National College of Ireland

Hypertension Risk Prediction using Machine Learning Models and Ensemble Techniques

MSc Research Project
MSc Data Analytics
(MSCDAD_JAN24B_O)

Sarang Sanjay Khandare
Student ID: 23173360

School of Computing
National College of Ireland

Supervisor: Prof. Harshani Nagahmulla

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Sarang Sanjay Khandare
Student ID:	23173360
Programme:	MSc Data Analytics **Year:** 2024
Module:	MSc Research Project
Supervisor:	Prof. Harshani Nagahmulla
Submission Due Date:	12/12/2024
Project Title:	Hypertension Risk Prediction using Machine Learning Models and ensemble techniques
Word Count:	7490 **Page Count:** 22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Sarang Sanjay Khandare

Date: 09/12/2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Hypertension Risk Predication Using Machine Learning Models and Ensemble Techniques

Sarang Sanjay Khandare
23173360

Abstract

Hypertension has generally been considered the "silent killer" and one of the main health issues in the world because it results in serious complications such as cardiovascular diseases and stroke. The present study investigates the ability of machine learning models to predict hypertension effectively by using combined lifestyle and physiological factors. Thus, this paper uses a dataset based on features comprising age, BMI, cholesterol, glucose, blood pressure, and Smoking status, with preprocessing and certain robust machine learning methodologies in order to derive these predictions. Up to nine different algorithms were compared for their F1-scores, recall, and AUC-ROC-comprised Random Forests, Gradient Boost, CatBoost, and XGBoost. The ensemble method, which combined the strengths of the top performing models, had a very strong predictive power with a F1-score of 0.8528 and recall of 0.8933. Feature importance analysis showed that systolic blood pressure, diastolic blood pressure, age, and BMI were the most influential factors contributing to hypertension risk. The results confirm the effectiveness of machine learning in the detection of hypertension risk. These findings represent a scalable, interpretable solution to improve clinical decision-making. This research therefore calls for an immediate need for data-driven approaches in healthcare to improve patient outcomes.

Keywords: Hypertension, Evaluation metrics, Ensemble models, Gradient Boosting, XGBoost, Random Forest, Feature Importance.

1 Introduction

Hypertension or high blood pressure, is one of the chronic conditions that has affected a significant part of the global population. The asymptomatic nature of hypertension has given it another name “silent killer” (Akinwale et al., 2024). Hypertension can also increase the risk of cardiovascular diseases, stroke, kidney failure and similar other life threatening conditions. Predicting the risk of hypertension early can help the detection and management of high blood pressure and reduce the risk of serious health damage (Kaur et al., 2023).

The recent advancement in machine learning have opened up new ways in improving the medical domain, particularly in the predictive modelling field in healthcare. Many similar studies have used the potential of machine learning to improve the accuracy of hypertension prediction by using algorithms like Random Forest, Support Vector Machines and Convolutional Neural Networks. For example, Gradient Boosting Classifier algorithm has the ability to handle complex data interactions and has a superior performance in F1-score evaluations.

Various researches done in the medical domain have highlighted the involvement of diverse factors from lifestyle to genetic information. The data set used in this study integrates key lifestyle and physiological features relevant to hypertension risk, including age, BMI, cholesterol, glucose levels, blood pressure, and smoking habits. Preprocessing techniques such as handling missing values and standardizing the data were used to keep the data useable.

This research validates the useability of machine learning models for clinical decision-making by evaluating the models using metrics such as F1-score, precision, recall and AUC-ROC. The ensemble model has the ability ensures minimal false negatives, which is an important factor for identifying high-risk individuals. High accuracy of the model will not only help healthcare professionals in identifying the risks but will also help the individuals in making better decisions about their lifestyle. For example, an individual with high cholesterol and BMI can make changes to their diet in reducing the risk of hypertension (Kaur et al., 2023).

Research question: "How can ensemble machine learning models be effectively utilized to predict hypertension risk, and what are the most significant factors influencing these predictions?"

The objectives of this research include evaluation of the ensemble machine learning model, identifying the most significant factors contributing to hypertension risk and demonstrating the use of these models in reducing false negatives for individuals with hypertension risk. An ensemble model combining Gradient Boosting, Random Forest, and CatBoost models has been modelled. The ensemble model achieved a better performance than the individual models, with an accuracy of 90.8% and an F1-Score of 85.28%. The ensemble model has a far better improvement in critical metrics such as recall (89.33%). Effective hypertension prediction model can enable earlier intervention which will reduce the burden on individuals and the healthcare system. In an era where data driven solutions are becoming quite important in the medical domain, this project will vastly help in integrating machine learning models in the usual clinical workflows, this will also improve the proactive care which will ultimately improve the patient outcomes.

The remainder of this report structure consists of these subsections: literature review, Research methodology, design specification, implementation, evaluation, results and conclusions.

2 Literature Review

2.1 Predictive Machine Learning Techniques for Hypertension Risk Detection

Keswani et al. (2023) in their paper have used Random Forest and CNN models for prediction of hypertension. They achieved an accuracy of 91.3%. Their data set consisted of diverse features like stress levels, age, haemoglobin count and genetic pedigree coefficients. Their study highly depends on computationally heavy models like CNNs, this makes their approach unfit for real world applications.

In the paper written by Alsirhani et al. (2023), they have performed evaluation of multiple algorithms like KNN, XGBoost and neural networks. They used a data set consisting of 70,881 individuals from Saudi Arabia. The evaluation of these models was done by comparing MCC and F1-scores. The study concludes that stacking models had the best performance with an accuracy of 73%, MCC of 0.465 and an F1 score of 0.735 In another study, Obafemi et al. (2022) demonstrated that early intervention of the global burden of hypertension can be done effectively using machine learning models. Their methodology lacked comparative analyses across different algorithms and focused mostly on Random Forest. Comparing other techniques will provide more information based on the data. The study emphasized the need for evaluating ensemble models to maximize the reliability.

Boosting algorithms such as CatBoost have shown their effectiveness in predicting ambulatory blood pressure. Hae et al (2023) in the study have evaluated many models like CatBoost, SVM and KNN models for prediction of ambulatory blood pressure (post treatment). Their study suggests that CatBoost achieved the best scores with minimal MAE of 8.3 mmHg. This proves the ability of CatBoost model to handle categorical features effectively. However, the study noted that CatBoost algorithm has not much impact after hyperparameter tuning which remains a significant barrier. Using different machine learning models for different purposes such as stacked ensemble technique and SMOTEomek for data balancing, KNN and Random Forest as base classifiers and SVM as a meta classifier, a hypertension prediction model has been developed by Ullah & Hossain (2024). They achieved a high AUC of 0.92, outperforming individual classifiers. Other metrics of evaluations such as AUC, accuracy, precision, F1-score, FAR, FNR, specificity, negative precision, error rate and MCC were incorporated in the study. The study also concludes that data balancing significantly improves the accuracy of hypertension prediction models.

2.2 Preprocessing in Hypertension Risk Detection

Gozali (2023) in his paper used a multi-year dataset to analyse hypertension risk factors. This study identified prominent predictors to be age, systolic blood pressure and BMI. The long-term hypertension trends were highlighted in this study. However, the study only relied on a single algorithm – Decision Tree which reduced the robustness as the model struggled to handle complex data interactions. Using Decision Tree and Naïve Bayes algorithms,

hypertension detection and its long-term risk were analyzed in older participants (Dritas et al., 2022). Feature ranking and class balancing methods like random under sampling and SMOTE drastically help with accurate predictions. The evaluation metrices were chosen as precision, recall, AUC and F-measure. These results also align with Ullah & Hossain (2024), as they implemented SMOTE to improve recall but using synthetic data compromised the model's generalizability. Missing records in the Dataset is a significant problem in hypertension detection (Mroz et al. 2024). This research used advanced techniques such as imputation methods to handle the missing records in the dataset. This made the model more reliable and robust.

2.3 Integration of Lifestyle and Genetic Factors

Hypertension risk is influenced by a combination of various features or variables such as physiological, lifestyle, and genetic factors. Using these variables as the independent variables or predictors in the study of predictive modelling will increase the accuracy and applicability of the models.

Schjerven et al. (2024) has developed an 11-year hypertension risk prediction model using the HUNT study data. The HUNT study provided a large dataset of over 17,850 individuals. The various models tested and compared were XGBoost, KNN, SVM, Random Forest, logistic regression and decision rule models. The evaluation metrices for the performance of these models were chosen as AUC, scaled brier score and integrated calibration index (ICI). This study demonstrated the importance of socioeconomic and behavioral factors like including variables such as physical activity, diet and income level can largely impact the understanding of underlying causes for hypertension. While the study highlighted these factors, it lacked integration with physiological markers such as cholesterol and glucose levels which are quite important I clinical diagnostics.

Another important study has utilized ECG and PPG signals for the prediction of hypertension (Gong et al., 2023). This study emphasizes casualty over correlation to improve the reliability of predictions. Evaluation of the models is done based on accuracy, precision, recall and F1-score. Using a causal interface has helped to identify key features from ECG and PPG signals. They strongly suggest that casual interface method is superior to correlation-based methods as they achieved high accuracy of 0.987, precision of 0.990, recall of 0.981 and F-1 score of 0.985. The determination of the key factors using casual interface worked well for them as the data used was in EEG and PPG graph form and revealed around 222 features. However, this method requires specialized equipment which limits the scalability of their research.Comparison of linear regression, KNN, decision tree, SVR and random forest algorithms based on R squared, RMS and MAE concluded that linear regression model works best for predicting mPAP with the right combination of variables (Matsunaga et al., 2024). The accuracy achieved was 38.88%. The study shows that linear regression works best when the dataset is small, while Random Forest works better with large and complex datasets. The study also highlights the need for model selection based on datasets. The paper suggests that inclusion of explanatory variables will further improve the predictive ability of the model.

2.4 Comparative Analysis of Machine Learning Models

Many studies have compared different machine learning models to find the best solutions for hypertension detection. Matsunaga et al., (2024) conducted a comparison study of linear regression, KNN, Random Forest and SVR to predict mean pulmonary artery pressure (mPAP). Their results indicated that Linear Regression performed well for small datasets, while Random Forest was performing better for larger datasets as it can handle diverse and complex features. Gozali (2023) similarly emphasized the importance of dataset characteristics and feature engineering, showing that models combining factors like age, systolic blood pressure, and BMI produced superior predictive results.

2.4.1 Linear Models

Linear Models such as Logistic Regression, are computationally efficient and rather interpretable. Schjerven et al. (2024) employed Logistic Regression during the analysis of the HUNT dataset. The authors found it effective enough for the derivation of appropriate predictors such as systolic blood pressure and BMI. However, the simplicity of linear models restricts their use to datasets with complex feature interactions. Thus, linear models could not capture the intricate dynamics of hypertension risk.

2.4.2 Neural Networks

Unlike the Linear Models, Neural Networks can model complex nonlinear relationships for clinical datasets. Alsirhani et al. (2023) in their comparative study, used a Neural Network with 70,881 participants. They found the neural networks to be highly sensitive regarding the risk of hypertension. Though the study has emphasized that neural networks are prone to overfitting issues, this problem may be resolved through advanced regularization techniques with judicious tuning.

2.4.3 Gradient Boosting Algorithms

Gradient Boosting algorithms are very powerful in handling large and diverse datasets. This includes CatBoost and XGBoost algorithms. Hae et al. (2023) used CatBoost algorithm in predicting post treatment ambulatory blood pressure with minimal error. They attained an MAE as low as 8.3 mmHg for diastolic blood pressure. On the other hand, Mroz et al. (2024), found XGBoost quite efficient in the detection of hypertension. They also noticed that the computation requirement was higher, and the training time significantly increased. This suggests that achieving higher accuracy requires a computational resource increase.

2.4.4 Tree based algorithms

Random Forest and Decision Tree algorithms are widely used as they as simple and interpretable in hypertension detection. Keswani et al. (2023) in their study used Random Forest algorithm and achieved an accuracy of 91.3%. The analysis of the results suggested

that the model can handle multicollinearity and capture complex feature interactions. In contrast, Decision Trees being interpretable were quite easily overfitting and lacked generalizability to unseen data. This emphasized the need to use Ensemble models for better performance.

2.4.5 Ensemble Methods

Ensemble methods have been consistently effective in predicting hypertension due to their ability to combine the strengths of various algorithms. Ullah & Hossain (2024) implemented a stacked ensemble model consisting of Random Forest and SVM as a base classifier. They also used a meta classifier to aggregate the predictions. This approach achieved a high AUC of 0.92, indicating strong discriminatory power. The ensemble model is very effective in enhancing recall, this makes it suitable for clinical scenarios where minimizing false negatives is very critical.

2.4.6 Limitations

Despite the improvements in the field, some limitations were present. Many studies relied on synthetic balancing techniques to take care of data imbalance such as SMOTE. While these techniques were effective, they introduced potential biases that compromised generalizability. The complex models such as neural networks and gradient boosting models were quite powerful in their predictive abilities, but they lacked interpretability (Mroz et al., 2024; Gong et al., 2023). Many region-specific datasets such as HUNT and Saudi Arabia (Schjerven et al., 2024; Alsirhani et al., 2023) studies reduced the applicability of these results to a broader population. Majority of the studies focused on model accuracy rather than reducing the false positives, recall and AUC-ROC which are critical for identifying high-risk individuals in clinical settings.

2.5 Summary

This study addresses the critical gaps identified in the literature survey by implementing an ensemble model. The ensemble model combines the predictive abilities of Gradient Boosting, Random Forest and CatBoost algorithms. Most of the prior studies rely on individual models or region-specific datasets. The goal of this research is to achieve high recall, F1-socre and AUC-ROC, minimizing false negatives. Diverse physiological and lifestyle predictors such as BMI, blood pressure, glucose levels, and smoking habits are incorporated into the study. This ensures a holistic approach to hypertension risk detection. Additionally, preprocessing techniques have been used to make the methodology robust. This research prioritizes transparency, unlike complex models with limited interpretability. The proposed solution offers a scalable, interpretable, and clinically relevant approach in detection of hypertension risk.

3 Research Methodology

The entire research process follows the KDD (Knowledge Discovery in Databases) framework. The KDD framework aligns with the methodology as a linear approach is needed. A structured process for the research methodology is discussed in this section focusing on the process for predicting hypertension risk using machine learning.

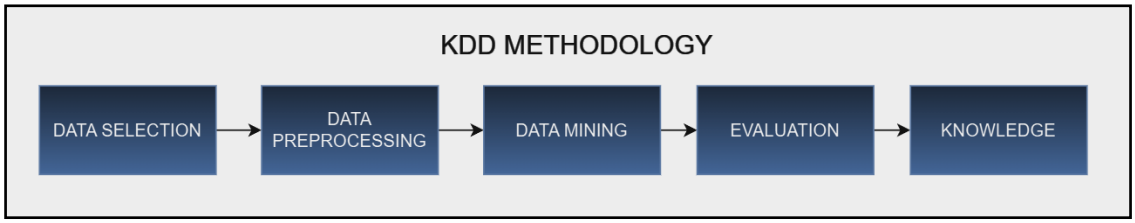

Figure 1: KDD Methodology

3.1 Data Collection and Understanding

The dataset used in this research was taken from the open source Kaggle repository. The dataset had a number of relevant features necessary for hypertension risk. The features in the dataset were mostly clinical and some lifestyle factors of individuals. The features in the dataset were age, gender, BMI, cholesterol levels, glucose levels, systolic and diastolic blood pressure, smoking status, and diabetes history.

3.2 Exploratory Data Analysis

The exploratory data analysis was performed to check for missing values, data types, unique values. Other statistical observations were derived and plotted in the form of heatmaps, box plots and histograms to check for distribution of the data.

3.3 Data Pre-processing

The next step was pre-processing the data to handle missing values, standardize the numerical features, and prepare it for machine learning algorithms. The missing values were replaced with the mean of each feature using mean imputation technique (Gozali, 2023). The data imbalance in the dataset was an issue. This imbalance was not addressed separately due to its clinical nature which ensured the model reflects real-world class distributions.

Standardization: The numerical features in the data were standardized. This made sure that the features had zero mean and unit variance. Standardization of the data features helped in better stabilizing and converging model like Logistic regression and Gradient Boosting.

Handling Missing Data: The missing values in the data were filled using mean imputation technique. This technique avoided data loss and kept the statistical properties of the dataset as it is.

Table 1: Preprocessing justification

Method	Justification	Reference
Standardization of the Data	Standardization improves model performance by scaling features to a uniform range.	Kaur et al. (2023)
Not Handling Data Imbalance	Clinical datasets often reflect natural class distributions. Retaining the imbalance ensures the models learn real-world patterns, avoiding biases.	Alsirhani et al. (2023)
Replacing Missing Values with Mean	Using mean imputation ensures data completeness with minimal computational cost. It is a simple and effective approach for handling missing clinical data.	Mroz et al. (2024)
Why Outliers Were Not Handled	Outliers are less of a concern in clinical datasets where the focus is on general patterns of hypertension risk, not extreme cases. Additionally, handling them may distort the natural variance of clinical data.	Kaur et al. (2023)

3.4 Model Selection and Training

The pre-processed data was then utilized in the machine learning models. Nine machine learning models were tested: Random Forest, CatBoost, Support Vector Machine (SVC), K-Nearest Neighbors (KNN), Artificial Neural Network (ANN), Gradient Boosting, Logistic Regression, Naïve Bayes, and XGBoost. Each model was trained using 5- fold cross-validation to ensure robustness. The standard 80% training and 20% testing data split was done. The baseline Machine learning models were evaluated. Hyperparameter tuning of these machine learning models was done to find the best parameters for each model. This was done using GridSearchCv for enhanced performance.

Table 2: Model selection Justification

Method	Justification	Reference
Random Forest	Random Forest is robust, interpretable, and handles multicollinearity well, making it suitable for hypertension prediction.	Keswani et al. (2023)
CatBoost	CatBoost efficiently handles categorical data and minimizes overfitting. It is particularly effective for high-dimensional datasets.	Hae et al. (2023)
SVC (Support Vector Classifier)	SVC excels in high-dimensional spaces and is effective for binary classification tasks, providing robust results when kernel functions are appropriately tuned.	Ullah & Hossain (2024)
KNN (K-Nearest Neighbors)	KNN is simple and effective for small datasets but performs less effectively in larger datasets due to high computational costs.	Matsunaga et al. (2024)
ANN (Artificial Neural Network)	ANN captures complex nonlinear relationships but is prone to overfitting, requiring careful regularization and tuning.	Alsirhani et al. (2023)
Gradient Boosting	Gradient Boosting models complex feature interactions effectively and achieves high accuracy in hypertension	Mroz et al. (2024)

	prediction.	
Logistic Regression	Logistic Regression provides interpretability and computational efficiency, making it a baseline model for clinical predictions.	Schjerven et al. (2024)
Naïve Bayes	Naïve Bayes is computationally efficient but assumes feature independence, making it less suitable for complex datasets.	Dritas et al. (2022)
XGBoost	XGBoost is highly effective for structured data and achieves excellent accuracy, but it requires significant computational resources.	Hae et al. (2023)

3.5 Ensemble Methodology

The best three models were then combined using the ensemble technique. Soft voting was used. The strengths of CatBoost, XGBoost, and Gradient Boosting models were combined in the ensemble model. Ensemble models were used as these models are known for their strength in reducing variance and bias. (Ullah & Hossain, 2024).

Table 3: Ensemble Methodology Justification

Method	Justification	Reference
Choosing F1-Score and AUC-ROC as Evaluation Metrics	F1-Score balances precision and recall, while AUC-ROC evaluates discriminatory power, making them essential for clinical models where false negatives are critical.	Ullah & Hossain (2024)
Feature Importance from Random Forest Model	Random Forest provides feature importance ranking, this identified the key predictors.	Gozali (2023) and Keswani et al. (2023)
Why Ensemble Soft Voting Was Used	Ensemble soft voting combines the strengths of different classifiers by averaging their predicted probabilities. This approach improves overall model robustness and performance, especially for balancing precision and recall.	Ullah & Hossain (2024) and Keswani et al. (2023)

3.6 Evaluation Metrics

The evaluation metrics used for evaluation of these models were chosen after conducting proper literature survey and its effectiveness in clinical settings.
The evaluation metrics were:

- **F1-Score**: For balancing precision and recall.
- **AUC-ROC**: To assess the model's ability to distinguish between classes.
- Confusion matrices were also used to evaluate misclassifications and emphasize minimizing false negatives which are critical in healthcare scenarios.

4 Design Specification

4.1 Framework Overview

The methodological framework for this research consists of the following stages (Fig.1):

Understanding the data: The collected data was loaded into python as a data frame and then it was analysed to check for missing values, the data types, the unique values and the summary of the data. Data visualizations such as histograms were plotted to check the distribution of the data (Keshwani et al., 2023). Box plots were used to check for outliers. Correlation matrix was used to understand the relations between the variables.

Data Preprocessing: The preprocessing of the data involved steps like standardizing the dataset using scalar in python (Mroz et al., 2024), Filling the missing values in the rows using mean imputation and splitting the data into 80-20 train test split. The data was then used in testing the performances of the machine learning models.

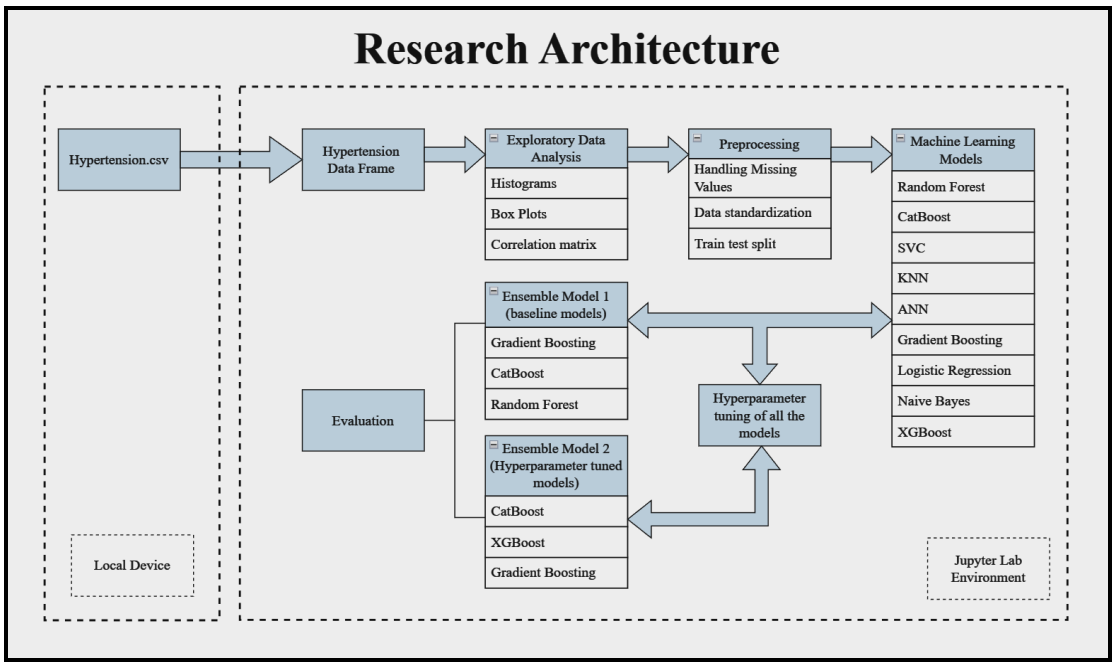

Figure 1: Research Architecture

Model Training and Tuning: Nine different machine learning models were tested, and these models were then enhanced using hyperparameter tuning for improved performance (Matsunaga et al., 2024). The Nine machine learning models were then evaluated using evaluation metrices like F1-score and AUC. The various evaluation metrics of these models were also plot for a better comparison of model performances. The machine learning algorithms are listed below:

1. **Random Forest:** It is a robust tree-based algorithm capable of handling multicollinearity properly. Random Forest was also used to find the feature

importance along with correlation matrix. Random Forest is easily interpretable; hence it was used for this clinical application.

2. **Gradient Boosting:** Gradient Boosting captures the complex feature interactions and offers better accuracy while minimizing overfitting.
3. **CatBoost:** CatBoost model works best on datasets with minimal preprocessing, and it is computationally efficient.
4. **XGBoost:** XGBoost is an optimized gradient boosting algorithm which is designed for scalability, speed and performance.
5. **Support Vector Classifier (SVC):** SVC is a simpler model that performs well in high dimensional spaces and works best with smaller datasets.
6. **K-Nearest Neighbors (KNN):** KNN is another simple algorithm that relies on proximity metrics. KNN works best on standardized datasets but can be computationally heavy for bigger datasets.
7. **Logistic Regression:** Logistic regression is a simple linear algorithm offering the best interpretability.
8. **Artificial Neural Networks (ANN**): ANN is a deep learning model which is capable of capturing complex non-linear data.
9. **Naïve Bayes:** Naïve Bayes is a probabilistic model that assumes feature independence, making it computationally efficient.

Ensemble Model: After evaluating the hyperparameter tuned models, the top three models with the highest F1-score and AUC were taken into consideration for the ensemble model. The ensemble model was built using baseline versions of CatBoost, XGBoost, and Gradient Boosting models using soft voting. Soft voting combines the predicted probabilities of individual models (Hae et al., 2023). This technique made sure to utilize the predictive abilities of more number of models for better results.

4.2 Requirements

4.2.1 Hardware:

A computational device with at least 16GB RAM and a multicore processor.

4.2.2 Software: Python programming environment with libraries including:

1. scikit-learn for preprocessing, training, and evaluation.
2. CatBoost, XGBoost, and TensorFlow for specialized models.
3. Matplotlib and Seaborn for visualization.

5 Implementation

5.1 Tools and Libraries used

The implementation phase of the research was done in the python programming language using the python libraries for data manipulation, preprocessing, machine learning models and visualization.

Data Manipulation and Preprocessing: Pandas was used for reading, cleaning, and manipulating the dataset. StandardScaler function was used from sklearn library for standardizing the numerical data in the dataset. The Train-Test Split function was used from the sklearn library for splitting the data into 80-20 train-test split (Ullah & Hussain, 2024).

Machine Leaning Models: Sklearn library was utilized for implementing machine learning models like Random Forest, Gradient Boosting, Logistic regression, SVC, KNN, Naïve Bayes and ANN (MLP). The xgboost library was used in modelling the XGBoost model and the catboost library was used to model the CatBoost model.

Hyperparameter Tuning: The GridSearchCV from sklearn library was used for hyperparameter tuning to find the best parameters for each model (Gozali, 2023).

Evaluation and Metrics: Sklearn library was used to perform 5-fold cross-validation to evaluate model performance and to plot accuracy, precision, recall, F-1 score, AUC-ROC, classification reports and confusion matrices.

Visualization: Matplotlib and Seaborn libraries were utilized for all the visualizations in this research like plotting ROC curves, histograms, and bar plots of performance metrices (Keswani et al., 2023).

5.2 Data Understanding

The data used in this research is sourced from Kaggle which a open source data repository (https://www.kaggle.com/datasets/khan1803115/hypertension-risk-model-main). The data includes a number of health and lifestyle features. Each row corresponds to a different individual. The dataset contains 4240 rows each representing an individual with different features. The features in the dataset are represented as columns. There are 13 features in the dataset listed in Figure 2 representing the head of the data.

	male	age	currentSmoker	cigsPerDay	BPMeds	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	Risk
0	1	39	0	0.0	0.0	0	195.0	106.0	70.0	26.97	80.0	77.0	0
1	0	46	0	0.0	0.0	0	250.0	121.0	81.0	28.73	95.0	76.0	0
2	1	48	1	20.0	0.0	0	245.0	127.5	80.0	25.34	75.0	70.0	0
3	0	61	1	30.0	0.0	0	225.0	150.0	95.0	28.58	65.0	103.0	1
4	0	46	1	23.0	0.0	0	285.0	130.0	84.0	23.10	85.0	85.0	0

Figure 2: Data Head

In the male column 1 represents the gender being male and 0 being a female. The age column corresponds to the age of the individual. The currentSmoker column represents if the individual is a smoker or not. The cigsPerDaY column denotes the average number of cigarettes smoked by the individual. The BPMeds column depict if the person is on any medication for Blood Pressure. The diabetes column represents the diabetes status. The totChol column shows the actual total cholesterol levels. The sysBP and diaBP represent the systolic Blood Pressure and the diastolic Blood Pressure respectively. The BMI and heartrate columns show the BMI index and heartRate of the person. The glucose column represents the glucose level of the individual. The Risk column depicts if the person has a risk of hypertension or not; 1 representing the risk of hypertension. While performing the Exploratory Data Analysis, 388 Missing values were found in the glucose column, 29 in cigsPerday column, 53 in BPMeds, 50 in totChol, 19 in BMI and 1 in heartRate.

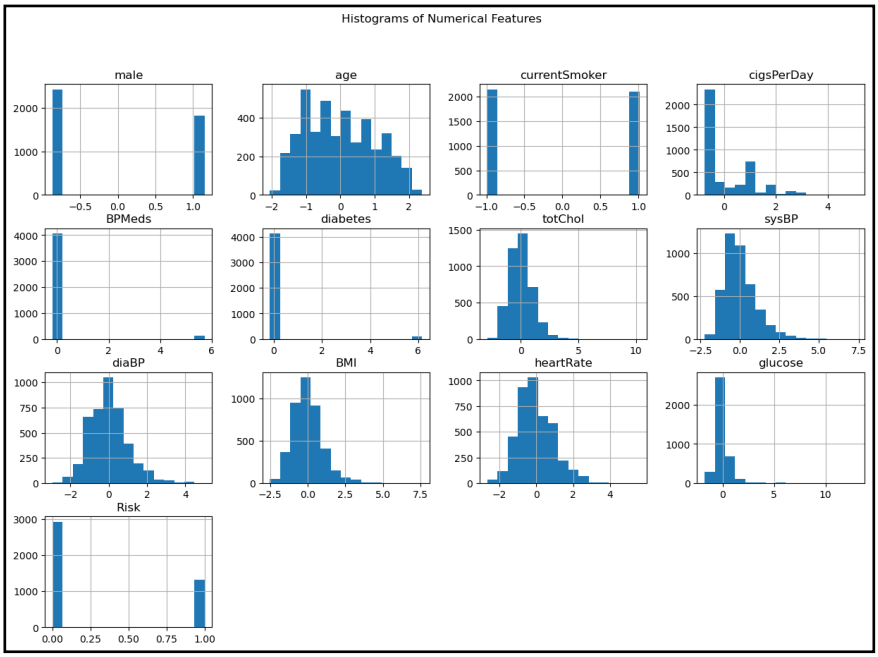

Figure 3: Histograms of numerical features

The Figure 3 depicts the distribution of the features in the form of histograms. High imbalance can be seen in some of the features like BPMeds, diabetes and Risk. Features such as age, currentSmoker, totChol, sysBP, diaBP, BMI, heartRate are more or less normally distributed with slight skewness (Gozali, 2023).

The Correlation Matrix (Fig. 4) provided insights into the key predictors of hypertension Risk. The Systolic Blood pressure, Diastolic Blood Pressure, age and BMI factors had strong correlations with the Risk variable. This indicated that these factors are important in hypertension risk detection (Schjerven et al., 2024). Other features such as cholesterol levels, Blood pressure medication and smoking status had lesser correlations to the risk variable, but they were useful factors for prediction of hypertension risk (Mroz et al., 2024).

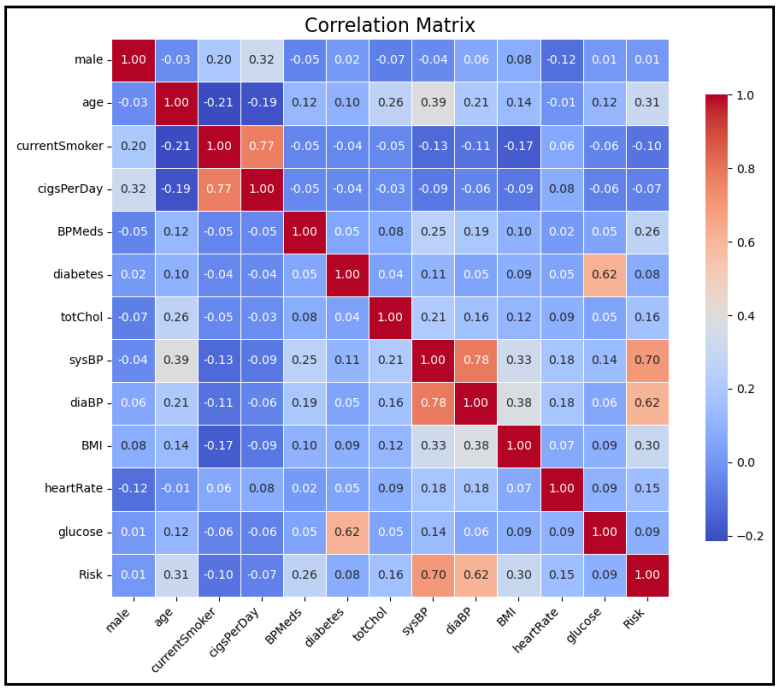

Figure 4: Correlation Matrix

5.3 Data Preprocessing

The data collected was moderately simple, so the data did not require a lot of preprocessing. The missing values were handled using mean imputation. This adds the mean of the column as a value. This helps in keeping the distribution normal (Abrar et al.,2021). This complete data was standardized to bring the variance to unit. Standardizing the data also helped in improving the performances of the machine learning models. The standardized dataset was then split into training and testing sets with an 80-20 split. The training set consisted of 3392 rows and the test set was made of 848 rows (You et al.,2024).

5.4 Machine Learning Models

The pre-processed data was then used in machine learning models. Nine machine learning models were tested using 5-fold cross-validation where the models were trained on the training dataset and then tested on the test data. The machine learning models were: Random Forest, CatBoost, Support Vector Machine (SVC), K-Nearest Neighbors (KNN), Artificial Neural Network (ANN), Gradient Boosting, Logistic Regression, Naïve Bayes, and XGBoost

(Qin, 2023). These machine learning models were evaluated based on F-1 scores and AUC-ROC (Ramadhan et al., 2023).

Similarly, hyperparameter tuning of these nine models were done. These hyperparameter tuned models were also evaluated using F-1 scores and AUC-ROC. Other evaluation metrics that were also used were accuracy, precision, recall, classification report and confusion matrix (Siopis et al., 2022). After proper evaluation of these individual models, ensemble technique was used to combine the strengths of these models (Schjerven et al., 2024; Mroz et al., 2024). Two ensemble models were built where the first model combined the top three best performing default models. And the second ensemble model was built combining the top three hyperparameter tuned models. These two ensemble models were critically evaluated to conclude the best performing model.

6 Evaluation

6.1 Experiment 1: Machine Leaning Models on default settings

Experiment 1 involved nine baseline machine learning models. Each model was evaluated using the evaluation metrices. From the Nine baseline models, Gradient Boosting, Random Forest and CatBoost model were found to be the top three best performing models. Apart from these three models XGBoost, SVC, Logistic Regression, ANN and KNN showed almost similar performances, but their F-1 scores were slightly lesser (Tahsin et al., 2021). The evaluation metrics of these models are denoted in table 4.

Table 4: Baseline model evaluation

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC Score
Gradient Boosting	0.9045	0.8209	0.8696	0.8445	0.9586
Random Forest	0.9021	0.8036	0.8893	0.8443	0.9528
CatBoost	0.9009	0.8073	0.8775	0.8409	0.9568
XGBoost	0.8915	0.8084	0.8340	0.8210	0.9502
Logistic Regression	0.8962	0.8700	0.7668	0.8151	0.9463
SVC	0.8892	0.8354	0.7826	0.8082	0.9474
ANN	0.8561	0.7510	0.7747	0.7626	0.9134
KNN	0.8656	0.8174	0.7075	0.7585	0.9115
Naive Bayes	0.8031	0.8583	0.4071	0.5523	0.9268

Feature Importance Analysis: The feature importance was derived from the Random Forest model. This gave us a clear view of the relevant features (Yagmurcu & Arslan, 2024) in prediction of hypertension risk. The Figure 5 depicts that Systolic Blood Pressure (sysBP), Diastolic Blood Pressure (diaBP) have the highest level of importance. Followed by BMI, Age, Cholesterol level, glucose level, heartRate, BPMeds, cigsPerDay, gender, current

smoking status and lastly diabetes. This helps us answer the question about what factors are key in hypertension prediction (Abrar et al., 2021).

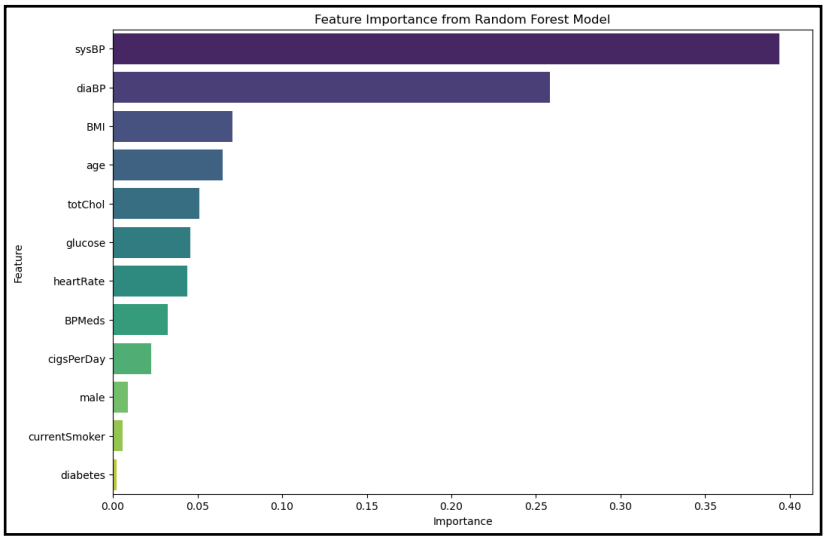

Figure 5: Feature Importance

6.2 Experiment 2: Hyperparameter Tuned Models

Hyperparameter tuning was done on the nine baseline models. A grid Search was used for a proper search of parameters over a manually specified hyperparameter grid (Yagmurcu & Arslan, 2024). Cross-validation was applied within grid search to ensure that the chosen hyperparameters were generalized to unseen data. 5-fold cross-validation along with hyperparameter tuning made the models more robust (Schjeven et al., 2024). Using Grid search, the best parameters were found that gave the best evaluation results. The best hyperparameters are given in table 5.

Table 5: Best hyperparameters for models

Model	Best Hyperparameters
Random Forest	max_depth: None, min_samples_split: 5, max_features: sqrt, n_estimators: 50, min_samples_leaf: 1
CatBoost	depth: 8, iterations: 100, l2_leaf_reg: 3, learning_rate: 0.01
SVC	C: 100, gamma: 0.001, kernel: rbf
KNN	metric:Manhattan, n_neighbors: 9, weights: distance
ANN	activation: relu, alpha: 0.001, hidden_layer_sizes: (128, 64), learning_rate: adaptive, solver: sgd
Gradient Boosting	learning_rate: 0.01, max_depth: 3, min_samples_leaf: 2, min_samples_split: 2, n_estimators: 200
Logistic Regression	C: 1, max_iter: 100, penalty: l2, solver: liblinear
Naïve Bayes	var_smoothing: 1e-09
XGBoost	colsample_bytree: 1.0, learning_rate: 0.1, max_depth: 3, min_child_weight: 3, n_estimators: 50, subsample: 0.9

After comparing the default models with the tuned models, it's clear that Gradient Boosting, CatBoost, and XGBoost were the top-ranked models (Table 6) in which the differences are truly negligible. The CatBoost model shows the highest F1-score of 0.8475 and a recall of 0.8893, indicating how well this model can balance between false positives and false negatives (Ramadhan et al., 2023). For the XGBoost model, F1-score 0.8453 was obtained along with an AUC-ROC of 0.9568. After tuning, the Gradient Boosting model was still fairly consistent on an F1-score of 0.8441 and AUC-ROC of 0.9541 (Qin, 2023).

Table 6: Hyperparameter tuned model evaluation

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC Score
CatBoost	0.9045	0.8094	0.8893	0.8475	0.9542
XGBoost	0.9033	0.8087	0.8854	0.8453	0.9568
Gradient Boosting	0.9033	0.8132	0.8775	0.8441	0.9541
Random Forest	0.8974	0.8051	0.8656	0.8343	0.9506
ANN	0.9033	0.8548	0.8142	0.8340	0.9503
SVC	0.8986	0.8615	0.7866	0.8223	0.9475
Logistic Regression	0.8962	0.8700	0.7668	0.8151	0.9463
KNN	0.8726	0.8341	0.7154	0.7702	0.9270
Naive Bayes	0.8031	0.8583	0.4071	0.5523	0.9268

On the other hand, Random Forest model after hyperparameter tuning only had a slight improvement in recall and F1-score compared to baseline model (Mroz et al., 2024). For simpler models like Logistic regression and SVC, hyperparameter tuning had minimal effect on their performance (Abrar et al., 2021). This can be seen in their evaluation metrics being almost the same.

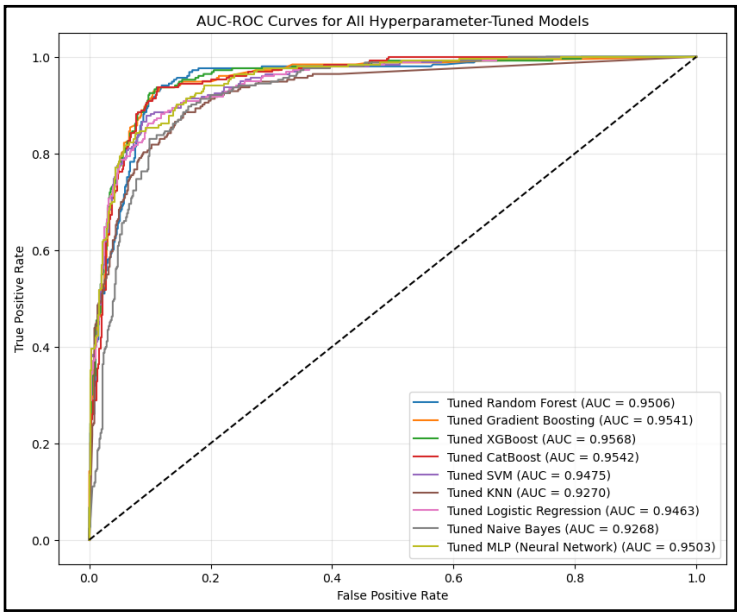

Figure 6: AUC-ROC curves of hyperparameter tuned models

6.3 Experiment 3: Ensemble Models

In experiment 3, two ensemble models were developed. Both the models used soft voting approach, where predictions were based on the probability results of individual models (Keswani et al., 2023). The ensemble Model 1 was a combination of the baseline models CatBoost, Gradient Boosting and Random Forest. The ensemble model 2 was a combination of hyperparameter tuned XGBoost, Gradient Boosting and XGBoost models (Qin, 2023).

Ensemble Model 1: The Ensemble model 1 resulted in 90.8% accuracy, a F1-score of 0.8528, while the AUC-ROC score was 0.9569 as depicted in Fig 7. The results are enough evidence that ensemble model 1 classified the data successfully (Mroz et al., 2024). The Model was focused on minimizing false negatives and hence managed to achieve a recall value of 0.8933. The confusion matrix in ensemble model tells us that a balanced classification for precision and recall is observed.

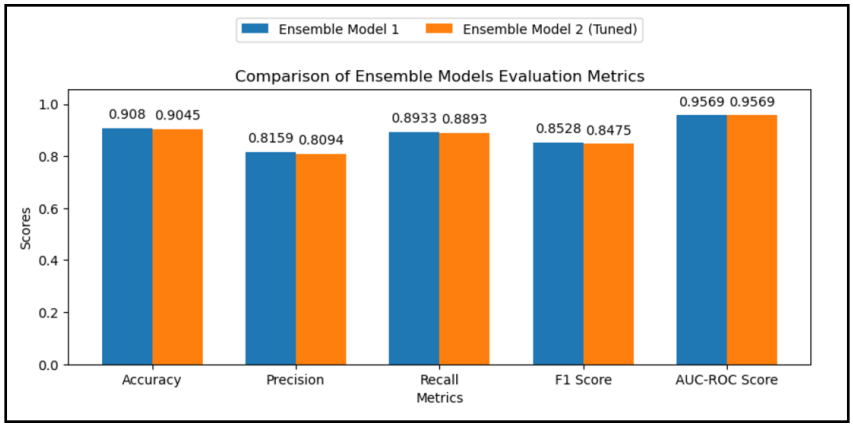

Figure 7: Comparison of ensemble models

Ensemble Model 2: The hyperparameter tuned ensemble model, combined the strengths of XGBoost, CatBoost and Gradient Boosting, achieving a slightly lower accuracy of 90.45% (Fig. 7) than ensemble model 1. The F1-score for ensemble model 2 was 0.8475 and the AUC-ROC was 0.9569. Overall ensemble model 2 had no significant improvements than ensemble model 1.

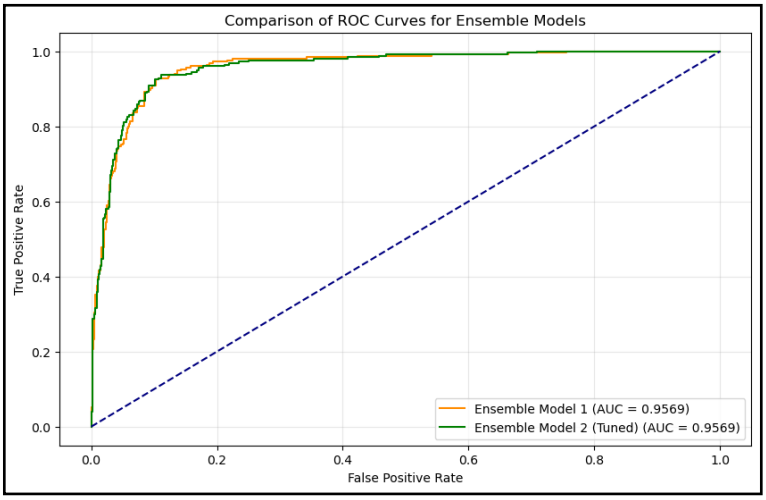

Figure 8: AUC-ROC curves of ensemble models

Comparison of Ensemble Models: Both the ensemble models displayed strong predictive capabilities with AUC-ROC scores of 0.9569, which suggests good discrimination between classes (Schjerven et al., 2024). The ensemble Model 1 which was made of the baseline individual models Random Forest, CatBoost and Gradient Boosting performed barely better in terms of F1-score and recall. This suggests that model 1 is effective in minimizing false negatives. The ensemble models 2 which was a combination of the hyperparameter tuned CatBoost, Gradient Boosting and XGBoost models performed almost similar to model 1 in terms of recall and precision but had a slightly lower F1-score. Suggesting that ensemble model of hyperparameter tuned models had no significant impact on the results (Abrar et al., 2021). The confusion matrices (Fig. 9) of both the models show that false positives and false negatives were under control. The confusion matrices suggest that both the models are reliable in practical applications.

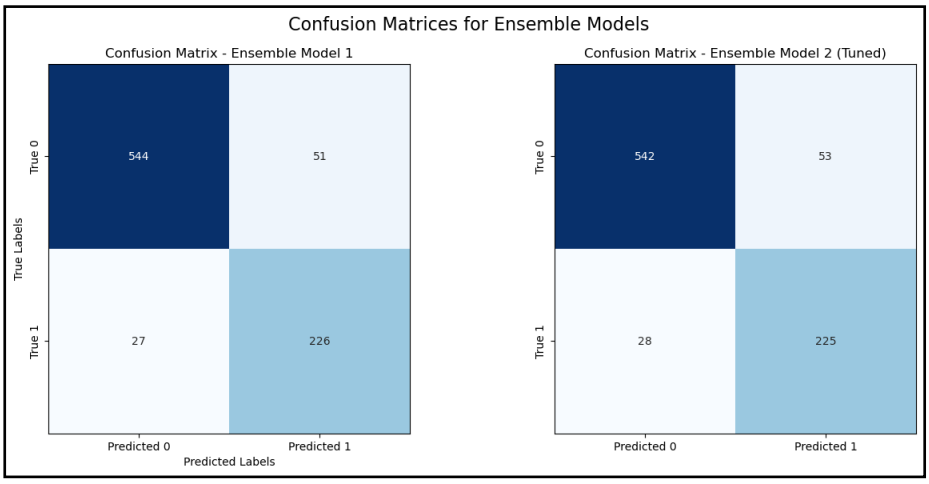

Figure 9: Confusion Matrices of ensemble models

6.4 Discussion

The evaluation of the baseline machine learning models revealed that Gradient Boosting, Random Forest and CatBoost models were the top performing models based on their F1-scores, recall and AUC-ROC metrics. Besides these models showing strong predictive power, models like Logistic regression and Naïve Bayes lacked in capturing the complexities in the data. This can be seen in their F1-scores and recall metrics. The inability of baseline models to effectively handle the imbalanced data emphasizes the need for ensemble models (Ullah & Hossain, 2024).

Hyperparameter tuning enhanced the performances of several models where CatBoost, XGBoost and Gradient Boosting had slight improvements in their F1-score and AUC-ROC (Tahsin et al.,2021). The hyperparameter tuned CatBoost model had the best evaluation metric with F1-score of 0.8475 and recall of 0.8893. This indicates that hyperparameter tuning was effective for models requiring fine adjustments such as boosting algorithms (Mroz et al., 2024). On the other hand, Simple models such as Logistic regression and SVC showed less improvements after tuning. Random Forest model being a robust model only had marginal improvements proving the model was already optimized in default settings (Yagmurcu & Arslan, 2024).

The ensemble modelling approach, combining the strengths of multiple algorithms proved effective in achieving balanced predictive abilities. Ensemble model 1, which was the combination of baseline Gradient Boosting, Random Forest, and CatBoost, depicted the best overall performance with an F1-score of 0.8528 and recall of 0.8933. This model effectively minimized false negatives, a critical factor in clinical datasets. The ensemble model 2 which utilized the hyperparameter tuned versions of CatBoost, Gradient Boosting and XGBoost offered comparable performance. Although, ensemble model 2 did not significantly outperform ensemble model 1.

The results align with the study done by Ullah & Hossain, (2024), where they suggest that ensemble technique is effective in minimizing false negatives and improving recall. However, the marginal improvements for hyperparameter tuning contrasts with literature emphasizing its importance in models like gradient boosting (Hae et al., 2023).

7 Conclusion and Future Work

Apart from exploring the applicability of machine learning models in predicting hypertension risk, this research also identified the most significant contributing factors for high blood pressure. This study demonstrated the potential of predictive analytics in healthcare, specifically in hypertension risk evaluation.

Nine machine learning models were implemented, and the performances of these models were evaluated using metrics such as F1-score, recall and AUC-ROC.
The comparison of these metrics (Table 4) show that Gradient Boosting, Random Forest and CatBoost models had good performances. Gradient Boosting was the best performing model with a F1-score of 0.8445. All the nine models were hyperparameter tuned by finding the best parameters using grid search. Among which CatBoost had the best results (Table 6) with a F1-score of 0.8475 and a recall of 0.8893. Despite the improvements in individual performances of these models, some models were unaffected like Random Forest. This tells us that the default configuration of Random Forest is well adjusted to the dataset.

The performance for the Ensemble models outperformed the individual machine learning models. The F1-score for Ensemble Model 1 was 0.8528, and the recall was 0.8933. On the other hand, the second ensemble model combined the tuned models had similar results but did not outperform the first ensemble model. Figure 7 showcases the scores for ensemble models. It seems that the hyperparameter tuning had very little impact on the ensemble models. Besides this, the feature importance contribution from Random Forest models was analysed to find that systolic blood pressure, diastolic blood pressure, BMI, and age are strong contributors in the prediction of hypertension risk. The dataset used in this research was moderately sized. A larger dataset with more features, such as genetic information, would further enhance the power of this study. Other advanced ensemble techniques could be explored to further improve model predictions. This research provides a scalable and clinically relevant solution that may help improve the healthcare system.

Acknowledgement

I sincerely thank Prof. Harshani Nagahamulla for providing her invaluable guidance and support throughout the course of this research. Project.

References

Mroz, T., Griffin, M., Cartabuke, R., et al. (2024). Predicting hypertension control using machine learning. *PLoS ONE*, 19(3), e0299932. https://doi.org/10.1371/journal.pone.0299932

Matsunaga, T., Kono, A., Nishio, M., et al. (2024). Development and web deployment of prediction model for pulmonary arterial pressure in chronic thromboembolic pulmonary hypertension using machine learning. *PLoS ONE*, 19(4), e0300716. https://doi.org/10.1371/journal.pone.0300716

Gupta, T. (2023). Development of User Friendly Based Home Health Monitoring System for The Prediction of Hypertension Using Machine Learning Algorithm. *2023 3rd Asian Conference on Innovation in Technology (ASIANCON)*. IEEE. https://doi.org/10.1109/ASIANCON58793.2023.10269848

Keswani, A., Jain, T., & Sharma, V. S. (2023). Predicted Blood Pressure using Machine Learning. *2023 3rd International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)*. IEEE. https://doi.org/10.1109/SMARTGENCON60755.2023.10442670

Gong, K., Chen, Y., & Ding, X. (2023). Causal Inference for Hypertension Prediction. *2023 45th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. https://doi.org/10.1109/EMBC.2023.10442691

Obafemi, A. S. (2022). A Predictive Model for Predicting Blood Pressure Levels Using Machine Learning Techniques. MSc Research Project, National College of Ireland.

Kaur, S., Bansal, K., & Kumar, Y. (2023). Machine Learning based Approaches for Accurately Diagnosis and Detection of Hypertension Disease. *2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*. IEEE. https://doi.org/10.1109/UPCON59197.2023.10434428

Hae, H., Kang, S.-J., Kim, T. O., et al. (2023). Machine Learning-Based Prediction of Post-Treatment Ambulatory Blood Pressure in Patients with Hypertension. *Blood Pressure*, 32(1), 2209674. https://doi.org/10.1080/08037051.2023.2209674

ALSIRHANI, N., ALSHAMMARI, A., AHMED, H., et al. (2023). Machine Learning-Based System for Predicting Hypertension in North Saudi Arabia. *2023 3rd International Conference on Computing and Information Technology (ICCIT)*. IEEE. https://doi.org/10.1109/ICCIT2023.10442689

Schjerven, F. E., Ingeström, E. M. L., Steinsland, I., & Lindseth, F. (2024). Development of risk models of incident hypertension using machine learning on the HUNT study data. *Scientific Reports*, 14, 5609. https://doi.org/10.1038/s41598-024-56170-7

Ullah, S. M. A., & Hossain, A. B. M. A. (2024). Hypertension prediction using stacked ensemble model from imbalanced clinical data. *2024 International Conference on Advances in Computing, Communication, Electrical, and Smart Systems (iCACCESS)*. IEEE. https://doi.org/10.1109/iCACCESS.2024.1234567

Gozali, A. A. (2023). Hypertension multi-year prediction and risk factors analysis using decision tree. *2023 10th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*. IEEE. https://doi.org/10.1109/ICITACEE.2023.10277644

Ramesh, Y., Rehan, M., Kodipalli, A., et al. (2023). Diabetes, hypertension and stroke prediction using computational algorithms. *2023 World Conference on Communication & Computing (WCONF)*. IEEE. https://doi.org/10.1109/WCONF.2023.10235026

Abbas, S., Sampedro, G. A., Krichen, M., et al. (2024). Effective hypertension detection using predictive feature engineering and deep learning. *IEEE Access*, 12, 89055. https://doi.org/10.1109/ACCESS.2024.3418553

Zhao, H., Ma, Z., & Sun, Y. (2019). A hypertension risk prediction model based on BP neural network. *2019 International Conference on Networking and Network Applications (NaNA)*. IEEE. https://doi.org/10.1109/NaNA.2019.00085

Shanmugapriya, G., Arjun, M., Bhargavaa, G., & Sidhaarth, A. (2024). Blood Pressure Prediction using Machine Learning Algorithms. *2024 4th International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*. IEEE. https://doi.org/10.1109/ICACITE60783.2024.10616781

Behki, P., Agrawal, A., & Das, D. (2023). Classification by Machine Learning Algorithms for Predicting Heart Disease. *2023 International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIHI)*. IEEE. https://doi.org/10.1109/ICAIHI57871.2023.10489638

Dritsas, E., Alexiou, S., & Moustakas, K. (2022). Efficient Data-driven Machine Learning Models for Hypertension Risk Prediction. *2022 International Conference on Innovations in Intelligent Systems and Applications (INISTA)*. IEEE. https://doi.org/10.1109/INISTA55318.2022.9894186

J., Li, J., Li, X., Gao, J., Wu, J., & Ye, J. (2023). Risk-prediction model for incident hypertension in patients with obstructive sleep apnea based on SpO2 signals. *45th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. https://doi.org/10.1109/EMBC40787.2023.10340756:contentReference[oaicite:0]{index=0}.

Qin, Y. (2023). Research on hypertension prediction model based on XGBoost algorithm. *2023 IEEE 3rd International Conference on Data Science and Computer Application (ICDSCA)*. IEEE. https://doi.org/10.1109/ICDSCA59871.2023.10393110:contentReference[oaicite:0]{index=0}.

Abrar, S., Loo, C. K., & Kubota, N. (2021). A multi-agent approach for personalized hypertension risk prediction. *IEEE Access*, 9, 75090–75105. https://doi.org/10.1109/ACCESS.2021.3074791:contentReference[oaicite:0]{index=0}.

Ramadhan, N. G., Maharani, W., Adiwijaya, & Gozali, A. A. (2023). Prediction of hypertension in the upcoming year: Feature correlation analysis and handling imbalanced based on Random Forest. *8th International Conference on Informatics and Computing (ICIC)*. IEEE. https://doi.org/10.1109/ICIC60109.2023.10381983.

Yağmurcu, M., & Arslan, S. (2024). Feature selection using automatic programming methods in hypertension risk prediction. *8th International Artificial Intelligence and Data Processing Symposium (IDAP)*, IEEE. https://doi.org/10.1109/IDAP64064.2024.10711046.

Tahsin, T., Mumenin, K. M., Pinki, F. T., Tuli, A. B., Sikder, S., Rahman, M. A., Bulbul, A. A. M., & Awal, M. A. (2021). GWO-XGB: Grey Wolf Optimization-based eXtreme Gradient Boosting for Hypertension Prediction in Bangladesh. *2021 International Conference on Electronics, Communications and Information Technology (ICECIT)*, IEEE. https://doi.org/10.1109/ICECIT.2021.9658593.

Siopis, N., Triantafyllidis, A., Alexiadis, A., Votis, K., Gerovasilis, G., & Tzovaras, D. (2022). HSmartBPM: A modular web platform for tailored management of hypertension. *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE.