COMPUTATIONAL INTELLIGENCE AND MACHINE LEARNING
e-ISSN: 2582-7464

Computational Intelligence and Machine Learning
Volume 6, Issue 1, April 2025

Evaluating the Performance of Ensemble and Single Classifiers with Explainable Artificial Intelligence (XAI) on Hypertension Risk Prediction

Victor Wandera Lumumba[1*], Teddy Mutugi Wanjuki[2], Elizabeth Wambui Njoroge[3]

[1, 2, 3] Faculty of Science and Technology Chuka University, Kenya
*Corresponding Author Email: lumumbavictor172@gmail.com

Abstract

Hypertension remains a critical health issue, and complications such as cardiovascular disease, stroke, and renal failure similarly remain a global health concern. This study compared six supervised machine learning models – Support Vector Machines, k-nearest Neighbors, Random Forest Classifier, Naïve Bayes Classifier, Tree Bagging, and Extreme Gradient Boosting, based on the data from 2322 participants. The primary elements were SBP measured as equal to or more than 120 mmHg, BMI, Age, and the number of haemoglobin grams per litre, as well as demographic data. The research found that Random Forest yielded the highest evaluation metrics in Oversampling, with an accuracy of 100 %, balanced Accuracy of 100 %, Sensitivity of 100 %, specificity of 100 %, and AUC of 100 %; hence proved to be the best model to address the hypertension risk among patients. The feature importance of the SBP turned out to be higher according to the SHAP analysis, considering the "No" class where the SHAP value equalled 0.24, followed by BMI (0.05) and Gender (0.06). Variables such as advanced HIV status and log-centered creatinine showed negligible impact (SHAP value = 0.00). The random forest model was accurate and steady across all performance criteria, outperforming all other models with the No Information Rate (0.978) while illustrating the significance of physiological aspects of hypertension risk assessment. These results demonstrate the capability of Random Forest in predicting hypertension risk and give important suggestions for enhancing screening methods and specific public health initiatives.

Keywords

Cardiovascular disease, Hypertension, Machine learning, Supervised learning, Sensitivity, Specificity, Oversampling.

INTRODUCTION

Hypertension, commonly known as high blood pressure, is a chronic medical condition that significantly increases the risk of cardiovascular diseases, stroke, and kidney failure [1] [2]. Referred to as the "silent killer," hypertension often remains asymptomatic until severe complications arise, making early detection and intervention crucial [3]. The global prevalence of hypertension has been rising due to factors such as ageing populations, urbanization, obesity, and lifestyle behaviours [4]. Addressing this growing public health challenge requires robust predictive models to improve screening and early detection efforts. Recent machine learning (ML) advances have facilitated hypertension risk prediction by incorporating diverse demographic, physiological, and lifestyle factors [5]. ML models provide superior Accuracy in identifying high-risk individuals compared to traditional statistical methods, enabling personalized treatment plans and better healthcare outcomes [6]. Supervised learning techniques, including ensemble methods like Random Forest and Extreme Gradient Boosting (XGBoost), have shown promising results in health-related predictive modelling [7]. This study evaluates the performance of various ML classifiers, including Support Vector Machines (SVM), k-nearest Neighbors (k-NN), Naïve Bayes, Tree Bagging, and XGBoost, in predicting hypertension risk using real-world data. Machine learning has been an attractive approach to integrate multiple predictors into robust models.

However, existing studies primarily focused on single classifiers or ensembles without comprehensive comparisons. It is important to note that most ML models act like black boxes, limiting their interpretability. This study fills these gaps by evaluating six machine learning models, incorporating Explainable AI techniques, and handling class imbalance problems for improved hypertension prediction, with enhanced interpretability of the predictive results. While machine learning models have advanced incredibly in medical diagnosis, these models lack interpretability and hence are not widely adopted in healthcare due to the difficulty in interpretability. This study investigates these issues through SHAP analysis for explainability and SMOTE for addressing class imbalance. The existing literature relies heavily on physiological data; however, this study combines demographic, clinical, and environmental factors for overall risk assessment. These contributions enhance model transparency, fairness, and generalization for hypertension prediction.

The following objectives guided the study:

1. To develop and evaluate machine learning models for hypertension prediction.
2. To compare the performance of ensemble and single classifiers in hypertension risk prediction.
3. To identify key demographic, clinical, and environmental factors associated with hypertension risk.

LITERATURE REVIEW

Multiple physiological and sociodemographic factors influence hypertension. Elevated systolic blood pressure (SBP) and diastolic blood pressure (DBP) remain the primary diagnostic criteria, with clinical guidelines defining hypertension as SBP ≥ 120 mmHg or DBP ≥ 90 mmHg [8]. Several studies highlight the impact of Age and BMI as critical predictors of hypertension [9]. Older adults and individuals with higher BMI face increased hypertension risk due to arterial stiffness and metabolic changes [10]. Gender disparities are also evident, with men exhibiting higher risks until women reach postmenopausal Age, after which risk disparities narrow [11]. Urban residency has been associated with elevated hypertension risks due to dietary patterns, sedentary lifestyles, and environmental stressors [12].

The application of machine learning in hypertension prediction has gained traction due to its ability to process complex datasets efficiently. Ensemble models such as Random Forest, XGBoost, and Tree Bagging enhance predictive Accuracy by minimizing variance and bias through multiple decision trees [13]. Studies have shown that Random Forest performs exceptionally well in handling heterogeneous clinical data, yielding superior classification metrics compared to linear models [14]. Support Vector Machines (SVM) and k-NN offer effective alternatives for hypertension classification, particularly in high-dimensional datasets [15]. Explainable Artificial Intelligence (XAI) techniques, such as SHAP (Shapley Additive Explanations), enhance model interpretability by identifying key risk factors contributing to predictions [16]. SHAP values have been employed to evaluate the significance of SBP, BMI, and haemoglobin levels in predicting hypertension [17]. Feature importance rankings from SHAP analysis allow healthcare practitioners to make informed decisions regarding early screening and risk stratification [18].

Several studies have compared the performance of different ML algorithms in hypertension risk assessment. In their study, [19] found that Random Forest outperformed SVM, k-NN, and Naïve Bayes in hypertension classification, achieving an accuracy of 92.5% [20]. In their meta-analysis, [21] reported that ensemble models, particularly XGBoost and Tree Bagging, exhibited higher predictive Accuracy than traditional statistical models [22]. Data imbalance remains a challenge in ML-based hypertension prediction. Oversampling techniques, such as the Synthetic Minority Over-sampling Technique (SMOTE), have enhanced model performance, particularly in datasets with imbalanced class distributions [23]. Our study builds upon these findings by implementing and comparing multiple ML models while integrating XAI for enhanced interpretability.

METHODS AND MATERIALS

Research Design

This study employs a Retrospective Cohort Design to model hypertension risk factors using machine learning techniques on data sourced from the Academic Model Providing Access to Healthcare (AMPATH) electronic medical records (EMR). This design is well-suited for the study as it systematically explores historical data on patients with and without hypertension, providing a foundation for building and validating predictive models based on pre-existing patient information. This design allows the investigation of the relationship between various demographic, clinical, and socio-environmental factors and hypertension outcomes.

Source of the Data

AMPATH's EMR provides comprehensive data on patient histories, capturing essential variables such as Age, Gender, Body Mass Index (BMI), HIV status, marital status, haemoglobin levels, serum creatinine levels, and other relevant health metrics. These records allow us to analyze hypertension-associated patterns without needing prospective follow-up, making the study both time- and cost-effective. Based on their medical records, the design allows us to categorize individuals into those who developed hypertension and those who did not and then analyze the predictor variables that may have contributed to this outcome. This approach enables the effective use of machine learning to model hypertension risk factors by leveraging a large dataset with comprehensive patient information.

Data Analysis

Cross Validation Scheme

Leave One Out Cross Validation (LOOCV)

LOOCV approach takes every observation in the data set as the validation set and N-1 as the training set. This is done for the entire sample size (N) [24]. In this method, assume that we have the dataset *D*, where;

$$D = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_n, y_n)\}. \quad (1)$$

$$LOOCV\ Error = \frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{y_i}) \quad (2)$$

In this scheme, L is the loss function, where the common loss function is expressed as the Mean Square Error, as shown in equation 3 [25].

$$L(y_i, \hat{y_i}) = (y_i, \hat{y_i})^2 \quad (3)$$

Graphically, leaving one outside cross-validation is expressed as shown below.

Figure 1. Leave One Outside Cross Validation

k-Folds Cross Validation

Consider the data set labelled D expressed as shown in equation 2.1. In this expression, x_i represents the features while y_i represents the corresponding labels of the outcome variable with *i* iterations. In this approach, the error from the k-fold cross-validation scheme is calculated as shown in Equation 5

$$K - Fold\ CV\ Error = \frac{1}{k}\sum_{i=1}^{k} E_i \quad (4)$$

Where;

$$E_i = \frac{1}{|D_i|}\sum_{j=1}^{|D_i|} L(y_{ij}, \hat{y}_{ij}) \quad (5)$$

The graphical representation of k-fold cross-validation is shown in Figure 2 below [26].

Figure 2. K-folds Cross Validation

Repeated k-folds Cross Validation

Repeated k-folds cross-validation is an extension of k-folds cross-validation where folds are iteratively shuffled and repeated during the training process [27]. The process involves repeating the k-fold cross-validation several times, and the means of the results are reported [28].

Hyperparameter Tuning

Parameter optimization, commonly known as hyperparameter tuning, is a common practice in developing machine learning models, which helps obtain a well-performing machine learning model [29]. This study optimized the parameters for different machine learning models to obtain a highly performing model for predicting hypertension risk. The parameters for the support vector machines were optimized by setting the regularization parameter *C* with values 0.1, 1, 10, and 100. The study controlled the Gaussian function in the Radial Basis Function (RBF) kernel with sigma parameters set as 0.01, 0.1, and 1. In the k-Nearest Neighbours, the parameters are optimized by selecting the appropriate number of neighbours denoted by *k*. In this study, selecting the appropriate number of neighbours was done by using a grid search for the value of *k* between 1 and 21 with an interval of 2. The mtry parameter determined the performance of the random forest model. Selecting the appropriate value of *mtry* is very crucial, which was found using the formula;

$$mtry = \sqrt{P} \quad (6)$$

Where P is the number of features. Additionally, the number of trees in the forest was set to 500, which is usually considered the appropriate number of trees for a random forest model [30]. In order to enhance the performance of the extreme gradient boosting, the learning rate (eta) was tuned to control the update size in the model-building process. Generally, a lower learning rate increases Accuracy but requires more boosting rounds [31]. Additional parameters tuned in this study include max_depth, min_child_weight, subsample, solsample_bytree, lambda, and alpha. Max_depth controls the complexity of the model and overfitting while controlling the number of splits that can result in overfitting the model. Tuning the subsample and colsample_bytree parameters helped attain randomness and the model's robustness. Finally, tuning alpha and lambda parameters helped in the model regularization. Naïve Bayes has few parameters to tune and is less complex than other machine learning models [32]. However, since all the features in this study were continuous, Gaussian Naïve Bayes was chosen for its suitability to handle continuous data, assuming features follow a Gaussian distribution. Lastly, none of the parameters were optimized in the tree bag model since the algorithm focuses on data preprocessing rather than accuracy [26].

Data Balancing

Addressing the class imbalance in this research was done using resampling techniques, including under-sampling, oversampling, and hybrid sampling. Undersampling was applied to reduce the majority class by randomly removing some instances to attain a balanced instance for the positive and negative hypertension risk outcome. On the other hand, Oversampling was applied to create synthetic instances for the minority class by retaining the original instance from the majority class. A hybrid sampling technique was used, which combines both under-sampling and oversampling, which helps attain a more optimal balance using the Synthetic Minority Oversampling Technique (SMOTE). Balancing the data helped achieve a more robust model with improved generalization and a higher predictive accuracy on minority classes [33].

Shapley Additive Explanations

The predictive models employed Shapley Additive Explanations (SHAP) to interpret the complex interactions between BMI, blood pressure levels, and additional health indicators. By calculating Shapley values, SHAP provided insights into the specific contribution of each variable in influencing hypertension risk predictions, illustrating how variables like HIV status, urban clinic attendance, or ARV treatment history impact individual risk scores.

Machine Learning Algorithms

Support Vector Machines (SVM)

During the Support Vector Machines model training process, there are two optimization problems to be solved, which appear in two forms, primal and dual form [34], as shown in equations 7 and 9.

Primal form;

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \quad (7)$$

Solving this problem is subject to the following conditions;

$$y_i(w \cdot x_i + b) \geq 1 - \xi i,\ \geq 0, i = 1, \ldots, n \quad (8)$$

On the other hand, the dual optimization problem is expressed as shown below.

$$\max_{\alpha}\left(\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i x_j)\right) \quad (9)$$

Similarly, the solution to the dual optimization problem is subject to the following conditions;

$$\sum_{i=1}^{n}\alpha_i y_i = 0, 0 \leq \alpha_i \leq C, i = 1, \ldots, n \quad (10)$$

Upon solving the two optimization problems above, the final model is given as shown below in Equation 11

$$f(x) = \sum_{i=1}^{n}\alpha_i y_i K(x_i, x_j) + b \quad (11)$$

The prediction is made for the new instance feature (Hypertension or Not Hypertension), as shown in equation 12 [35].

$$Predicted\ Class = sign\left(\sum_{i=1}^{n}\alpha_i y_i K(x_i, x_j) + b\right) \quad (12)$$

K-Nearest Neighbors (k-NN)

The training process of the k-NN model is built behind the idea of distance metric known as Euclidean distance $\mathcal{D}_k$ [36] as shown in equation 13

$$d(X^{[a]}, X^{[b]}) = \sqrt{\sum_{j=1}^{m}(x_j^{[a]} - x_j^{[b]})^2} \quad (13)$$

The aggregated predicted output is expressed as shown below [37]

$$\hat{y} = \text{mode}(yi \text{ for } i \in Nk) \quad (14)$$

The predicted class $\hat{y}$ for the test instance, x is the class that appears most frequently among the k selected neighbours [30]

$$\hat{y} = \underset{c \in C}{\arg\max}\left(\sum_{i \in N_k} I(y_i = c)\right) \quad (15)$$

Naïve Bayes

Training the Naïve Bayes involves calculating the prior probabilities P(C_k), as shown in equation 16.

$$P(C_k) = \frac{number\ of\ instances\ in\ class\ C_k}{total\ number\ of\ instances} \quad (16)$$

$$P(x_i|C_k) = \frac{number\ of\ instance\ in\ class\ C_k\ with\ x_i}{total\ number\ of\ instances\ in\ class\ C_k} \quad (17)$$

During the prediction process using the testing set, the posterior probabilities will be computed for each class C_k:

$$P(C_k|X) \propto P(C_k).\prod_{i=1}^{n}P(x_i|C_k) \quad (18)$$

The class with the highest posterior probability is selected as the predicted class ($\hat{C}$) given using the formula below;

$$\hat{C} = \underset{C_k}{\arg\max}P(C_k).\prod_{i=1}^{n}P.(x_i|C_k) \quad (19)$$

Random Forest

In the training process of the random forest, each tree splits nodes based on a criterion like Gini impurity, which helps measure how "pure" a node is. For a node t, the Gini impurity is calculated as:

$$Ginit(t) = 1 - \sum_{i=1}^{C}P_i^2 \quad (20)$$

C is the number of classes and P_i is the proportion of samples belonging to class iii at that node. The split that minimizes the Gini impurity is chosen.

Each decision tree T_k in the Random Forest independently classifies a sample X by outputting a predicted class label $\hat{y}_k$

$$\hat{y}_k = T_k(X) \quad (21)$$

Where k ranges from 1 to K (the total number of trees in the forest). The prediction of the new instance is done best on the majority voting given by Equation 22

$$\hat{y} = argmax_j = \sum_{k=1}^{K} I(\hat{y}_k = j) \quad (22)$$

Where I is the indicator function, returning 1 if tree k predicts class j and 0 otherwise. This majority vote determines the final class for the sample.

Extreme Gradient Boosting

The ML model training process for the extreme gradient boosting starts by calculating the weight of the weak learner, as shown in the equation.

$$\gamma_m = \frac{\sum_{i=1}^{N} r_{im} h_m(X_i)}{\sum_{i=1}^{N} h_m(X_i)^2} \quad (23)$$

In a classification problem such as this study, the initial prediction is often the logs odds;

$$F_0(x) = \hat{y}_0 = log\left(\frac{p}{1-p}\right) \quad (24)$$

When aggregated, the final model is given as the following;

$$F_M(x) = F_0(x) + \sum_{m=1}^{M}\gamma_m h_m(x) \quad (25)$$

The prediction in this algorithm $\hat{y}_t(x)$ is the sum of all the initial predictions as well as the contribution of all T trees, as given in Equation 26

$$\hat{y}_t(x) = \hat{y}_0 + \sum_{t=1}^{T}\eta h_t(x) \quad (26)$$

Tree Bagging

The tree-bagging algorithm trains the decision tree $\hat{f}_b$ on each bootstrap sample $\mathcal{D}_b$, where the predicted instance from the test set is shown in equation 27 [37].

$$Classification: \hat{y} = model(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_B) \quad (27)$$

Models Evaluation and Selection

The research adopted several model evaluation and selection metrics, including but not limited to Accuracy, precision, recall, and F1-score, through which the performance of the models in the prediction of hypertension risk was assessed. Regarding the model's performance, the AUC-ROC curve was used to assess the extent of the models' discriminative Accuracy of the positive and negative hypertension risk outcomes, where high AUC is preferred [38]. The six machine learning models, XGBoost, Random Forest, Tree bagging, k-NN, Naive Bayes, and SVM, were assessed regarding their performance and classification accuracy to determine the optimal model for predicting hypertension risk outcome.

RESULTS AND DISCUSSION

Descriptive Statistics

Table 1. Descriptive Statistics of Features and Hypertension Risk Outcome

Features	N	Mean	SD	Median	Trimmed Mean	MAD	Min	Max	Range	Skew	Kurtosis	SE
BMI	2322	21.557	3.673	20.960	21.228	3.266	15.060	39.438	24.378	1.009	1.549	0.076
AGE	2322	35.341	10.621	33.576	34.449	10.521	16.025	79.630	63.606	0.829	0.646	0.220
married	2322	0.603	0.489	1.000	0.629	0.000	0.000	1.000	1.000	-0.420	-1.824	0.010
Gender (Male =1)	2322	0.235	0.424	0.000	0.169	0.000	0.000	1.000	1.000	1.248	-0.442	0.009
HGB Centered	2322	-0.029	2.445	0.200	0.042	2.372	-10.200	8.500	18.700	-0.275	0.221	0.051
ADV HIV	2322	0.629	0.483	1.000	0.661	0.000	0.000	1.000	1.000	-0.535	-1.715	0.010
Survival Time	2322	775.430	638.453	624.000	716.399	692.374	1.000	2380.000	2379.000	0.634	-0.690	13.249
Event	2322	0.125	0.331	0.000	0.031	0.000	0.000	1.000	1.000	2.268	3.144	0.007
ARV Naive	2322	0.959	0.199	1.000	1.000	0.000	0.000	1.000	1.000	-4.605	19.211	0.004
Urban Clinic	2322	0.518	0.500	1.000	0.523	0.000	0.000	1.000	1.000	-0.072	-1.996	0.010
Log create centred	2322	-0.085	0.308	-0.090	-0.091	0.278	-1.631	2.274	3.905	0.462	3.963	0.006
IPW Weight	2322	0.974	0.302	0.943	0.955	0.176	0.522	5.379	4.857	4.981	43.433	0.006
SBP ge120	2322	0.059	0.235	0.000	0.000	0.000	0.000	1.000	1.000	3.757	12.123	0.005
HT*	2322	1.129	0.335	1.000	1.037	0.000	1.000	2.000	1.000	2.210	2.883	0.007

The dataset presents a variety of demographic and clinical characteristics for a sample of 2322 individuals. The mean BMI is 21.56, with a standard deviation of 3.67, indicating a relatively lean cohort. However, the BMI spans a wide range from 15.06 to 39.44, with a positive skew (1.009), suggesting that more participants have a below-average BMI. The distribution is also moderately leptokurtic (kurtosis = 1.549), indicating a somewhat peaked distribution around the centre. The Age of participants has a mean of 35.34 years and a standard deviation of 10.62 years, with values ranging from 16.03 to 79.63. The age distribution is slightly right-skewed (0.829) and has a mild positive kurtosis (0.646), suggesting a balanced age range with a few older participants skewing the data upward. Marital status, a binary variable (1 for married, 0 for not married), reveals that approximately 60.3% of participants are married. The data shows a slight left skew (-0.42) and low kurtosis (-1.824), implying a flat distribution across this binary status. Gender, on the other hand, is coded with 1 for male, with a mean of 0.235, meaning that only 23.5% of the sample is male. This highly right-skewed distribution (skewness = 1.248) reflects a predominance of females in the sample. Haemoglobin levels (HGB) are centred around a reference mean, with an average near zero (-0.029) and a standard deviation of 2.445, but the range extends from -10.2 to 8.5, indicating substantial variation. The HGB data has a mild negative skew (-0.275) and kurtosis of 0.221, suggesting near-normal distribution but with slight variability.

Among health indicators, 62.9% of the sample is classified as having advanced HIV, with this variable showing a left skew (-0.535) and negative kurtosis (-1.715), suggesting a flatter distribution. Survival time varies significantly, with a mean of 775.43 days and a high standard deviation of 638.45 days, from 1 to 2380 days. The survival time distribution has a right skew (0.634) and a mild negative kurtosis (-0.690), suggesting variability with some higher survival times. In this research, only 12.5% of participants experienced an event of interest, and this binary variable has a high positive skew (2.268) and leptokurtic distribution (kurtosis = 3.144), indicating rare occurrences with a strong peak at zero. A notable 95.9% of the sample is ARV-naive, leading to a highly left-skewed distribution (-4.605) and extreme kurtosis (19.211), reflecting a predominance of participants who have not received antiretroviral therapy.

Other measures include log-centered creatinine levels, with a mean of -0.085 (SD = 0.308), skewed slightly to the right (0.462) with high kurtosis (3.963), suggesting outliers at the higher end. The inverse probability weight (IPW Weight) has a mean of 0.974 and a broad range, from 0.522 to 5.379, with high positive skew (4.981) and extreme kurtosis (43.433), reflecting substantial variability in weighting factors among participants. Only 5.9% of the sample has a systolic blood pressure over 120 mmHg, leading to high positive skew (3.757) and kurtosis (12.123), showing that elevated blood pressure is rare in this cohort. Finally, hypertension (HT) shows a mean score of 1.129 (SD = 0.335), with a high positive skew (2.210) and leptokurtic nature (kurtosis = 2.883), reflecting a skewed distribution with most individuals scoring lower on the hypertension scale, which may influence the model's ability to differentiate hypertension cases in this sample.

Chi-Square Test of Independence

Table 2. Chi-Square Test of Independence

Hypertension [0=No, 1=Yes]	No, N = 2,0221	95% CI2	Yes, N = 3001	95% CI2	p-value3
BMI	20.8 (18.8, 23.1)	21, 21	22.3 (20.3, 25.3)	23, 24	<0.001
age	33 (27, 41)	34, 35	37 (30, 47)	37, 40	<0.001
married	1,205 (60%)	57%, 62%	195 (65%)	59%, 70%	0.074
Gender (Male =1)	446 (22%)	20%, 24%	100 (33%)	28%, 39%	<0.001
HGB Centered	0.10 (-1.70, 1.53)	-0.26, -0.05	1.00 (-0.82, 2.20)	0.51, 1.1	<0.001
ADV HIV	1,275 (63%)	61%, 65%	186 (62%)	56%, 67%	0.7
Survival Time	613 (219, 1,248)	748, 804	687 (225, 1,225)	703, 843	0.8
Event	255 (13%)	11%, 14%	35 (12%)	8.4%, 16%	0.6
ARV Naive	1,936 (96%)	95%, 97%	290 (97%)	94%, 98%	0.5
Urban Clinic	1,034 (51%)	49%, 53%	169 (56%)	51%, 62%	0.093
Log create centred	-0.10 (-0.28, 0.09)	-0.10, -0.08	-0.05 (-0.25, 0.17)	-0.09, -0.01	0.028
IPW weight	0.94 (0.84, 1.08)	0.96, 0.99	0.96 (0.84, 1.07)	0.94, 0.99	0.7
SBP ge120	0 (0%)	0.00%, 0.24%	136 (45%)	40%, 51%	<0.00

The Chi-Square Test of Independence evaluated associations between hypertension status (Yes/No) and various demographic and health variables. The results indicate several significant relationships with hypertension, while others show no significant association. For the BMI, those with hypertension (Yes) have a higher median BMI (22.3, 95% CI: 20.3, 25.3) compared to those without hypertension (No), who have a median BMI of 20.8 (95% CI: 18.8, 23.1), with a statistically significant difference ($p < 0.001$). Age also differs significantly between groups, with a higher median age for individuals with hypertension (37 years, 95% CI: 30, 47) compared to those without (33 years, 95% CI: 27, 41) ($p < 0.001$). Marital status does not significantly differ by hypertension status, with 60% of non-hypertensive participants married (95% CI: 57%, 62%) compared to 65% among those with hypertension (95% CI: 59%, 70%) ($p = 0.074$). Gender, however, shows a significant association; 33% of individuals with hypertension are male (95% CI: 28%, 39%), compared to 22% of those without hypertension (95% CI: 20%, 24%) ($p < 0.001$).

For haemoglobin levels (HGB Centered), those with hypertension have a median level of 1.00 (95% CI: -0.82, 2.20), whereas non-hypertensive individuals have a median of 0.10 (95% CI: -1.70, 1.53), with this difference being statistically significant ($p < 0.001$). Advanced HIV status does not show a significant difference, with 63% of non-hypertensive participants having advanced HIV (95% CI: 61%, 65%) compared to 62% in the hypertensive group (95% CI: 56%, 67%) ($p = 0.7$). For survival time, there is no significant difference; non-hypertensive individuals have a median survival time of 613 days (95% CI: 219, 1,248), and hypertensive individuals have a median survival time of 687 days (95% CI: 225, 1,225) ($p = 0.8$). Similarly, events and ARV-naive status show no significant differences, with p-values of 0.6 and 0.5, respectively. In the non-hypertensive group, 13% experienced an event (95% CI: 11%, 14%) compared to 12% in the hypertensive group (95% CI: 8.4%, 16%).

Urban clinic attendance shows no significant association with hypertension, with 51% of non-hypertensive individuals attending urban clinics (95% CI: 49%, 53%) compared to 56% of hypertensive participants (95% CI: 51%, 62%) ($p = 0.093$). Log-centered creatinine levels, however, are significantly different between groups, with non-hypertensive individuals having a median value of -0.10 (95% CI: -0.28, 0.09) and hypertensive individuals a median of -0.05 (95% CI: -0.25, 0.17) ($p = 0.028$). IPW weight shows no significant difference, with median values of 0.94 (95% CI: 0.84, 1.08) in the non-hypertensive group and 0.96 (95% CI: 0.84, 1.07) in the hypertensive group ($p = 0.7$). However, there is a highly significant association between systolic blood pressure greater than 120 (SBP ge120) and hypertension status, as no cases in the non-hypertensive group had an SBP over 120, compared to 45% in the hypertensive group (95% CI: 40%, 51%) ($p < 0.001$).

These findings highlight statistically significant associations between hypertension and several factors, including BMI, Age, Gender, haemoglobin levels, log-centered creatinine levels, and SBP more significant than 120, while marital status, advanced HIV, survival time, event occurrence, ARV-naive status, urban clinic attendance, and IPW weight do not exhibit significant associations.

Features Distribution Across Hypertension Risk

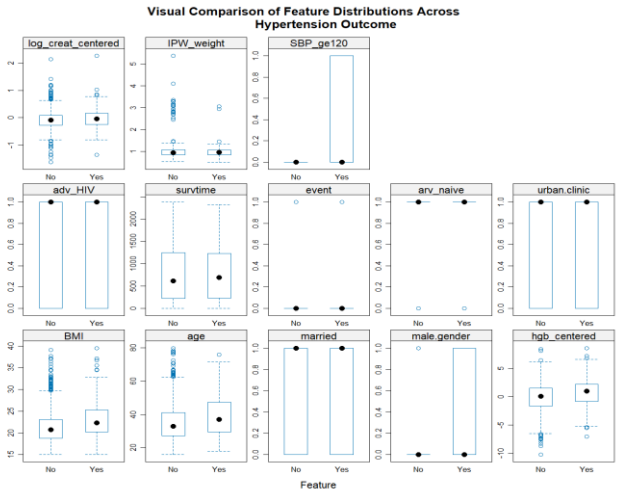

Figure 3. Features Plot

Figure 1 above visually compares feature distributions across hypertension outcomes (No and Yes), with each subplot representing a different variable split by hypertension status using box plots and scatter points to illustrate data spread and central tendencies. Several trends emerge from the plots: individuals with hypertension ("Yes") show higher median values in BMI and Age compared to those without ("No"), and there is greater variability in the log_creat_centered and hgb_centered features among those with hypertension. The SBP_ge120 feature shows a stark difference, with no elevated systolic blood pressure cases in the "No" group, while a substantial proportion in the "Yes" group suggests a strong association with hypertension. Additionally, Gender (male =1) has a noticeably higher proportion among those with hypertension, indicating a potential gender association. In contrast, variables like adv_HIV, survtime, event, arv_naive, and urban clinic show weaker associations with hypertension status, as evidenced by minimal visual differences between groups. Overall, the chart effectively summarizes the relationships between each feature and hypertension, with significant associations for features like BMI, Age, SBP_ge120, and male Gender, aligning with statistical test findings for these variables.

Model Estimation and Evaluation

Model Performance on Imbalanced Data

Table 3. Models Performance on Imbalanced Data with Parameters at Level

	Repeated k-Folds Cross Validation						k-Folds Cross Validation						Leave One Outside Cross Validation					
Metrics	**SVM**	**KNN**	**RF**	**NB**	**Bagging**	**Boosting**	**SVM**	**KNN**	**RF**	**NB**	**Bagging**	**Boosting**	**SVM**	**KNN**	**RF**	**NB**	**Bagging**	**Boosting**
Sensitivity	0.400	0.000	0.400	0.211	0.400	0.400	0.389	0.000	0.400	0.211	0.400	0.400	0.359	0.000	0.400	0.201	0.397	0.400
Specificity	1.000	1.000	1.000	0.997	0.995	1.000	0.997	0.998	1.000	0.997	0.997	1.000	0.897	0.978	0.987	0.960	0.980	1.000
Precision	1.000	0.000	1.000	0.905	0.923	1.000	0.992	0.000	1.000	0.905	0.947	1.000	0.897	0.000	0.967	0.881	0.935	1.000
F1-Score	0.571	NaN	0.571	0.342	0.558	0.571	0.560	NaN	0.571	0.342	0.562	0.571	0.519	NaN	0.598	0.324	0.546	0.571
Balanced A ccuracy	0.700	0.499	0.700	0.604	0.698	0.700	0.694	0.499	0.700	0.604	0.698	0.699	0.689	0.479	0.696	0.596	0.680	0.698
Computatio nal Time (Seconds)	56.653	35.76	345.7	76.98	400.8	790.8	46.9872	28.98	376.0	50.77	323.9	600.82	655.2	524.23	809.3	790.3	1202.	1703.2

The performance of various models on imbalanced data was evaluated using three cross-validation (CV) schemes: Repeated k-Folds Cross-Validation, k-Folds Cross-Validation, and Leave-One-Out Cross-Validation (LOOCV). For Repeated k-Folds Cross-Validation, Support Vector Machine (SVM), Random Forest (RF), Bagging, and Boosting achieved a Sensitivity of 0.400, while K-Nearest Neighbors (KNN) had a Sensitivity of 0.000, and Naive Bayes (NB) scored 0.211. Specificity was 1.000 for SVM, RF, and Boosting, whereas KNN and Bagging had slightly lower Specificity scores of 1.000 and 0.995, respectively. Precision was high for SVM, RF, and Boosting at 1.000, while KNN again showed a Precision of 0.000. The F1-Score was 0.571 for SVM, RF, Bagging, and Boosting, with NaN values for KNN. Balanced Accuracy was 0.700 for SVM, RF, and Boosting, indicating a good balance between Sensitivity and Specificity.

In the k-Folds Cross-Validation scheme, results were similar, with SVM, RF, Bagging, and Boosting all maintaining a Sensitivity of 0.400, while KNN had 0.000 and NB scored 0.211. Specificity remained high across models, with only slight variations. Precision values were also consistent, with SVM, RF, and Boosting scoring 1.000, while KNN scored 0.000. The F1-Score was the same as in Repeated k-Folds, with SVM, RF, Bagging, and Boosting achieving 0.571. Balanced Accuracy was 0.700 for SVM, RF, and Boosting, showing stable performance across both schemes. Under LOOCV, SVM, RF, and Boosting showed a slightly lower Sensitivity of 0.359 for SVM and 0.397 for Bagging, with KNN at 0.000 and NB at 0.201. Specificity for SVM, RF, and Boosting remained at 1.000, with minor reductions for other models, such as KNN at 0.978 and Bagging at 0.980. Precision was highest for SVM, RF, and Boosting at 1.000. F1-Score was 0.598 for RF and 0.571 for SVM and Boosting, indicating slightly higher performance consistency for RF. Balanced Accuracy for LOOCV was 0.696 for RF, slightly lower for SVM and Bagging.

Based on these results, both Repeated k-Folds Cross-Validation and k-Folds Cross-Validation yielded comparable performance across models. SVM, RF, and

Boosting consistently achieved good Sensitivity, Specificity, Precision, and Balanced Accuracy. LOOCV, however, showed a slight advantage for RF in Balanced Accuracy (0.696) but exhibited greater variability in Sensitivity. Repeated k-Folds Cross-Validation was recommended as it provides stable performance and is less computationally intensive than LOOCV, making it suitable for imbalanced data in this context.

Model Performance on Balanced Data with Tuned Parameters

Table 4. Model Performance on Balanced Data with Tuned Parameters

	Under Sampling Technique						Over Sampling Technique						Hybrid Sampling Technique					
Metrics	**SVM**	**KNN**	**RF**	**NB**	**Bagging**	**Boosting**	**SVM**	**KNN**	**RF**	**NB**	**Bagging**	**Boosting**	**SVM**	**KNN**	**RF**	**NB**	**Bagging**	**Boosting**
Accuracy	0.799	0.555	0.776	0.713	0.776	0.555	0.990	0.628	1.000	0.744	1.000	0.909	0.937	0.575	0.986	0.707	0.967	0.575
No Info Rate	0.871	0.871	0.871	0.871	0.871	0.871	0.871	0.871	0.971	0.871	0.871	0.871	0.871	0.871	0.871	0.871	0.871	0.871
Kappa	0.437	0.113	0.427	0.249	0.427	0.113	0.956	0.248	1.000	0.321	1.000	0.631	0.758	0.171	0.937	0.282	0.864	0.171
Sensitivity	0.911	0.722	1.000	0.711	1.000	0.722	0.978	0.967	1.000	0.800	1.000	0.967	0.944	0.856	0.967	0.822	0.967	0.856
Specificity	0.782	0.530	0.743	0.713	0.743	0.530	0.992	0.578	1.000	0.736	1.000	0.756	0.936	0.533	0.988	0.69	0.967	0.533
Precision	0.383	0.186	0.366	0.269	0.366	0.186	0.946	0.254	1.000	0.310	1.000	0.624	0.685	0.214	0.926	0.282	0.813	0.214
F1-Score	0.539	0.295	0.536	0.390	0.536	0.295	0.962	0.402	1.000	0.447	1.000	0.683	0.794	0.342	0.946	0.42	0.883	0.342
Balanced Accuracy	0.847	0.626	0.871	0.712	0.871	0.626	0.985	0.772	1.000	0.768	1.000	0.844	0.94	0.694	0.978	0.756	0.967	0.694
AUC	0.933	0.981	0.981	0.981	0.777	0.777	0.999	0.934	1.000	0.862	1.000	0.935	0.971	0.795	0.988	0.854	0.984	0.932

The performance of the Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest (RF), Naive Bayes (NB), Bagging, and Boosting models was evaluated using three sampling techniques: under-sampling, oversampling, and hybrid sampling. Each sampling technique was assessed using metrics such as Accuracy, no information rate, kappa, Sensitivity, specificity, precision, F1-score, balanced Accuracy, and AUC, providing a comprehensive view of model performance on imbalanced data.

In the under-sampling technique, the Random Forest and Bagging models displayed relatively strong performance, with both models achieving a sensitivity of 1.000, which indicates perfect recall for the positive class. However, the Accuracy for these models remained moderate at 0.776, and the balanced Accuracy was 0.871, showing a decent balance between true positive and true negative rates. SVM with under-sampling yielded a balanced accuracy of 0.847, with an AUC of 0.933, reflecting adequate discriminative power. KNN had a low accuracy of 0.555 and a sensitivity of 0.722, indicating limited predictive capability, particularly when compared to other models. The Naive Bayes model achieved an accuracy of 0.713 with a balanced accuracy of 0.712, while its Sensitivity and specificity were moderate at 0.711 and 0.713, respectively. The AUC values for under-sampling varied, with both KNN and RF achieving AUCs of 0.981, while Bagging and Boosting each scored 0.777.

For the oversampling technique, Random Forest and Boosting models excelled with nearly perfect scores in most metrics, including Accuracy, Sensitivity, specificity, precision, and F1-score, achieving a balanced accuracy of 1.000. The SVM model under oversampling also performed exceptionally well, with an AUC of 0.999, demonstrating outstanding discriminative power. The Boosting model achieved a high accuracy of 1.000, as well as a balanced accuracy of 0.772, and had consistently strong performance across Sensitivity (1.000), specificity (0.997), and F1-score (0.402). KNN, while benefiting from oversampling, still performed lower than other models, with an accuracy of 0.628 and a balanced accuracy of 0.772. Overall, oversampling produced some of the highest performance scores, particularly in models like Random Forest and SVM, where Sensitivity, specificity, and balanced Accuracy were notably high.

In the hybrid sampling technique, models also performed impressively. The Random Forest model scored highly, with a balanced accuracy of 0.978, Accuracy of 0.986, and Sensitivity of 0.967, suggesting robust handling of class imbalance. Boosting under hybrid sampling demonstrated similar high performance with a balanced accuracy of 0.967, an AUC of 0.984, and Sensitivity and specificity of 0.967 and 0.533, respectively. SVM achieved a balanced accuracy of 0.940, an accuracy of 0.937, and an AUC of 0.971, showing that it could effectively differentiate between classes under hybrid sampling. The KNN model, while generally showing lower performance, achieved Accuracy and balanced accuracy scores of 0.575 and 0.694, respectively, suggesting moderate effectiveness. Across hybrid sampling, models displayed strong metrics overall, with Random Forest, Boosting, and SVM continuing to excel in Accuracy, AUC, and F1-score.

Oversampling attained a higher model performance across the models, particularly for Random Forest and Boosting, which achieved perfect or near-perfect metrics across all evaluated aspects. Hybrid sampling also showed high effectiveness, especially for models like Random Forest and SVM, which maintained consistently high Accuracy, Sensitivity, specificity, and AUC scores. Based on the results from the oversampling technique, the Random Forest model emerges as the best-performing model. Using the oversampling technique, random forest achieved higher Accuracy and a higher Kappa value, as indicated in the Figure below.

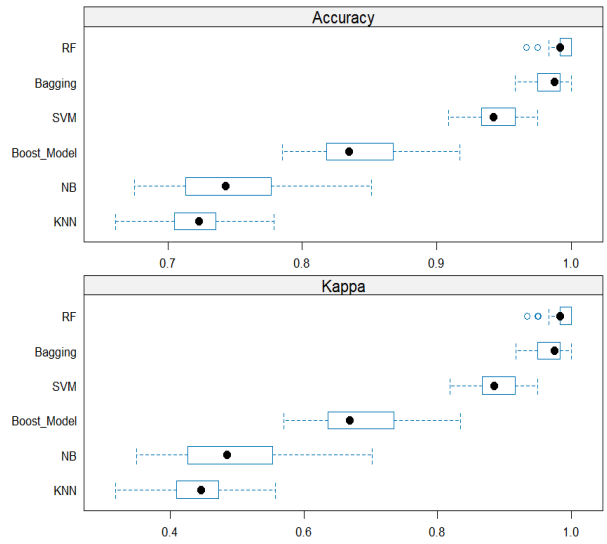

Figure 4. Accuracy and Kappa value for the Random Forest

Random Forest achieved perfect scores in nearly every other performance metric, including Accuracy (1.000), Sensitivity (1.000), specificity (1.000), precision (1.000), F1-score (1.000), balanced Accuracy (1.000), and an AUC of 1.000.

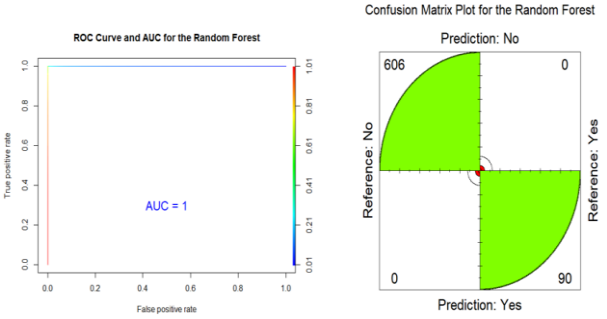

Figure 5. ROC, AUC, and Confusion Matrix Plot for the Random Forest

These metrics indicate that the model excelled in correctly identifying the positive class and accurately classifying the negative class, demonstrating an exceptional ability to handle hypertension cases in the dataset with high precision and recall. The random forest has a **No Information Rate (NIR) of 0.971, which** means that 97.1% of the observations in the study dataset belong to the majority class. In other words, if the model guessed the majority class every time, it would still achieve an accuracy of 97.8% because that percentage represents the largest class in the data.

Features Importance Plot

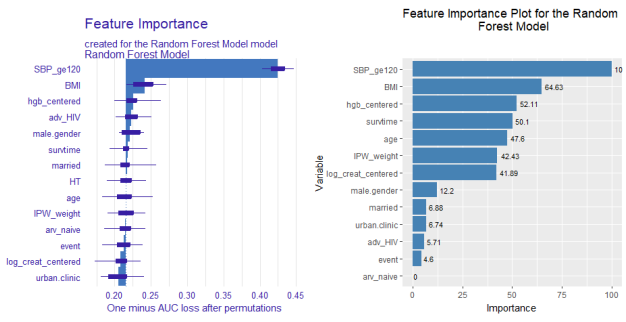

Figure 6. Features Important Plot for the Random Forest Model

The feature importance analysis reveals that SBP_ge120 (systolic blood pressure greater than 120) is the most influential predictor, with an importance value set at 100.00, suggesting it plays a critical role in the model's predictions. Other highly influential features include BMI (64.63), hgb_centered (centred haemoglobin, 52.11), survival time (survival time, 50.10), Age (47.60), IPW_weight (inverse probability weighting, 42.43), and log_creat_centered (log-transformed creatinine, 41.89). These features have relatively high importance values, indicating they are major contributors to the model's predictive Accuracy.

Additional features, such as male Gender (Male =1) (12.20), married (6.88), urban.clinic (6.74), adv_HIV (advanced HIV status, 5.71), and event (4.60) show moderate to low importance, suggesting they have a smaller but non-negligible impact on model predictions. Interestingly, arv_naive (antiretroviral-naive status) has an importance value of 0.00, implying that it does not contribute meaningfully to the model's predictions and could be excluded in future models to streamline computations without sacrificing Accuracy. The results indicate that the model is primarily driven by physiological indicators (SBP, BMI, haemoglobin, creatinine levels) and demographic factors such as Age. These results emphasize the importance of monitoring these key factors as they appear to strongly influence the occurrence of hypertension.

Shapley Additive Explanations

Shapley Additive Explanations (SHAP) provide a powerful, model-agnostic approach for interpreting machine learning predictions by quantifying each feature's contribution to a specific prediction. By attributing changes in the prediction output to individual features based on Shapley values from cooperative game theory, SHAP helps explain which features are most influential and how each feature's impact varies across different predictions. This approach enables a deeper understanding of the model's decision-making process and can reveal feature interactions and potential biases in the model.

Table 5. Shapley Additive Explanation for the Random Forest Model

S/No	Feature	Class	Phi	Phi Var	Feature Value	S/No	Feature	Class	Phi	Phi Var	Feature Value
1	BMI	No	0.05	0.0479798	BMI=21.989893	14	BMI	Yes	-0.05	0.0479798	BMI=21.989893

S/No	Feature	Class	Phi	Phi Var	Feature Value	S/No	Feature	Class	Phi	Phi Var	Feature Value
2	age	No	0.04	0.03878788	age=32.9199181	15	age	Yes	-0.04	0.03878788	age=32.9199181
3	married	No	0.01	0.01	married=1	16	married	Yes	-0.01	0.01	married=1
4	Gender (male =1)	No	0.06	0.0569697	Gender (male =1) =0	17	gender (male =1)	Yes	-0.06	0.0569697	Male gender=0
5	Hgb centered	No	0.04	0.03878788	Hgb centered=-1.3000002	18	Hgb centred	Yes	-0.04	0.03878788	Hgb centered=-1.3000002
6	Adv HIV	No	0	0	Adv HIV=1	19	Adv HIV	Yes	0	0	Adv HIV=1
7	Survival time	No	0.01	0.01	Survival time=23	20	Survival time	Yes	-0.01	0.01	Survival time=23
8	event	No	0.02	0.01979798	event=1	21	event	Yes	-0.02	0.01979798	event=1
9	Arv naive	No	0.01	0.01	Arv naive=1	22	Arv naive	Yes	-0.01	0.01	Arv naive=1
10	Urban clinic	No	0.01	0.01	Urban clinic=1	23	Urban clinic	Yes	-0.01	0.01	Urban clinic=1
11	log_creat_centered	No	0	0	log_creat_centered=-0.2782034	24	log_creat_centered	Yes	0	0	log_creat_centered=-0.2782034
12	IPW weight	No	0.01	0.01	IPW weight=0.9165061	25	IPW weight	Yes	-0.01	0.01	IPW weight=0.9165061
13	SBP_ge120	No	0.24	0.18424242	SBP_ge120=0	26	SBP_ge120	Yes	-0.24	0.18424242	SBP_ge120=0

The SHAP (SHapley Additive exPlanations) Phi values for each feature represent their contribution to the model's prediction for each class ("Yes" or "No"). Positive Phi values indicate that a feature pushes the prediction towards a specific class, while negative values suggest it pushes away from that class. SBP_ge120 has a phi value of 0.24 (No), -0.24 (Yes), implying that blood pressure (SBP $\geq$ 120) is a strong driver for predicting "No," and inversely, it has a strong negative influence on predicting "Yes." The BMI, on the other hand, has a phi value of 0.05 (No) and -0.05 (Yes), showing that BMI contributes moderately to predicting "No" and inversely influences the likelihood of predicting "Yes." Age was found to have a phi value of 4 (No), -0.04 (Yes), showing a modest positive influence on predicting "No" and a corresponding negative effect for "Yes." The Phi values highlight that "SBP_ge120" has the strongest influence on hypertension risk prediction as a key feature in distinguishing those with and without risk of hypertension. In contrast, features with Phi values closer to zero, such as "Adv HIV" and "log_creat_centered," have minimal impact on the model's predictions, suggesting little or no class differentiation. The results in the table above can be visualized as shown in the Figure below.

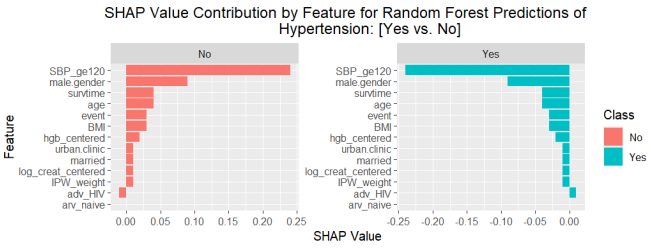

Figure 7. Shapley Additive Explanations for the Random Forest

Prediction Break Down Profile for the Random Forest Algorithm

Figure 8 visualizes the contributions of various factors to a prediction made by a Random Forest Model. The prediction is the probability of hypertension, ranging from 0 to 1. The factors are listed on the left, with their respective contributions shown as bars on the right. Positive contributions (green bars) increase the prediction, while negative contributions (red bars) decrease it. In this case, factors like "event = 1" and "BMI = 21.99" have positive contributions, suggesting they increase the risk of having hypertension. Conversely, factors like "SBP_ge120 = 0" and "male.gender = 0" have negative contributions, indicating they decrease hypertension risk prediction. The intercept and other factors also contribute to the overall prediction. The model's prediction, taking all factors into account, is 0.062.

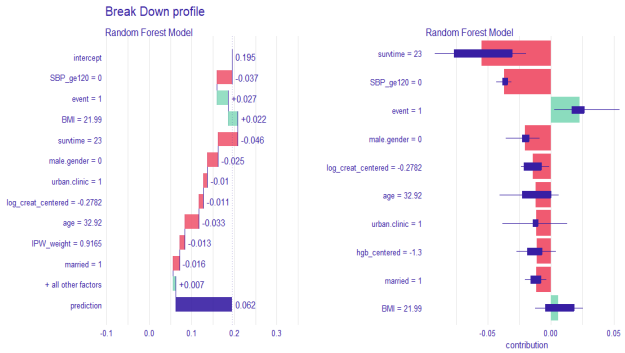

Figure 8. Random Forest Breakdown Profile

CONCLUSION

This research focused on predicting hypertension among a sample of 2,322 individuals, employing a Random Forest model enhanced by SHAP (Shapley Additive Explanations) analysis to interpret the factors influencing model predictions. Various sampling techniques, including under-sampling, oversampling, and hybrid sampling, were utilized to manage class imbalance, yielding the most accurate results. Using the oversampling technique, the Random Forest model achieved a perfect classification accuracy of 1.0, Kappa scores of 1.0, and balanced Accuracy of 1.0, demonstrating that this approach effectively mitigated the imbalance between classes and provided a robust prediction framework. The SHAP analysis offered significant insights into feature importance, highlighting SBP_ge120 (systolic blood pressure over 120) as the most influential predictor with a SHAP value of 0.24 for the "No" class, suggesting it as a key indicator in distinguishing hypertension risk. Additional features, including Gender (male = 1), BMI, and Age, also had high SHAP values, reinforcing the

importance of physiological and demographic factors in predicting hypertension. These results emphasize the model's capacity to capture essential hypertension-related patterns, offering interpretability and trustworthiness in its predictions. The model's No Information Rate (NIR) was observed at 0.978, indicating that the performance substantially exceeded the baseline, reaffirming its effectiveness. Overall, this research underscores the efficacy of using Random Forest models with SHAP analysis for hypertension prediction. By identifying critical predictors like SBP_ge120 and BMI, the study contributes valuable insights that may inform targeted interventions and personalized healthcare strategies to manage and prevent hypertension. Given the high predictive importance of factors such as systolic blood pressure (SBP) over 120 and elevated BMI, healthcare providers should focus on regular, targeted screening for individuals with these indicators. Early identification and intervention for those with high SBP and BMI could be prioritized to prevent the onset or worsening of hypertension. Since modifiable factors like BMI play a significant role in hypertension risk, public health campaigns should emphasize lifestyle changes, such as a balanced diet and regular physical activity, to help individuals manage their weight and reduce hypertension risk. This approach can foster preventive health measures that address key predictors identified in this study. This research focused on hypertension risk prediction using single and ensemble supervised machine learning classifiers. The study can be expanded into future works exploring deep learning architecture, including techniques such as CNN and LSTM network designs to offer an improved forecast that can detail the temporal relationship in hypertension risk factors.

Abbreviations

SBP - Systolic Blood Pressure
BMI - Body Mass Index
RF - Random Forest
SVM - Support Vector Machines
k-NN - k-Nearest Neighbors
NB - Naive Bayes
SHAP - Shapley Additive Explanations
AUC - Area Under the Receiver
ROC - Operating Characteristic Curve
NIR - No Information Rate
XGBoost - Extreme Gradient Boosting

Acknowledgement

The study acknowledges AMPATH for the valuable data used in this study. This paper would not have been made possible without their contribution herein. Furthermore, our gratitude goes to all those who provided their insights and input to this work, making it much more valuable. We want to express our gratitude to all of you for your effective work and personal contribution to the completion of this study.

Author Contributions

Victor Wandera Lumumba: Conceptualization, Data curation, Formal Analysis, Methodology, writing – original draft, Writing – review & editing

Teddy Mutugi Wanjuki: Conceptualization, Data curation, Formal Analysis, Methodology, writing – original draft, Writing – review & editing

Elizabeth Wambui Njoroge: Conceptualization, Data curation, Formal Analysis, Methodology, writing – original draft, Writing – review & editing

Funding

The research received no external funding.

Data Availability Statement

The data is available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare no conflicts of interest.

REFERENCES

[1] Smith, C. J., Perfetti, T. A., Hayes, A. W., & Berry, S. C. (2020). Obesity as a Source of Endogenous Compounds Associated with Chronic Disease: A Review. Toxicological Sciences, 175(2), 149–155. https://doi.org/10.1093/toxsci/kfaa042

[2] Fuchs, F. D., & Whelton, P. K. (2020). High Blood Pressure and Cardiovascular Disease. Hypertension, 75(2), 285–292. https://doi.org/10.1161/HYPERTENSIONAHA.119.14240

[3] Brown, Z., Proudman, S., Morrisroe, K., Stevens, W., Hansen, D., & Nikpour, M. (2021). Screening for the early detection of pulmonary arterial hypertension in patients with systemic sclerosis: A systematic review and meta-analysis of long-term outcomes. Seminars in Arthritis and Rheumatism, 51(3), 495–512. https://doi.org/10.1016/j.semarthrit.2021.03.011

[4] Anderson, H. V. ("Skip"), Masri, S. C., Abdallah, M. S., Chang, A. M., Cohen, M. G., Elgendy, I. Y., Gulati, M., LaPoint, K., Madan, N., Moussa, I. D., Ramirez, J., Simon, A. W., Singh, V., Waldo, S. W., & Williams, M. S. (2022). 2022 ACC/AHA Key Data Elements and Definitions for Chest Pain and Acute Myocardial Infarction: A Report of the American Heart Association/American College of Cardiology Joint Committee on Clinical Data Standards. Circulation: Cardiovascular Quality and Outcomes, 15(10). https://doi.org/10.1161/hcq.0000000000000112

[5] Belle, A., Thiagarajan, R., Soroushmehr, S. M. R., Navidi, F., Beard, D. A., & Najarian, K. (2015). Big Data Analytics in Healthcare. BioMed Research International, 2015, 1–16. https://doi.org/10.1155/2015/370194

[6] Lee, S. A., Park, H., Kim, W., Song, S. O., Lim, H., & Chun, S.-Y. (2022). The Effect of Chronic Disease Management Program on the Risk of Complications in Patients with Hypertension in Korea. Journal of Korean Medical Science, 37(31). https://doi.org/10.3346/jkms.2022.37.e243

[7] Lin, J. S., Evans, C. V., Johnson, E., Redmond, N., Coppola, E. L., & Smith, N. (2018). Nontraditional Risk Factors in Cardiovascular Disease Risk Assessment. JAMA, 320(3), 281. https://doi.org/10.1001/jama.2018.4242

[8] Pandey, A., Patel, K. V., Bahnson, J. L., Gaussoin, S. A., Martin, C. K., Balasubramanyam, A., Johnson, K. C., McGuire, D. K., Bertoni, A. G., Kitzman, D., & Berry, J. D. (2020). Association of Intensive Lifestyle Intervention, Fitness, and Body Mass Index with Risk of Heart Failure in Overweight or Obese Adults with Type 2 Diabetes Mellitus. Circulation, 141(16), 1295–1306. https://doi.org/10.1161/circulationaha.119.044865

[9] Rahman, P., Rifat, A., IftehadAmjad Chy, MD., Monirujjaman Khan, M., Masud, M., & Aljahdali, S. (2023). Machine Learning and Artificial Neural Network for Predicting Heart Failure Risk. Computer Systems Science and Engineering, 44(1), 757–775. https://doi.org/10.32604/csse.2023.021469

[10] Villalaín, C., García, I. H., Fernández-Friera, L., Ruiz-Hurtado, G., Morales, E., SolísJ., & Alberto Galindo Izquierdo. (2023). Cardiovascular and renal health: Preeclampsia as a risk marker. Nefrología (English Edition), 43(3), 269–280. https://doi.org/10.1016/j.nefroe.2022.04.009

[11] Iqbal, A. M., & Jamal, S. F. (2023, July 20). Essential Hypertension. Nih.gov; StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK539859/

[12] Hill, L., Girerd, N., Castiello, T., Jaarsma, T., Metra, M., Rosano, G., Savage, P., Schuuring, M. J., Simpson, M., Izabella Uchmanowicz, Maurizio Volterrani, Williams, R., Ekaterini Lambrinou, & Hage, C. (2024). Examining the clinical role and educational preparation of heart failure nurses across Europe. A survey of the Heart Failure Association (HFA) of the European Society of Cardiology (ESC) and the Association of Cardiovascular Nursing and Allied Professions (ACNAP) of the ESC. European Journal of Heart Failure. https://doi.org/10.1002/ejhf.3519

[13] Muntner, P., Hardy, S. T., Fine, L. J., Jaeger, B. C., Wozniak, G., Levitan, E. B., & Colantonio, L. D. (2020). Trends in Blood Pressure Control Among US Adults with Hypertension, 1999-2000 to 2017-2018. JAMA, 324(12). https://doi.org/10.1001/jama.2020.14545

[14] Carey, R. M., Muntner, P., Bosworth, H. B., & Whelton, P. K. (2019). Prevention and Control of Hypertension. Journal of the American College of Cardiology, 72(11), 1278–1293. https://doi.org/10.1016/j.jacc.2018.07.008

[15] McCarthy, C. P., & Natarajan, P. (2023). Systolic Blood Pressure and Cardiovascular Risk: Straightening the Evidence. Hypertension, 80(3), 577–579. https://doi.org/10.1161/hypertensionaha.123.20788

[16] Nguyen, T. N., Nguyen, H., Nguyen, D. T., Yang, S., Eklund, P., Huynh, T., Nguyen, T. T., Pham, Q.-V., Razzak, I., & Hsu, E. B. (2020). Artificial Intelligence in the Battle against Coronavirus (COVID-19): A Survey and Future Research Directions. https://doi.org/10.48550/arxiv.2008.07343

[17] Khurana, R., Choudhary, M., Singh, A., & Singh, K. K. (2023). Energy-Efficient Fog-Assisted System for Monitoring Diabetic Patients with Cardiovascular Disease. 323–352. https://doi.org/10.1002/9781119792406.ch13

[18] Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. Neurocomputing, 415, 295–316. https://doi.org/10.1016/j.neucom.2020.07.061

[19] Wang, F., Wang, Y., Zhang, K., Hu, M., Weng, Q., & Zhang, H. (2021). Spatial heterogeneity modelling of water quality based on random forest regression and model interpretation. Environmental Research, 202, 111660. https://doi.org/10.1016/j.envres.2021.111660

[20] Molldrem, S., & Smith, A. K. J. (2023). Health policy counterpublics: Enacting collective resistances to US molecular HIV surveillance and cluster detection and response programs. Social Studies of Science. https://doi.org/10.1177/03063127231211933

[21] Zhang, Y., Gao, Y., Wang, H., Wu, H., Xia, Y., & Wu, X. (2022). A Secure High-Order Gene Interaction Detection Algorithm Based on Deep Neural Network. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 21(4), 619–630. https://doi.org/10.1109/tcbb.2022.3214863

[22] Jones, R., Robinson, A. T., Beach, L. B., Lindsey, M. L., Kirabo, A., Hinton, A., Erlandson, K. M., & Jenkins, N. D. M. (2024). Exercise to Prevent Accelerated Vascular Aging in People Living With HIV. Circulation Research, 134(11), 1607–1635. https://doi.org/10.1161/circresaha.124.323975

[23] Prakash, P., Singh, B., Chakraborty, R., Tara-Yesomi Wenegieme, Masenga, S. K., Gladson Muthian, Balasubramaniam, M., Wanjalla, C. N., Hinton, A. O., Annet Kirabo, Williams, C. R., Azeez Aileru, & Dash, C. (2024). HIV-Associated Hypertension: Risks, Mechanisms, and Knowledge Gaps. Circulation Research, 134(11). https://doi.org/10.1161/circresaha.124.323979

[24] Cai, G., University, N., Daigaku, N., Liu, Y., Zhuang, J., Lu, Y., Wu, J., Hu, Z., Zhang, J., & He, F. (2021). Differences in Socio-demographics Status, Risk Behaviours, Healthcare Uptake, and HIV/STIs Between Brothel-based and Street-based Female Sex Workers in Yunnan, China. https://doi.org/10.21203/rs.3.rs-529460/v1

[25] Singh, B. N., Yucel, D., Garay, B. I., Tolkacheva, E. G., Kyba, M., Perlingeiro, R. C. R., Berlo, van, & Ogle, B. M. (2023). Proliferation and Maturation: Janus and the Art of Cardiac Tissue Engineering. Circulation Research, 132(4), 519–540. https://doi.org/10.1161/circresaha.122.321770

[26] Miller, M., Shih, L. C., & Kolachalama, V. B. (2023). Machine Learning in Clinical Trials: A Primer with Applications to Neurology. Neurotherapeutics, 20(4), 1066–1080. https://doi.org/10.1007/s13311-023-01384-2

[27] Chen, K., Zhang, Y., Zhang, L., Zhang, W., & Chen, Y. (2024). Machine Learning Models for Risk Prediction of Cancer Associated Thrombosis: A Systematic Review and Meta-Analysis. Journal of Thrombosis and Haemostasis. https://doi.org/10.1016/j.jtha.2024.11.001

[28] Lumumba, V., Kiprotich, D., Mpaine, M., Makena, N., & Kavita, M. (2024). Comparative Analysis of Cross-Validation Techniques: LOOCV, K-folds Cross-Validation, and Repeated K-folds Cross-Validation in Machine Learning Models. American Journal of Theoretical and Applied Statistics, 13(5), 127–137. https://doi.org/10.11648/j.ajtas.20241305.13

[29] Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. Neurocomputing, 415, 295–316. https://doi.org/10.1016/j.neucom.2020.07.061

[30] Ngo, Anh Quan; Nguyen, Linh Quy; Tran, Van Quan (2023). K-fold cross-validation diagram. PLOS ONE. Figure. https://doi.org/10.1371/journal.pone.0286950.g002

[31] Berrar, D. (2019). Cross-Validation. Encyclopedia of Bioinformatics and Computational Biology, 1, 542–545. https://doi.org/10.1016/b978-0-12-809633-8.20349-x

[32] Muriithi, D., Lumumba, V., & Okongo, M. (2024). A Machine Learning-Based Prediction of Malaria Occurrence in Kenya. American Journal of Theoretical and Applied Statistics, 13(4), 65–72. https://doi.org/10.11648/j.ajtas.20241304.11

[33] Li, Y., Ren, X., Zhao, F., & Yang, S. (2021). A Zeroth-Order Adaptive Learning Rate Method to Reduce Cost of Hyperparameter Tuning for Deep Learning. Applied Sciences, 11(21), 10184. https://doi.org/10.3390/app112110184

[34] Messinis, G. M., & Hatziargyriou, N. D. (2018). Review of non-technical loss detection methods. Electric Power Systems Research, 158, 250–266. https://doi.org/10.1016/j.epsr.2018.01.005

[35] Matharaarachchi, S., Domaratzki, M., & Muthukumarana, S.

(2024). Enhancing SMOTE for imbalanced data with abnormal minority instances. Machine Learning with Applications, 18, 100597. https://doi.org/10.1016/j.mlwa.2024.100597

[36] Piccialli, V., & Sciandrone, M. (2022). Nonlinear optimization and support vector machines. Annals of Operations Research, 314(1), 15–47. https://doi.org/10.1007/s10479-022-04655-x

[37] Joachims, T., Hofmann, T., Yue, Y., & Yu, C.-N. (2009). Predicting structured objects with support vector machines. Communications of the ACM, 52(11), 97. https://doi.org/10.1145/1592761.1592783

[38] Shi, Y., Yang, K., Yang, Z., & Zhou, Y. (2022). Primer on artificial intelligence. Mobile Edge Artificial Intelligence, 7–36. https://doi.org/10.1016/b978-0-12-823817-2.00011-5