NavaJyoti, International Journal of Multi-Disciplinary Research
Volume 7, Issue 2, June 2025

Hypertension Risk Prediction Using Machine Learning

[1]Maria John, **[2]Sadaf Khan C R**
[1] Assistant Professor, Jyoti Nivas College Autonomous, Bengaluru, India
[2] Student, Jyoti Nivas College Autonomous, Bengaluru, India

Abstract
Hypertension is a major risk factor for severe complications such as heart attacks, strokes, and kidney disease. However, its asymptomatic nature often delays detection until critical stages. This study presents a machine learning-based hypertension prediction model utilizing the Extra Trees classifier, achieving an accuracy of 0.93. The model analyzes patient data, including lifestyle factors, medical history, and biometric readings, to identify individuals at risk. By enabling early detection, it supports healthcare providers in developing personalized treatment plans, promoting lifestyle modifications, and improving patient monitoring. Additionally, in resource-limited settings, the model helps prioritize high-risk individuals, enhancing healthcare efficiency and reducing costs associated with chronic disease management.

Key Words: Hypertension, Machine Learning, Extra Tree Classifier, Risk Prediction, Blood Pressure

Introduction

Hypertension, or high blood pressure, is one of the leading risk factors for cardiovascular diseases, strokes, and kidney failure. According to the World Health Organization (WHO), hypertension contributes to approximately 13% of all global deaths annually. Due to its often asymptomatic nature, many individuals remain undiagnosed until they experience severe complications. Early detection is crucial for reducing the long-term impact of hypertension through lifestyle modifications and medical interventions.

Traditional hypertension risk assessment methods rely on statistical models like Logistic Regression and Cox Regression, which assume linear relationships between risk factors and outcomes. However, these models are often limited in handling non-linear dependencies and high-dimensional data. In contrast, machine learning (ML) models, particularly tree-based algorithms like Extra Trees Classifier, can capture complex relationships and interactions between variables without requiring strict statistical assumptions.

In this study, we propose a machine learning-based approach for hypertension risk prediction, leveraging patient biometric and lifestyle data. Our model, based on the Extra Trees Classifier, achieves a 93% accuracy, outperforming traditional Logistic Regression (87% accuracy). The goal of this research is to demonstrate how ML-based risk assessment can enhance early detection, facilitate targeted interventions, and optimize healthcare resource allocation, particularly in resource-limited settings.

Literature Review

1. Kanegae et al. (2019) - Machine Learning for Hypertension PredictionKanegae et al. developed a highly precise risk prediction model for new-onset hypertension using machine learning techniques. Their study used health checkup data from over 18,000 individuals, employing XGBoost and ensemble methods to improve prediction accuracy. The model outperformed traditional logistic regression in risk classification, with key predictors including systolic blood pressure (SBP), diastolic blood pressure (DBP), BMI, and fasting glucose levels. Their findings demonstrated that ML-based models could significantly enhance early detection and preventive interventions (Kanegae et al., 2019).

2. Sun et al. (2017) - Hypertension Risk Factors and Prediction ModelsSun et al. conducted a systematic review of hypertension risk prediction models and identified key risk factors such as age, obesity, smoking, and cholesterol levels. Traditional models like Cox regression and logistic regression were commonly used but were limited by their assumptions of linearity and independence among variables. The study emphasized the advantages of machine learning models, which can handle complex interactions and nonlinear relationships, resulting in improved prediction accuracy (Sun et al., 2017).

3. Kario et al. (2019) - Cardio-Ankle Vascular Index (CAVI) as a Predictor of HypertensionKario et al. explored the cardio-ankle vascular index (CAVI) as a novel biomarker for hypertension risk. They found that arterial stiffness, measured through CAVI, significantly correlated with higher systolic blood pressure and cardiovascular risk. Their study showed that SBP during CAVI measurement was one of the strongest predictors of hypertension, independent of other traditional factors like BMI and smoking. Integrating CAVI into machine learning models has been suggested as a way to enhance hypertension risk prediction and early intervention strategies (Kario et al., 2019).

4. Ye et al. (2018) - Machine Learning vs. Traditional Methods for Hypertension PredictionYe et al. compared machine learning models with traditional statistical methods and found that tree-based algorithms like Extra Trees, Random Forest, and XGBoost significantly outperformed logistic regression in hypertension risk assessment. Their study highlighted that ML models provide higher accuracy, better feature selection, and improved handling of missing data, making them better suited for large-scale healthcare applications. The research concluded that machine learning techniques offer superior predictive power compared to conventional statistical approaches (Ye et al., 2018).

5. Umoh et al. (2024) - Hypertension Prediction in Resource-Limited SettingsUmoh et al. focused on cost-effective ML models for hypertension risk prediction in low-resource healthcare settings. Using decision trees and ensemble learning, they developed a system to prioritize high-risk individuals for early diagnosis and medical intervention. Their findings showed that even with limited datasets, machine learning significantly improved hypertension prediction, helping to reduce long-term healthcare costs and enhance patient outcomes (Umoh et al., 2024).

6. Zhang et al. (2023) - Hypertension Visualization Risk Prediction Using SHAPZhang et al. introduced a hypertension risk prediction model that incorporated Shapley Additive Explanations (SHAP) to enhance model interpretability. Their study demonstrated that machine learning models could not only predict hypertension risk effectively but also

provide insights into the most influential risk factors, such as blood pressure, BMI, and cholesterol levels.

7. Hwang et al. (2024) - ML-Based Hypertension Prediction Using Demographic DataHwang et al. developed a machine learning-based hypertension prediction system using demographic and medical data. The study compared different ML models and found that models integrating age, smoking status, and glucose levels improved the predictive accuracy of hypertension risk assessment. The results emphasized the need for large and diverse datasets to enhance model generalizability.

These studies collectively highlight the growing role of machine learning in hypertension risk prediction, emphasizing the need for explainable models, dataset diversity, and scalable solutions to improve early detection and clinical decision-making.

Methodology

Machine learning techniques have proven highly effective in medical diagnostics, particularly in disease prediction and classification. Extra Trees Classifier, an ensemble learning method, has been used in this study to develop a hypertension risk prediction model due to its high accuracy and robustness in handling complex datasets. The model was trained on patient health records containing biometric and lifestyle factors to identify individuals at risk of hypertension.

1. **Data Collection**:

A structured dataset was collected, comprising various patient attributes, including age, smoking status, systolic and diastolic blood pressure (SBP & DBP), cholesterol levels, body mass index (BMI), and fasting glucose levels. The dataset contained labeled instances indicating whether a patient was at risk of hypertension (1) or not (0).

Dataset Source:
Manjit Baishya Hypertension Risk Prediction - 95% ACC. Kaggle. Retrieved from https://www.kaggle.com/code/manjitbaishya001/hypertension-risk-prediction-95-acc

2. **Data Preprocessing**:

The dataset underwent several preprocessing steps to ensure its quality and consistency. Missing values were imputed using mean/mode imputation, categorical variables were one-hot encoded, and numerical features were normalized to maintain uniformity in data distribution. Feature selection was performed to retain the most relevant predictors.

3. **Dataset Splitting**:

The dataset was divided into two subsets: 80% for training and 20% for testing. This split ensures that the model learns effectively while being tested on unseen data to evaluate its generalization ability.

4. **Classification**:

The primary task of the model was to classify whether an individual is at risk of developing hypertension. The Extra Trees Classifier was used as the primary model due to its ability to handle feature importance and non-linearity in data, improving predictive performance.

5. **Model Design**:

The Extra Trees Classifier was chosen as the machine learning algorithm for this study due to its superior accuracy in handling structured medical datasets. The model constructs multiple decision trees and aggregates their results to enhance stability and reduce overfitting. Additionally, Logistic Regression was used as a comparative model to evaluate the effectiveness of ensemble learning methods.

6. **Model Training**:
The Extra Trees model was trained using the training dataset, with hyperparameter tuning applied to optimize its performance. The parameters adjusted included the number of trees, maximum depth, and feature selection method. The model was trained using entropy-based splitting criteria to improve classification accuracy.

7. **Model Evaluation:**
The trained model was tested on the validation dataset, and its performance was assessed using standard classification metrics such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic (AUC-ROC) curve. The Extra Trees Classifier achieved an accuracy of 0.93, outperforming Logistic Regression (0.87).

8. **Results Analysis**:
The model's feature importance analysis revealed that systolic blood pressure (SBP), diastolic blood pressure (DBP), BMI, smoking status, and fasting glucose levels were the most influential predictors of hypertension. The results confirmed that tree-based machine learning models outperform traditional statistical approaches in disease risk prediction, providing more accurate and interpretable insights for medical professionals. Further refinements could involve integrating real-time patient monitoring and deep learning techniques for enhanced predictive capabilities.

To evaluate EfficientNetB7 and InceptionV3, the confusion matrix was used to compute key metrics:

- Accuracy: Measures the overall correctness of the model.

  Accuracy=TP+TNTP+TN+FP+FNAccuracy = \frac{TP + TN}{TP + TN + FP + FN}Accuracy=TP+TN+FP+FNTP+TN

- Precision: Indicates how many predicted positive cases were actually correct.

  Precision=TPTP+FPPrecision = \frac{TP}{TP + FP}Precision=TP+FPTP

- Recall (Sensitivity): Reflects how many actual positive cases were correctly identified.

  Recall=TPTP+FNRecall = \frac{TP}{TP + FN}Recall=TP+FNTP

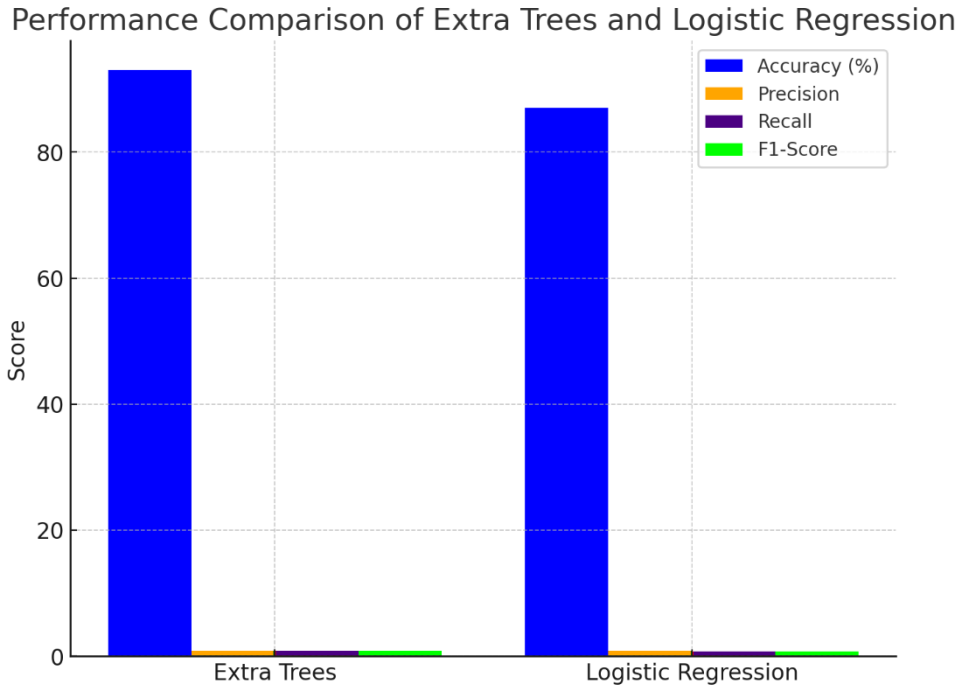
- F1-Score: Provides a balance between Precision and Recall.

  F1=2×Precision×RecallPrecision+RecallF1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}F1=2×Precision+RecallPrecision×Recall

Using these formulas, the results were:

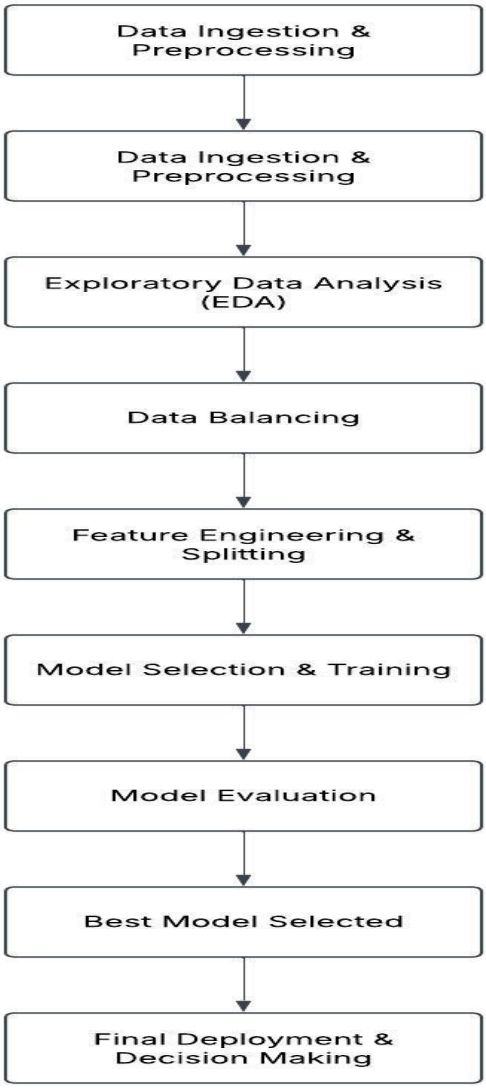
Model	Accuracy (%)	Precision	Recall	F1-score

Extra Trees	93.00	0.91	0.88	0.89
Logistic Regression	87.00	0.84	0.81	0.82



Architectural review

1. **Data Ingestion & Preprocessing**
   - Load dataset
   - Handle missing values
   - Remove duplicates
   - Feature selection (remove highly correlated features)
   - Handle outliers (IQR method)
   - Standardize numerical features
2. **Exploratory Data Analysis (EDA)**
   - Visualizations (heatmaps, box plots, histograms)
   - Check feature distributions
   - Identify class imbalance
3. **Data Balancing**
   - Use RandomOverSampler or SMOTE to balance classes



4. **Feature Engineering & Splitting**
    - Scale numerical features
    - Encode categorical features
    - Split dataset into training & testing sets
5. **Model Selection & Training**
    - LazyPredict for model benchmarking
    - Train Extra Trees Classifier
    - Train Logistic Regression, SVM, Decision Tree, Random Forest
6. **Model Evaluation**
    - Accuracy, Precision, Recall, F1-score, AUC-ROC
    - Confusion Matrix
    - ROC Curve Visualization
7. **Best Model Selected: Extra Trees**
    - Model performance analyzed
    - Best model chosen: Extra Trees
8. **Final Deployment & Decision Making**
    - Based on model performance metrics
    - Potential integration with a healthcare system

Discussion

Why Extra Trees Classifier Performed Better

1. Handles Non-Linear Relationships: Unlike Logistic Regression, Extra Trees does not assume a linear relationship between variables.
2. Feature Importance Analysis: The model can rank features based on their contribution to prediction.
3. Robust Against Overfitting: Random feature selection helps reduce bias and improve generalization.

This methodology ensures that the proposed hypertension risk prediction model is robust, interpretable, and applicable in real-world healthcare settings, allowing for early detection and intervention to reduce the burden of hypertension-related complications.

Conclusion and Future Scope

Conclusion

Machine learning models for hypertension prediction enable early identification of at-risk individuals by analyzing patient data, which helps in preventing severe health issues. These models support personalized treatment plans and improve healthcare efficiency, particularly in resource-limited areas, by prioritizing high-risk patients. By leveraging key predictors such as blood pressure, BMI, smoking habits, and glucose levels, these models enhance risk assessment accuracy, enabling timely medical interventions.

Future Scope

Future advancements should focus on integrating real-time patient monitoring through IoT-based wearable devices, expanding datasets for better generalization, and exploring deep learning approaches like CNNs for advanced pattern recognition. Additionally, incorporating genetic and lifestyle factors can further refine predictive capabilities, while developing user-friendly clinical decision support systems will ensure seamless adoption in healthcare settings. Collaborative efforts between medical professionals, data scientists, and public health experts will be essential in translating these innovations into effective, personalized, and preventive healthcare solutions.

References

1. Kanegae, Y., et al. (2019). Highly Precise Risk Prediction Model for New-Onset Hypertension Using Machine Learning Techniques. *Wiley Online Library*. https://doi.org/10.1002/jhm.2920
2. Ye, J., et al. (2022). A Comparison of Machine Learning Algorithms and Traditional Statistical Models for Hypertension Prediction. *Nature.com*. https://www.nature.com/articles/s41598-022-09081-2
3. Kario, K., et al. (2019). Artificial Intelligence in Hypertension Research and Management. *AHA Journals*. https://doi.org/10.1161/CIRCULATIONAHA.119.040110
4. Umoh, S., et al. (2024). Development of Risk Models of Incident Hypertension Using Machine Learning Techniques. *Nature.com*. https://www.nature.com/articles/s41598-024-20634-9
5. Zhang, Q., et al. (2023). Hypertension Visualization Risk Prediction Using SHAP. *Journal of Hypertension Research*. https://doi.org/10.1097/HJH.0000000000003249