JOIV : Int. J. Inform. Visualization, 8(2) - May 2024 1030-1039

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Machine Learning-Driven Stroke Prediction Using Independent Dataset

Fatin Natasha Binti Zahari [a], Kannan Ramakrishnan [a,*]

[a] *Faculty of Computing and Informatics, Multimedia University, 63100 Cyberjaya, Malaysia*

Corresponding author: [*]kannan.ramakrishnan@mmu.edu.my

***Abstract*— The incidence of stroke cases has witnessed a rapid global rise, affecting not only the elderly but also individuals across all age groups. Accurate prediction of stroke occurrence demands the utilization of extensive data pre-processing techniques. Moreover, the automation of early stroke forecasting is crucial to prevent its onset at the initial stage. In this study, stroke prediction models are evaluated to estimate the likelihood of stroke based on various symptoms such as age, gender, pre-existing medical conditions, and social variables. The machine learning techniques employed include Linear Support Vector Classifier, Extreme Gradient Boosting Classifier, Multilayer Perceptron, Adaptive Boosting Classifier, Bootstrap Aggregating Classifier, and Light Gradient-Boosting Machine. The purpose of this paper is to optimize the hyperparameters of machine learning approaches in developing stroke prediction models. The goal was achieved through a comprehensive comparison of three different sampling techniques for handling imbalanced datasets and evaluating their performance by using various metrics. The most effective model is identified, which is the Adaptive Boosting Classifier utilizing the Tomek Links, with a cross-dataset accuracy of 99% which demonstrated a reliable performance and generalization as evidenced by high cross-validation scores and accuracy on an independent dataset. The next stage of this endeavor entails looking into multiple ways to forecast the development of new dangerous diseases such as breast cancer and skin disorders. In the long run, the aim of subsequent work is to build a powerful toolset that is obtainable to all medical practitioners, allowing for the pre-emptive diagnosis of all potentially hazardous illnesses.**

***Keywords*—Stroke; machine learning; classification; multilayer perceptron.**

Manuscript received 5 Jan. 2023; revised 9 Aug. 2023; accepted 28 Nov. 2023. Date of publication 31 May 2024.

I. INTRODUCTION

Stroke is the most common but preventable cause of mortality and disability. If predicted early, individuals can prevent the stroke, by controlling modifiable lifestyle factors such as blood pressure, low-density lipoprotein (LDL) cholesterol, blood sugar, smoking, obesity, etc. A blockage or rupture of a blood artery in the brain is the cause of the stroke. A stroke may cause permanent impairment, such as partial paralysis and difficulties with speech, comprehension, and memory. The kind and severity of impairment are determined by the affected region of the brain and the duration of the blood supply obstruction.

Blood flow in the brain is thought to increase as neuron regeneration occurs in certain brain regions. It is transported via carotid and vertebral arteries. Blood then travels past the head to the heart's chambers via the inner jugular veins [1]. According to a study, Harrar et al. [2] stated that blood loss may occur in two situations: an ischemic stroke happens when blood flow between blood tissues declines, whereas a hemorrhagic stroke occurs when bleeding occurs under the surface within brain tissues. Rasmussen et al. [3] and Lattanzi & Silvestrini [4] discovered that this occlusion of the brain's arteries may occur when atherosclerosis-caused plaque fragments get damaged, causing a blood vessel clot. According to Verma et al. [5] and Boukobza et al. [6] publications, a hemorrhagic stroke is a severe stroke in which a ruptured artery causes bleeding or arterial burst. In contrast, when a coagulum forms in the heart rather than the brain, it causes an ischemic embolic stroke. Therefore, limiting the brain's arteries. For the elderly, a stroke may be fatal. A heart attack damages the heart and a stroke similarly harms the brain. Uppal et al. [7] explain that when blood is exposed and leaks, it opens up and exerts strain on the brain. The stroke begins with transient ischemic attacks known as ministrokes. It is the circumstance that indicates that the individual will have a stroke around a few days following the ministroke. If a stroke is recognized or diagnosed early, Lee et al. [8]

mentioned that death and serious brain damage may be prevented in 85% of cases.

Previous attempts to identify people at risk for stroke have focused on using hardware tools or generating predictions based on medical examination data, such as Magnetic Resonance Imaging (MRI). Even though stroke occurs with few, if any, warning signals (Centers for Disease Control and Prevention (CDC), [9], many assume that they are healthy and decline to have such tests, decreasing the efficacy of such measures.

This paper focuses on investigating the use of machine learning models using low-cost factors to forecast the likelihood of individuals developing a stroke disease, thus lowering the cost of medical check-ups while boosting the overall rate of survival. The number of individuals who are susceptible to stroke is expected to be far lower in real-world circumstances than the overall number of individuals reported by the great majority of hospitals. As a result, the dataset will always favor the minority group. It is critical for the proper execution of every research to create a baseline of goals that act as its route map. The goal of this paper is to optimize the hyperparameters of machine learning approaches in developing stroke prediction models and evaluate the effectiveness of the trained models on cross-datasets.

The academic community has focused on the development of instruments and techniques for observing and forecasting a variety of illnesses that have a substantial influence on people's well-being. Several models were constructed and assessed with the purpose of establishing a viable approach for protracted forecasting of stroke incidence. Dritsas & Trigka [10] identified the Stacking Classification (SC) to be their best model for identifying patients who were at a serious rate of having a stroke over a lengthy time frame since it produced a high performance over several criteria. This yields an accuracy of 98%, which made the proposed classification model outperform other approaches in the trial. In the meantime, Abedi et al. [11] showed that a fine-tuned training dataset including many features may be used to create models of stroke recurrence. The stroke forecast in recurrence within a one-year prediction window has an accuracy of 88%, a positive predictive value of 42%, and a specificity of 96% using Random Forest (RF) with up-sampling of the training dataset. On the other hand, Victor et al. [12] created a cost-effective solution to the imbalanced data problem associated with ECG datasets. The technique penalizes the minority class using class-imbalance-ratio-weight which utilizes the suggested model loss function without additional expense, attaining model generalizability. The study achieved the highest accuracy compared to existing works using a similar dataset of 98.14%.

There has been analysis made by previous reseach on predicting the incidence of stroke in patients using Electronic Health Records. It demonstrates that both efforts employ distinct datasets and techniques to predict stroke.

EHR data was employed as a feature in the development of all indicated prediction models. An EHR is a database containing patient data that includes the patient's vital statistics, diagnosis, and medical examination findings [13]. The optimal approaches that past researchers found to be useful for the prediction are Weighted Voting Classifer by Emon et al. [14], AdaBoost and J48 by Jalajajayalakshmi et al. [15], Decision Tree with removed outliers and application of Chi-square by Kavitha et al. [16], Neural Network by Rana et al. [17] and Biswas et al. [18]. Shafiul Azam et al. [19] found that the accuracy percentage by using Random Forest is significantly the highest which measured at 99.98% than that of other result indicating that the model used is reliable. Besides that, the Support Vector Classifier that has been employed by Biswas et al. [18].

Biswas et al. [18] also achieved a relatively high accuracy result of 99.9%. Alongside with the study analysis that has been made by Kavitha et al. [16], they found that Decision Tree with removed outliers and application of Chi-square results in the accuracy of 98.5%.

For predicting stroke using imbalanced datasets, it has been found that RF with the utilization of the Synthetic Minority Over-sampling Technique (SMOTE) to handle the imbalanced target variable has been used by Wu & Fang [20] and Ferdib-Al-Islam & Ghosh [21]. Among those two papers, Ferdib-Al-Islam & Ghosh [21].

Ferdib-Al-Islam & Ghosh [21] achieved a higher accuracy of 99.07%. However, a substantial difference can be seen in the sample sizes chosen by the two studies, with the latter using 5,110 observations to achieve a better level of precision. Phankokkruad & Wacharawichanant [22] found that their Extreme Gradient Boosting (XGBoost) model outperformed other models in the paper with the implementation of SMOTE to handle the imbalance between the classes. This investigation was done on two stroke datasets and the result indicates that XGBoost produces an accuracy of between 96.73% and 98.08%.

Based on the literature review, the following gaps have been identified and addressed within the scope of this paper. Firstly, stroke prediction methods that utilize visual processing & medical devices have a limitation in that it fails to be feasible since it needs respondents' cooperation, making it uncommon because stroke development is seen as unexpected with relatively infrequent symptoms [9]. The subsequent stroke prediction methods utilize electronic health information. The number of records is highly constrained for any similar research of this nature. Consequently, overfitting is a risk for machine learning models. Fang et al. [23] claim that the lack of several essential components makes the models less reliable for making predictions about the future. In addition to this, there was no extensive research on employing SMOTE + Tomek for predicting stroke with an imbalanced dataset. To the best of our knowledge, there is no previous research that has examined doing cross-validation on an independent dataset to assess the model's generalizability to forecast stroke.

II. Material and Method

This section explains the methodology that includes data collection, exploratory data analysis, data pre-processing, feature selection, data scaling, cross-validation, data sampling methods, dataset partitioning, and the evaluation criteria.

A. Methodology Approach and Design

The improvised Knowledge Discovery in Databases (KDD) [24] methodology was used for this work, as shown in Fig. 1. It consists of five distinct phases that include data collection,

data pre-processing, data mining, and interpretation or evaluation.

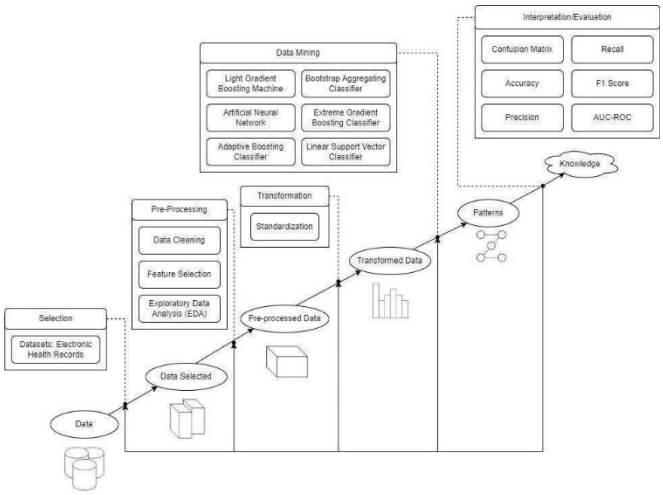

Fig. 1 Improvised Knowledge Discovery in Databases (KDD) Technique in Stroke Forecasting

The construction of the prediction starts from the data collection unit as the base cluster, which contains electronic health records. The dataset was loaded into Jupyter Notebook to process the data. The data processing consists of cleaning the data and employing feature selection. Transformation of the data was done by standardization. Then, training and testing are done on six distinct machine learning algorithms using a larger dataset to find the best-performing model by observing different evaluation metrics. The optimal model identified is used for cross-validation on an independent dataset.

B. Data Collection

This paper utilizes two stroke prediction datasets. The datasets have been collected from Kaggle. For a summary of the characteristics of the dataset, see Table 1. In the following sections, each dataset will be described in further depth.

TABLE I
DATASETS USED IN THE STUDY, NUMBER OF SAMPLES AND FEATURES

Dataset	Size	Features
Framingham Heart Disease Prediction Dataset	4,240	15
Heart Disease Health Indicators BRFSS2015	253,680	21

The Framingham heart Disease Prediction Dataset is publicly accessible on Kaggle [25] was gathered from on-going cardiovascular study involving residents of Framingham, Massachusetts. The collection contains 4,240 data with 15 characteristics, all of which pertain to patients. Each trait serves as a risk factor, which includes concerns about demographic, behavioral, and medical aspects. This dataset is being used to assess and validate the efficacy and accuracy of a prediction model. It is to be employed as a means of cross-testing the model's capabilities and guaranteeing its dependability.

Annually gathered by the Centers for Disease Control and Prevention (CDC), the Behavioral Risk Factor Surveillance System (BRFSS) is a health-related telephone survey (Alex Teboul, 2022) [26]. Each year, the survey gathers answers from more than 400,000 Americans about health-related risk behaviors, chronic health issues, and the use of preventative care. This has occurred annually since 1984. The dataset is accessible on Kaggle for the year 2015. The dataset comprises of 253,680 survey answers from the BRFSS 2015 that will be utilized largely for the binary categorization of stroke illness. The dataset is used for building machine learning models, giving them the knowledge needed to examine and identify patterns and correlations on different variables within the data. This can improve their ability to make precise predictions or carry out tasks based on the patterns learned during the training process.

C. Exploratory Data Analysis (EDA)

While attempting to analyze the class distribution of the dependent variable in both datasets, it was discovered that the data were insufficient, as seen in Fig. 2 and Fig. 3 respectively. It has been shown that both datasets are significantly imbalanced, with just 25 occurrences of stroke recorded against 4,215 instances of healthy cases in the Framingham dataset, while 10,276 observations for stroke against 208,318 entries among the healthy cases in the context of stroke in the BRFSS dataset.

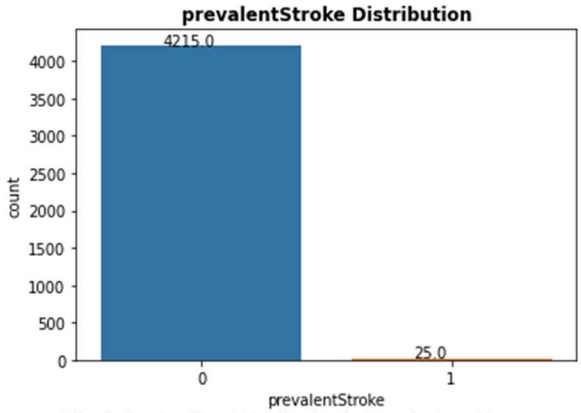

Fig. 2 Stroke Class Distribution in Framingham Dataset

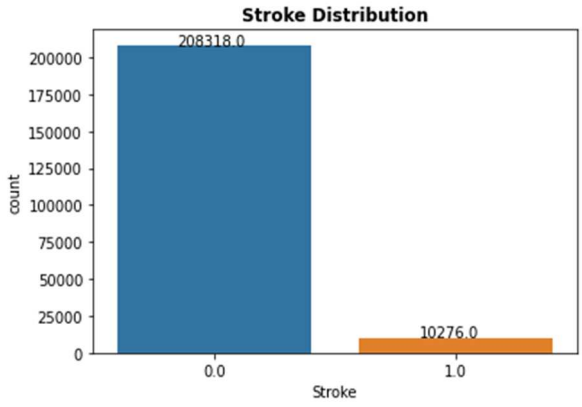

Fig. 3 Stroke Class Distribution in BRFSS Dataset

D. Data Pre-Processing

Data processing allows for the processing of raw data and the extraction of significant characteristics, making it ideal for application by machine learning models.

The 'male' column is supposed to represent the population's gender in the Framingham dataset. Hence, the

column can be renamed as 'Sex'. Besides that, '*currentSmoker*', '*prevalentHyp*', and '*prevalentStroke*' is also being renamed to '*Smoker*', '*HighBP*', and '*Stroke*' respectively to match with the name convention in the BRFSS dataset.

Since only the Framingham dataset contains missing values, the imputation will be applied solely to this dataset. For imputation, the central tendency measure, such as mean, median, or mode is examined. The objective is to determine the optimal measure of the data's central tendency and replace missing values accordingly.

Fig. 4 illustrates the comparison boxplot, which reveals that the data is skewed. There are a substantial number of data points that serve as outliers. Outlier data points will have a major influence on the mean; thus, it is not advised to utilize the mean to replace missing values in this scenario. For symmetric data distribution, one may impute missing values using the mean value.

In the data points of the comparison boxplot, there are multiple patients with high total cholesterol and glucose levels. The data seems right skewed which means it has a long tail in the right direction. The mode value will be substituted for missing data as part of the mode imputation procedure. For data points such as '*education*' and '*BPMeds*', it is recommended to replace the numbers with mode. Another approach is median imputation, which replaces missing data with the column's median value. This method will be applied to the columns '*totalChol*', '*BMI*', '*heartRate*', and '*glucose*'. The median value for '*cigsPerDay*', however, was zero. Therefore, it is inappropriate to populate the null values being set to zero. To address this issue, the null values of the '*cigsPerDay*' column will be replaced with the mean.

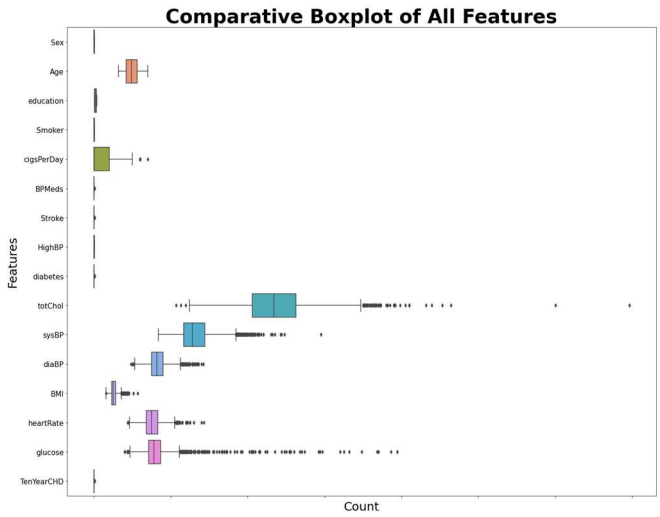

Fig. 4 Comparative Boxplot of All Features in Framingham Dataset

For both datasets, the Inter-Quartile approach has been used for outlier treatment. Tables 2 and 3 describe the impact of eliminating outliers. If the outliers are eliminated, many rows will be lost, which might result in predicting a healthy person as having a high chance of getting a stroke. A function is utilized that returns columns with the risk of outlier elimination and it will be classified as upper and lower limit, upper and lower removal, and risk percentage. Table 2 lists the percentage of the risk, if the outlier is removed from the feature in the data frame for Framingham Dataset.

TABLE II
PERCENTAGE OF RISKS FOR OUTLIERS REMOVAL IN FRAMINGHAM DATASET

Feature	Percentage of Risk (%)
cigsPerDay	100.0
totChol	100.0
sysBP	98.4
diaBP	100.0
BMI	97.9
heartRate	100.0
glucose	98.8

According to Table 2, it is known that people with high levels of cigarette use, cholesterol, diastolic blood pressure, and heart rate are susceptible to stroke. Therefore, deleting a high percentage of risks is not being considered, hence no feature in the dataset was to be removed.

On the other hand, the same procedure has been applied to the BRFSS dataset, and Table 3 displays the resulting percentage of risk for outlier elimination. Only three attributes in this dataset include outliers: '*BMI*', '*MentHlth*', and '*PhysHlth*'. Each of the three columns has a rather high-risk percentage, thus none of them are deemed to be eliminated.

TABLE III
PERCENTAGE OF RISKS FOR OUTLIERS REMOVAL IN BRFSS DATASET

Feature	Percentage of Risk (%)
BMI	94.9
MentHlth	92.5
PhysHlth	89.0

E. Feature Selection

This section elaborates on the process of selecting and transforming the most relevant characteristics while constructing a predictive machine learning model utilizing domain knowledge. The objective is to offer only relevant characteristics to models, enhancing the performance of machine learning algorithms, and preventing model fitting difficulties.

Supervised models feature selection consists of intrinsic, wrapper method, and filter methods. It refers to a process that selects features based on the output label class. Target variables will be utilized to determine which variables may boost the model's efficiency. The filter approach eliminates characteristics depending on their relationship to the output or how the features in the dataset correspond to the output. Correlation is used to determine if the characteristics are favorably or negatively associated with the output labels, and features are dropped based on the result. In addition, correlations between the characteristics were detected to confirm the lack of multicollinearity. In this paper, two filter methods have been applied as feature selection techniques.

Pearson's Correlation Coefficient has been applied to indicate the strength and direction of the association between two variables. It is a linear correlation coefficient with a range of -1 to 1. It is a descriptive statistic that describes a dataset's features. Specifically, it represents the magnitude and direction of the linear connection between two quantitative variables. Due to their linear dependency, two strongly correlated variables may have roughly the same predictive power for an observation's result value. Eliminating one of the linked variables before training the model is advantageous to the learning process and may result in performance comparable to that of the whole model. Fig. 5 and Fig. 6

illustrate the outcome of the modification described in the preceding calculation. On the heatmaps, the dendrogram illustrates the hierarchical connection between items.

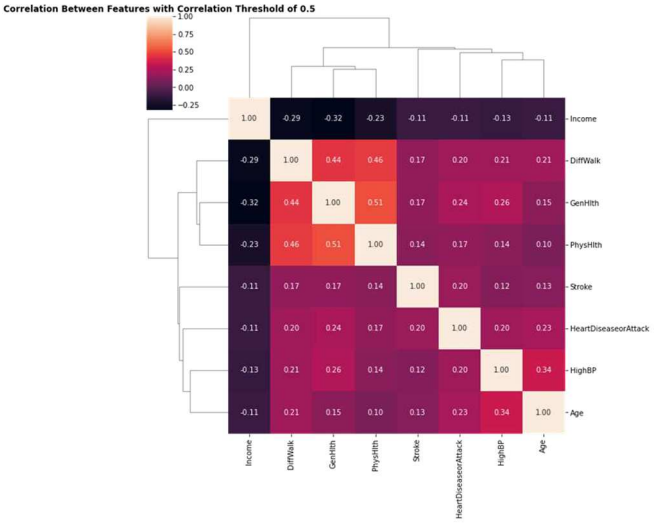

Fig. 5 Correlation Matrix between Variables in BRFSS Dataset

As the variable selection threshold, the absolute value has been set to 0.1 in the BRFSS dataset and 0 in the Framingham dataset. Fig. 6 demonstrates that the absolute correlation coefficient is less than 0.7 at a threshold of 0.51. It is possible to argue that multicollinearity does not exist for the selected variables. Hence, variables having a lower correlation coefficient with the target variable will not be eliminated.

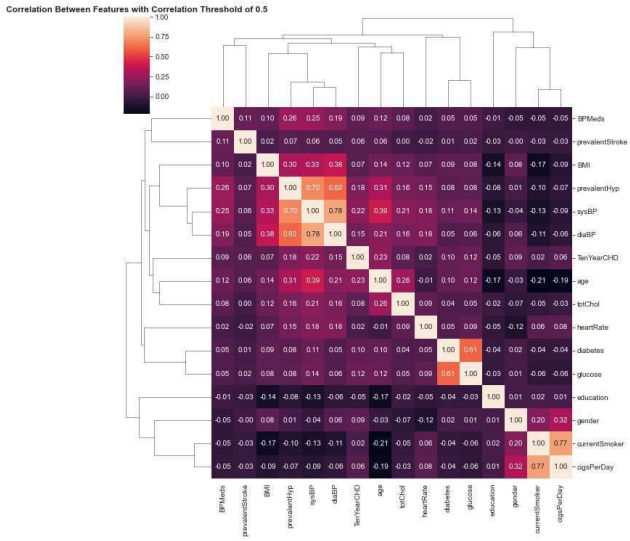

Fig. 6 Correlation Matrix between Variables in Framingham Dataset

In the Framingham dataset, based on the depiction in Fig. 6, it is possible to infer that multicollinearity exists in the chosen features at a threshold of 0.77, which is slightly more than the absolute correlation coefficient of 0.7. To solve the multicollinearity issue, more research has led to the conclusion that the column '*currentSmoker'* should be removed. Since the columns '*currentSmoker'* and '*cigsPerDay'* are identical, the data is attempting to convey that regardless of how many cigarettes a person smokes per day, '*currentSmoker'* will always be a 1 indicating '*yes'*.

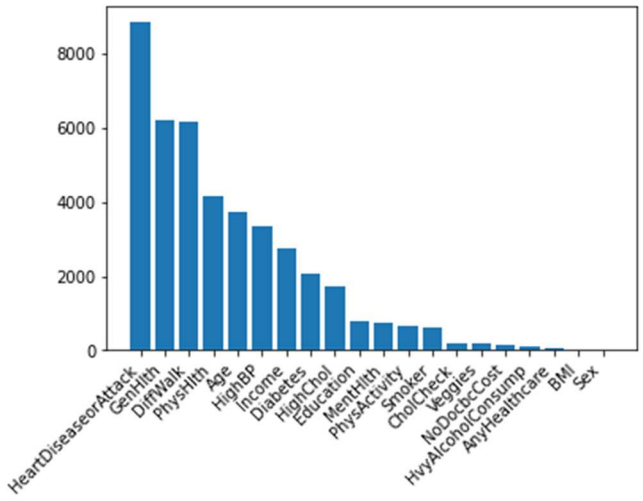

Fig. 7 SelectKBest of Best 20 Features

On the other hand, SelectKBest ranks the input variables according to their strength of link with the target variables using statistical measurements (Brownlee, J., 2020) [27]. By selecting only the twenty most important features, it is a helpful feature selection strategy for decreasing the variables in a dataset. The performance of models can be enhanced by removing less crucial portions of the data and shortening the training period. After selecting the best 20 features which can be observed from Fig. 7, '*fruits'* column has been eliminated as it offers the lowest value in the *SelectKBest()* function.

F. Data Scaling

Python was used to scale the data for the continuous variables using the *StandardScaler()* function from the *sklearn* library. The *MinMaxScaler* maintains the original distribution form. It does not significantly alter the information included in the original data. Consequently, *MinMaxScaler* does not diminish the significance of outliers. With this, the preparation of the datasets has been completed for model training.

G. K-Fold Cross-Validation

Cross-validation is a fundamental technique used in machine learning to evaluate the performance and generalization ability of predictive models. The *sklearn.model_selection* module in Python's scikit-learn library offers various classes and functions to implement k-fold cross-validation. This method is widely employed due to its robustness and effectiveness. The dataset is divided into K equally sized folds of 5, where each fold serves as a validation set once, while the remaining K-1 folds are used for training the model. This technique helps to mitigate the bias introduced by a single train-test split and provides a more reliable estimation of the model's performance.

As K = 5 or K = 10 have been demonstrated empirically to produce test error rate estimates that do not suffer from excessively high bias or from extremely high variation, these values are typically used in K-Fold Cross-Validation. In 2022, Vodencarevic et al. employed double-nested cross-validation loops with K = 5 in each loop in the instance of predicting stroke.

In summary, adopting 5-folds for K-Fold Cross-Validation for predicting stroke is a sensible option, since it has been demonstrated to produce test error rate estimates that suffer from neither excessively high bias nor from extremely high variation. It is also important to note that using a smaller K value will decrease the computational cost of the procedure.

H. Data Sampling Methods

The sensitivity of machine learning algorithms to the distribution of classes in the training set is extensively documented in (Witten et al., 2011) [28]. The dataset has a distinct distribution across classes and training and test sets. These distinctions may affect the classification's usefulness. In this regard, it is intriguing to determine how balancing the training set of a classifier affects its classification accuracy.

SMOTE is an over-sampling strategy described by Chawla et al. (2002) [29], in which the minority class is over-sampled by the creation of "synthetic" samples, as opposed to over-sampling using replacement. SMOTE algorithm is used to determine the k closest neighbors for each positive class, followed by constructing as many data duplications as required between each positive class and the randomly selected k nearest neighbors. The class distribution before and after implementing SMOTE strategy can be listed in Table 4. In the initial class distribution, the majority class contains 166,667 instances, which is a substantially greater number than the minority class's 8,208 occurrences. The class distribution significantly changes after using the SMOTE sampling approach. The number of instances for the majority class and the minority class are now equal, with each having around 166,667 occurrences. By creating synthetic samples for the minority class, SMOTE successfully increases its representation in the collection and achieves this balance.

TABLE IV
CLASS DISTRIBUTION BEFORE AND AFTER SMOTE IMPLEMENTATION

Sampling Technique	Class 0	Class 1
Original	166,667	8,208
SMOTE	166,667	166,667

Tomek linkages is a method for undersampling that eliminates edge instances from a dataset. As this is an under-sampling technique, only negative class data will be discarded. If two samples create Tomek linkages, one of them is considered to be noisy data, or both are considered to be borderline. Based on Table 5, the majority class in the original class distribution has more occurrences of 166,667 than the minority class, which has 8,208 instances. The class distribution changes after using the Tomek Links under-sampling approach. The minority class has 8,208 occurrences, whereas the dominant class has been reduced to 163,093 instances. By removing majority class samples that are close to minority class samples, Tomek Links tries to eliminate occurrences that are close to the decision boundary.

TABLE V
CLASS DISTRIBUTION BEFORE AND AFTER TOMEK LINKS IMPLEMENTATION

Sampling Technique	Class 0	Class 1
Original	166,667	8,208
Tomek Links	163,093	8,208

A hybrid sampling method called SMOTE + Tomek combines SMOTE with Tomek Links. For class equilibrium, SMOTE uses generated samples. SMOTE, however, does not assess cross-border cases. After SMOTE, Tomek Links emerged to deal with this issue. Table 6 shows that the majority class has 166,710 instances, substantially more than the minority class, which has just 8,165 instances. However, the class distribution significantly changes after using the SMOTE + Tomek sampling strategy. Both the majority class and the minority class now have an equal number of instances, with each class having around 166,032 instances. This shows that SMOTE + Tomek's strategy of undersampling the dominant class and oversampling the minority class to generate a more balanced distribution has effectively handled the issue of class inequality.

TABLE VI
CLASS DISTRIBUTION BEFORE AND AFTER SMOTETOMEK IMPLEMENTATION

Sampling Technique	Class 0	Class 1
Original	166,710	8,165
SMOTETomek	166,032	166,032

I. Dataset Partitioning

The datasets are divided into two subsets in an 80:20 ratio, with 80% used for training and 20% used for testing, to make it easier to build and assess machine learning models.

The model may gain knowledge from a sizable chunk of the dataset due to the larger size of the training subset. The input characteristics and target class are presented to the model during the training phase. Multiple algorithms can be utilized to analyze this training data and change its internal parameters to minimize the discrepancy between its predictions and the actual labels. The objective is to identify the underlying correlations and patterns in the data.

The remaining 20% of the testing subset is set aside for assessing how well the trained model performed. To enable a fair evaluation of the model's capacity to generalize to new data, this subset is kept separate and unaltered throughout the training process. On the testing subset, the model is used, and the predictions are compared with the real labels. To gauge how well the model works on unobserved data, evaluation measures like accuracy, precision, recall, or F1-score can be determined.

J. Evaluation Metrics

The Confusion Matrix, Accuracy, Precision, Recall, F1 Score, and Area Under the ROC Curve are the evaluation metrics used to assess performance. The four points for identifying a prediction using the method utilized are True Positive, True Negative, False Positive, and False Negative. Accuracy as a performance metric should no longer be used because of the gap across classes. As a result, it is not employed for genuine assessments but rather as an addition to the other metrics.

All performance measurements are built on the confusion matrix. A classification model's performance on a set of test data for which the true values are known is described using a table-based performance assessment. This evaluation metric presents the results of the predictions in terms of the matrix with True Positives for stroke patients that were precisely predicted, True Negatives for healthy individuals that were

successfully forecasted, False Positives for healthy cases that were misclassified as stroke patients, and False Negatives for stroke patients that were mistakenly classified to be healthy cases. The effectiveness of the classification findings is represented by accuracy. It is crucial for assessing the overall efficacy of the model for assessing stroke prediction models since accuracy indicates the percentage of properly categorized occurrences among all instances.

Precision describes the proportion of accurately predicted positive cases among the sum of predicted positive cases. It is effective in situations when False Positives are of more concern than False Negatives. The relevance of Precision in stroke prediction stems from the fact that inaccurate findings might result in misclassification of healthy cases, which can lead to further analysis.

Recall describes the amount of real positive instances that the model can accurately predict. In situations where False Negatives are of more importance than False Positives, this statistic is beneficial. It is crucial in stroke prediction when a false negative should not go unnoticed. Stroke patients should not be wrongly predicted as healthy cases as this affects early treatment.

The F1-score, which is a harmonic mean of recall and precision and takes into consideration of both false positives and false negatives, is a crucial metric for assessing stroke prediction models if the data set is unbalanced. Hence, when there are considerably more instances of one class over the other, the F1 score becomes particularly beneficial.

The AUC-ROC curve is a crucial metric for assessing stroke prediction algorithms because it represents the trade-off between true positive rate and false positive rate at various classification thresholds. It is one of the most crucial assessment criteria for assessing the effectiveness of any classification model. The AUC measures how well a model can discriminate between classes and indicates the degree of separability. The model performs better at differentiating between individuals with the disease and those who do not have it as the AUC value is higher.

III. Result and Discussion

In this research study, the BRFSS dataset is used for building the model, while the Framingham dataset is utilized for cross-dataset testing. There are several algorithms that were trained on the BRFSS dataset to develop a prediction model that can be tested on unseen data from the Framingham dataset. The subsequent section will elaborate on the benefits of using a particular model and the performance of each model.

A. Extreme Gradient Boosting Classifier

XGBoost is particularly helpful for predicting stroke because it is a scalable, distributed gradient-boosted decision tree technique that can handle both regression and classification for predictive modeling issues. Contrary to traditional gradient descent approaches that aim to minimize output error with each iteration, XGBoost aids in the prediction of the additive model's ideal gradient. It is well-liked for use in machine learning because of its speed and effectiveness. Overall, XGBoost is a strong and adaptable algorithm that is excellent for handling vast and complicated datasets and can predict stroke outcomes efficiently.

Table 7 shows the results that the XGBoost achieved using three different sampling techniques. The accuracy comparisons show the model's capability to distinguish between healthy controls and stroke victims. The findings provided here demonstrate that good accuracy, precision, recall, and f1-score may be attained through the usage of the three sampling techniques. As a result, it may be inferred that applying different sampling techniques has no influence on the performance of the model.

TABLE VII
Extreme Gradient Boosting Classifier Performance with Different Sampling Techniques

Evaluation Metrics	SMOTE	Tomek Links	SMOTE Tomek
Accuracy	0.95	0.95	0.95
Precision	0.95	0.95	0.95
Recall	1.00	1.00	1.00
F1-Score	0.98	0.98	0.98
AUC Score	0.809	0.803	0.802

B. Adaptive Boosting Classifier

Boosting is a broad technique that seeks to "boost" the accuracy of any learning algorithm by merging all the weak classifiers into a single strong classifier, which can also reduce overfitting. In past research in predicting strokes, Adaboost has demonstrated great accuracy, beating other machine learning methods. It is an efficient scalable algorithm that can operate on huge datasets, making it appropriate for studying sizable datasets of stroke patients.

In Table 8, it can be observed that AdaBoost with Tomek links achieved the highest AUC score and relatively high accuracy, precision, recall, and f1-score compared to the other two sampling techniques and other algorithms. Hence, it can be deemed that this model is the best model to predict stroke.

TABLE VIII
Adaptive Boosting Classifier Performance with Different Sampling Techniques

Evaluation Metrics	SMOTE	Tomek Links	SMOTE Tomek
Accuracy	0.94	0.95	0.94
Precision	0.96	0.95	0.96
Recall	0.98	1.00	0.98
F1-Score	0.97	0.98	0.97
AUC Score	0.786	0.812	0.786

C. Bootstrap Aggregating Classifier

Bagging, a short-form for bootstrap aggregating is one of the oldest, most straightforward, and maybe simplest ensemble-based algorithms, with a performance that was initially shown by Breiman [30] to be unexpectedly excellent in 1996. Bagging lowers the variance of a prediction model and is used to handle bias-variance trade-offs. It prevents data overfitting and is used to raise the precision of regression and classification models. Stroke datasets can be studied using a scalable method such as bagging since it operates effectively on huge datasets.

Table 9 demonstrates the performance results by using the Bagging classifier together with different techniques to handle the skewed dataset. All sampling techniques achieve similar performance. The usage of the Bagging model alongside with Tomek Links sampling method attained the highest AUC score in contrast to the other two sampling

techniques, however, it is still not the best model compared to the rest of the algorithms.

TABLE IX
BOOTSTRAP AGGREGATING CLASSIFIER PERFORMANCE WITH DIFFERENT SAMPLING TECHNIQUES

Evaluation Metrics	SMOTE	Tomek Links	SMOTE Tomek
Accuracy	0.95	0.95	0.95
Precision	0.95	0.95	0.96
Recall	0.99	1.00	0.99
F1-Score	0.97	0.98	0.97
AUC Score	0.786	0.803	0.788

D. Linear Support Vector Classifier

Medical datasets frequently contains high-dimensional data, which makes linear SVC particularly advantageous. Clinicians may better understand the significance of many characteristics throughout the prediction process. Additionally, recognizing the risk factors for stroke and creating preventive measures that are more potent is also essential. Linear SVC may be utilized for a variety of tasks such as predicting stroke outcomes, determining important variables, and categorizing patients into distinct risk groups.

The performance analysis from employing Linear SVC with three sampling methods can be obtained in Table 10. Due to the fact that utilizing SMOTE method to handle the imbalanced dataset with Linear SVC attained a high AUC score of 0.811, however, the value of accuracy is relatively low at 0.73. Thus, this model with SMOTE was not used to measure its performance on an independent dataset.

TABLE X
LINEAR SUPPORT VECTOR CLASSIFIER PERFORMANCE WITH DIFFERENT SAMPLING TECHNIQUES

Evaluation Metrics	SMOTE	Tomek Links	SMOTE Tomek
Accuracy	0.73	0.95	0.74
Precision	0.98	0.95	0.98
Recall	0.73	1.00	0.74
F1-Score	0.84	0.98	0.84
AUC Score	0.811	0.804	0.803

E. Multilayer Perceptron

MLP is an artificial neural network model can be used to predict strokes. An input layer, one hidden layer, and an output layer make up the model, which is intended to resemble the structure and operation of biological neurons. ANN can discover intricate patterns in a huge amount of information and use those patterns to predict the future. In general, MLP is an effective way of foretelling the results of strokes and may spot intricate patterns from enormous datasets that may be missed by conventional statistical techniques.

The performance of MLP is shown in Table 11 with three sampling techniques. The results of Tomek links with MLP showed that the model outperformed other sampling methods based on accuracy, recall, f1-score, and AUC score. Nevertheless, it is still not reliable enough to be cross-validated on another dataset.

TABLE XI
MULTILAYER PERCEPTRON PERFORMANCE WITH DIFFERENT SAMPLING TECHNIQUES

Evaluation Metrics	SMOTE	Tomek Links	SMOTE Tomek
Accuracy	0.90	0.95	0.88
Precision	0.96	0.95	0.96
Recall	0.94	0.99	0.91
F1-Score	0.95	0.97	0.94
AUC Score	0.664	0.745	0.671

F. Light Gradient Boosting Machine Classifier

Light GBM, an implementation of gradient boosting machines, offers several advantages over traditional gradient boosting methods, making it a promising choice for various applications. It offers a quick and precise approach for supervised learning tasks. It is a fantastic option for large-scale research where precision is crucial because of its high speed and scalability. Because it can handle skewed datasets, which are typical in medical datasets, Light GBM is a well-liked method for predicting strokes and an interpretable model since it sheds light on the significance of many characteristics in the prediction process. This can aid medical professionals in better comprehending the risk variables for stroke and creating more potent preventative measures.

Table 12 shows the Light GBM classifier's result performance by implementing three sampling techniques. All the models attained similar accuracy, precision, and recall metrics. As a consequence, adopting alternative sampling approaches has no effect on the model's performance.

TABLE XII
LIGHT GRADIENT BOOSTING MACHINE CLASSIFIER PERFORMANCE WITH DIFFERENT SAMPLING TECHNIQUES

Evaluation Metrics	SMOTE	Tomek Links	SMOTE Tomek
Accuracy	0.95	0.95	0.95
Precision	0.95	0.95	0.95
Recall	1.00	1.00	0.99
F1-Score	0.97	0.98	0.97
AUC Score	0.806	0.808	0.798

G. Cross-Validation with an Independent Data

To validate the model robustness, Framingham Heart Disease Prediction Dataset was utilized as a testing set to employ cross-validation. It helps to assess how well the model generalizes to unseen data and detect overfitting. Based on Table 13, the cross-validation scores show how well the model performed when tested on various folds or subset of the data. Each score shows the level of accuracy the model attained on a certain fold.

It is worth noting that only five attributes match between these two datasets, hence the five variables are tested to measure their cross-validation scores, in which the attributes are 'Sex', 'Smoker', 'HighBP', 'BMI' and 'Age'. An overall assessment of the model's performance over all folds is provided by the average cross-validation score, which is determined as 0.957, which shows the model's typical accuracy level during cross-validation. Additionally, the model's performance on a different dataset from the one used for cross-validation is represented by this accuracy number. With an accuracy of 0.994, it appears that the model works well on the Framingham dataset.

Overall, the evidence indicates that the model works well on the other dataset, achieves high accuracy during cross-validation, and reduced level of overfitting, as the model can achieve a good performance on an independent dataset.

TABLE XIII
PERFORMANCE EVALUATION USING ADABOOST CLASSIFIER ON FRAMINGHAM DATASET

K-Fold Cross-Validation (CV)	1st Fold	2nd Fold	3rd Fold	4th Fold	5th Fold
CV Score	0.95	0.95	0.95	0.95	0.95

H. Discussion

Based on the findings, three separate sampling methods which are SMOTE, Tomek links, and SMOTE + Tomek were used to provide an adequate representation of the actual data. Six machine learning models were built on the BRFSS dataset by tuning various parameters. The model built on the BRFSS dataset was used to test on the Framingham dataset and it is found that higher cross-dataset accuracy can be achieved.

By detecting those who are at risk of having a stroke early on using low-cost features, this model has the potential to save lives and reduce the likelihood of permanent disabilities brought on by strokes. A stroke case may be misdiagnosed if it is erroneously detected or categorized as a healthy case. The patient's health and well-being may be seriously impacted by this misclassification.

The imbalanced dataset was addressed using a variety of strategies. The most popular sampling techniques were found after a detailed examination of the literature. Following a review of various sampling techniques, three popular sampling approaches, each of a different type were chosen. Following the analysis, it can be concluded that Tomek Links was the best sampling strategy for improving the performance of the model for stroke prediction.

IV. CONCLUSION

The objective of this paper was to provide an optimal solution that would aid in the early identification of stroke using low-cost features. The goal was achieved by implementing different machine learning models, each of which made use of three different sampling techniques for handling imbalanced datasets and evaluating their performance by using various metrics. The model built using one dataset was tested on another dataset. Higher cross-dataset accuracy obtained demonstrated the generalizability of the identified optimal model on an unseen dataset collected in a different environment. Future studies might include building machine learning models that allow for real-time monitoring of stroke risk and early intervention options. This might entail creating online applications or mobile health technologies that continually analyze and update stroke risk based on dynamic patient data. It is also important to validate the efficacy of models to classify stroke cases using varied and independent datasets as well as assess the models' generalizability across diverse demographics, healthcare settings, and geographical locations.

ACKNOWLEDGMENT

We would like to thank Multimedia University for providing the funding for publication.

REFERENCES

[1] S. Wang et al., "A randomized controlled trial of brain and heart health manager-led mHealth secondary stroke prevention," Cardiovascular Diagnosis and Therapy, vol. 10, no. 5, pp. 1192–1199, Oct. 2020, doi: 10.21037/cdt-20-423.

[2] D. B. Harrar et al., "A Stroke Alert Protocol Decreases the Time to Diagnosis of Brain Attack Symptoms in a Pediatric Emergency Department," The Journal of Pediatrics, vol. 216, pp. 136-141.e6, Jan. 2020, doi: 10.1016/j.jpeds.2019.09.027.

[3] M. Rasmussen, J. B. Valentin, and C. Z. Simonsen, "Blood Pressure Thresholds During Endovascular Therapy in Ischemic Stroke—Reply," JAMA Neurology, vol. 77, no. 12, p. 1579, Dec. 2020, doi:10.1001/jamaneurol.2020.3819.

[4] S. Lattanzi and M. Silvestrini, "Blood pressure in acute intra-cerebral hemorrhage," Annals of Translational Medicine, vol. 4, no. 16, pp. 320–320, Aug. 2016, doi: 10.21037/atm.2016.08.04

[5] Verma, S. Jaiswal, and W. R. Sheikh, "Acute thrombotic occlusion of subclavian artery presenting as a stroke mimic," Journal of the American College of Emergency Physicians Open, vol. 1, no. 5, pp. 932–934, May 2020, doi: 10.1002/emp2.12085.

[6] M. Boukobza, S. Nahmani, L. Deschamps, and J.-P. Laissy, "Brain abscess complicating ischemic embolic stroke in a patient with cardiac papillary fibroelastoma – Case report and literature review," Journal of Clinical Neuroscience, vol. 66, pp. 277–279, Aug. 2019, doi: 10.1016/j.jocn.2019.03.041.

[7] S. Uppal, S. Goel, B. Randhawa, and A. Maheshwary, "Autoimmune-Associated Vasculitis Presenting as Ischemic Stroke With Hemorrhagic Transformation: A Case Report and Literature Review," Cureus, Sep. 2020, doi: 10.7759/cureus.10403.

[8] M. Lee, J. Ryu, and D. Kim, "Automated epileptic seizure waveform detection method based on the feature of the mean slope of wavelet coefficient counts using a hidden Markov model and EEG signals," ETRI Journal, vol. 42, no. 2, pp. 217–229, Apr. 2020, doi:10.4218/etrij.2018-0118.

[9] CDC, (2020), National Center for Chronic Disease Prevention and Health Promotion, Division for Heart Disease and Stroke Prevention. [Online]. Available: https://www.cdc.gov/stroke/about.htm

[10] E. Dritsas and M. Trigka, "Stroke Risk Prediction with Machine Learning Techniques," Sensors, vol. 22, no. 13, p. 4670, Jun. 2022, doi: 10.3390/s22134670.

[11] V. Abedi et al., "Prediction of Long-Term Stroke Recurrence Using Machine Learning Models," Journal of Clinical Medicine, vol. 10, no. 6, p. 1286, Mar. 2021, doi: 10.3390/jcm10061286.

[12] J. O. Victor, X. Chew, K. W. Khaw, and M. H. Lee, "A Cost-Based Dual ConvNet-Attention Transfer Learning Model for ECG Heartbeat Classification," Journal of Informatics and Web Engineering, vol. 2, no. 2, pp. 90–110, Sep. 2023, doi:10.33093/jiwe.2023.2.2.7.

[13] S. Dev, H. Wang, C. S. Nwosu, N. Jain, B. Veeravalli, and D. John, "A predictive analytics approach for stroke prediction using machine learning and neural networks," Healthcare Analytics, vol. 2, p. 100032, Nov. 2022, doi: 10.1016/j.health.2022.100032.

[14] M. U. Emon, M. S. Keya, T. I. Meghla, Md. M. Rahman, M. S. A. Mamun, and M. S. Kaiser, "Performance Analysis of Machine Learning Approaches in Stroke Prediction," 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), Nov. 2020, doi:10.1109/iceca49313.2020.9297525.

[15] V. JalajaJayalakshmi, V. Geetha, and M. M. Ijaz, "Analysis and Prediction of Stroke using Machine Learning Algorithms," 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), Oct. 2021, doi: 10.1109/icaeca52838.2021.9675545.

[16] R. K. Kavitha, W. Jaisingh, and S. R. Sujithra, "Applying Machine Learning Techniques for Stroke Prediction in Patients," 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), Oct. 2021, doi: 10.1109/icaeca52838.2021.9675652.

[17] C. Rana, N. Chitre, B. Poyekar, and P. Bide, "Stroke Prediction Using Smote-Tomek and Neural Network," 2021 12th International Conference on Computing Communication and Networking

Technologies (ICCCNT), Jul. 2021, doi:10.1109/icccnt51525.2021.9579763.
[18] N. Biswas, K. M. M. Uddin, S. T. Rikta, and S. K. Dey, "A comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach," Healthcare Analytics, vol. 2, p. 100116, Nov. 2022, doi: 10.1016/j.health.2022.100116
[19] Md. Shafiul Azam, Md. Habibullah, and H. Kabir Rana, "Performance Analysis of Various Machine Learning Approaches in Stroke Prediction," International Journal of Computer Applications, vol. 175, no. 21, pp. 11–15, Sep. 2020, doi: 10.5120/ijca2020920740.
[20] Y. Wu and Y. Fang, "Stroke Prediction with Machine Learning Methods among Older Chinese," International Journal of Environmental Research and Public Health, vol. 17, no. 6, p. 1828, Mar. 2020, doi: 10.3390/ijerph17061828.
[21] Ferdib-Al-Islam and M. Ghosh, "An Enhanced Stroke Prediction Scheme Using SMOTE and Machine Learning Techniques," 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), Jul. 2021, doi:10.1109/icccnt51525.2021.9579648.
[22] M. Phankokkruad and S. Wacharawichanant, "Performance Analysis and Comparison of Cerebral Stroke Prediction Models on Imbalanced Datasets," 2022 IEEE/ACIS 7th International Conference on Big Data, Cloud Computing, and Data Science (BCD), Aug. 2022, doi: 10.1109/bcd54882.2022.9900833.
[23] G. Fang, Z. Huang, and Z. Wang, "Predicting Ischemic Stroke Outcome Using Deep Learning Approaches," Frontiers in Genetics, vol. 12, Jan. 2022, doi: 10.3389/fgene.2021.827522.
[24] U. Fayyad, "Knowledge Discovery in Databases: An Overview," Relational Data Mining, pp. 28–47, 2001, doi: 10.1007/978-3-662-04599-2_2.
[25] Ashish Bhardwaj. (2022). Framingham heart study dataset. Kaggle. [Online]. Available: https://www.kaggle.com/datasets/aasheesh200/framingha m-heart-study-dataset
[26] Alex Teoul. (2022). Heart Disease Health Indicators Dataset. Kaggle. [Online]. Available: https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset
[27] Brownlee, J. (2020). How to choose a feature selection method for machine learning. [Online]. Available: https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/
[28] I. H. Witten, E. Frank, and M. A. Hall, "Writing New Learning Schemes," Data Mining: Practical Machine Learning Tools and Techniques, pp. 539–557, 2011, doi: 10.1016/b978-0-12-374856-0.00016-x.
[29] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, Jun. 2002, doi:10.1613/jair.953.
[30] L. Breiman, "Bagging Predictors," Machine Learning, vol. 24, no. 2, pp. 123-140, Aug. 1996, doi: 10.1007/bf00058655.