hide class-specific misclassification trends, especially in medical screening tasks. Sensitivity is essential for identifying hypertension patients, as false negatives can delay treatment and increase cardiovascular risk. Specificity measures the model's ability to reliably identify normotensive patients, reducing unnecessary follow-up or treatment. Precision reduces false-positive rates, which is critical for preventing patient concern and resource abuse. The F-measure strikes a balance between precision and sensitivity, providing a reliable indicator of uneven decision costs. Finally, the area under the ROC Curve (AUC) is a threshold-independent measure of discrimination, making it particularly ideal for comparative model evaluation in clinical scenarios.

IV. Results and Discussion

AUC, accuracy, F-measure, precision, sensitivity, and specificity were among the measures used to assess how well different models predicted hypertension. However, for assessing how well a model predicts hypertension in this work, the Area Under the Curve (AUC) is an essential statistic. With a remarkable AUC of 99.9%, the GB classifier performs better than any other model, suggesting superior discrimination and a lower risk of false negatives. An AUC of 97.9% was attained with the robust RF. The AUCs for the ANN and CN2 rule inducer were 96.5% and 96.6%, respectively. The AUCs for the AdaBoost and DT classifiers were 95.6% and 95.4%, respectively. When it comes to forecasting hypertension, ensemble approaches, specifically RF and GB, perform better than individual models.

In terms of accuracy, the GB successfully classified both hypertensive and non-hypertensive cases, with the greatest accuracy of 98.7%. AdaBoost demonstrated its efficacy in raising prediction rates with a 95.7% accuracy rate. Despite its 94.5% accuracy, DT's shortcomings as compared to ensemble approaches point to overfitting. Although the RF and CN2 rule inducer models achieved 92.1% and 92.3% accuracy, respectively, their performance might require more fine-tuning. At 89.3%, the ANN accuracy was the lowest. This finding is supported by the F-measure results, which show that the GB achieved an impressive 98.7%. This measure, which achieves a compromise between recall and precision, shows that the model successfully detects real positive cases with a low false positive rate. In contrast, the CN2 rule inducer and Tree model both obtained F-measures of 92.3% and 94.5%, respectively, indicating that they are less successful than GB but still dependable.

Metrics for precision and sensitivity provide additional insight into the models' effectiveness. At 98.7%, the GB classifier's high precision and sensitivity guarantee that it not only correctly detects hypertension patients but also reduces false positives. On the other hand, the ANN classifier demonstrated consistent metrics for both sensitivity and precision. However, all models maintained consistently high specificity scores, with the GB classifier reaching 98.7%. This implies that the classifiers are successful in correctly detecting negative situations, which is important in medical settings because incorrect diagnoses can result in needless worry and treatment.

However, for each performance indicator (AUC, accuracy, F-measure, precision, sensitivity, and specificity), we will compute 95% confidence intervals. This will provide a range of values within which we can be certain about the genuine performance. To calculate these intervals, we shall use bootstrapping or other appropriate statistical approaches. Near-perfect AUC and accuracy scores, especially when cross-validated, may represent dataset simplicity rather than genuine clinical predictive power. Balanced class distributions and low feature complexity can enhance perceived performance, emphasizing the significance of careful interpretation.

In the other hand, the dataset size is an important factor in affecting the stability and generalizability of machine learning models. Smaller datasets have more variance and sensitivity to noise, which can lead to overly optimistic or unstable performance predictions. Larger datasets, particularly those drawn from real-world clinical settings, bring heterogeneity, missingness, and measurement variability, which frequently impair apparent model performance. In the current study, the relatively small and balanced synthetic dataset (n = 1985) allows for efficient learning and contributes to near-ceiling performance measures. As dataset sizes grow, learning curves are predicted to plateau at lower, more realistic performance levels. The published results should be taken as methodological benchmarks rather than approximations of real-world clinical accuracy.

Table III explores the performance measures of the model's different machine learning classifiers, including accuracy, precision, recall, F1-score, and AUC for effective hypertension prediction. Table IV compares the proposed model's performance to previous research, emphasizing methodological improvements. Fig. 8 depicts a visual comparative examination of classifier performance, assisting in the identification of the best models. Together, these elements provide a full knowledge of the classifier's strengths and weaknesses in predicting hypertension risk.

TABLE III. Comparison of ML Performance Classifiers on the Training Dataset

Model	AUC	Accuracy	F-measure	Precision	Sensitivity	specificity
AdaBoost	96.4 %	96.4%	96.4%	96.4%	96.4%	96.4%
Tree	96.5%	95.6%	95.6%	95.6%	95.6%	95.6%
ANN	96.9%	89.8%	89.8%	89.8%	89.8%	89.8%
CN2 rule inducer	97.4%	93.5%	93.5%	93.5%	93.5%	93.5%
RF	98.6%	93.8%	93.8%	93.8%	93.8%	93.9%
GB	100%	99.2%	99.2%	99.2%	99.2%	99.2%

TABLE IV. COMPARISON OF PREVIOUS STUDIES WITH THE PROPOSED MODEL

Study	Model	Performance Metrices
Wang *et al.* [12]	Ensemble model (Adaptive Boosting, Logistic Regression)	Accuracy: 0.812, Sensitivity: 0.806, Specificity: 0.818, AUC: 0.901, Balanced Accuracy: 0.741, AUC: 0.824
Mroz *et al.* [15]	Unspecified ML models	AUC: 0.76, Sensitivity: about 61.5%, Specificity: about 75.7%
Li *et al.* [17]	Logistic Regression, Random Forest, XGBoost, Linear SVM	XGBoost; AUC: 0.710, Accuracy: 0.664, Sensitivity: 0.652, Specificity: 0.671
Mohammadi *et al.* [18]	Logistic Regression, Recurrent Neural Networks	AUC-ROC: 0.719, outperformed last-BP baseline (0.634)
Ballinger *et al.* [19]	Semi-supervised LSTM	AUC: 0.8086 for high blood pressure detection
Proposed	GB, CN2 rule inducer, AdaBoost, DT, RF, and ANN	GB AUC: 1.00, Accuracy: 99.2%

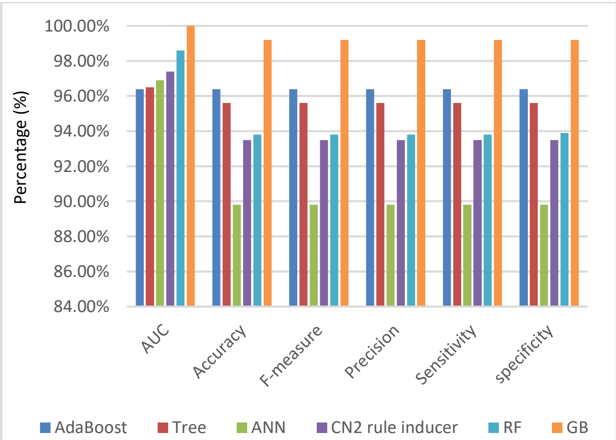

Fig. 8. Comparative analysis of different classifiers' performances.

Although the GB classifier had an AUC near 1.0 and an accuracy of 99.2% on the synthetic Kaggle dataset, these results should be regarded with caution. Even with stratified 10-fold cross-validation, a balanced class distribution, no missing data, and simplified feature interactions can lead to optimistic performance predictions. As a result, the current findings should be seen as a methodological standard, with external validation on real-world datasets required prior to clinical application.

However, the performance values presented in Table IV for prior studies were acquired from diverse datasets (mainly large electronic health record cohorts and longitudinal epidemiological studies). Table IV should be viewed as a qualitative benchmark demonstrating that our GB-based model achieves performance on the synthetic Kaggle dataset that is comparable to, but not directly equivalent to, models trained on real-world clinical data. However, Comparisons with earlier studies are qualitative in nature, because changes in datasets, population characteristics, and study design prevent direct performance benchmarking.

Additionally, this paper also expands on the issue of misclassification patterns using confusion matrices. This examination now focuses on specific areas where models struggle, outlining the types of errors—such as false positives and false negatives—and their implications for clinical practice. By combining these elements, we expect to gain a better understanding of model performance and practical issues in hypertension risk prediction.

V. CONCLUSION AND FUTURE WORK

Using a dataset from Kaggle for training and validation, the main objective of this study was to develop a ML model that can distinguish between hypertension and non-hypertensive cases. Metrics including recall, accuracy, specificity, precision, and F1-score were used to evaluate the performance of a number of classifiers, including AdaBoost, GB, RF, DT, CN2 rule inducer, and Artificial Neural Networks (ANN). The GB classifier's remarkable accuracy of 99.2% was emphasized in the paper, making it the most successful model among those assessed and demonstrating the substantial potential of ensemble approaches in medical classification tasks. This model gives doctors a trustworthy way to recognize hypertension, which is essential for prompt patient monitoring and care. However, Future studies will concentrate on using real-time data from wearable devices and testing the model with other datasets. Additionally, there is a focus on broadening the dataset to incorporate variables such as physical activity levels, eating patterns, and socioeconomic status. Furthermore, investigating hybrid models that combine deep learning with conventional techniques may improve predictive abilities, and tweaking hyperparameters through methods such as Bayesian optimization may further improve model performance. Longitudinal research and better model interpretability will be necessary to foster trust among medical providers. Improving patient engagement and encouraging better self-management of hypertension through the development of an intuitive application for personalized risk assessments and practical advice could lead to a more successful hypertension detection system and better patient treatment.

In the other hand, Models which are trained using real-world electronic health records or longitudinal cohorts often perform worse because to noise, missing data, and population heterogeneity. As a result, the current findings should be viewed as controlled benchmarking outcomes rather than direct markers of clinical awareness. Finally, the suggested models rely on static feature snapshots and do not account for temporal changes in blood pressure, medication effects, or illness development. Future studies will look into longitudinal modeling methodologies.

Despite their high performance, the presented models have several drawbacks. The use of a synthetic, class-balanced Kaggle dataset limits clinical generalizability and may result in too optimistic performance predictions. Furthermore, the lack of longitudinal and temporal modeling limits the ability to understand illness development and therapy effects. While interpretability is discussed, thorough explainable AI assessments have not been fully implemented. These limitations emphasize the need for external validation on real-world datasets, integration of explainable AI approaches, and adoption of longitudinal modeling frameworks prior to clinical deployment.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization, Areen Arabiat and Hamza Abu Owida; methodology, Areen Arabiat and Muneera Altayeb; software, Areen Arabiat; validation, Areen Arabiat and Hamza Abu Owida; formal analysis, Areen Arabiat and Hamza Abu Owida; investigation, Hamza Abu Owida and Muneera Altayeb; data curation, Areen Arabiat and Hamza Abu Owida; writing—original draft preparation, Hamza Abu Owida, Areen Arabiat, and Muneera Altayeb; writing—review and editing, all authors; visualization, Areen Arabiat and Muneera Altayeb; supervision, Hamza Abu Owida and Muneera Altayeb. All authors reviewed the results and approved the final version of the manuscript.

Acknowledgment

The author wishes to thank Al Ahliyya Amman University for their support.

References

[1] World Health Organization. (September 2025). Hypertension—fact sheet. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/hypertension

[2] S. Goorani, S. Zangene, and J. D. Imig, "Hypertension: A continuing public healthcare issue," *International Journal of Molecular Sciences*, vol. 26, no. 1, 123, 2025.

[3] G. B. Ehret and M. J. Caulfield, "Genes for blood pressure: An opportunity to understand hypertension," *European Heart Journal*, vol. 34, no. 13, pp. 951–961, 2013.

[4] M. H. Forouzanfar, P Liu, G. A. Roth *et al.*, "Global burden of hypertension and systolic blood pressure of at least 110 to 115 mm Hg, 1990–2015," *JAMA*, vol. 317, no. 2, pp. 165–182, 2017.

[5] R. D. Brook and S. Rajagopalan, "Particulate matter, air pollution, and blood pressure," *Journal of the American Society of Hypertension*, vol. 3, no. 5, pp. 332–350, 2009.

[6] N. I. Parikh, M. J. Pencina, T. J. Wang *et al.*, "A risk score for predicting near-term incidence of hypertension: The framingham heart study," *Annals of Internal Medicine*, vol. 148, no. 2, pp. 102–110, 2008.

[7] D. H. Syllos, V. F. Calsavara, I. M. Bensenor *et al.*, "Validating the framingham hypertension risk score: A 4-year follow-up from the Brazilian longitudinal study of the adult health (ELSA-Brasil)," *Journal of Clinical Hypertension*, vol. 22, no. 5, pp. 850–856, 2020.

[8] M. Satoh, Y. Tatsumi, S. Nakayama *et al.*, "Self-measurement of blood pressure at home using a cuff device for change in blood pressure levels: Systematic review and meta-analysis," *Hypertens. Res.*, vol. 48, pp. 574–591, 2025.

[9] D. Kostova, G. Spencer, A. E. Moran *et al.*, "The cost-effectiveness of hypertension management in low-income and middle-income countries: A review," *BMJ Global Health*, vol. 5, no. 9, Art. no. e002213, 2020.

[10] L. J. Appel, C. M. Champagne, D. W. Harsha *et al.*, "Effects of comprehensive lifestyle modification on blood pressure control," *JAMA*, vol. 289, no. 16, pp. 2083–2093, 2003.

[11] K. Shameer, M. A. Badgeley, R. Miotto *et al.*, "Translational bioinformatics in the era of real-time biomedical, health care and wellness data streams," *Briefings in Bioinformatics*, vol. 18, no. 1, pp. 105–124, 2016.

[12] C. C. Wang, T. W. Chu, and J. S. R. Jang, "Next-visit prediction and prevention of hypertension using large-scale routine health checkup data," *PLoS ONE*, vol. 19, no. 11, e0313658, 2024.

[13] F. E. Schjerven, F. Lindseth, and I. Steinsland, "Prognostic risk models for incident hypertension: A PRISMA systematic review and meta-analysis," *PLoS ONE*, vol. 19, no. 3, e0294148, 2024.

[14] J. Du, X. Chang, C. Ye *et al.*, "Developing a hypertension visualization risk prediction system utilizing machine learning and health check-up data," *Scientific Reports*, vol. 13, 18953, 2023.

[15] T. Mroz, M. Griffin, R. Cartabuke *et al.*, "Predicting hypertension control using machine learning," *PLoS One*, vol. 19, no. 3, e0299932, 2024.

[16] M. M. Islam, M. J. Alam, M. Maniruzzaman *et al.*, "Predicting the risk of hypertension using machine learning algorithms: A cross-sectional study in Ethiopia," *PLoS ONE*, vol. 18, no. 8, e0289613, 2023.

[17] J. Li, Z. Song, Q. Xu *et al.* (June 2025). Machine learning prediction of incident hypertension using baseline biomarkers: Evidence from the CHARLS cohort. *Research Square*. [Online]. Available: https://www.researchsquare.com/article/rs-6824239/v1

[18] R. Mohammadi, S. Jain, S. Agboola *et al.*, "Learning to identify patients at risk of uncontrolled hypertension using electronic health records data," *AMIA Jt Summits Transl. Sci. Proc.*, vol. 2019, pp. 533–542, 2019.

[19] B. Ballinger, J. Hsieh, A. Singh *et al.*, "DeepHeArt: Semi-supervised sequence learning for cardiovascular risk prediction," in *Proc. Thirty-Second AAAI Conf. on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conf. and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, 2018, pp. 2079–2086.

[20] N. Terranova, K. Venkatakrishnan, and L. J. Benincosa, "Application of machine learning in translational medicine: Current status and future opportunities," *AAPS J.*, vol. 23, no. 4, 74, 2021.

[21] E. V. Bernstam, P. K. Shireman, F. Meric-Bernstam *et al.*, "Artificial intelligence in clinical and translational science: Successes, challenges and opportunities," *Clin. Transl. Sci.*, vol. 15, no. 2, pp. 309–321, 2022.

[22] D. I. Kasartzian and T. Tsiampalis, "Transforming cardiovascular risk prediction: A review of machine learning and artificial intelligence innovations," *Life*, vol. 15, no. 1, 94, 2025.

[23] Kaggle. Hypertension risk prediction dataset. [Online]. Available: https://www.kaggle.com/datasets/miadul/hypertension-risk-prediction-dataset

[24] D. O. Melinte, A. M. Travediu, and D. N. Dumitriu, "Deep convolutional neural networks object detector for real-time waste identification," *Applied Sciences*, vol. 10, no. 20, 7301, 2020.

[25] S. Li, X. Yang, P. Cui *et al.*, "Machine-learning-algorithm-based prediction of land use/land cover and land surface temperature changes to characterize the surface urban heat island phenomena over Harbin, China," *Land*, vol. 13, no. 8, 1164, 2024.

[26] S. S. Hussain and S. S. H. Zaidi, "AdaBoost ensemble approach with weak classifiers for gear fault diagnosis and prognosis in DC motors," *Applied Sciences*, vol. 14, no. 7, 3105, 2024.

[27] Z. Azouz, B. H. S. Asli, and M. Khan, "Evolution of crack analysis in structures using image processing technique: A review," *Electronics*, vol. 12, no. 18, 3862, 2023.

[28] Z. M. Çınar, A. A. Nuhu, Q. Zeeshan *et al.*, "Machine learning in predictive maintenance towards sustainable smart manufacturing in industry 4.0," *Sustainability*, vol. 12, no. 19, 8211, 2020.

[29] M. Altayeb, A. Arabiat, and A. Al-Ghraibah, "Detection and classification of pneumonia using the Orange3 data mining tool," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 14, no. 6, pp. 6894–6903, 2024.

[30] A. M. Arabiat, "Intelligent model for detecting GAN-generated images based on multi-classifier and advanced data mining techniques," *International Journal of Electrical and Electronic Engineering & Telecommunications*, vol. 14, no. 3, pp. 147–157, 2025.

[31] M. Altayeb and A. Arabiat, "A sustainable system for predicting appliance energy consumption based on machine learning," *Journal of Environmental Management*, vol. 382, 125434, 2025.

[32] C. Wang, S. Xu, and J. Yang, "AdaBoost algorithm in artificial intelligence for optimizing the IRI prediction accuracy of asphalt concrete pavement," *Sensors*, vol. 21, no. 17, 5682, 2021.

[33] W. Ni, L. Zhao, L. Zhang *et al.*, "Coupling progressive deep learning with the AdaBoost framework for landslide displacement

rate prediction in the Baihetan Dam Reservoir, China," *Remote Sensing*, vol. 15, no. 9, 2296, 2023.
[34] Z. Wang and K. Gai, "Decision tree-based federated learning: A survey," *Blockchains*, vol. 2, no. 1, pp. 40–60, 2024.
[35] B. Charbuty and A. M. Abdulazeez, "Classification based on decision tree algorithm for machine learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 20–28, 2021.
[36] S. Agatonovic-Kustrin and R. Beresford, "Basic concepts of Artificial Neural Network (ANN) modeling and its application in pharmaceutical research," *Journal of Pharmaceutical and Biomedical Analysis*, vol. 22, no. 5, pp. 717–727, 2000.
[37] M. A. Ghorbani, R. Khatibi, V. Karimi *et al.*, "Learning from multiple models using artificial intelligence to improve model prediction accuracies: Application to river flows," *Water Resour. Manage*, vol. 32, pp. 4201–4215, 2018.
[38] W. Toghuj and Y. Alraba'nah, "A two-stage approach for aircraft detection with convolutional neural network," *International Journal of Electrical & Computer Engineering*, vol. 14, no. 4, 4627, 2024.
[39] P. Tsikas, A. Chassiakos, V. Papadimitropoulos *et al.*, "BIM-based machine learning application for parametric assessment of building energy performance," *Energies*, vol. 18, no. 1, 201, 2025.
[40] M. S. Alzboon, M. Alqaraleh, and M. S. Al-Batah, "Diabetes prediction and management using machine learning approaches," arXiv preprint, arXiv: 2506.11501, 2025.
[41] E. Naranjo, N. Ulloa, K. M. V. Vallejo *et al.*, "Prediction and validation of compressive strength of metakaolin-based mortars using machine learning," *Asian Journal of Civil Engineering*, vol. 26, pp. 3423–3451, 2025.
[42] F. Heymann, R. Bessa, M. Liebensteiner *et al.*, "Scarcity events analysis in adequacy studies using CN2 rule mining," *Energy and AI*, vol. 8, 100154, 2022.
[43] A. Arabiat and M. Altayeb, "Driving behavior analytics: An intelligent system based on machine learning and data mining techniques," *Bulletin of Electrical Engineering and Informatics*, vol. 14, no. 3, pp. 2055–2065, 2025.
[44] N. Alshdaifat, H. A. Owida, Z. Mustafa *et al.*, "Automated blood cancer detection models based on EfficientNet-B3 architecture and transfer learning," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 36, no. 3, pp. 1731–1738, 2024.
[45] A. Arabiat, M. Hassan, and O. Almomani, "WEKA-based machine learning for traffic congestion prediction in Amman City," *IAES International Journal of Artificial Intelligence*, vol. 13, no. 4, pp. 4422–4434, 2024.
[46] M. M. Abualhaj, S. N. Al-Khatib, A. A. Abu-Shareha *et al.*, "Spam detection boosted by firefly-based feature selection and optimized classifiers," *International Journal of Advances in Soft Computing*, vol. 17, no. 3, pp. 16–33, 2025.
[47] E. Brati, A. Braimllari, and A. Gjeçi, "Machine learning applications for predicting high-cost claims using insurance data," *Data*, vol. 10, no. 6, 90, 2025.
[48] Q. Shambour, M. Al-Zyoud, and O. Almomani, "Quantum-inspired hybrid metaheuristic feature selection with SHAP for optimized and explainable spam detection," *Symmetry*, vol. 17, no. 10, 1716, 2025.
[49] R. Alazaidah, H. A. Owida, N. Alshdaifat *et al.*, "A comprehensive analysis of eye diseases and medical data classification," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 22, no. 6, pp. 1422–1430, 2024.
[50] A. M. Arabiat and Y. G. Eljaafreh, "Intrusion detection in wireless sensor networks using ML based classification of denial of service (DoS) attacks," *Journal of Communications*, vol. 20, no. 4, pp. 501–514, 2025.
[51] E. Tapia-Mendez, M. Hernandez-Sandoval, S. Salazar-Colores *et al.*, "A novel deep learning approach for precision agriculture: Quality detection in fruits and vegetables using object detection models," *Agronomy*, vol. 15, no. 6, 1307, 2025.
[52] D. Ribeiro, D. Tavares, E. Tiradentes *et al.*, "Performance evaluation of YOLOv11 and YOLOv12 deep learning architectures for automated detection and classification of immature macauba (acrocomia aculeata) fruits," *Agriculture*, vol. 15, no. 15, 1571, 2025.
[53] Ü. Ş. Ertuğrul and H. Kodaz, "Discrete wavelet transform-based data fusion with ResUNet model for liver tumor segmentation," *Electronics*, vol. 14, no. 13, 2589, 2025.
[54] M. Niksirat, J. Tayyebi, S. F. Javadi *et al.*, "Developing a model to predict the effectiveness of vaccination on mortality caused by COVID-19," *Mathematics*, vol. 13, no. 11, 1816, 2025.