A Comparative Analysis of Machine Learning Techniques for Hypertension Risk Prediction and Diagnostic Classification

D P Singh
Amity University Uttar Pradesh, Greater Noida Campus

Abstract— **Hypertension is a significant contributor to cardiovascular diseases, necessitating early prediction and precise diagnostic classification for effective clinical intervention and prevention strategies. This research conducts an in-depth comparative analysis of multiple machine learning (ML) models for predicting hypertension risk and classifying diagnoses. The performance of various supervised learning algorithms—such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Gradient Boosting, and Multilayer Perceptron (MLP) is assessed using a clinically validated hypertension dataset.**

The evaluation of the models is conducted using essential performance indicators, including Accuracy, Precision, Recall, F1-score, and the Area Under the Curve (AUC), to ensure a comprehensive assessment. Prior to modeling, the dataset undergoes pre-processing steps such as feature selection, normalization, and imputation of missing values to improve model performance. The findings indicate that ensemble techniques like CatBoost, Gradient Boosting, Random Forest, and AdaBoost surpass conventional classifiers in terms of both predictive accuracy and diagnostic consistency.

The research emphasizes the capability of advanced machine learning models to assist healthcare professionals in making timely, data-informed decisions for managing hypertension. It also highlights the critical role of model interpretability and clinical relevance in effectively implementing ML-based diagnostic tools in practical medical environments.

Keywords: **Classification Models, Comparative Analysis, Diagnostic Classification, Health Informatics, Hypertension, Machine Learning, Medical Diagnosis, Predictive Analytics, Risk Prediction.**

I. INTRODUCTION

Hypertension, or high blood pressure, is a major global public health concern due to its strong association with life-threatening conditions such as cardiovascular disease, stroke, kidney failure, and other serious outcomes. According to the World Health Organization (WHO), an estimated 1.28 billion adults aged 30 to 79 are affected by hypertension, with many cases going undiagnosed or inadequately managed. In a 2018 survey, only 59.5% of 502,079 individuals with hypertension were aware of their condition [11]. The detection of hypertension is complicated by variability in blood pressure (BP) measurements and its often-hidden nature. WHO ranks hypertension as the third leading cause of death globally, responsible for one in every eight deaths [9]. Currently, around 1.3 billion people, including 116 million in the United States, are affected [25]. Hypertensive individuals face a 2–4 times higher risk of developing heart disease, peripheral vascular disease, and stroke increasing the economic burden due to medical expenses[1]. Additionally, hypertension is a significant contributor to illness, disability, and mortality worldwide [29,36]. To combat this, the World Hypertension League has led global awareness campaigns since 2005, emphasizing regular screening. Between 2015 and 2019, 37,110 individuals participated in World Hypertension Day screenings, from which 20,206 were included in a current study after excluding previously diagnosed cases. Among these, 4,192 (20.75%) were newly identified as hypertensive. By 2025, the global burden is projected to rise to 1.56 billion adults aged 30 to 79, with nearly two-thirds residing in low- and middle-income countries [32].

Hypertension (HTN), defined by blood pressure readings exceeding normal levels, is a significant public health concern affecting adults globally [20]. If left untreated, it significantly increases the risk of cardiovascular diseases, coronary heart disease, stroke, kidney damage, and other severe health complications [8]. HTN is a leading cause of premature death worldwide, affecting over one in four men and one in five women [21]. Given its widespread prevalence and association with chronic kidney disease, HTN poses a major global health challenge [2, 3, 5]. As a key risk factor for cardiovascular conditions, it also contributes to rising healthcare costs and lost productivity [4].

Predictive modeling for assessing the risk of developing hypertension (HTN) can help identify key risk factors associated with HTN, provide accurate forecasts of future HTN risk [22], and pinpoint high-risk individuals who could benefit from medical interventions and healthier habits to prevent HTN [6,7,19]. Over time, several models have been developed to predict HTN risk in the general population, using either modern machine learning methods or traditional regression-based techniques [13].Previous research has primarily focused on traditional linear models, such as logistic regression (LR) and the Cox proportional hazard model, to identify risk factors strongly associated with hypertension (HTN) [17,33,34]. Additionally, various studies have employed machine learning techniques to assess the accuracy of HTN predictions. One study explored multiple algorithms, with Random Forest (RF), K-Nearest Neighbors (KNN), Decision Trees (DT), and Naive Bayes (NB) models yielding promising results—particularly RF, which achieved an accuracy of 80.12% [35]. A later study used RF, CatBoost, MLP Neural Network, and LR, with RF achieving 92% accuracy [31]. Another study, based on medical data, applied SVM, C4.5, RF, and XGBoost, with XGBoost reaching 94.36% accuracy [18]. However, a study using RF, LR, ANN, and XGBoost on Ethiopian data reported a lower accuracy of 88.81% for XGBoost [39], which falls below the accuracy reported in previous studies. Thus, there is a need to improve HTN prediction performance.

Due to the widespread nature of hypertension (HTN), early detection and accurate identification of its critical risk factors are essential for timely prevention and effective management. Arterial hypertension remains the leading modifiable risk factor for cardiovascular disease worldwide. Despite advances in its prevention and treatment, the global incidence and prevalence of HTN and related cardiovascular complications remain high—primarily due to persistent gaps in detection, prevention, and management strategies [14,28]. In under-sampling experiments, the highest sensitivity scores were recorded, albeit often accompanied by reduced specificity and overall accuracy. Notably, the XGBoost model under the under-sampling scheme achieved the highest sensitivity but showed low specificity and poor overall accuracy. The best overall performance was obtained using the Random Forest model, yielding a sensitivity of 0.818, specificity of 0.629, accuracy of 0.681, and an AUC of 0.816. In recent years, artificial intelligence (AI)—which involves computer systems performing tasks that typically require human intelligence—has emerged as a powerful tool in healthcare, especially for managing complex clinical conditions [15, 37]. Among its branches, machine learning (ML) has attracted particular interest for transforming medical diagnostics by enabling early detection, risk prediction, and classification of hypertension (HTN). These algorithms utilize large datasets to uncover hidden patterns and generate predictive insights that often surpass traditional diagnostic methods.

This research focuses on improving the prediction of hypertension (HTN) risk by implementing an optimized algorithm combined with a stacking ensemble model. Utilizing a Kaggle dataset that includes demographic information, risk factors, knowledge-based responses, and multiple blood pressure readings, the study systematically compares the performance of several machine learning techniques—including Support Vector Machines (SVM), Random Forest (RF), Gradient Boosting Machines (GBM), Extreme Gradient Boosting (XGBoost), and K-Nearest Neighbors (KNN). These models are evaluated based on key performance metrics such as Accuracy, Precision, Recall, F1 Score, and Area Under the ROC Curve (AUC-ROC) to assess their diagnostic and predictive effectiveness. The objective is to determine the most accurate approach for early detection of hypertension, ultimately aiming to reduce healthcare costs and enhance patient outcomes, while providing actionable insights for healthcare professionals, policymakers, and researchers in preventive medicine.

II. RELATED WORK

Hypertension often referred to as the “silent killer,” represents a major global public health concern due to its strong association with cardiovascular diseases and elevated mortality rates. Numerous studies have employed machine learning (ML) techniques to predict hypertension risk and support diagnostic classification. These studies have leveraged diverse datasets, feature sets, and algorithmic approaches to improve the accuracy and reliability of hypertension detection.

Artificial intelligence, particularly machine learning, has emerged as a powerful tool in healthcare, enabling data-driven decision-making for a variety of clinical conditions. ML models excel at identifying complex patterns within data and generating highly accurate predictions about individual health outcomes—without the need for explicit programming. Singh [42] discussed the application of machine learning techniques in the healthcare sector to optimize resources and enhance operational efficiency. In a separate study, Singh [41] evaluated multiple algorithms to predict breast cancer recurrence. Among the models tested, AdaBoost and CatBoost achieved outstanding results, with an accuracy of 97.36%, a precision of 97.61%, a sensitivity of 95.34%, a specificity of 98.59%, and an F1 score of 96.47%.

Singh [40] identified logistic regression as the most effective algorithm for predicting diabetes, achieving an accuracy of 76.63% and an ROC-AUC score of 0.82. In another study, Singh [44] found that the random forest model demonstrated exceptional performance in predicting cardiovascular disease, achieving an accuracy of 98.53%. Additionally, Singh [43] evaluated various algorithms for predicting lung cancer and determined that the extreme gradient boosting (XGBoost) model performed the best, achieving an accuracy of 98.38%, a precision of 98.38%, and an F1-score of 99.17%. Shoaib et al. [27] utilized logistic regression and decision tree algorithms to forecast hypertension by leveraging demographic and lifestyle data, achieving a favorable trade-off between model accuracy and interpretability. In a similar vein, Aldhyani et al. [30] employed support vector machines (SVM) and k-nearest neighbors (KNN) for hypertension prediction using medical datasets, attaining strong classification outcomes, especially with effective data pre-processing. Nour et al. [23] utilized the PPG-BP dataset to build and evaluate multiple classifiers—including the C4.5 decision tree, random forest, linear discriminant analysis, and linear support vector machine for identifying hypertensive individuals.

With recent progress, ensemble methods have gained traction. For example, Patel et al. [16] applied Random Forest and Gradient Boosting models to evaluate hypertension risk in Indian populations, and their results indicated that ensemble classifiers considerably outperformed traditional models in terms of sensitivity and F1-score. Additionally, Krittanawong et al. [12] explored artificial neural networks (ANNs) and deep learning approaches, emphasizing their potential for enhancing diagnostic accuracy, particularly when large volumes of data are available. Alaa and van der Schaar [10] investigated personalized risk prediction by enhancing survival models with machine learning approaches. Their study highlights the value of integrating time-dependent features and longitudinal datasets in modeling chronic conditions, which is especially significant for managing hypertension.

Beyond clinical metrics, several studies have also considered socio-economic and behavioral factors for a more comprehensive prediction model. For instance, Mandal and Saha [26] incorporated health behavior information into machine learning algorithms such as Naïve Bayes, SVM, and Random Forest, resulting in improved prediction accuracy and earlier identification of individuals at high risk. Furthermore, recent reviews [24] have thoroughly examined the use of machine learning in cardiovascular and hypertension-related conditions, emphasizing that no single algorithm consistently outperforms others across all use cases. Rather, optimal outcomes are often achieved through model tuning and hybrid methodologies.

Although there is substantial research in this area, few studies have conducted a detailed, side-by-side comparison of multiple ML algorithms using the same dataset specifically for hypertension prediction and classification. This shortfall underpins our study, in which we systematically assess various machine learning techniques—including Logistic Regression, Decision Trees, Random Forest, XGBoost, Support Vector Machines, and Multi-layer Perceptron—to determine the most effective models for both prediction and classification in the context of hypertension diagnosis.

III. PROPOSED METHODOLOGY

Our proposed approach seeks to determine the most effective algorithm for prediction of hypertension risk by assessing the performance of fifteen different machine-learning models. [38,40,41,42,43,44].

3.1. Random Forest: Random Forest is an ensemble of Decision Trees trained on random subsets of data and features.

Model: For T trees, each tree outputs a prediction $\hat{y}_t$. The final prediction is determined by majority voting: $\hat{y} = mode(\{\hat{y}_1, \hat{y}_2, \hat{y}_3, \dots\dots \hat{y}_T\})$.

Feature Randomness: Each tree uses a random subset of features at each split to reduce correlation between trees.

3.2. Logistic Regression: Logistic Regression models the probability of a binary outcome using a logistic (sigmoid) function. For a given input feature vector $x = [x_1, x_2, x_3 \dots\dots x_n]$

Model: $P\left(x = \frac{1}{X}\right) = \sigma(W^T X + p)$,where $\sigma(z) = \frac{1}{1+e^{-z}}$ (sigmoid function), $W = [w_1, w_2, w_3 \dots\dots w_n]$ are the model weights, b is the bias term, $P\left(y = \frac{0}{x}\right) = 1 - P(y = \frac{1}{x})$.

Decision Rule: $\hat{y} = \begin{cases} 1 & If\ P(y = \frac{1}{x}) \geq 0.5 \\ 0 & Otherwise \end{cases}$

3.3. Decision Tree: Decision Tree recursively splits the data based on feature values to maximize a purity criterion (e.g., Gini index or Information Gain).

Splitting Criterion: For a node with N samples:

Gini Index: $G = 1 - \sum_{i=1}^{k} p_i^2$, where P_i is the proportion of samples belonging to class i.

Information Gain: $= H_{Parent} - \sum_i \frac{N_i H_i}{N}$, where $H = -\sum p_i \log_2 p_i\ (Entropy)$

Decision Rule: Traverse the tree according to feature splits until a leaf node is reached. The class label of the leaf node is the prediction.

3.4. Support Vector Machine (SVM): SVM finds a hyperplane that maximizes the margin between two classes in a feature space.

Model: For a feature vector x: $f(x) = w^T x + b$, Where: w is the weight vector, b is the bias.

Optimization Problem: $\min_{w,b} \frac{1}{2} \|w\|^2$ subject to $y_i(w^T x_i + b) \geq 1, \forall\ i$.

Decision Rule: $\hat{y} = \begin{cases} 1 & Iff(x) \geq 0 \\ -1 & Otherwise \end{cases}$

Kernel Trick: To handle non-linear data, SVM uses kernels $K(X_i, X_j)$ to map data into a higher-dimensional space.

3.5. K-Nearest Neighbors (KNN): KNN predicts the label of a sample based on the labels of its k nearest neighbors in the feature space.

Model: Given a distance metric d(x_i,x_j), identify the k nearest neighbors of a query point x_q.

Decision Rule: $\hat{y} = mode(\{\hat{y}_1, \hat{y}_2, \hat{y}_3, \dots\dots \hat{y}_k\})$. Where are the labels of the k nearest neighbors.

Euclidean: $\sqrt{\sum_n \left(x_{i,n} - x_{j,n}\right)^2}$

3.6. Naive Bayes (Gaussian): Model Assumptions: Features are conditionally independent given the class. Each feature follows a Gaussian (normal) distribution.

Formula: For a feature vector $X = [x_1, x_2, x_3 \dots\dots x_n]$ the posterior probability is:

$P(C_k|X) = \frac{P(C_k) \prod_{i=1}^{n} P(x_i|C_k)}{P(X)}$, where $P(x_i|C_k)$ is modelled as: $P(x_i|C_k) = \frac{1}{\sqrt{2\pi\sigma_{k,i}^2}} \exp\left(-\frac{\left(x_i - \mu_{k,i}\right)^2}{2\pi\sigma_{k,i}^2}\right)$

With $\mu_{k,i}$ and $\sigma_{k,i}^2$ being the mean and variance of feature i for class k.

3.7. Naive Bayes (Bernoulli): Model Assumptions: Features are binary. Features are conditionally independent given the class.

Formula: For binary features X: $P(C_k|X) = \frac{P(C_k) \prod_{i=1}^{n} P(x_i|C_k)^{x_i} (1 - P(x_i|C_k))^{1-x_i}}{P(X)}$

3.8. XGBoost (Extreme Gradient Boosting): Model: XGBoost is an ensemble method based on decision trees optimized with gradient boosting.

Objective Function: $\mathcal{L} = \sum_{i=1}^{n} l\,(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$ where: $l(y_i, \hat{y}_i)$ is the loss function (e.g., log loss for classification).

$\Omega(f_k) = \Upsilon T + \frac{1}{2}\lambda \|w\|^2$ a regularization term to penalize tree complexity.

Prediction: $\hat{y}_i = \sum_{k=1}^{K} f_k(X_i)$ where f_k is the k-th decision tree.

3.9. Cat Boost: Model: CatBoost is a gradient-boosting algorithm specifically designed for categorical features.

Objective Function: Same as XGBoost, but with additional handling of categorical features using "ordered boosting."

Handling Categorical Features: Transform categorical features into numeric representations using statistics derived from the training data.

3.10. AdaBoost Classifier: Model: AdaBoost combines weak learners (typically decision stumps) iteratively.

Prediction: $H(X) = sign(\sum_{t=1}^{T} \alpha_t h_t(X))$, where: $h_t(X)$: Weak classifier at iteration t. α_t : Weight of h_t, determined by its accuracy.

Update Rule: Weights for misclassified samples are increased: $w_i, t + 1 = w_i, t$ exp $(\alpha_t . 1_{y_{i \neq h_t(x_i)}})$

3.11. Extra Trees Classifier (Extremely Randomized Trees): Model: An ensemble of randomized decision trees.

Split Criteria: Unlike standard decision trees, Extra Trees randomly selects Features for split. Split thresholds within the feature range.

Prediction: $H(X) = \frac{1}{T}\sum_{t=1}^{T} h_t(X)$, where T is the number of trees.

3.12. Multi-Layer Perceptron (MLP): Objective: MLP is a specific type of feedforward neural network composed of multiple layers, including one or more hidden layers.

Mathematical Formulation: Layer-Wise Computation: For layer : $z^{(l)} = \sigma\left(W^{(l)}\alpha^{(l-1)} + b^{(l)}\right)$

where: $z^{(l)}$ is the pre-activation output, $a^{(l-1)}$ is the output of the previous layer, $W^{(l)}$ and $b^{(l)}$ are weights and biases for layer σ(·) is the activation function.

3.13. Stochastic Gradient Descent (SGD) Classifier: SGD minimizes a loss function L(w) using iterative updates with a learning rate η: Objective: $L(w) = \frac{1}{n}\sum_{i=1}^{n} l\,(y_i, f(x_i; w))$

Where: $l(y_i, f(x_i; w))$: The loss for the i-th sample. $f(x_i; w)$: The predicted output for x_i based on weights w.

Update Rule: $w^{t+1} = w^t - \eta \nabla l\left(y_i, f(x_i; w)\right)$, where: η: Learning rate. $\nabla l\left(y_i, f(x_i; w)\right)$ Gradient of the loss with respect to weights w, computed for a single sample i.

Common loss functions: Hinge loss for SVM: $l\left(y, f(x; w)\right) = \max\,(0{,}1 - w. f(x, w))$

Logistic loss for logistic regression: $l\left(y, f(x; w)\right) = \log\,(1 + \exp\,(-y. f(x, w)))$.

3.14. Bagging Classifier: Bagging (Bootstrap Aggregating) is an ensemble technique that combines the predictions of multiple base estimators (e.g., decision trees).

Objective: The final prediction f(x) is the aggregate of individual predictions from B base models: $f(x) = \frac{1}{B}\sum_{b=1}^{B} f_b\,(x)$. Where: $f_b(x)$ The prediction of the b-th base model. B: Total number of base models.

Workflow: Randomly sample m datasets with replacement from the original dataset. Train a base model $f_b(x)$ on each sampled dataset. Aggregate predictions by: Majority vote for classification. Averaging for regression.

3.15. Gradient Boosting

Gradient Boosting is an iterative process where weak learners (usually decision trees) are added sequentially to minimize a loss function. Its mathematics is based on functional gradient descent.

Loss Function: Let $L(y, \hat{y})$be the loss function to be minimized (e.g., log loss for classification, MSE for regression). The goal is to find a function F(x) such that the predictions F(x) minimize this loss: $F(x) = arg \min_{F(x)} \sum_{i=1}^{n} L(y_i, F(x_i))$

Additive Model: Gradient Boosting builds an additive model: $F_m(x) = F_{m-1}(x) + \alpha. h_m(x)$

where $F_{m-1}(x)$ is the prediction from the previous iteration, $h_m(x)$ is the new decision tree (or weak learner), and α is the learning rate.

Gradient Descent: The new learner $h_m(x)$ is fitted to the negative gradient of the loss function with respect to the current predictions: $h_m(x) = arg \min_{F(x)} \sum_{i=1}^{n} \left[-\frac{\partial L(y_i, F_{m-1}(x))}{\partial F_{m-1}(x)} - h(x_i)\right]^2$

Final Prediction: After M iterations, the final prediction is: $\hat{y} = F_m(x) = \sum_{m=1}^{M} \alpha h_m(x)$.

IV. CONFUSION MATRIX IN MACHINE LEARNING

It enables a deeper understanding of the model's recall, accuracy, precision, and overall ability to distinguish between classes by showing the frequency of predicted outcomes on the test dataset[40,41,42,43,44].

4.1 Accuracy: $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$, where TP= True positives, TN= True negatives, FP= False positives and FN= False negatives.

4.2 Precision: It is calculated as the proportion of true positive predictions to the total positive predictions made by the model.

$$Precision = \frac{TP}{TP + FP}$$

4.3 Recall: It is determined by dividing the count of true positives (TP) by the sum of true positives and false negatives (FN).

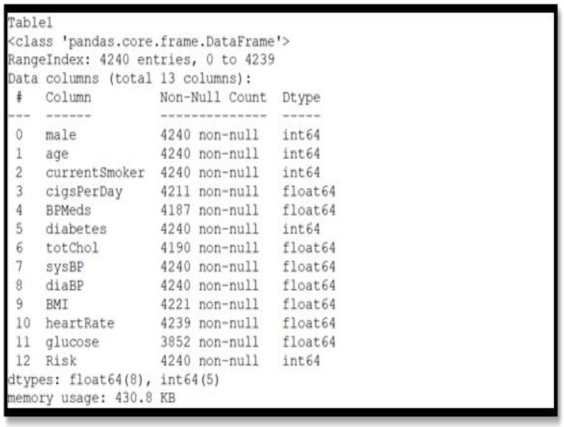
$$Recall = \frac{TP}{TP + FN}$$

4.4 Specificity: Specificity is a crucial metric for assessing classification models, especially in binary situations, as it gauges the model's ability to correctly identify negative instances, also referred to as the True Negative Rate.

$$Specificity = \frac{TN}{TP + FP}$$

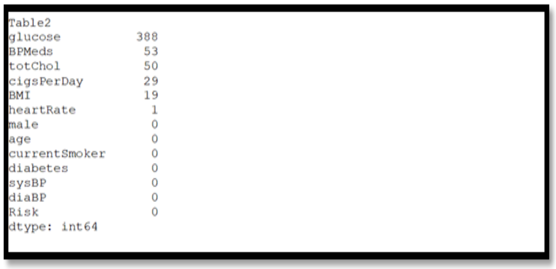
V. DATA CLEANING AND FEATURE ENGINEERING

To test hypertension, we collected data from(https://www.kaggle.com/datasets/hypertension-risk-model-main/code), The dataset consists of 4,240 entries across 13 columns, including five integer-type and eight float-type attributes and employed them to construct models for detection, forecasting, and categorization of hypertension risk. Table 1 is showing attributes of the dataset.



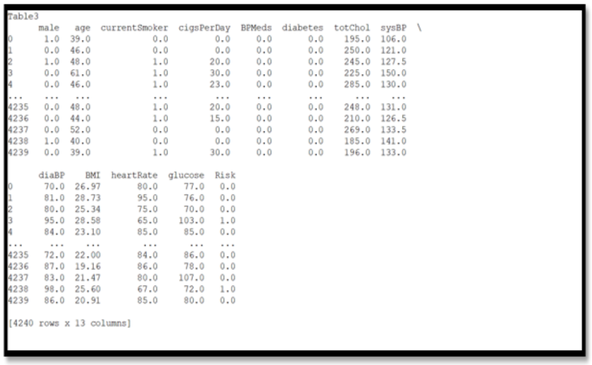
```
Table1
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4240 entries, 0 to 4239
Data columns (total 13 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   male           4240 non-null   int64
 1   age            4240 non-null   int64
 2   currentSmoker  4240 non-null   int64
 3   cigsPerDay     4211 non-null   float64
 4   BPMeds         4187 non-null   float64
 5   diabetes       4240 non-null   int64
 6   totChol        4190 non-null   float64
 7   sysBP          4240 non-null   float64
 8   diaBP          4240 non-null   float64
 9   BMI            4221 non-null   float64
 10  heartRate      4239 non-null   float64
 11  glucose        3852 non-null   float64
 12  Risk           4240 non-null   int64
dtypes: float64(8), int64(5)
memory usage: 430.8 KB
```

The columns male, age, current Smoker, diabetes, SYSBP, DIABP, and Risk are fully populated, while CIGSPERDAY, BPMEDS, TOTCHOL, BMI, heartrate, and glucose contain missing values to varying extents. Notably, glucose has the most missing data, with 388 entries lacking values. The dataset's attributes encompass demographic information (male, age), lifestyle factors (current Smoker, CIGSPerDay), health conditions (diabetes, BPMeds), clinical metrics (TOTCHOL, SYSBP, DIABP, BMI, heartrate, glucose), and an outcome variable (Risk). It occupies approximately 430.8 KB of memory. Table2 represents details of missing values:



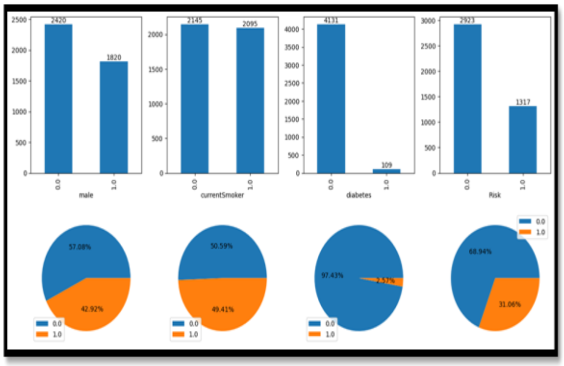
```
Table2
glucose          388
BPMeds            53
totChol           50
cigsPerDay        29
BMI               19
heartRate          1
male               0
age                0
currentSmoker      0
diabetes           0
sysBP              0
diaBP              0
Risk               0
dtype: int64
```

To handle missing values in the dataset, we can use an imputer that employs the median strategy. The median is a robust measure of central tendency, which works well when the data may contain outliers or skewed distributions. By filling missing values with the median, we ensure that the imputation does not introduce bias that could arise from extreme values. Which shown in table3:

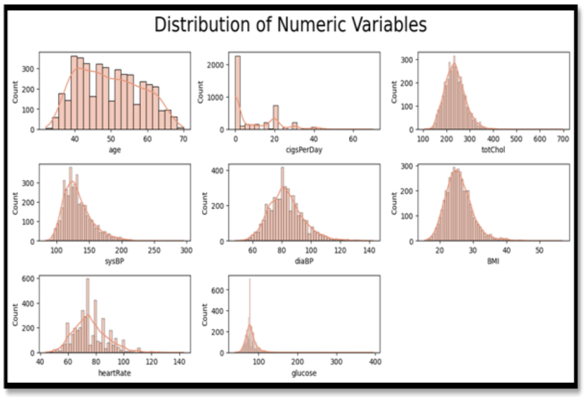


```
Table3
      male   age  currentSmoker  cigsPerDay  BPMeds  diabetes  totChol  sysBP  \
0      1.0  39.0            0.0         0.0     0.0       0.0    195.0  106.0
1      0.0  46.0            0.0         0.0     0.0       0.0    250.0  121.0
2      1.0  48.0            1.0        20.0     0.0       0.0    245.0  127.5
3      0.0  61.0            1.0        30.0     0.0       0.0    225.0  150.0
4      0.0  46.0            1.0        23.0     0.0       0.0    285.0  130.0
...    ...   ...            ...         ...     ...       ...      ...    ...
4235   0.0  48.0            1.0        20.0     0.0       0.0    248.0  131.0
4236   0.0  44.0            1.0        15.0     0.0       0.0    210.0  126.5
4237   0.0  52.0            0.0         0.0     0.0       0.0    269.0  133.5
4238   1.0  40.0            0.0         0.0     0.0       0.0    185.0  141.0
4239   0.0  39.0            1.0        30.0     0.0       0.0    196.0  133.0

      diaBP    BMI  heartRate  glucose  Risk
0      70.0  26.97       80.0     77.0   0.0
1      81.0  28.73       95.0     76.0   0.0
2      80.0  25.34       75.0     70.0   0.0
3      95.0  28.58       65.0    103.0   1.0
4      84.0  23.10       85.0     85.0   0.0
...     ...    ...        ...      ...   ...
4235   72.0  22.00       84.0     86.0   0.0
4236   87.0  19.16       86.0     78.0   0.0
4237   83.0  21.47       80.0    107.0   0.0
4238   98.0  25.60       67.0     72.0   1.0
4239   86.0  20.91       85.0     80.0   0.0

[4240 rows x 13 columns]
```

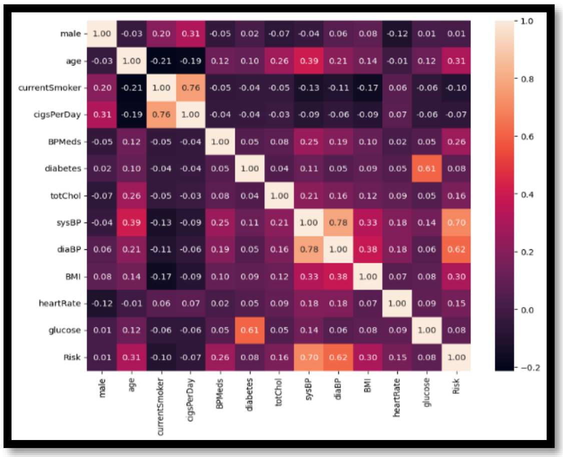
The dataset contains 4,240 entries across the following columns: "male," "current smoker," "diabetes," and "risk." Among these entries, there are 2,420 males and 1,820 females. Regarding smoking status, 2,095 are smokers and 2,145 are non-smokers. For diabetes, 109 individuals are diabetic, while 4,131 are non-diabetic. In terms of risk categories, 2,923 individuals fall under the low-risk category, while 1,317 are categorized as high-risk, indicating the presence of cardiovascular issues.



These attributes used in data analysis are related to health, particularly in predicting health risks such as cardiovascular diseases, diabetes, and kidney-related conditions.



The dataset shows various relationships among the attributes. Around 42.9% of the sample is male, and approximately half of the individuals smoke. Smokers exhibit considerable variation in cigarettes smoked per day. Only about 3% of individuals are on blood pressure medication, with higher blood pressure correlating with medication usage. There is a potential link between higher glucose levels and diabetes, as well as between cholesterol levels and blood pressure. BMI, which is in the overweight range on average, is associated with a higher risk of heart disease. Additionally, heart rate tends to increase with blood pressure, and age is positively correlated with heart disease risk.



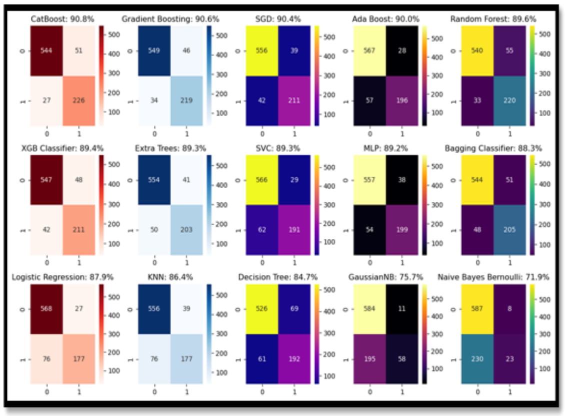
	male	age	currentSmoker	cigsPerDay	BPMeds	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	Risk
male	1.00	-0.03	0.20	0.31	-0.05	0.02	-0.07	-0.04	0.06	0.08	-0.12	0.01	0.01
age	-0.03	1.00	-0.21	-0.19	0.12	0.10	0.26	0.39	0.21	0.14	-0.01	0.12	0.31
currentSmoker	0.20	-0.21	1.00	0.76	-0.05	-0.04	-0.05	-0.13	-0.11	-0.17	0.06	-0.06	-0.10
cigsPerDay	0.31	-0.19	0.76	1.00	-0.04	-0.04	-0.03	-0.09	-0.06	-0.09	0.07	-0.06	-0.07
BPMeds	-0.05	0.12	-0.05	-0.04	1.00	0.05	0.08	0.25	0.19	0.10	0.02	0.05	0.26
diabetes	0.02	0.10	-0.04	-0.04	0.05	1.00	0.04	0.11	0.05	0.09	0.05	0.61	0.08
totChol	-0.07	0.26	-0.05	-0.03	0.08	0.04	1.00	0.21	0.16	0.12	0.09	0.05	0.16
sysBP	-0.04	0.39	-0.13	-0.09	0.25	0.11	0.21	1.00	0.78	0.33	0.18	0.14	0.70
diaBP	0.06	0.21	-0.11	-0.06	0.19	0.05	0.16	0.78	1.00	0.38	0.18	0.06	0.62
BMI	0.08	0.14	-0.17	-0.09	0.10	0.09	0.12	0.33	0.38	1.00	0.07	0.08	0.30
heartRate	-0.12	-0.01	0.06	0.07	0.02	0.05	0.09	0.18	0.18	0.07	1.00	0.09	0.15
glucose	0.01	0.12	-0.06	-0.06	0.05	0.61	0.05	0.14	0.06	0.08	0.09	1.00	0.08
Risk	0.01	0.31	-0.10	-0.07	0.26	0.08	0.16	0.70	0.62	0.30	0.15	0.08	1.00

VI. PREDICTION OF HYPERTENSION

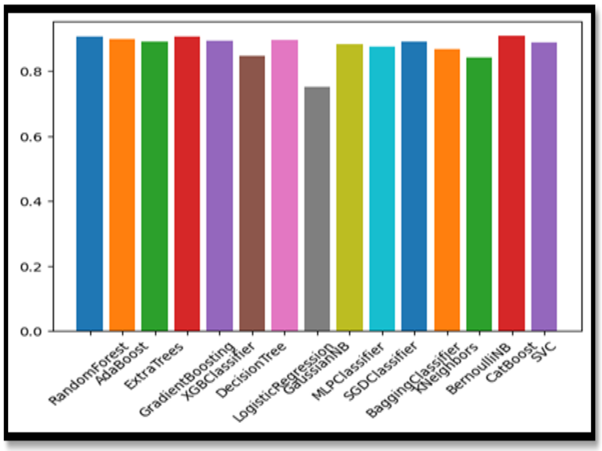
We assessed the performance of fifteen machine learning algorithms and then evaluated their potential to function as clinical decision support systems for predicting hypertension.

6.1. Examine the effectiveness of Machine Learning models by Confusion matrix:

To evaluate the model's performance, the data was divided into training, validation, and test sets. The data was standardized to ensure uniformity, a crucial step for many machine learning algorithms. During model development, clinical features were randomly selected from 80% of the patients to create the training dataset.



CatBoost and Gradient Boosting emerged as the best performers. Among these, CatBoost demonstrated slightly better overall performance, particularly due to its higher F1-score for class 1.0.AdaBoost and Random Forest delivered solid results but fell slightly short compared to CatBoost and Gradient Boosting in terms of performance. Models such as Gaussian Naive Bayes, Naive Bayes Bernoulli, and Decision Tree exhibited low performance, especially for class 1.0. These models should either be avoided or further optimized.



6.2. Selection of Models for Detecting, Forecasting and Categorizing Hypertension Risk:

CatBoost, with an accuracy of 0.908 and an F1 score of 0.909, stands out as the best-performing model, demonstrating excellence across all metrics. Its gradient boosting mechanism and strong handling of categorical variables make it highly effective for predicting hypertension risk. The Gradient Boosting Classifier, achieving an accuracy and F1 score of 0.906, delivers nearly comparable performance, balancing the trade-off between false positives and false negatives. Similarly, the Random Forest Classifier, with an accuracy of 0.904 and an F1 score of 0.906, provides consistent results by utilizing an ensemble learning technique that combines multiple decision trees for increased robustness and accuracy. Accuracy , Precision , Sensitivity , Specificity and F1 score of all the models are shown in table4:

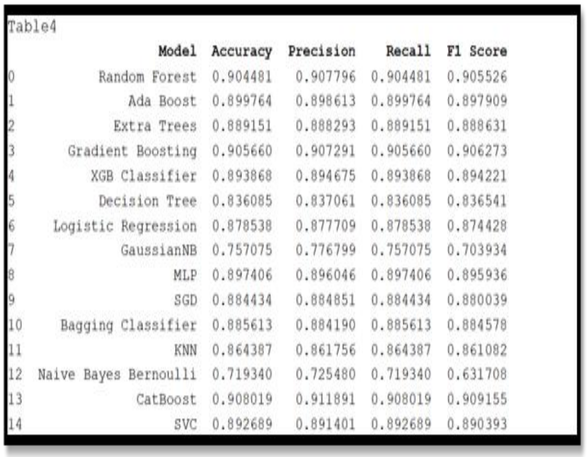
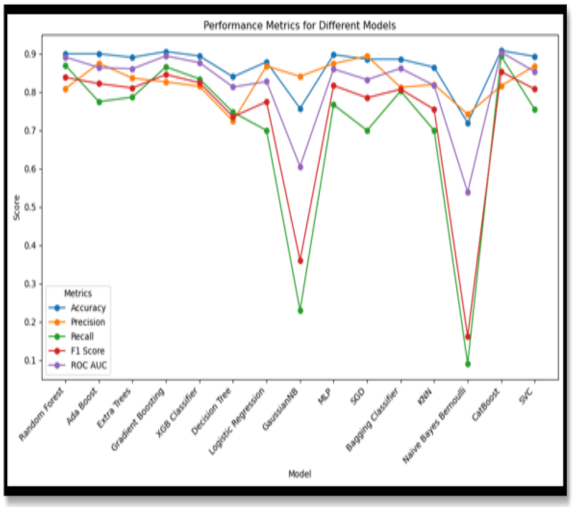
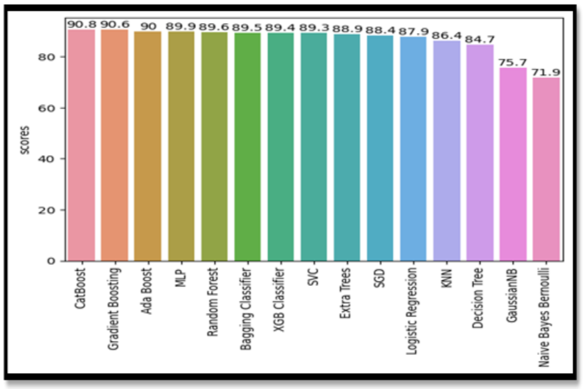


Table4

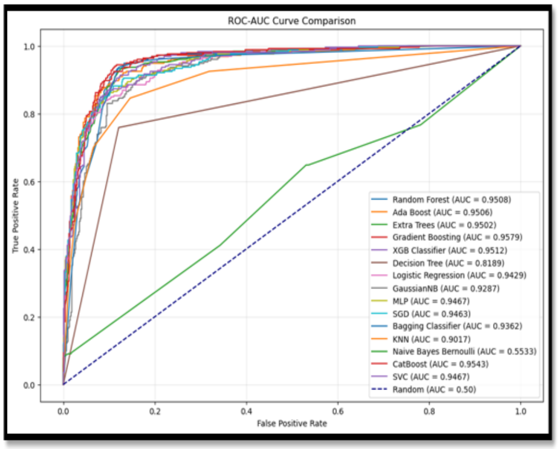
	Model	Accuracy	Precision	Recall	F1 Score
0	Random Forest	0.904481	0.907796	0.904481	0.905526
1	Ada Boost	0.899764	0.898613	0.899764	0.897909
2	Extra Trees	0.889151	0.888293	0.889151	0.888631
3	Gradient Boosting	0.905660	0.907291	0.905660	0.906273
4	XGB Classifier	0.893868	0.894675	0.893868	0.894221
5	Decision Tree	0.836085	0.837061	0.836085	0.836541
6	Logistic Regression	0.878538	0.877709	0.878538	0.874428
7	GaussianNB	0.757075	0.776799	0.757075	0.703934
8	MLP	0.897406	0.896046	0.897406	0.895936
9	SGD	0.884434	0.884851	0.884434	0.880039
10	Bagging Classifier	0.885613	0.884190	0.885613	0.884578
11	KNN	0.864387	0.861756	0.864387	0.861082
12	Naive Bayes Bernoulli	0.719340	0.725480	0.719340	0.631708
13	CatBoost	0.908019	0.911891	0.908019	0.909155
14	SVC	0.892689	0.891401	0.892689	0.890393





VII. RESULT

In this study, we assessed 15 machine learning algorithms based on 13 clinical features extracted from electronic medical records.

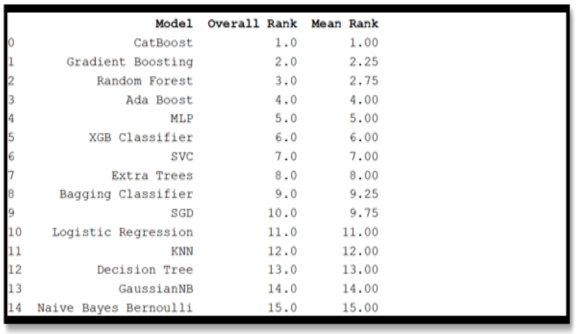


CatBoost emerges as the top-performing model, delivering the highest overall performance with an accuracy of 90.80%, recall of 89.32%, F1 score of 85.28%, and an impressive ROC AUC of 95.43%. Its ability to effectively identify true positive cases makes it highly suitable for forecasting and categorizing hypertension risk. Gradient Boosting closely follows with an accuracy of 90.33%, recall of 86.95%, F1 score of 84.29%, and an ROC AUC of 95.79%. While it slightly trails CatBoost, its competitive metrics establish it as a strong alternative for similar use cases.

Random Forest provides solid performance with an accuracy of 89.74% and balanced metrics across precision, recall, and ROC AUC (95.08%). Its robustness and versatility make it well-suited for general scenarios and handling noisy data effectively.

AdaBoost prioritizes precision, achieving 87.50%, but has a lower recall of 77.47%. This model is particularly advantageous when minimizing false positives is more critical than capturing all true positives.

Both CatBoost and Gradient Boosting stand out as the most effective models for identifying at-risk patients, due to their high recall and ROC AUC scores. Their predictive accuracy and reliability make them ideal choices for healthcare risk forecasting and categorization.



	Model	Overall Rank	Mean Rank
0	CatBoost	1.0	1.00
1	Gradient Boosting	2.0	2.25
2	Random Forest	3.0	2.75
3	Ada Boost	4.0	4.00
4	MLP	5.0	5.00
5	XGB Classifier	6.0	6.00
6	SVC	7.0	7.00
7	Extra Trees	8.0	8.00
8	Bagging Classifier	9.0	9.25
9	SGD	10.0	9.75
10	Logistic Regression	11.0	11.00
11	KNN	12.0	12.00
12	Decision Tree	13.0	13.00
13	GaussianNB	14.0	14.00
14	Naive Bayes Bernoulli	15.0	15.00

The CatBoost algorithm stand out as the leading models, outperforming the others.

VIII. CONCLUSION

Machine-learning algorithms present a promising approach to improving hypertension risk management, offering substantial benefits in detection, forecasting, and categorization. The analysis reveals that certain algorithms, such as CatBoost, Gradient Boosting, Random Forest and AdaBoost, have shown strong performance in predicting hypertension risk with high accuracy, precision, recall, F1 score and ROC AUC. However, the variability in results across different datasets and the complexity of model interpretability remain challenges that need addressing. Despite these challenges, machine learning holds significant potential for enhancing healthcare outcomes by providing more personalized and timely interventions for patients at risk of hypertension. Continued research and development in this field, alongside collaboration with healthcare professionals, will be vital to refining these models and ensuring their widespread, effective implementation in clinical practice.

Abbreviations List

AI – Artificial Intelligence

ANN – Artificial Neural Network

AUC – Area Under the Curve

BMI – Body Mass Index

BP – Blood Pressure

CNN – Convolutional Neural Network

DT – Decision Tree

ECG – Electrocardiogram

EHR – Electronic Health Records

F1-score – F1 Measure (Harmonic Mean of Precision and Recall)

HTN – Hypertension

KNN – K-Nearest Neighbors

LGBM – Light Gradient Boosting Machine

LR – Logistic Regression

LSTMs – Long Short-Term Memory Networks

MIMIC – Multi-parameter Intelligent Monitoring in Intensive Care

ML – Machine Learning

MLP – Multilayer Perceptron

NB – Naïve Bayes

NHANES – National Health and Nutrition Examination Survey

NHIC – National Health Insurance Corporation

PPG – Photoplethysmography

RF – Random Forest

ROC–AUC – Receiver Operating Characteristic – Area Under the Curve

SVM – Support Vector Machine

WHO – World Health Organization

DECLARATIONS

Ethics Approval and Consent to Participate: Not applicable. This study utilized publicly available datasets and did not involve human participants, clinical trials, or experimental interventions requiring ethical approval.

Consent for Publication: Not applicable. This manuscript does not include individual-level data or identifying information requiring explicit consent for publication.

Availability of Data and Materials: The datasets used and analyzed in this study are publicly available from Kaggle and other open-access sources. Further details regarding data availability can be provided upon reasonable request.

Competing Interests: The author declares no competing financial or non-financial interests in relation to this study.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' Contributions: Dr. D. P. Singh is the sole author of this study. He conceptualized the research, conducted data analysis, implemented machine learning models, interpreted the results, and wrote the manuscript.

ACKNOWLEDGMENT

I express my deep gratitude to Amity University Uttar Pradesh, Greater Noida Campus, for providing an excellent research environment and necessary resources for conducting this study. I sincerely appreciate the support and encouragement from my colleagues, whose insightful discussions and constructive feedback have been invaluable in shaping this research. I am also thankful to the researchers and scholars whose previous works have significantly contributed to the advancement of machine learning applications in hypertension risk detection, forecasting, and classification. Their contributions have served as a strong foundation for this study.

Conflicts of Interest: The author declares no conflicts of interest regarding this research.

Author Information:

D. P. Singh, Professor

dr.dps97@gmail.com

Amity University Uttar Pradesh, Greater Noida Campus

ID: https://orcid.org/0000-0001-9494-4296

REFERENCES

[1]. Rocha E. Fifty Years of Framingham Study Contributions to Understanding Hypertension. Rev. Port. Cardiol. 2001, 20, 795–796,

https://doi.org/10.1038/sj.jhh.1000949 PMID: 11582630

[2]. Kearney P.M.; Whelton M.; Reynolds K.; Muntner P.; Whelton P.K.; He J. Global Burden of Hypertension: Analysis of Worldwide Data; 2005; Vol. 365; ISBN 1992932069, https://doi.org/10.1016/S0140- 6736(05)17741-1 PMID: 15652604

[3]. Lawes C.M.M.; Hoorn S. Vander; Rodgers A. Global Burden of Blood-Pressure-Related Disease, 2008; Vol. 371, https://doi.org/10.1016/S0140-6736(08)60655-8 PMID: 18456100

[4]. Lloyd-Jones D.; Adams R.J.; Brown T.M.; Carnethon M.; Dai S.; De Simone G.; et al. Executive Summary: Heart Disease and Stroke Statistics-2010 Update: A Report from the American Heart Association. Circulation 2010, 121, https://doi.org/10.1161/CIRCULATIONAHA.109.192666 PMID: 20177011

[5]. Forouzanfar M.H.; Afshin A.; Alexander L.T.; Biryukov S.; Brauer M.; Cercy K.; et al. Global, Regional, and National Comparative Risk Assessment of 79 Behavioural, Environmental and Occupational, and Metabolic Risks or Clusters of Risks, 1990–2015: A Systematic Analysis for the Global Burden of Disease Study 2015. Lancet 2016, 388, 1659–1724, https://doi.org/10.1016/S0140-6736(16)31679-8 PMID: 27733284

[6]. Usher-Smith J.A.; Silarova B.; Schuit E.; Moons K.G.M.; Griffin S.J. Impact of Provision of Cardiovascular Disease Risk Estimates to Healthcare Professionals and Patients: A Systematic Review. BMJ Open 2015, 5, https://doi.org/10.1136/bmjopen-2015-008717 PMID: 26503388

[7]. Lopez-Gonzalez A.A.; Aguilo A.; Frontera M.; Bennasar-Veny M.; Campos I.; Vicente-Herrero T.; et al. Effectiveness of the Heart Age Tool for Improving Modifiable Cardiovascular Risk Factors in a Southern European Population: A Randomized Trial. Eur. J. Prev. Cardiol. 2015, 22, 389–396, https://doi.org/10.1177/2047487313518479 PMID: 24491403

[8]. Kalantari S.; Khalili D.; Asgari S.; Fahimfar N.; Hadaegh F.; Tohidi M.; et al. Predictors of Early Adulthood Hypertension during Adolescence: A Population-Based Cohort Study. BMC Public Health 2017,17, https://doi.org/10.1186/s12889-017-4922-3 PMID: 29183297

[9]. Whelton P.K.; Carey R.M.; Aronow W.S.; Casey D.E.; Collins K.J.; Dennison Himmelfarb C.; 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: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Pr. Circulation 2018, 138, e484–e594, https://doi.org/10.1161/CIR.0000000000000596 PMID: 30354654

[10]. A. M. Alaa and M. van der Schaar, "Forecasting individualized disease trajectories using interpretable deep learning," Nature Communications, vol. 10, no. 1, pp. 1–10, 2019.

[11]. Beaney, T., Burrell, L.M., Castillo, R.R., Charchar, F.J., Cro, S., Damasceno, A., Kruger, R., Nilsson, P.M., Prabhakaran, D., Ramirez, A.J., Schlaich, M.P., Schutte, A.E., Tomaszewski, M., Touyz, R., Wang, J.G., Weber, M.A., Poulter, N.R., the MMM Investigators: May Measurement Month 2018: a pragmatic global screening campaign to raise awareness of blood pressure by the International Society of Hypertension. European Heart Journal 40(25), (2019). https:// doi. org/ 10. 1093/eurhe artj/ ehz300

[12]. C. Krittanawong et al., "Deep learning for cardiovascular medicine: a practical primer," European Heart Journal, vol. 40, no. 25, pp. 2058–2073, 2019.

[13]. Chowdhury M.Z.I.; Yeasmin F.; Rabi D.M.; Ronksley P.E.; Turin T.C. Prognostic Tools for Cardiovascular Disease in Patients with Type 2 Diabetes: A Systematic Review and Meta-Analysis of C-Statistics. J. Diabetes Complications 2019, 33, 98–111, https://doi.org/10.1016/j.jdiacomp.2018.10.010 PMID:30446478

[14]. Collaborators, G..R.F.: Global burden of 87 risk factors in 204 countries and territories, 1990-2019: a systematic analysis for the global burden of disease study 2019. The Lancet 396(10258),1223–1249 (2020)

[15]. Topol, E.: High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine 25(1), 44–56 (2019).https:// doi. org/ 10. 1038/ s41591- 018- 0300-7

[16]. S. Patel, M. Patel, and H. G. Patel, "Comparative analysis of machine learning approaches for hypertension prediction," International Journal of Computer Applications, vol. 182, no. 45, pp. 1–6, 2019.

[17]. Ghosh S.; Kumar M. Prevalence and Associated Risk Factors of Hypertension among Persons Aged 15–49 in India: A Cross-Sectional Study. BMJ Open 2019, 9, https://doi.org/10.1136/bmjopen-2019-029714 PMID: 31848161

[18]. Chang W.; Liu Y.; Xiao Y.; Yuan X.; Xu X.; Zhang S.; et al. A Machine-Learning-Based Prediction Method for Hypertension Outcomes Based on Medical Data. Diagnostics 2019, 9, https://doi.org/10.3390/diagnostics9040178 PMID: 31703364

[19]. Chowdhury M.Z.I.; Naeem I.; Quan H.; Leung A.A.; Sikdar K.C.; O'Beirne M.; et al. Summarising and Synthesising Regression Coefficients through Systematic Review and Meta-Analysis for Improving Hypertension Prediction Using Metamodelling: Protocol. BMJ Open 2020, 10, https://doi.org/10.1136/ bmjopen-2019-036388 PMID: 32276958

[20]. Katherine T., Mills A.S.& J.H. The Global Epidemiology of Hypertension _ Nature Reviews Nephrology. Nat. Rev. Nephrol. 2020, 16, 223–237, https://doi.org/10.1038/s41581-019-02442

[21]. KCDC Korea Centers for Disease Control and Prevention. Press Release. [Internet], Http://Knhanes. Cdc.Go.Kr. 2020.

[22]. Chowdhury M.Z.I.; Turin T.C. Precision Health through Prediction Modelling: Factors to Consider before Implementing a Prediction Model in Clinical Practice. J. Prim. Health Care 2020, 12, 3–9, https://doi.org/10.1071/HC19087 PMID: 32223844

[23]. M. Nour, K. Polat, Automatic classification of hypertension types based on personal features by machine learning algorithms, Mathematical Problems in Engineering (2020).

[24]. A. Ahmad et al., "Machine learning methods for heart disease prediction: A review and current status," Computer Methods and Programs in Biomedicine, vol. 207, pp. 106193, 2021.

[25]. CDC. Hypertension Prevalence in the U.S. | Million Hearts®. In: Centers for Disease Control and Prevention [Internet]. 22 Mar 2021. Available: Https://Millionhearts.Hhs.Gov/Data-Reports/Hypertension_prevalence.Html.

[26]. D. Mandal and S. Saha, "Prediction of hypertension using machine learning algorithms in a primary health care setup," Informatics in Medicine Unlocked, vol. 23, pp. 100540, 2021.

[27]. M. Shoaib, S. Rauf, and T. A. Khan, "Machine learning-based risk prediction for hypertension using behavioral and physiological attributes," Journal of Healthcare Engineering, vol. 2021, pp. 1–9, 2021.

[28]. Parati G, Stergiou GS, Bilo G, Kollias A, Pengo M, Ochoa JE, Agarwal R, Asayama K, Asmar R, Burnier M, De La Sierra A, Giannattasio C, Gosse P, Head G, Hoshide S, Imai Y, Kario K, Li Y, Manios E, Mant J, McManus RJ, Mengden T, Mihailidou AS, Muntner P, Myers M, Niiranen T, Ntineri A, O'Brien E, Octavio JA, Ohkubo T, Omboni S, Padfield P, Palatini P, Pellegrini D, Postel-Vinay N, Ramirez AJ, Sharman JE, Shennan A, Silva E, Topouchian J, Torlasco C, Wang JG, Weber MA, Whelton PK, White WB, Mancia G; Working Group on Blood Pressure Monitoring and Cardiovascular Variability of the European Society of Hypertension. Home blood pressure monitoring: methodology, clinical relevance and practical application. 2021 Sep 1;39(9):1742-1767. doi: 10.1097/HJH.0000000000002922. PMID: 34269334; PMCID: PMC9904446.

[29]. Sorato M.M.; Davari M.; Kebriaeezadeh A.; Sarrafzadegan N.; Shibru T. Societal Economic Burden of Hypertension at Selected Hospitals in Southern Ethiopia: A Patient-Level Analysis. BMJ Open 2022, 12, https://doi.org/10.1136/bmjopen-2021-056627 PMID: 35387822

[30]. T. H. Aldhyani, A. M. Alshebami, and S. S. Alzahrani, "Soft computing model to predict chronic diseases," Computers, Materials & Continua, vol. 67, no. 2, pp. 1933–1948, 2021.

[31]. Zhao H.; Zhang X.; Xu Y.; Gao L.; Ma Z.; Sun Y.; et al. Predicting the Risk of Hypertension Based on Several Easy-to-Collect Risk Factors: A Machine

Learning Method. Front. Public Heal. 2021, 9, https://doi.org/10.3389/fpubh.2021.619429 PMID: 34631636

[32]. Belay D.G.; Fekadu Wolde H.; Molla M.D.; Aragie H.; Adugna D.G.; Melese E.B.; et al. Prevalence and Associated Factors of Hypertension among Adult Patients Attending the Outpatient Department at the Primary Hospitals of Wolkait Tegedie Zone, Northwest Ethiopia. Front. Neurol. 2022, 13, https://doi.org/10.3389/fneur.2022.943595 PMID: 36034276

[33]. Chowdhury M.Z.I.; Naeem I.; Quan H.; Leung A.A.; Sikdar K.C.; OBeirne M.; et al. Prediction of Hypertension Using Traditional Regression and Machine Learning Models: A Systematic Review and Meta- Analysis. PLoS One 2022, 17, https://doi.org/10.1371/journal.pone.0266334 PMID: 35390039

[34]. Chowdhury M.Z.I.; Leung A.A.; Sikdar K.C.; O'Beirne M.; Quan H.; Turin T.C. Development and Validation of a Hypertension Risk Prediction Model and Construction of a Risk Score in a Canadian Population. Sci. Rep. 2022, 12, https://doi.org/10.1038/s41598-022-16904-x PMID: 35896590

[35]. Khongorzul D.; Kim M. Comparison of Feature Selection Methods Applied on Risk Prediction for Hypertension. KIPS Transactions on Software and Data Engineering, 11 (3), 107–114, https://doi.org/10. 3745/KTSDE.2022.11.3.107

[36]. Mehta R.; Mantri N.; Goel A.; Gupta M.; Joshi N.; Bhardwaj P. Out-of-Pocket Spending on Hypertension and Diabetes among Patients Reporting in a Health -Care Teaching Institute of the Western Rajasthan. J. Fam. Med. Prim. Care 2022, 11, 1083, https://doi.org/10.4103/jfmpc.jfmpc_998_21 PMID: 35495832

[37]. Rajpurkar, P., Chen, E., Banerjee, O., Topol, E.J.: AI in health and medicine. Nature Medicine 28(1), 31–38 (2022). https:// doi. org/10. 1038/ s41591- 021- 01614-0

[38].Singh, D. P.,Jassi J.S., Sunaina, 2023. Exploring the Significance of Statistics in the Research: A Comprehensive Overview, European Chemical Bulletin 12(Special Issue 2):2089-2102

[39]. Islam M.M.; Alam M.J.; Maniruzzaman M.; Ahmed N.A.M.F.; Ali M.S.; Rahman M.J.; et al. Predicting the Risk of Hypertension Using Machine Learning Algorithms: A Cross Sectional Study in Ethiopia. PLoS One 2023, 18, https://doi.org/10.1371/journal.pone.0289613 PMID: 37616271

[40]. Singh, D. P. (2024). An extensive examination of machine learning methods for identifying diabetes. Tuijin Jishu/Journal of Propulsion Technology, 45(2).

[41]. Singh, D. P. (2024). An extensive analysis of machine learning models to predict breast cancer recurrence. Tuijin Jishu/Journal of Propulsion Technology, 45(2).

[42]. Singh, D. P. (2024). A notable utilization of machine learning techniques in the healthcare sector for optimizing resources and enhancing operational efficiency. European Journal of Biomedical and Pharmaceutical Sciences, 11(7), 212–224. http://www.ejbps.com

[43]. Singh, D. P. (2024). An extensive analysis of machine learning techniques for predicting the onset of lung cancer. Tuijin Jishu/Journal of Propulsion Technology, 45(4).

[44]. Singh, D. P. (2024). Comprehensive analysis of machine learning models for cardiovascular disease detection and diagnosis. Tuijin Jishu/Journal of Propulsion Technology, 45(4).