

Machine Learning-Based Prediction of Hypertension Risk: A Comparative Study of Classification Algorithms

Sivakoti Sai Giridhar¹, Suhas K. C², and S. Sathyanarayanan^{*3}

Department of Mathematics and Computer Science
Sri Sathya Sai University for Human Excellence,
Navanihal, Okali Post, Kamalapur
Kalaburagi, Karnataka - 585313

Corresponding author: S. Sathyanarayanan³, E-mail: sathyanarayanan.s@sssuhe.ac.in / sathyanarayanan.brn@gmail.com

Abstract—Hypertension is a significant preventable risk factor for cardiovascular disease (CVD), stroke, renal failure and premature mortality. Identification of a person who is at high risk can facilitate preventive care, follow-up of the person, and screening in public health care. This study proposes a comparative machine-learning approach to hypertension-risk prediction with the National Health and Nutrition Examination Survey (NHANES) database. Under a common experimental framework, five supervised classifiers—Logistic Regression, Random Forest, Gradient Boosting, XGBoost and CatBoost—were trained using a consistent pipeline of data processing steps, including: Median imputation, standard scaling, clinically guided feature engineering, randomized hyperparameter optimization, class-imbalance handling with SMOTE (this was only applied to the training set). Hypertension status was considered in terms of Stage 2 in which those with an SBP greater than or equal to 140 mmHg or DBP greater than or equal to 90 mmHg were deemed hypertensive. Due to the imbalance of the data set, the model comparison focused on recall, precision, F1-score, Cohen's Kappa, and ROC-AUC instead of just the accuracy score. XGBoost model showed the best test F1 score and test Recall, whereas the test Recall of Random Forest model also showed good stability on cross validation. The results indicated that the ensemble and boosting models are better at characterizing nonlinear physiological and demographic correlations than a linear baseline model. Still, the low positive-class recall suggests that optimization of the thresholds and external validation, calibration, and explainability analysis are necessary prior to use in clinical screening. The proposed framework is then suggested as a research pipeline for decision support which is repeatable and not designed to supplant clinical diagnosis. [16]-[18]

Index Terms—Hypertension prediction, machine learning, NHANES, XGBoost, Random Forest, CatBoost, SMOTE, clinical decision support, preventive healthcare.

I. INTRODUCTION

Hypertension is a disease where blood pressure is high for long periods of time. It is also one of the greatest public health problems as it can cause damage to the cardiovascular system, vascular deterioration, myocardial infarction, stroke, renal disease and premature death. Frequently patients have no symptoms for many years, adding to the medical gravity of the condition called hypertension. This makes it possible that high blood pressure remains undiagnosed until organ damage has become evident or while doing an unrelated checkup. [19]-[22]

From normal blood pressure to clinical hypertension, it is not often a sudden change. Factors that affect it include age, body composition, physical inactivity, diet, metabolic state, alcohol use and tobacco consumption, genetic predisposition and other socio-demographic factors. These risk factors are not on their own. The effects of these together may be nonlinear, such as age might enhance the effect of obesity, or lifestyle variables the clinical impact of BMI (body mass index). These interactions encourage the use of data-driven learning methods to predict hypertension. [16], [17]

A traditional diagnosis relies on blood pressure readout at doctor visits. While direct measure is usually the gold standard of their diagnosis, a single measurement only gives a snapshot of the patient's situation. In reality, early detection may be impacted by white coat hypertension, masked hypertension, infrequent medical visits and/or irregular screening programs.

The predictive model cannot be used as a method of diagnosing hypertension, but it can be used to aid screening by identifying individuals at high-risk for hypertension based on their profile of demographic and physiological characteristics. [18]

With the rise of electronic healthcare data and large population surveys, machine learning (ML) approaches are beginning to emerge which are able to estimate disease risk. Machine learning models can be trained to identify complex relationships within structured health data, and can be used to make risk predictions that could help inform clinical care, public health planning, and digital health systems. The most important goal in hypertension prediction is to be able to reliably identify individuals who are at risk of developing hypertension, rather than simply being high overall accuracy. This is important especially when the medical dataset may be imbalanced, that is, have more healthy people than people attending the hospital for the target medical condition. [23], [24]

This paper re-formats and combines the project work in the standard IEEE-style research paper. The study is based on a common set of experimental design parameters and compares five machine learning algorithms. The most significant part is a clinically appropriate and easily replicable pipeline, involving preprocessing, class-balancing, feature engineering, model optimization, and the clinically relevant evaluation metrics for hypertension-risk prediction. This study fulfills this need by evaluating several supervised learning algorithms on a standard-

ized pipeline of preprocessing and evaluation metrics and evaluating these metrics in the context of imbalanced data and with clinical caution. [25], [26]

II. OBJECTIVES AND CONTRIBUTIONS

The main aim of this research is to develop and test a comparative classification model for the prediction of hypertension risk from structured public-health information. The following objectives are the specific ones: To build a preprocessing pipeline that is able to manage missing data, inconsistent feature scales and class imbalance in NHANES-derived health records. To use the thresholds for systolic and diastolic BP to define a clinically relevant binary target for hypertension. To compare a linear baseline model with tree-based ensemble and boosting techniques, such as Random Forest, Gradient Boosting, XGBoost and CatBoost. To assess the models using metrics that are appropriate with imbalanced healthcare data such as recall, precision, F1-score, Kappa, and ROC-AUC. To detect the most dominant physiology and demographic factors associated with the classification of hypertension. [27]

This paper makes four main contributions. First, it is using uniform preprocessing and evaluation procedure for all the models, facilitating fair comparison. Second, it specifies the target for prediction based on clinically relevant BP thresholds, and assesses performance based on screening-oriented metrics. Third, it only operates with class imbalance in the training pipeline which reduces the risk of data leakage. Fourth, it offers a clear basis for adjusting the thresholds, calibrating the system, carrying out explainable AI analysis and external validation prior to any real-world decision-support usage. [24], [25]

III. RELATED WORK

Recently, machine learning has increasingly been utilized to predict chronic diseases, a good example of which is the ability to predict hypertension risk from structured clinical information, survey data, and lifestyle information. For modeling nonlinear interactions among age, body composition, lifestyle factors and comorbid conditions, Logistic Regression is a good baseline to use, whereas Random Forest, Gradient Boosting, XGBoost or CatBoost are appropriate. [28]-[31]

Previous studies that attempt to predict hypertension based on population data like NHANES demonstrate that noninvasive demographic and physiological variables can be used as complementary indicators for risk stratification with varying performance, depending on the chosen class distribution, target definition, feature selection, validation strategy, and threshold selection. In addition, systematic reviews highlight the fact that numerous studies in the literature do not adequately address issues of minority class recall, external validation, calibration, or clinical utility. [26], [27]

To this existing literature, the present work has two contributions: the first was to compare multiple algorithms based on the same pipeline; the second was to interpret the results based on metrics more suitable for imbalanced screening in the healthcare domain.

Comparing against existing work: This paper also employs a preprocessing routine that tends to be common across the five classifiers, as opposed to studies that report on a single model or accuracy results only, before comparing results with those classifiers, through the use of recall, precision, f1 score, kappa and roc-auc. It also explicitly presents the class imbalance problem and applies SMOTE just to the training set, and does not give the impression of a diagnosis replacement.

Research gap and novelty: The main gap which the paper addresses is the fair model comparison in imbalanced screening conditions. The innovation is not only in the clinically relevant target labeling, but also in incorporating leakage-aware preprocessing, class balancing, and the random tuning of different parameters, in one reproducible predictive hypertension risk pathway, together with the multi-metric interpretation. [32], [33]

IV. DATASET DESCRIPTION

This study utilizes data collected in the National Health and Nutrition Examination Survey (NHANES). NHANES is appropriate for this work because it has individual-level health data such as questionnaire responses, health examination data, and demographic data. This experiment has a total number of observations ($N = 5,903$) each of which represents a health profile of a particular individual. [2], [19]

The DV is the hypertension status. Hypertensive was defined as those participants who had their SBP ≥ 140 or DBP ≥ 90 . This definition is in line with hypertension stage 2 criteria. Those who did not fall into this category were classified as non-hypertensive. It is seen that there is a clear imbalance in the target distribution, where about 84.45% of the individuals are classified as normal and about 15.55% as hypertensive. Clinically realistic and algorithmically hard imbalance.

Demographic and physiological metrics are included, such as age, gender indicators, ethnicity indicators, BMI, body weight, smoking-related measures, alcohol-related measures and physical-activity indicators. Engineered features were also created to improve the ability of the models to represent nonlinear patterns. They range from BMI categories, age groups, squared age terms, log transformed weight and interaction style variables.

TABLE I
DATASET AND TARGET SUMMARY

Item	Description
Data source	National Health and Nutrition Examination Survey (NHANES)
Observations	5,903 individual health profiles
Target class	Hypertension status: 1 = hypertensive, 0 = normal
Clinical rule	SBP ≥ 140 mmHg or DBP ≥ 90 mmHg
Class distribution	Normal: 84.45%, Hypertensive: 15.55%
Main predictors	Age, BMI, weight, demographic variables, lifestyle indicators
Engineered features	BMI category, age group, squared age, log weight, interaction variables

V. DATA PREPROCESSING

The machine learning process has a significant problem due to the presence of missing values and outliers in the raw data, mixed scales, and class imbalance in the survey's data, among other issues, which requires data preprocessing. A well-constructed preprocessing pipeline can strengthen the reliability of the model and avoid the model capturing false patterns. [31]

A. Missing Value Treatment

A simple analysis of data revealed missing data in specific attributes, particularly in lifestyle attributes and a few physical measurements. If all records were deleted if there were missing values, the sample size will also decrease and this could lead to bias. Hence, missing values were imputed with the median value. Physiological variables are thus more frequently reported as median than mean. For instance, if these values are extreme (e.g. weight or BMI) the mean value will be affected whereas the median (central value) will not be affected. [31]

B. Target Variable Construction

Clinically accepted thresholds were used to create the binary target. The person was considered hypertensive if either SBP was ≥ 140 mmHg or DBP was ≥ 90 mmHg. This is the rule that reads a continuous blood-pressure measurement and turns it into a target to be classified as binary. It is crucial to specify the target as it defines what the model is supposed to learn and how to act. [16]-[18]

C. Feature Engineering

Medically meaningful representations were introduced by applying feature engineering. BMI was sorted into clinically meaningful categories and age was recategorized into age groups. Some features were squared and log transformed to account for the nonlinear growth patterns. The reason is risk is not always in direct proportion. The effect of age may be more pronounced in the older age groups, for instance, and the effect of BMI on blood pressure may be different at various levels of BMI. [16], [17]

D. Scaling and Class Balancing

Standard scaler was used to standardise continuous variables. Algorithms like Logistic Regression require scaling, and stable optimization in gradient-based settings. While tree-based methods are less sensitive to scale, a consistent preprocessing strategy is important for making the models comparable.

The dataset was imbalanced with SMOTE performed on it. SMOTE generates minority class examples by interpolating between existing examples of the minority class. This combats the model's tendency to disregard the cases in which the individuals present with hypertension and it helps the model adopt a more balanced classification boundary. In order to avoid information leakage to the test set, SMOTE had to be used only on the training dataset. [10], [23]

VI. EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) was conducted to gain insight of the dataset's structure before modelling. The class distribu-

TABLE II
PREPROCESSING PIPELINE

Step	Problem Addressed	Method Used
Data cleaning	Missing and inconsistent values	Median imputation
Target labeling	Clinical binary class definition	SBP ≥ 140 or DBP ≥ 90
Feature scaling	Unequal numerical ranges	StandardScaler
Feature engineering	Nonlinear risk patterns	BMI category, age group, transformations
Class balancing	Minority hypertensive class	SMOTE on training set
Validation	Overfitting control	3-fold randomized cross-validation

tion, age-risk pattern, BMI/weight relationship, and the correlation between original and engineered variables were the items examined. The importance of EDA is that it gives evidence towards preprocessing decisions and to check if relationships between the selected features and hypertension are plausible.

The first big observation was that it was a class imbalance problem. The hypertensive cases are fewer in number compared with the majority (normal class). That result showed that accuracy is not enough to assess model performance. A classifier that performs well when asked to predict the majority class is generally considered accurate but he would not be a clinically accurate classifier if he fails to identify high-risk individuals.

There was a strong association between age and prevalence of Hypertension. The proportion of cases hypertensive was higher in older individuals than in younger individuals. This trend is physiologically consistent, as for older age, vascular stiffness, changes in metabolism and lifetime exposure to other risk factors tend to rise. BMI and body weight also were significantly associated with hypertension. Those with greater BMI were more likely to have increased hypertension risk, thus highlighting the importance of body composition as a predictor. As is seen in Fig. 1, prevalence of hypertension increases with age, especially in middle and older age groups. The visual pattern facilitates the incorporation of age (and age-related variables) into the model.

BMI is highly correlated with weight, and age and some lifestyle factors add information. The correlations alone are not enough to predict, but they do show useful relationships which lead to the training of the models.

The relationship between engineered variables is shown in Fig. 3. The correlations between BMI, weight, and the transformations indicate a high level of correlation between these elements, suggesting that feature engineering brings in structured data. While redundancy in some models does not make much sense, in others it can be valuable to build useful decision rules with highly correlated features.

There are also graphics to visualize the numerical measures in pairs (Fig. 4). There is no separation between the classes by only a single feature: the hypertensive and non-hypertensive

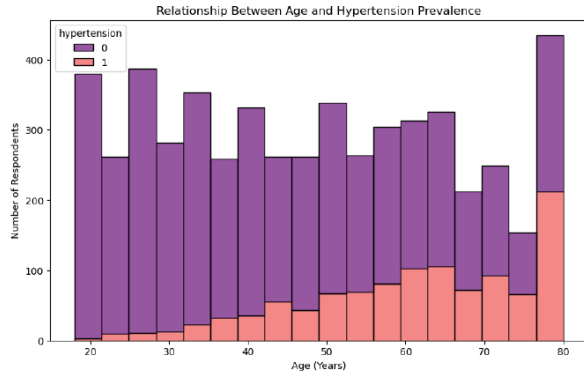


Fig. 1. Relationship between age and hypertension prevalence.

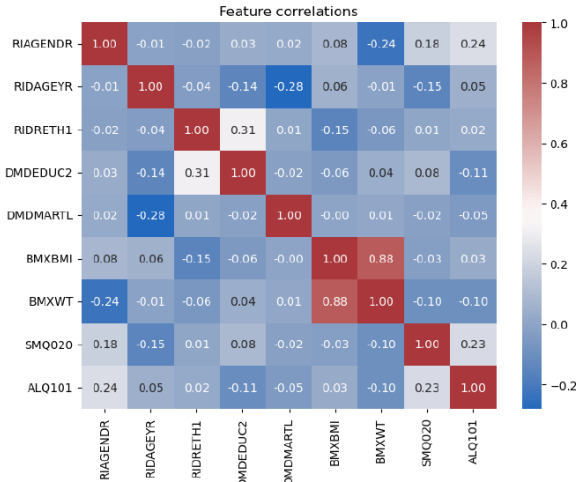


Fig. 2. Correlation matrix of original features.

classes overlap. This overlap is sufficient reason for using multivariate machine learning algorithms. [29]-[31]

VII. METHODOLOGY

The methodology involves preparing the data, training the model, hyperparameter optimization, and evaluation with multiple metrics. Observed differences in performance has been achieved by training all models with the same data split as well as preprocessing logic, thus eliminating the possibility that differences are due to varying experimental settings. [25], [28]

A. Classification Algorithms

5 classification algorithms were chosen. Logistic Regression was chosen as the baseline model because it is simple, interpretable and commonly used in clinical modeling. Random Forest was added as an ensemble based on ‘bagging’ which is similar to boosting but with the advantage of being able to capture nonlinear interactions while minimizing variance. Gradient Boosting was added as a representation of sequential error correction. XGBoost was chosen because of its regularized boost-

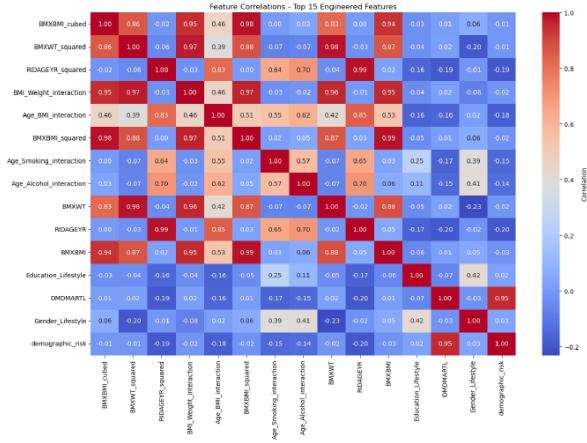


Fig. 3. Correlation matrix of engineered features.

ing approach and high performance on tabular data. It was decided to use CatBoost due to its resistance to prediction shift and robustness in dealing with categorical-style features.

B. Hyperparameter Optimization

Hyperparameter tuning was carried out by using Randomized-SearchCV with 3-fold cross-validation. This technique involves trying out various sets of hyperparameters with cross validation over a specified set of hyperparameters. The number of estimators, max tree depth, learning rate, subsampling rate, regularization strength and splitting criteria were tuned depending on the model. But F1 score was given importance during tuning as it balances precision and recall. [28]

C. Evaluation Metrics

The chosen evaluation metrics are related to the clinical screening goal. Under class imbalance, accuracy can be misleading as it is the percentage of correct predictions. The reliability of positive predictions is precision. Recall is the percentage of patients that the model correctly identifies as having hypertension. The F1-score gives a harmonic average of the precision and the recall. Cohen’s Kappa is a measure of agreement that accounts for chance agreement. ROC-AUC does a threshold-independent evaluation of how well the model segregates classes. [32], [33]

Recall is particularly of importance in the context of a hypertension screening. Failure to identify an individual with hypertension can occur and result in a delay in treatment and the development of potentially serious cardiovascular consequences. A false positive, albeit a nuisance, can often be resolved by subsequent measurement. Thus, the accuracy of the best model is not the highest, but the model that offers the optimum combination for clinical utility. [26], [27]

VIII. IMPLEMENTATION WORKFLOW

Workflow steps were designed to avoid data leakage and to maintain the comparability of the implementation. The dataset was initially split into training and test sets. Imputation, scaling and

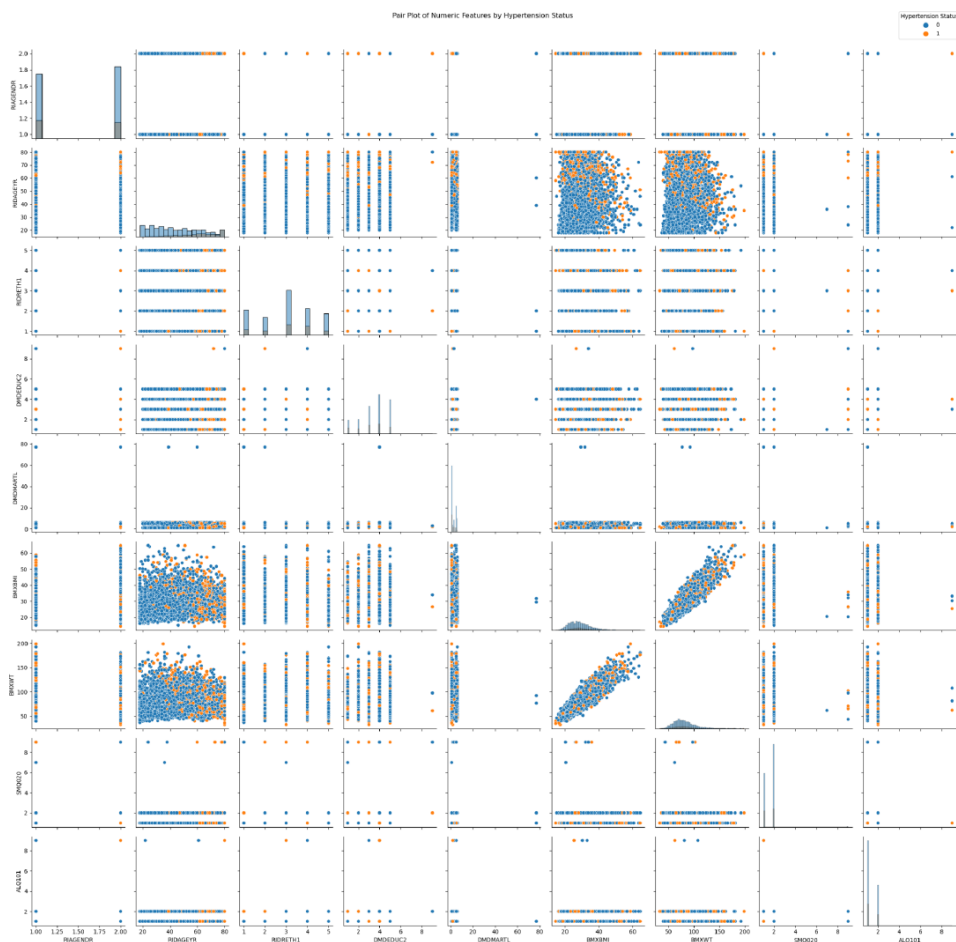


Fig. 4. Pair plot of selected numerical features.

SMOTE were then applied to the training part alone. The same learned preprocessing transformation was used for the test set, but the SMOTE technique was not used for test data. This design is significant due to the fact that the synthetic examples shouldn't skew the unseen evaluation set. [23]

There are six key stages to the overall workflow. The raw demographic, examination and questionnaire variables were loaded, and examined in the first phase. In the second phase, clinically meaningful labels were derived from the blood pressure measurements, systolic and diastolic. The third step was to fill in the missing values and to derive various features. In the fourth stage, SMOTE was used to balance the training set. Each algorithm was optimised using a randomized cross validation in the fifth stage. The optimized models were run on the hold-out test set with the selected metrics in the final stage. [32], [33]

This threshold determines the classification of predicted probabilities into classes and thus affects clinical outcomes. Precision may be enhanced with a high threshold but it may underestimate the number of patients with true hypertension. A lower threshold can improve recall, but also contribute to increased

false positives. With early screening, it is better to use threshold tuning, which should include sensitivity first, as follow-up blood-pressure measurement can confirm the suspected cases. In the present study, the optimized classification results are reported; however, the selection of the thresholds for clinical use needs to take into account the cost of false negatives and the resources available for screening. [24]-[27]

To ensure reproducibility, the selection of pre-processing options should be reported. Robustness to outliers led to the choice of median imputation. Scaling up to standard was applied for consistency with algorithms. SMOTE was only applied during the training process as we don't want to artificially boost the accuracy of the test sets. Randomized search was chosen over exhaustive grid search to search broadly over ranges of hyperparameter choices with lower computational demands. The framework can be extended to larger data sets or other algorithms. [28], [31]

IX. MODEL INTERPRETATION CONSIDERATIONS

In healthcare, it is crucial to see into model behaviours. A model without an explanation of the risk label will be less believable to

TABLE III
EXPERIMENTAL WORKFLOW FOR REPRODUCIBLE TRAINING

Stage	Operation
1	Load NHANES records and select demographic, clinical, and lifestyle predictors.
2	Create hypertension label using SBP and DBP threshold rule.
3	Apply median imputation, scaling, and feature engineering.
4	Split data and apply SMOTE only to the training subset.
5	Tune LR, RF, GBC, XGB, and CATB using randomized 3-fold cross-validation.
6	Evaluate the tuned models on the independent test set using accuracy, recall, precision, F1-score, Kappa, and ROC-AUC.

a clinician or the patient. Medically meaningful behavior is suggested from the patterns of features that were observed during EDA as well as during model analysis, though these patterns are not the primary focus of the present study. The age, BMI, and weight factors emerged as important factors repeatedly, mirroring known clinical associations with hypertension.

To use the model for operations, explanation methods should be integrated into the trained model. Global explanations can indicate what variables have the strongest influence in predicting the population. Local explanations can be provided to explain why a specific subject was selected for their high-risk score. For instance, if someone is elderly and has a high BMI (body mass index) level and low levels of physical activity, an individual could be flagged. These explanations can even turn an output from a machine learning model into preventative advice. [34], [35]

Interpretability also is useful to identify model failure. Explanation tools can expose the problem when a model is very dependent on an irrelevant variable or a biased variable. This is particularly relevant in public-health settings where demographic factors may relate to social or structural factors. Future systems need to be predictive, offer some form of fairness analysis, be evaluated for calibration, and have the system reviewed by a clinician. [24], [35]

X. RESULTS

The experimental results show that ensemble methods that use trees perform well in the screening-oriented task, compared to the linear baseline, but give moderate absolute positive-class recall. Overall, in this experiment, XGBoost had the best test F1 score (0.2579) and test recall (0.1814), indicating that it was the strongest relative performer. Random Forest had the best CV F1 score of 0.7582, indicating consistent model-selection performance. The highest accuracy was obtained using Logistic Regression (0.7972), and the precision (0.5800), indicating that more numbers of hypertensive cases were missed by Logistic Regression, as can be seen from its recall (0.1283). [32], [33]

Thus, the current experimental comparison of XGBoost on the test set does not necessarily indicate that XGBoost is the best model, but rather the best model among the current model set in the experimental comparison. The findings call for more

research to be conducted on thresholds tuning, probability calibration, the cost-sensitive learning and external validation. The optimized model results are summarized in Table IV and the behavior of the ROC curves across classifiers is shown in Fig. 5.

XI. DISCUSSION

Results show that hypertension prediction based on the structured survey variables is a nonlinear and imbalanced classification problem. XGBoost and Random Forest model performances are relatively better, indicating that interactions between age, BMI, weight, and lifestyle factors are better captured by tree-based ensembles than by Logistic Regression. Meanwhile, the small recalls indicate that the existing feature set combined with the default decision thresholds are not adequate for direct clinical screening.

This is a meaningful scientific discovery: A model might look like it was accurate, but it may not capture many positive cases. Hence, the choice of a model for healthcare screenings should be based on the clinical objective, false-negative cost, calibration and validation quality (but not only on accuracy).

According to the feature analysis, age, BMI, and weight are always significant predictors. This is physiologically reasonable and reasonable for the modeling pipeline to be valid. Age captures changes at the physiological level and the exposure to risk factors. BMI and weight are indicators of body composition and metabolic load. The engineering of variables like age groups and BMI categories imparts clinical interpretability, and the conversion of continuous variables into relevant risk bands.

The other lesson, much more important, relates to the requirement for metric-based model selection. Imbalanced healthcare data could encourage a model to overly predict the dominant class. Recall and F1-score can better depict the utility of a screening test, however. Safety for clinical use is not guaranteed for the model with the highest accuracy. Although Logistic Regression achieved a small margin of improvement in terms of accuracy, XGBoost was chosen as it was more effective in identifying positive cases.

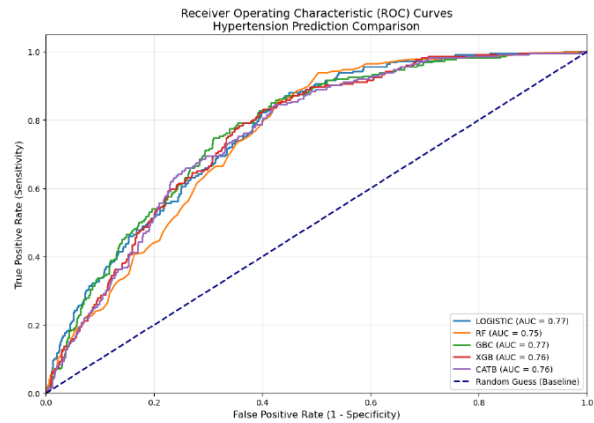


Fig. 5. Receiver Operating Characteristic curves for hypertension prediction models.

TABLE IV
COMPARATIVE PERFORMANCE OF OPTIMIZED MODELS

Model	Best CV F1	Accuracy	Recall	Precision	F1-score
XGBoost (XGB)	0.752897	0.780465	0.181416	0.445652	0.257862
Random Forest (RF)	0.758196	0.786977	0.159292	0.480000	0.239203
CatBoost (CATB)	0.756879	0.783256	0.150442	0.453333	0.225914
Logistic Regression	0.745714	0.797209	0.128319	0.580000	0.210145
Gradient Boosting	0.754706	0.786977	0.123894	0.474576	0.196491

Optimized parameter settings: XGB: subsample = 0.8, n_estimators = 150, max_depth = 3, learning_rate = 0.1; RF: n_estimators = 150, min_samples_split = 2, max_depth = None; CATB: random_strength = 10, n_estimators = 200, learning_rate = 0.1, depth = 6; LR: solver = liblinear, C = 0.0048329, penalty = l2; GBC: n_estimators = 100, max_depth = 3, learning_rate = 0.1.

XII. CLINICAL RELEVANCE AND DEPLOYMENT CONSIDERATIONS

A practical hypertension risk-prediction model can facilitate the preventive health-care approach by prioritizing individuals for repeated blood-pressure monitoring, counselling on lifestyle and physical examination. The current model, however, should be viewed as research level decision support. The deployment would involve external validation, external calibration, assessment of the subgroups for fairness, review by clinicians, and testing.

These techniques may also be integrated into provide local and global explanations to aid understanding of the contribution of age, BMI, weight and lifestyle to a risk score. The model is not to be used to deny care, make a substitution diagnosis or to make unsupported medical conclusions. The safest use is to support the triage and after confirmed blood pressure is taken.

XIII. PRACTICAL SCREENING SCENARIO

One of the potential applications of the proposed framework is to provide preliminary risk screening during community health camps, at university health centers, corporate wellness programs, or mobile health apps. In these contexts, direct measurement of blood pressure might only be possible on an occasional basis, whilst simple demographic and anthropometric data is able to be gathered in a swift manner.

These variables can be used by a trained model to generate a risk score which will inform decisions regarding repeated measurement or medical follow-up of individuals. It is strongly recommended to do no classification of a patient from the model output: the workflow is to classify the patient rather from the result of applying the model. Rather, the model is meant as a “triage layer.” Those who are deemed at high risk may be recommended to have their blood pressure monitored more frequently, discuss lifestyle modifications, and have a doctor’s review. Regular health checks should be encouraged in individuals who are predicted to be low risk as the health check is a model and not a disease predictor.

When the recall is less than perfect, this is a particularly important triage interpretation. If the prediction is positive, consider this a screening alert, rather than a diagnosis. Negative prediction does NOT constitute a permanent clearance. This way of deployment can minimize the possibility of negative outcomes and make use of the predictive model anyway.

The model can also be applied to the population level in public-

TABLE V
SUGGESTED USE OF MODEL OUTPUT IN SCREENING

Model Output	Suggested Action
High-risk score	Repeat blood-pressure measurement, check clinical history, advise lifestyle counseling, and recommend physician review.
Moderate-risk score	Schedule follow-up screening and monitor BMI, activity, diet, and related risk factors.
Low-risk score	Continue routine monitoring and preventive health education; do not use as permanent exclusion from risk.
Uncertain result	Request additional measurements or laboratory data before decision support is used.

health deployment. Health administrators can use this to target resources to screening and direct them to individuals with a higher expected probability of having a disease. For instance, if there is an older person with a higher BMI, they could be directed towards an awareness session, periodic blood-pressure camp, or digital reminder. Applications such as these demonstrate the potential of predictive modeling in supporting preventive healthcare beyond the one-to-one clinical decision-making.

XIV. QUALITY AND VALIDATION REQUIREMENTS

Prior to real-world application, a prediction model for hypertension needs to meet a number of validation conditions. To check that the model isn’t overfitting to the training data, it should be first validated internally. Cross validation is used to assess stability, but is not an external test. Secondly, performance should be tested on another dataset obtained from a different population or time period. Third, one would want to test the validity of the calibration to ensure that the forecast risk probability is equal to the observed hypertension rate.

Fourth, test for performance of subgroups. Model results can vary by age, gender or socio-demographic. Where there is poor performance for a sub group, deployment can result in unequal benefit or risk. Fifth, the decision threshold needs to be determined in conjunction with the clinician. The threshold value that optimizes a mathematical measure might not be the best screening policy threshold. [24], [27]

There should also be monitoring of model's post-deployment. The distribution of certain population characteristics may evolve over time, as may the link between the predictors and hypertension status. This is called Data Drift. A good model that is effective in development can lose its effectiveness over time if it is not reevaluated. Hence, any decision-making system should feature periodic training, audit logs and performance dashboards.

XV. COMPARATIVE INTERPRETATION OF ALGORITHMS

The five algorithms are different philosophies behind the models. Logistic Regression makes some basic assumptions about the relationship between the predictors and log-odds of hypertension. Good in cases where interpretability is paramount, but poor in cases of a highly interactive feature. Random forest includes multiple decision trees and takes a voting approach to the predictions, making it more flexible. This will decrease the variance and enable complex decision boundary.

Gradient Boosting and XGBoost will learn sequentially. Every tree tries to correct the errors made by the previously created tree. This is suitable for structured health data as there are subtle feature interactions. This can be further enhanced by a process of regularization and efficient optimization in XGBoost. When categories (or category-derived variables) are present, CatBoost proves to be a valuable tool.

The results of this study showed that XGBoost model with a trade-off between nonlinear learning and regularization proved to be the most effective model that can be applied in screening purposes. It is important to keep in mind that the comparative result must be evaluated as part of the entire context. There is no one best model to fit all medical datasets. In cases where explanation is more important than recall, a transparent model might be desired. When screening and/or high discrimination is desired, a boosting model might be more appropriate. The final clinical system can also be a composite system which utilise a simpler model to explain the cause of the disease and a more powerful ensemble for risk scoring.

XVI. LIMITATIONS

There are a number of limitations to the study. First, only variables that are available in the NHANES data set were included in this experiment, so important predictors like family history, medications, diet, stress, longitudinal blood-pressure patterns and laboratory markers may be missing or incomplete. Secondly, some variables may be self-reported which can lead to reporting bias. Third, the target label is based on BP targets and may not be repeated clinical confirmation. Fourth, while SMOTE is good for representation of minority classes during training, it doesn't mean it will lead to clinically adequate recall of unseen minority-class populations. Fifth, modest positive class-recall remains in the best model, and, thus, the present system would not be suitable for stand-alone clinical use. Sixth, external validation, calibration, and confidence intervals must be requested to establish robustness; and subgroup-fairness analysis is required to establish robustness. [25], [27]

XVII. FUTURE WORK

Future research needs to focus on enhance sensitivity with reasonable precision. Steps to be taken include optimizing thresholds, learning to be cost-sensitive, reporting probabilities in calibrated probability intervals, reporting confidence intervals by bootstrapping and/or repeated cross validation, and external validation on independent cohorts. Other individual factors like medication history, repeated blood pressure readings, dietary factors, family history, laboratory factors, and longitudinal clinical data can enhance discrimination. [16]-[18]

Explainable AI techniques like SHAP, LIME, partial-dependence analysis, and decision-curve analysis should be incorporated to better understand and measure the net effect of screening in the clinical context, with the goal of enhancing clinical interpretation. A prototype of a decision support system in the future should output the risk of being diagnosed as a result of the screening, a level of confidence about the result, features noted as important to look for, and a clear recommendation of what to do next, all without claiming to be a diagnosis. [34], [35]

XVIII. CONCLUSION

This paper provided a comparative study based on the machine-learning method with NHANES data to predict hypertension risk. The entire preprocessing pipeline was developed, ranging from median imputation, standard scaling, clinically relevant feature engineering, class balancing in training data through SMOTE, and randomised hyperparameter optimisation. Five algorithms were assessed using a shared framework and metrics suitable for imbalanced classification in healthcare. [23], [32], [33]

The XGBoost model performed best on the test set with an F1 score and a high recall rate, the Random Forest model was relatively stable across the cross validation, and the Logistic Regression model was interpretable and had a reasonable recall rate as a baseline. The findings corroborate the use of ensemble and boosting methods for identifying nonlinear hypertension-risk patterns but indicate the importance of tuning the threshold, calibration, more comprehensive predictors and external validation. The study should really be viewed as a research framework rather than a diagnostic system because it can be repeated to establish a screening-oriented research-based decision support system.

REFERENCES

- [1] World Health Organization, *Global Report on Hypertension: The Race Against a Silent Killer*. Geneva, Switzerland: WHO, 2023.
- [2] Centers for Disease Control and Prevention, National Center for Health Statistics, *National Health and Nutrition Examination Survey (NHANES): Questionnaires, Datasets, and Related Documentation*, 2024.
- [3] F. Lopez-Martinez, E. R. Nunez-Valdez, R. Gonzalez-Crespo, and V. Garcia-Diaz, "An artificial neural network approach for predicting hypertension using NHANES data," *Scientific Reports*, vol. 10, article 10620, 2020, doi: 10.1038/s41598-020-67640-z.

- [4] G. F. S. Silva, T. P. Fagundes, B. C. Teixeira, and A. D. P. Chiavegatto Filho, "Machine Learning for Hypertension Prediction: A Systematic Review," *Current Hypertension Reports*, vol. 24, no. 11, pp. 523-533, 2022, doi: 10.1007/s11906-022-01212-6.
- [5] M. Z. I. Chowdhury et al., "Prediction of hypertension using traditional regression and machine learning models: A systematic review and meta-analysis," *PLOS ONE*, vol. 17, no. 4, e0266334, 2022, doi: 10.1371/journal.pone.0266334.
- [6] R. I. Estiko et al., "Screening hypertension using non-laboratory risk factors with machine learning: A retrospective cross-sectional study in Indonesia," *BMJ Open*, vol. 15, no. 8, e092364, 2025.
- [7] K. Dwivedi, R. Lakshmanan, and R. Regunathan, "K-means Under-Sampling for Hypertension Prediction using NHANES Dataset," *International Journal of Performability Engineering*, vol. 17, no. 8, pp. 733-740, 2021, doi: 10.23940/ijpe.21.08.p9.733740.
- [8] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.
- [9] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785-794, doi: 10.1145/2939672.2939785.
- [10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321-357, 2002.
- [11] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.
- [12] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37-46, 1960.
- [13] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861-874, 2006.
- [14] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [15] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You? Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135-1144.
- [16] P. K. Whelton et al., "2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults," *Hypertension*, vol. 71, no. 6, pp. e13-e115, 2018, doi: 10.1161/HYP.0000000000000065.
- [17] T. Unger et al., "2020 International Society of Hypertension global hypertension practice guidelines," *Hypertension*, vol. 75, no. 6, pp. 1334-1357, 2020, doi: 10.1161/HYPERTENSIONAHA.120.15026.
- [18] B. Williams et al., "2018 ESC/ESH guidelines for the management of arterial hypertension," *Eur. Heart J.*, vol. 39, no. 33, pp. 3021-3104, 2018, doi: 10.1093/eurheartj/ehy339.
- [19] NCD Risk Factor Collaboration, "Worldwide trends in hypertension prevalence and progress in treatment and control from 1990 to 2019," *Lancet*, vol. 398, no. 10304, pp. 957-980, 2021, doi: 10.1016/S0140-6736(21)01330-1.
- [20] GBD 2019 Risk Factors Collaborators, "Global burden of 87 risk factors in 204 countries and territories, 1990-2019," *Lancet*, vol. 396, no. 10258, pp. 1223-1249, 2020, doi: 10.1016/S0140-6736(20)30752-2.
- [21] P. M. Kearney et al., "Global burden of hypertension: analysis of worldwide data," *Lancet*, vol. 365, no. 9455, pp. 217-223, 2005, doi: 10.1016/S0140-6736(05)17741-1.
- [22] K. T. Mills et al., "Global disparities of hypertension prevalence and control: a systematic analysis of population-based studies from 90 countries," *Circulation*, vol. 134, no. 6, pp. 441-450, 2016, doi: 10.1161/CIRCULATIONAHA.115.018912.
- [23] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263-1284, 2009, doi: 10.1109/TKDE.2008.239.
- [24] G. Van Calster et al., "Calibration: the Achilles heel of predictive analytics," *BMC Med.*, vol. 17, art. 230, 2019, doi: 10.1186/s12916-019-1466-7.
- [25] D. G. Altman and P. Royston, "What do we mean by validating a prognostic model?," *Stat. Med.*, vol. 19, no. 4, pp. 453-473, 2000, doi: 10.1002/(SICI)1097-0258(20000229)19:4<453::AID-SIM350>3.0.CO;2-5.
- [26] A. J. Vickers and E. B. Elkin, "Decision curve analysis: a novel method for evaluating prediction models," *Med. Decis. Making*, vol. 26, no. 6, pp. 565-574, 2006, doi: 10.1177/0272989X06295361.
- [27] E. W. Steyerberg et al., "Assessing the performance of prediction models: a framework for traditional and novel measures," *Epidemiology*, vol. 21, no. 1, pp. 128-138, 2010, doi: 10.1097/EDE.0b013e3181c30fb2.
- [28] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, pp. 281-305, 2012.
- [29] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5-32, 2001, doi: 10.1023/A:1010933404324.
- [30] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189-1232, 2001.
- [31] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York, NY, USA: Springer, 2009.
- [32] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proc. 23rd Int. Conf. Machine Learning*, 2006, pp. 233-240, doi: 10.1145/1143844.1143874.
- [33] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS ONE*, vol. 10, no. 3, e0118432, 2015, doi: 10.1371/journal.pone.0118432.
- [34] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nat. Mach. Intell.*, vol. 1, pp. 206-215, 2019, doi: 10.1038/s42256-019-0048-x.
- [35] J. Amann, A. Blasimme, E. Vayena, D. Frey, and V. I. Madai, "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective," *BMC Med. Inform. Decis. Mak.*, vol. 20, art. 310, 2020, doi: 10.1186/s12911-020-01332-6.