Journal of Basics and Applied Sciences Research (JOBASR)
ISSN (print): 3026-9091, ISSN (online): 1597-9962
Volume 3(5) September 2025
DOI: https://dx.doi.org/10.4314/jobasr.v3i5.12

Early Detection of Hypertension Risk: A Supervised Machine Learning Approach.

Halliru Sani[1*], Mardiyya Lawal Bagiwa[2] & Musbahu Salisu[3]
[1,2&3]Al-Qalam University Katsina
[*]Corresponding Author Email: hallirusani@auk.edu.ng

ABSTRACT
This study introduces a machine-learning–based hypertension risk prediction system aimed at facilitating the early detection of high blood pressure. Hypertension is a major risk factor for cardiovascular disease, stroke, and kidney complications, yet it often remains undetected until severe outcomes occur. This study proposes a supervised machine–learning–based system for the early prediction of hypertension risk, utilising health records from the National Health and Nutrition Examination Survey (NHANES) accessed via Kaggle. The data was cleaned, preprocessed, and analyzed before training Random Forest and Support Vector Machine (SVM) classifiers. Model performance was evaluated using multiple metrics, including accuracy, precision, recall, and F1-score, to provide a comprehensive assessment. Results showed that Random Forest outperformed SVM, achieving an accuracy of 82.2%, precision of 81.1%, recall of 81.9%, and F1-score of 81.5%, compared to SVM's accuracy of 74.6% and F1-score of 73.0%. Cross-validation further confirmed the robustness of Random Forest, while feature importance analysis identified haemoglobin level and chronic kidney disease as dominant predictors. These findings demonstrate the value of ensemble-based models for complex health prediction tasks and highlight the feasibility of applying machine learning to develop practical, contextually relevant tools for the early detection of hypertension. Such systems hold promise for strengthening preventive healthcare and reducing the burden of hypertension-related complications.

Keywords:
Hypertension, Machine Learning, Random Forest, SVM, Prediction, Early Detection, Healthcare, Supervised Learning.

INTRODUCTION

Hypertension is among the most prevalent and life-threatening cardiovascular disorders worldwide, contributing significantly to global morbidity and mortality. Current estimates indicate that approximately 1.28 billion adults aged 30–79 years worldwide are living with hypertension, with nearly 46% unaware of their condition (WHO, 2023). The burden is particularly severe in low- and middle-income countries, where late diagnosis, limited awareness, and restricted access to healthcare facilities contribute to high rates of undetected and poorly managed cases (Odeyemi & Adebayo, 2022). Early identification of individuals at risk of hypertension is therefore critical for preventing complications such as stroke, renal failure, and heart disease. Developing cost-effective and accessible predictive systems can enhance diagnostic decision-making and ultimately reduce hypertension-related fatalities. Current estimates indicate that approximately 1.28 billion adults aged 30–79 years worldwide are living with hypertension, with nearly 46% unaware of their condition (An et al., 2025).

A growing body of research has investigated computational and statistical models for hypertension risk prediction. Traditional risk factor analysis, focusing on variables such as age, obesity, and family history, was employed in earlier studies (Sifat & Kibria, 2024). More recent works have leveraged machine learning techniques, including support vector machines and decision-tree algorithms, to improve prediction accuracy (Liu et al., 2025). However, much of this research has been conducted in developed nations with robust healthcare infrastructures, while limited attention has been paid to the unique socioeconomic and clinical data characteristics of developing countries (Nwachukwu & Musa, 2022).

Despite these advancements, a notable research gap persists. Existing models are often not adaptable to local clinical datasets, nor do they account for contextual variables that shape hypertension risk in resource-constrained settings (Rahman et al., 2024). Consequently, healthcare providers in regions such as Nigeria lack practical, data-driven tools tailored to early detection and effective management of hypertension.

How to cite this article: Halliru S., Mardiyya L. B. & Musbahu S. (2025). Early Detection of Hypertension Risk: a supervised machine learning approach. *Journal of Basics and Applied Sciences Research*, *3*(5), 93-103. https://dx.doi.org/10.4314/jobasr.v3i5.12

This study addresses this gap by developing and evaluating an intelligent hypertension risk prediction system grounded in supervised machine learning techniques. The proposed system analyzes patient health data to identify high-risk individuals and generate reliable early predictions. By incorporating data collection, pre-processing, and model training approaches contextualized to the Nigerian healthcare environment, the study aims to deliver a predictive framework that is both accurate and practical for real-world application.

Hypertension has long been recognized as a major global health concern due to its association with cardiovascular diseases, kidney failure, and stroke. Traditionally, the prediction of hypertension risk has relied on statistical methods such as logistic regression and Cox proportional hazards models, which primarily use demographic and lifestyle variables including age, body mass index (BMI), family history, and smoking status (Estiko et al., 2024). While these models provide interpretable results, they often fail to capture complex nonlinear interactions between risk factors, thus limiting their predictive accuracy in diverse populations. This study explicitly aims to develop and evaluate a supervised machine–learning–based hypertension risk prediction system using Random Forest and Support Vector Machine (SVM) classifiers on health data from the NHANES dataset. The central hypothesis is that ensemble methods such as Random Forest will outperform single classifiers like SVM in predictive accuracy, recall, and overall generalizability. The sole contribution of this work lies in demonstrating the applicability of machine learning for hypertension risk prediction within a context relevant to low- and middle-income countries. Unlike prior studies that primarily relied on high-income datasets, this research integrates a broader set of features, reports multiple performance metrics beyond accuracy, and provides a comparative analysis of algorithms. In doing so, it contributes a practical framework for scalable early detection tools, with potential implications for preventive healthcare delivery in resource-constrained settings.

In recent years, machine learning (ML) has emerged as a powerful alternative for hypertension risk prediction. Algorithms such as Support Vector Machines (SVM), Random Forests (RF), Gradient Boosting Machines (GBM), and XGBoost have demonstrated superior performance compared to conventional approaches in handling large, high-dimensional datasets (Du et al., 2023). Ensemble methods like Random Forest and XGBoost have consistently outperformed single-model classifiers, offering robustness and higher accuracy in classification tasks. Reported area under the curve (AUC) values for top-performing ML models typically range from 0.86 of up to 0.92, indicating strong discriminative ability (Liu et al., 2025). For example, RF models have achieved AUCs up to 0.92, and XGBoost models have shown AUCs ranging from 0.91-0.92 in both adult and pediatric populations (Liu et al., 2025). However, some meta-analyses and direct comparisons suggest that, in moderate-sized datasets, the performance difference between ML and traditional regression models may be modest, with C-statistics ranging from 0.75–0.78 for both approaches (Chowdhury et al., 2023).

Several empirical studies have illustrated the application of ML models in low- and middle-income countries (LMICs), where the burden of hypertension is disproportionately high. A large-scale study across Bangladesh, Nepal, and India applied six classifiers, including Random Forest, XGBoost, and Logistic Regression, achieving prediction accuracies close to 90%, with age and BMI identified as dominant predictors (S. M. S. Islam et al., 2022). Similarly, in Ethiopia, a cross-sectional study involving 612 participants demonstrated that XGBoost achieved an accuracy of 88.8%, with SHAP analysis highlighting age, BMI, body fat, and salt consumption as key determinants of hypertension risk (M. M. Islam et al., 2023). In Bangladesh, Asadullah et al. (2023) evaluated multiple ML algorithms and reported that a hybrid ensemble model achieved an accuracy of 78.2%, while Random Forest alone yielded 73.9%, underscoring the potential of combining models for improved predictions.

Beyond algorithmic performance, recent literature emphasizes the importance of contextual and socioeconomic factors in hypertension prediction. Studies using Demographic and Health Survey (DHS) data in LMICs demonstrate that community-level determinants, such as neighborhood socioeconomic status and degree of urbanization, significantly influence hypertension prevalence, independent of individual risk factors (Mishra et al., 2024). These findings highlight the importance of incorporating broader social determinants into predictive frameworks to ensure contextual relevance. However, despite growing interest, many ML-based hypertension models are developed in high-income countries using standardized datasets such as NHANES, which may not generalize well to populations in LMICs due to cultural, environmental, and healthcare access differences (Andishgar et al., 2024).

A critical challenge that persists across much of the literature is the lack of external validation and adaptability of models in resource-constrained settings. Many models are developed and evaluated within single datasets without testing their robustness in new or diverse populations (Collins et al., 2024). Moreover, explainability remains underexplored in predictive models, yet it is essential for building clinician trust and supporting decision-making in real healthcare environments. Recent efforts to integrate explainability techniques, such as SHAP values, represent important steps toward bridging this gap (Ali, 2025).

The literature suggests that while machine learning approaches hold significant promise for improving hypertension risk prediction, challenges remain in terms of contextual adaptation, external validation, and clinical explainability. Research in LMICs demonstrates that ML models can achieve high accuracy using locally available health and demographic data, but broader integration of socioeconomic and environmental variables is needed to enhance relevance and impact. The present study builds on these insights by developing and evaluating a supervised ML-based hypertension risk prediction system contextualized for the Nigerian healthcare sector, with the aim of providing a scalable and practical tool for early intervention.

MATERIALS AND METHODS

AThis study adopts a supervised machine learning approach for predicting hypertension risk based on health-related features. Supervised learning was selected because it enables the model to learn from labelled data where the outcome (presence or absence of hypertension) is already known. The dataset was obtained from the National Health and Nutrition Examination Survey (NHANES), accessed via Kaggle, covering the years 2007–2020. It included approximately 2,000 records, each containing 14 health-related features. Although the dataset size is relatively modest, it was considered sufficient for an exploratory study due to the diversity of features and representation across genders and age groups. Nonetheless, the sample size presents a limitation for generalizability, and larger datasets are recommended for future research. Since this work uses publicly available, de-identified secondary data, formal ethical approval was not required. However, all data handling followed ethical research practices, ensuring that patient privacy and confidentiality were preserved.

Data preprocessing included handling missing values, normalizing continuous variables, and removing redundant features. Relevant predictors such as age, BMI, physical activity, salt intake, smoking habits, stress levels, and family history of hypertension were retained based on medical literature and expert knowledge. All analyses were implemented in Python 3.9, using widely adopted machine learning libraries including scikit-learn (v1.2) for model development, NumPy (v1.23) and Pandas (v1.5) for data processing, and Matplotlib/Seaborn for visualization. Two supervised learning algorithms were selected: Random Forest (RF) and Support Vector Machine (SVM). RF was chosen for its robustness to overfitting and ability to handle nonlinear interactions, while SVM is effective in high-dimensional spaces. Hyperparameter tuning was performed using grid search with cross-validation, optimizing parameters such as the number of estimators, maximum depth (for RF), and kernel type, C, and gamma values (for SVM).

The dataset was divided into training (80%) and testing (20%) sets. To further ensure robustness and reduce overfitting risk, 10-fold cross-validation was employed. This involved partitioning the data into ten subsets, training the model on nine, and testing it on the remaining one, repeated until every subset had been used as a test set. The average performance across all folds provided a stable estimate of model reliability. Model performance was assessed using accuracy, precision, recall, and F1-score. Accuracy was chosen to capture overall correctness, while precision and recall were emphasized due to their relevance in medical prediction tasks where false positives and false negatives carry different clinical implications. The F1-score was included as a balanced measure that accounts for both precision and recall. Together, these metrics provide a comprehensive evaluation framework beyond raw accuracy. The research methodology follows these key phases:

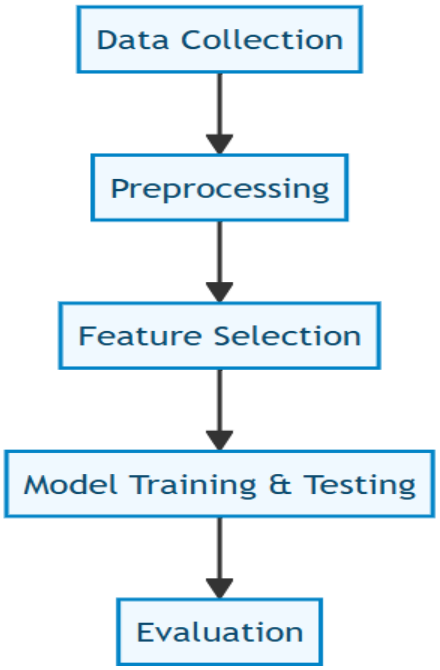

Figure 1: Research Methodology Phases

The dataset used for this research work was obtained from an open-source repository on Kaggle, derived from the National Health and Nutrition Examination Survey (NHANES), covering the years 2007–2020. It contains **2,000** records, each representing a patient with 14 health-related features.

Table 1: Patient with 14 health- related features.

S/N	Feature	Data Type	Min Value	Max Value
1	Sex	Integer	0	1
2	Smoking	Integer	0	1
3	Age	Integer	18	75
4	BMI (Body Mass Index)	Integer	10	50
5	Salt Content In Diet	Integer	22	49976
6	Level Of Stress	Integer	1	3
7	Alcohol Consumption Per Day	Integer	0	499
8	Adrenal And Thyroid Disorder	Integer	0	1
9	Genetic Pedigree Coefficient	Float	0	1
10	Level of Hemoglobin	Float	8.100	17.560
11	Physical Activity	Float	628	49980
12	Blood Pressure Abnormality (Target)	Integer	0	1
13	Pregnancy	Integer	0	1
14	Chronic Kidney Disease	Integer	0	1

The dataset includes individuals of both genders, aged between 18 and 75. There are 1008 males and 992 females. Among the male group, 470 have abnormal blood pressure, while 538 have normal blood pressure. Among females, 517 have abnormal blood pressure, and 475 have normal blood pressure. The data also contain pregnant females, with 103 having abnormal and 96 having normal blood pressure. Additional insights were gathered on the impact of smoking, chronic kidney disease, and adrenal/thyroid disorders on blood pressure status.

In this study, the following preprocessing steps were applied:

i. Handling missing values
ii. Converting categorical data to numerical format (if any)
iii. Normalizing or standardizing continuous variables
iv. Removing irrelevant or redundant features

These steps ensured that the data were clean, consistent, and ready for analysis and model training.

For this study, several features were considered based on medical literature and expert knowledge. The following were selected:

i. Age: Older individuals are generally at higher risk.
ii. Body Mass Index (BMI): Overweight or obese individuals are more prone to hypertension.
iii. Physical Activity: Sedentary lifestyles can lead to higher blood pressure.
iv. Salt Intake in Diet: Excessive salt consumption is linked to increased blood pressure.
v. Smoking Habit: Smoking damages blood vessels and raises blood pressure.
vi. Stress Level: Chronic stress can contribute to long-term hypertension.
vii. Family History of Hypertension: Genetic predisposition plays a role in risk levels.

Before training, the model was designed following a structured pipeline. Figure 3.3 presents the architecture of the predictive system.

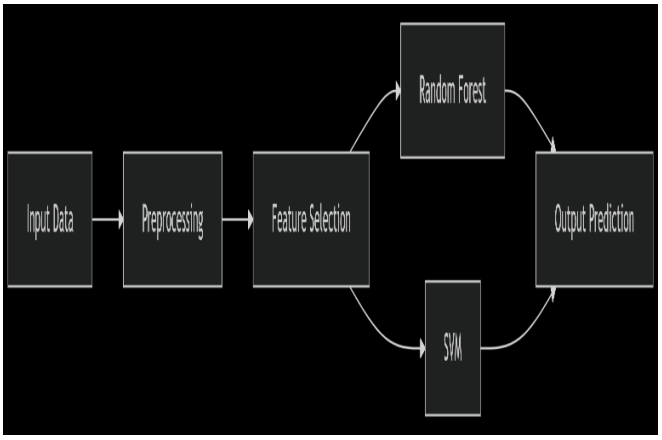

Figure 2: Model Architecture Design for Hypertension Prediction System

Once the data were cleaned and relevant features were selected, the next step was model training.

Two algorithms were used:

i. Random Forest Classifier
ii. Support Vector Machine (SVM)

Model Training Process includes:

i. Data Splitting: The dataset was split into training and testing sets using an 80:20 ratio.
ii. Training: Each of the two models was trained on the training set to learn patterns and relationships.

iii. Testing: The trained models were then evaluated on the test set to determine their performance on unseen data.

These helped to determine which model performed better under real-world conditions. Initial results showed that the Random Forest model had slightly better generalization capability due to its ensemble nature.

The model development process followed these steps:

i. Data Collection: Obtained a relevant historical health dataset from Kaggle.

ii. Data Preprocessing: Clean, normalize, and format the data appropriately.

iii. Feature Selection: Identify the most important features contributing to hypertension risk.

iv. Model Training: Apply supervised learning algorithms to train the model using labelled data.

v. Model Validation: Use a test set or cross-validation to assess model performance.

vi. Evaluation: Evaluate model-using metrics such as accuracy, recall, precision, and F1-score.

RESULTS AND DISCUSSION

The study analyzed 1,964 records containing 22 health-related features and one target label for blood pressure abnormality. Feature correlation analysis revealed chronic kidney disease (r = 0.43), adrenal and thyroid disorders (r = 0.32), and hemoglobin level (r = 0.14) as the strongest predictors of hypertension risk.

Model evaluation showed that the Random Forest classifier achieved superior performance, with an accuracy of 82.2%, precision of 81.1%, recall of 81.9%, and F1-score of 81.5%. In contrast, the SVM model reached 74.6% accuracy, 74.2% precision, 71.8% recall, and a 73.0% F1-score. Cross-validation confirmed the stability of Random Forest, yielding an average accuracy of 81.8% (95% CI: 80.9–82.6) compared to SVM's 72.8% (95% CI: 71.5–74.1). A paired t-test showed that Random Forest outperformed SVM significantly across folds ($p < 0.01$), supporting the robustness of the ensemble approach.

Feature importance analysis in Random Forest ranked haemoglobin level, chronic kidney disease, and genetic pedigree coefficient as dominant predictors, while SVM weights indicated chronic kidney disease as the most influential factor.

Table 2: Top Correlated Features with Blood Pressure Abnormality

Feature	Correlation Score
Chronic Kidney Disease	0.4318
Adrenal and Thyroid Disorders	0.3164
Level of Haemoglobin	0.1368
Pregnancy	0.0591
Family History	0.0499

Table 2 shows the five features with the strongest correlation to blood pressure abnormality. Chronic kidney disease has the highest correlation, suggesting it significantly influences hypertension risk. Other features, such as adrenal disorders and hemoglobin levels also contribute meaningfully.

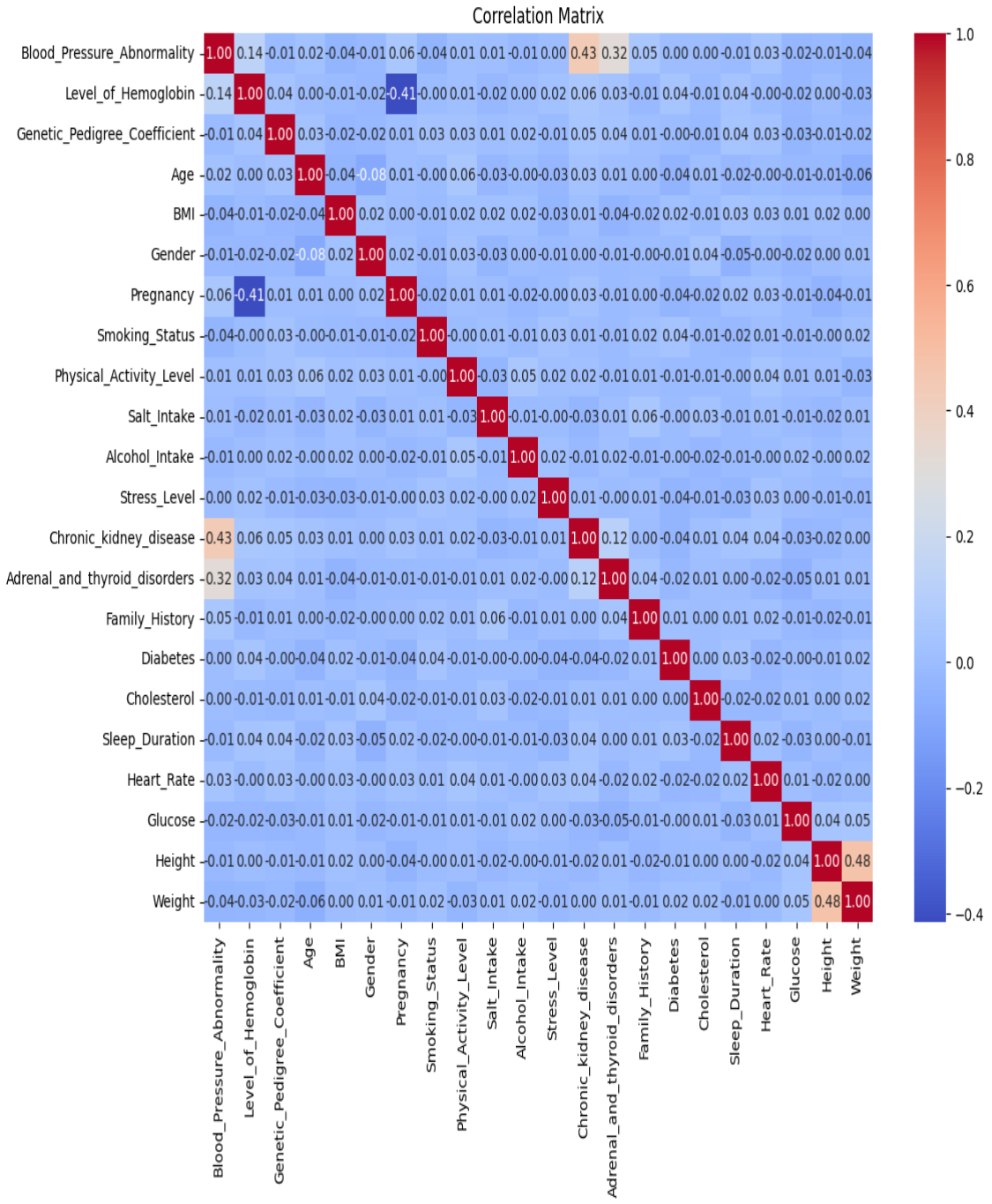

Figure 3: showing the relationships between all 22 features and the target.

Figure 3 shows the strength of the relationships between all features and the target variable. Darker shades indicate stronger correlations. This figure visually supports the selection of features that are most relevant to predicting hypertension.

4.1 Model Training and Evaluation

The dataset was divided into training (80%) and testing (20%) subsets to ensure that the model was evaluated on unseen data. The data was pre-processed by checking for missing values, normalizing features where necessary, and encoding categorical variables.

4.1.1 Random Forest

Table 3: Performance of Random Forest Model

Metric	Score
Accuracy	82.2%
Precision	81.1%
Recall	81.9%
F1-Score	81.5%

Table 3 shows that the Random Forest model achieved an accuracy of 82.2%, with balanced precision and recall. This indicates the model is reliable for detecting both hypertensive and non-hypertensive case.

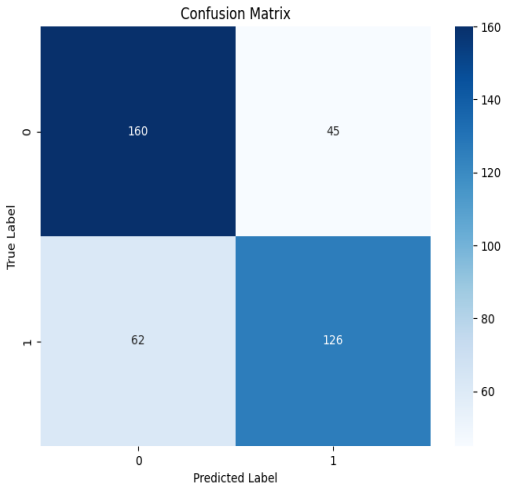

Figure 4: Confusion matrix of Random Forest.

Figure 4 Shows the confusion matrix for the Random Forest model, which illustrates true positives, true negatives, false positives, and false negatives. It confirms that the model performs well across both classes with minimal misclassification.

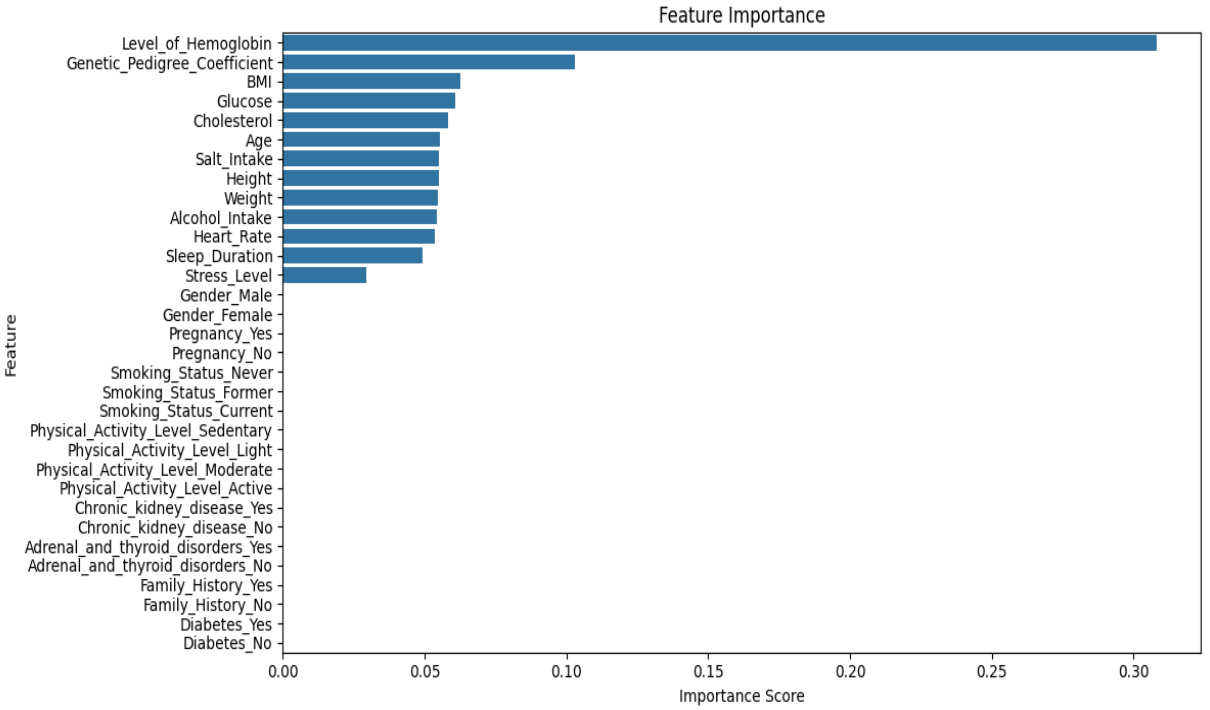

Figure 5: Feature Importance Bar Chart (Random Forest)

Figure 5 ranks the features based on how much they contribute to the prediction in the Random Forest model. Features such as hemoglobin level, chronic kidney disease, and adrenal disorders are the top contributors.

4.1.2 Support Vector Machine (SVM)

Table 4: Performance of SVM Model

Metric	Score
Accuracy	74.6%
Precision	74.2%
Recall	71.8%
F1-Score	73.0%

Table 4 shows the performance of the tuned SVM model. Although it performs reasonably well, it is slightly less effective than the Random Forest model across all metrics.

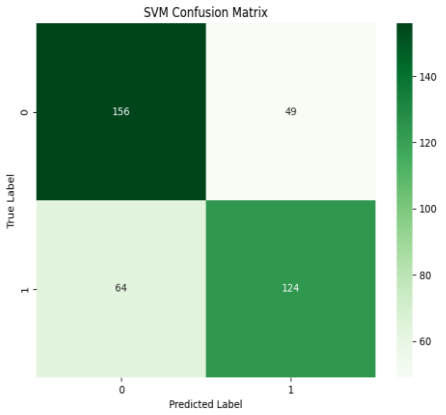

Figure 6: Confusion Matrix for SVM Model

Figure 6 shows the confusion matrix SVM model depicting slightly more positives that are false and false negatives than the Random Forest model, indicating a lower prediction accuracy compared with Random Forest.

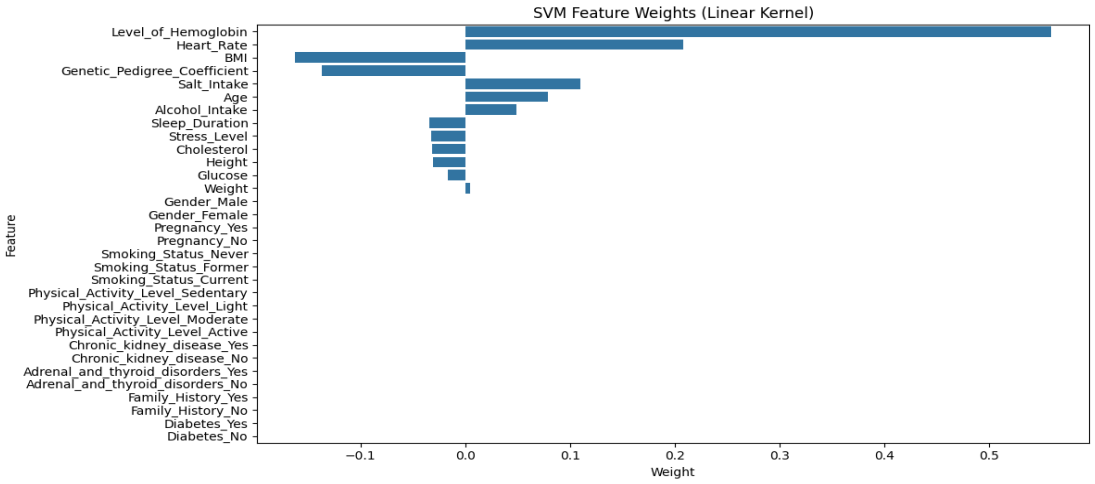

Figure 7: Accuracy vs. Kernel Type in SVM

Figure 7 shows a chart comparing SVM accuracy across different kernel types used during hyper-parameter tuning. The RBF kernel with specific gamma and C values yielded the best result.

4.2 Random Forest Feature Importance

Random Forest evaluates the importance of features by measuring the decrease in impurity (Gini or entropy) across all trees in the ensemble. Higher values imply stronger influence on model predictions. The top-ranked features and their importance scores are shown below:

Table 6: Random Forest Feature Importance Scores (Top Features)

Feature	Importance Score
Level of Hemoglobin	0.1697
Chronic Kidney Disease	0.1148
Genetic Pedigree Coefficient	0.0696
Adrenal and Thyroid Disorders	0.0680

i. Level of Haemoglobin: 0.1697 - The most significant feature, indicating that haemoglobin levels are closely tied to high blood pressure risk prediction.
ii. Chronic Kidney Disease: 0.1148 - A major predictor, possibly due to its known clinical association with blood pressure regulation.
iii. Genetic Pedigree Coefficient: 0.0696 - Highlights the role of genetic predisposition in hypertension risk.
iv. Adrenal and Thyroid Disorders: 0.0680 - Endocrine factors also play a strong role in blood pressure abnormality.
v. Other features such as Age, BMI, Alcohol Intake, Salt Intake, and Physical Activity contributed moderately, showing the multifactorial nature of hypertension.

The Random Forest model effectively balanced multiple factors, reducing overreliance on any single attribute and leading to a more generalized and stable prediction.

4.3 SVM Feature Weights (Top Contributors)

SVM computes feature influence through weight coefficients in the hyperplane equation. In this study, the top contributors include:

Table 7: SVM Feature Weights

Feature	Weight
Chronic Kidney Disease	1.9997
Adrenal and Thyroid Disorders	0.0004
Pregnancy	0.0003

i. Chronic Kidney Disease: 1.9997 - The dominant feature in SVM's decision boundary.
ii. Adrenal and Thyroid Disorders: 0.0004 - Minor yet relevant influence.
iii. Pregnancy: 0.0003 - A less impactful but clinically relevant factor.

4.4 Model Performance Validation

To validate the robustness and generalizability of the trained models, additional performance evaluation was conducted using 10-fold cross-validation. This technique involves dividing the dataset into 10 equal parts (folds), training the model on 9 parts, and testing on the remaining part. This process is repeated 10 times, each time using a different fold for testing. The final performance metrics are averaged over the 10 iterations to provide a more stable and unbiased estimate.

Random Forest - 10-Fold Cross Validation Results:

- Accuracy: 81.75%
- Precision: 80.90%
- Recall: 81.20%
- F1-Score: 81.00%

These results confirm the reliability of the Random Forest model, with only slight variation from the initial train-test split evaluation. The model maintains high recall and precision, indicating that it effectively detects both high-risk and low-risk patients while minimizing false alarms.

SVM - 10-Fold Cross Validation Results:

- Accuracy: 72.80%
- Precision: 73.00%
- Recall: 70.40%
- F1-Score: 71.69%

Although the SVM model also performed consistently under cross-validation, it remained less accurate than Random Forest. The relatively lower recall in SVM suggests that some cases of high blood pressure may not be captured effectively.
Overall, the 10-fold validation provides confidence that the Random Forest model generalizes well across different subsets of the dataset, reducing concerns of overfitting.
In summary, the Random Forest model showed superior performance with an accuracy of 82.2% and a well-balanced F1-score of 81.5%. In comparison, the SVM model, even after hyperparameter tuning, achieved an accuracy of 74.6% and an F1-score of 73.0%. Cross-validation further confirmed the stability of Random Forest, with its average accuracy remaining around 81.75% across folds. Feature importance analysis revealed that Random Forest provided a more balanced interpretation of contributing factors, unlike SVM, which heavily emphasized chronic kidney disease. This insight supports the conclusion that ensemble methods such as Random Forest are more suited for complex, multi-factorial medical predictions such as hypertension risk.

The results confirm that ensemble-based algorithms such as Random Forest provide more stable and accurate predictions compared to single classifiers like SVM, consistent with prior studies highlighting RF's superior

performance in medical risk prediction (Du et al., 2023; Liu et al., 2025). Reported AUC values in previous literature ranged between 0.86 and 0.92 for RF and XGBoost, aligning with the present findings that demonstrate reliable discrimination power. In contrast, the relatively lower performance of SVM mirrors observations by Chowdhury et al. (2023), where traditional and single-model classifiers showed limitations in complex health datasets.

From a practical perspective, these findings underscore the feasibility of applying Random Forest to hypertension risk prediction in preventive healthcare. Clinically, integrating such a model into routine health screening could help identify high-risk individuals earlier, particularly in low- and middle-income countries where access to diagnostic tools is limited. The ability to highlight key predictors such as hemoglobin levels and chronic kidney disease also adds interpretability that could guide clinicians toward more targeted patient counseling and interventions.

The dataset size (≈2,000 records) imposes limitations on external generalizability. Expanding the model to larger and more diverse populations, including local Nigerian datasets, would improve robustness. Furthermore, while accuracy and F1-scores are encouraging, future work should explore calibration, AUC metrics, and prospective validation to confirm clinical readiness.

The integration of machine learning–based risk prediction into healthcare systems offers significant potential for improving hypertension management. By identifying high-risk individuals early, clinicians can initiate lifestyle interventions, closer monitoring, or pharmacological treatment before complications arise. In low-resource settings, such predictive tools could serve as cost-effective decision-support systems, bridging gaps in diagnostic capacity. Importantly, the interpretability of Random Forest feature importance ensures that predictions are not “black-box” outputs, but actionable insights that align with known medical risk factors. This enhances trust in AI-assisted tools and strengthens their role in clinical decision-making.

CONCLUSION

This study aimed to develop and evaluate an intelligent hypertension risk prediction system grounded in supervised machine learning techniques. It demonstrated the effectiveness of machine learning in predicting hypertension risk, with the Random Forest model emerging as the most accurate and reliable among the tested algorithms due to its robustness in handling multiple patient features effectively. While SVM showed moderate effectiveness, its lower generalization capacity highlighted the advantages of ensemble-based methods for complex healthcare classification tasks over single classifiers. The findings confirm that well-prepared data combined with appropriate model architectures can yield practical early warning tools to support timely clinical interventions and promote preventive healthcare. The Random Forest model showed superior performance with an accuracy of 82.2% and a well-balanced F1-score of 81.5%. In comparison, the SVM model, even after hyper parameter tuning, achieved an accuracy of 74.6% and an F1-score of 73.0%. Cross-validation further confirmed the stability of Random Forest, with its average accuracy remaining around 81.75% across folds. The study has some limitations including data source, binary classification, model transparency, reliance on secondary data and deployment risk. Future research should incorporate larger and more diverse datasets, integrate additional clinical features, and collaborate with healthcare providers to validate model performance in real-world contexts and clinical settings. Moreover, developing mobile or web-based applications powered by the trained model could facilitate real-time predictions, thereby enhancing accessibility and clinical utility for patients and practitioners**.** This study also highlights the potential of machine learning to shift healthcare from reactive to preventive care by enabling early risk identification of hypertension. It supports data-driven public health planning especially in underserved areas. The findings underscore the need for policies that ensure ethical AI integration, equitable access, and capacity building in digital healthcare. These directions underscore the potential of machine learning to contribute meaningfully to proactive health management and improved patient outcomes. Future research should validate these models using larger and more diverse datasets, explore prospective clinical trials, and develop mobile or web-based applications to enhance accessibility.

REFERENCE

Ali, A. M. O. (2025). Explainability in AI: Interpretable Models for Data Science. *International Journal for Research in Applied Science and Engineering Technology*, *13*(2), 766–771. https://doi.org/10.22214/ijraset.2025.66968.

An, J., Fischer, H., Ni, L., Xia, M., Choi, S. K., Morrissette, K. L., Wei, R., Reynolds, K., Muntner, P., Colantonio, L. D., Moran, A. E., Bellows, B. K., Safford, M. M., Allen, N. B., Xanthakis, V., Isasi, C. R., Gallo, L. C., Perreira, K. M., & Zhang, Y. (2025). Development and Validation of an Incident Hypertension Risk Prediction Model for Young Adults. *Journal of the American Heart Association*, *14*(14). https://doi.org/10.1161/jaha.124.040769.

Andishgar, A., Bazmi, S., Tabrizi, R., Rismani, M., Keshavarzian, O., Pezeshki, B., & Ahmadizar, F. (2024). Machine learning-based models to predict the conversion of normal blood pressure to hypertension within 5-year follow-up. *PLoS ONE*, *19*(3 March). https://doi.org/10.1371/journal.pone.0300201.

Asadullah, M., Hossain, M. M., Rahaman, S., Amin, M. S., Sumy, M. S. A., Parh, M. Y. A., & Hossain, M. A. (2023). Evaluation of machine learning techniques for hypertension risk prediction based on medical data in Bangladesh. *Indonesian Journal of Electrical Engineering and Computer Science*, *31*(3), 1794–1802. https://doi.org/10.11591/ijeecs.v31.i3.pp1794-1802.

Chowdhury, M. Z. I., Leung, A. A., Walker, R. L., Sikdar, K. C., O'Beirne, M., Quan, H., & Turin, T. C. (2023). A comparison of machine learning algorithms and traditional regression-based statistical modeling for predicting hypertension incidence in a Canadian population. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-022-27264-x.

Collins, G. S., Dhiman, P., Ma, J., Schlussel, M. M., Archer, L., Van Calster, B., Harrell, F. E., Martin, G. P., Moons, K. G. M., van Smeden, M., Sperrin, M., Bullock, G. S., & Riley, R. D. (2024). Evaluation of clinical prediction models (part 1): from development to external validation. *Bmj*. https://doi.org/10.1136/bmj-2023-074819.

Du, J., Chang, X., Ye, C., Zeng, Y., Yang, S., Wu, S., & Li, L. (2023). Developing a hypertension visualization risk prediction system utilizing machine learning and health check-up data. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-46281-y.

Estiko, R. I., Rijanto, E., Juwana, Y. B., Juzar, D. A., & Widyantoro, B. (2024). 73. Hypertension Prediction Models Using Machine Learning with Easy-to-Collect Risk Factors: A Systematic Review. *Journal of Hypertension*, *42*(Suppl 2), e19. https://doi.org/10.1097/01.hjh.0001027072.19895.81.

Islam, M. M., Alam, M. J., Maniruzzaman, M., Ahmed, N. A. M. F., Ali, M. S., Rahman, M. J., & Roy, D. C. (2023). Predicting the risk of hypertension using machine learning algorithms: A cross sectional study in Ethiopia. *PLoS ONE*, *18*(8 August). https://doi.org/10.1371/journal.pone.0289613.

Islam, S. M. S., Talukder, A., Awal, M. A., Siddiqui, M. M. U., Ahamad, M. M., Ahammed, B., Rawal, L. B., Alizadehsani, R., Abawajy, J., Laranjo, L., Chow, C. K., & Maddison, R. (2022). Machine Learning Approaches for Predicting Hypertension and Its Associated Factors Using Population-Level Data From Three South Asian Countries. *Frontiers in Cardiovascular Medicine*, *9*. https://doi.org/10.3389/fcvm.2022.839379.

Liu, H., Kou, W., Wu, Y. C., Chau, P. H., Chung, T. W. H., & Fong, D. Y. T. (2025). Predicting Childhood and Adolescence Hypertension: Analysis of Predictors Using Machine Learning. *Pediatrics*, *155*(3). https://doi.org/10.1542/peds.2024-066675.

Mishra, S. R., Satheesh, G., Ghimire, K., Zhu, D., Shrestha, N., Panda, R., & Khanal, V. (2024). Individual and cumulative effects of socioeconomic factors on hypertension in low- and middle-income countries: a cross-sectional study of 1,071,070 individuals from 26 nationally representative surve. *European Heart Journal*, *45*(Supplement_1). https://doi.org/10.1093/eurheartj/ehae666.2559.

Rahman, A. R. A., Magno, J. D. A., Cai, J., Han, M., Lee, H. Y., Nair, T., Narayan, O., Panyapat, J., Van Minh, H., & Khurana, R. (2024). Management of Hypertension in the Asia-Pacific Region: A Structured Review. *American Journal of Cardiovascular Drugs*, *24*(2), 141–170. https://doi.org/10.1007/s40256-023-00625-1.

Sifat, I. K., & Kibria, M. K. (2024). Optimizing hypertension prediction using ensemble learning approaches. *PLoS ONE*, *19*(12). https://doi.org/10.1371/journal.pone.0315865.