Design of a Novel Enhanced Machine Learning Model for Early Prediction of Cerebral Stroke

Snehal Shinde[1] · Manish P Kurhekar [2] · Monali Gulhane* [3] · Nileshchandra K Pikle*[4]
·

Received: date / Accepted: date

Abstract A stroke can happen when the blood supply to a certain region in the brain's cortex is abruptly severed. Without adequate blood flow, brain cells will eventually die, and the extent of the damage will be inversely correlated with the area of the injured brain. Early symptom identification is essential for stroke prediction and encouraging healthy habits by giving helpful information. The solution to these issues is the development of a precise and effective early-stage prediction model employing analytical support in clinical decision-making with digitized patient information. Most research focuses on forecasting cardiovascular stroke, but the cerebral risk for stroke has gotten far less attention. The current study aims to progress and assess several machine learning models to create a framework for predicting the long-term potential of cerebral stroke. By conducting a thorough experimental assessment using two different methodologies, namely SMOTE and SMOTE ENN, the suggested work tried to address the issue of imbalanced data from the Kaggle dataset. On SMOTE-Balanced as well as SMOTE ENN-balanced datasets, several models such as K-nearest neighbour (KNN), logistic regression(LR), support vector machine(SVM), decision tree(DT), random forests, XG boosting, stacking, and ANN are trained. As demonstrated by the results, SMOTE ENN and the stacking classification approach achieved a remarkable 99.52% accuracy, with a recall of 99.30%, a precision of 99.4%, and an F1 measure of 99%. We found that a relatively simple data balancing technique combined with a supervised machine learning algorithm can be used to predict strokes with high accuracy and great practical potential.

Keywords Stroke · Data Balancing · Machine Learning · Data Analysis · Performance evaluation

1.Computer Science and Engineering, Indian Institute of Information Technology, Nagpur, India. ***sshinde@iiitn.ac.in***

2.Department of Computer Science and Engineering, Visvesvaraya National Institute of Technology, Nagpur, Maharashtra, India ***manishkurhekar@cse.vnit.ac.in***

3.Symbiosis Institute of Technology, Nagpur Campus, Symbiosis International (Deemed University), Pune, India. ***monali.gulhane4@gmail.com***

4.Computer Science and Engineering, Indian Institute of Information Technology, Nagpur, ***npikle@iiitn.ac.in***
*corresponding author

Address(es) of author(s) should be given

1 Introduction

Our ideas, feelings, and language are all products of the brain, which also directs our movements and serves as a memory bank. The brain controls many additional biological functions, including breathing and digestion. If your brain doesn't have enough oxygen, it can't work correctly. The arteries in your body deliver oxygen-rich blood to all of the many parts of your brain. Brain cells start dying minutes after losing the supply of oxygen when any form of restriction reduces the blood supply to the area of the body that receives it. This results in a stroke. Risk elements include diabetes, atherosclerosis, high blood pressure, homocysteine levels, and your genes.[1] Stroke is the main reason for disability area of the United States.[2] It ranks as the fifth largest reason for death in the United States.[3] It is anticipated that the number of strokes will more than double between 2010 and 2050, with the majority of strokes occurring in individuals over the age of 75.[4] World Stroke Organization[5] says that approx 13 million people suffer from a stroke every year, and about 5.5 million of them die from it. Because of this, it has a big effect on everything. Stroke affects not only the person who has it but also their friends, family, and workplace. Also, contrary to what most people think, it may occur to anyone of any age, no matter their gender or whatever is physical health. Efforts that are complementary in the fields of it are becoming clearer thanks to advances in brain engineering along with restorative medicine, how neuroprosthetic procedures can control devices, along with ultimately restore body function.[6] Despite the significant strides made over the years, finding novel therapies for arm recovery from function following stroke remains a top goal. Additionally, a stroke's symptoms might change, and the condition progresses swiftly. Sometimes, symptoms may appear very gradually, while other times, they may appear very suddenly. Typically, the most prevalent symptoms include paralysis or weakness in the legs or arms. A person's risk of stroke can be decreased by regularly monitoring their blood pressure, participating in regular exercise, maintaining a healthy weight, quitting smoking and consuming alcohol, and eating a diet low in sodium and saturated fat.[7] As observed, many countries have a shortage of stroke experts, and there is a substantial rate of increase in cases that are incorrectly diagnosed.

Predicting a stroke is essential, and treatment needs to start as quickly as possible to avoid irreversible harm or death. Machine learning (ML) approaches can now be used to anticipate the development of a stroke thanks to advances in technology in the medical profession. Since algorithms offer precise prediction and proper analysis, including them in ML is helpful.[8–10] The majority of earlier studies on stroke focused on several issues, including the prognosis of heart attacks.[11,12] Brain strokes have not been the subject of numerous scientific projects. In the past, many scholars utilized machine learning to forecast strokes. We will be particularly interested in the stroke within the context of this inquiry. Machine learning models have been employed in a sizable number of study inquiries on this condition.[13–17] To predict the stroke, a Kaggle dataset[18] that consists of 5110 patients with 11 features (10 independent and 1 dependent) is used. The synthetic minority over-sampling technique (SMOTE) and SMOTE ENN were used to balance the data, as balancing the class is essential to constructing effective algorithms in stroke prediction. After that, with the help of the SMOTE-balanced and SMOTE ENN-balanced datasets, several different models are constructed and configured.

In this paper, the following are our contributions:

1. To explore the data for analysis and perform data balancing with the help of SMOTE and SMOTE ENN.
2. In order to research and develop machine learning models, support vector machine (SVM), K-nearest neighbour (KNN), decision tree (DT), and logistic regression (LR) on both SMOTE[19] -Balanced and SMOTE ENN[20]-Balanced dataset.

3. To enhance the performance of the proposed system model using ensemble techniques such as random forest, XG boosting, stacking, and ANN.
4. To examine the predictive accuracy of these proposed techniques for estimating stroke risk survival.
5. To analyze these techniques with the state of art literature.
6. To provide a system for predicting stroke survival prediction that is precise and efficient.

The overview of the proposed paper is structured with the given flow as, Section 2 describes the clinical data involved in this study and state art. Section 3 provides details about the method and materials that are used in the proposed work. The detailed workflow of the system model is presented in section 4. Section 5 provides the detailed experimental settings, results, and comparative study. Section 6 gives the conclusions drawn from the proposed work

2 Literature Review

The literature review represents the efficient literature study on the survival prediction of a person suffering from a stroke patients.
Shoily, Tasfia Ismail, et al. in [2019],[13] constructed a dataset by compiling information about strokes from a number of different sources. There were a total of 1058 patients represented in our dataset; 412 are men, and 646 are women. In this work, the Naive Bayes classifier has achieved an accuracy of 85.6%. while DT, KNN, and RF all have achieved an accuracy of 99.8%. The recall, precision, and f-measure according to (NB)Naive Bayes are respectively 88.1%, 85.6%, and 86.1%. All of the DT, k-NN, and Random Forest models have exact precision, recall, and f-measure values, which are respectively 99.8%, 99.8%, and 99.8%.

The authors constructed nine models in [2019][14] on the real-time data that was collected during China's nationwide program for early detection and treatment of stroke in 2017. The suggested models assisted in determining each participant's individual degrees of stroke risk. The model with the random forest got the maximum level of precision i.e. 97.33%), in this observation boosting model that used decision trees obtained the maximum level of recall i.e. 99.94%. If the random forest model, which has a recall of 98.44%, were employed instead of the method that is now being utilized, the recall would be enhanced by around 2.8%, and each year, many thousands more people who have had a high risk of stroke can be recognized.

Sheng-Feng Sung et al. in [2021][21] analyzed 1361 patients stroke data. The FAST model has a value of 0.737 for its area under a highly accurate curve (AUPRC), whereas the BE-FAST-1 has a value of 0.710, as well as BE-FAST-2 models, had values of 0.562, respectively. The top three ML models' areas under receiver operating characteristic curves, or AUPRC, were as follows: 0.782 for LR using class weighting, 0.783 for LR to synthesized minority oversampling method (SMOTE), and 0.787 for the regression and classification tree to under-sampling. Logistic regression and random forest methods obtained greater results of AUPRC than other ML models, particularly when class weighting with SMOTE was used to address the issue of class imbalance. Age, body temperature, diastolic blood pressure well as pulse rate were additional crucial factors that were taken into consideration while establishing a stroke-alert trigger. These features were taken into consideration in addition to the presenting symptoms with the triage level. Performance utilizing forecasting algorithms for identifying patients who may have had a stroke was greatly enhanced by the use of ML approaches. These machine learning models may be included in an electronic triage system to provide clinical decision assistance in emergency department triage.

Data preprocessing was carried out by Sailasya et al. in [2021],[15] who took into account many Kaggle datasets[18] and dealt with issues such as handling imbalanced data, label encoding, and missing value management. Six different machine-learning techniques were used to analyze the dataset once it has been cleaned and prepared. When compared to the other methods, Naive Bayes classification achieved the highest accuracy, at 82%. They have created an HTML portal where users may input information to determine for themselves if they have suffered a stroke.

Tazin, Tahia and Alam et.al. in [2021][16] investigated the efficacy of many ML algorithms in accurately predicting stroke based on a number of physiological factors on the Kaggle Dataset.[18] It has been observed that in this work the maximum classification accuracy of 96 %, random forest classification surpasses the other methods evaluated. The research indicated that when cross-validation measures are applied to brain stroke predictions, the random forest method beats alternative approaches.

With the help of machine learning (ML), in [2022], Dritsas, Elias, and Trigka et.al.[17] built and tested a number of ML models on Kaggle dataset[18] to make a strong framework for predicting the long-term dangerous conditions for the occurrence of stroke. The main benefit of this research was the SMOTE data balancing methodology and a stacking strategy that provided good performance. Different metrics, including accuracy, AUC, recall, precision, and F-measure showed this. The results of the experiment showed that the stacking classification approach is better than the others, with an AUC of 98.9%, an F-measure, precision, recall, and accuracy of 97.4%, and an accuracy of 98%.

There are various ML-based sepsis survival prediction techniques. However, these techniques were shown to have quite restricted performances, with only SMOTE balancing technique for a dataset of research obtaining an acceptable degree of efficiency.

3 Materials and Methods

In this section, we provide detailed information about the dataset, conduct exploratory data analysis, and describe the data preprocessing steps undertaken in the proposed work.

3.1 Materials

This subsection presents essential details about the dataset used in the study, including its origin, size, and relevant features. It also provides the data preprocessing steps to prepare the data for further analysis. The dataset undergoes an exploratory data analysis to gain insights and understand its characteristics better. Summary statistics, data visualizations, and other relevant techniques are employed to explore the dataset thoroughly.

3.1.1 Dataset

Our study utilized a dataset obtained from Kaggle.[18] The adults in this data set were the primary focus of our analysis. There were 5,110 people involved, and their attributes are as follows: (1 as the target class and 10 as inputs to ML models).

1. **id:** unique identifier
2. **gender**[22]: “Male”, “Female” or “Other”. In this data set Female, 2907 Male 2074 patients. A more in-depth understanding of the ways in which Age influences stroke outcomes will benefit people of all ages and backgrounds.

3. **age:**[22] age of the patient. By the age of 25, women have a one-in-four chance of having a stroke in their lifetime, which is significantly higher than the risk for males.
4. **hypertension:**[23] 0 if a person has normal blood pressure and 1 if a person has high blood pressure. 9.7% of patients are hypertensive
5. **heart disease:**[24] 0 if a person has no heart problems, if a person had cardiac disease, for example. Heart disease damages the brain. The 5.40% patients have heart disease.
6. **ever married:**[25] This indicates if patient is married or not. 3353 patients are married and 1757 are unmarried.
7. **work type:**[26] Work status is represented by five categories: private 57.24%, self-employed 16.02%, children 13.44 %, govt job 12.85%, and never worked 0.43%.
8. **Residence type:**[27] 2596 patients are from urban and 2514 are from Rural.
9. **average glucose level:**[28] The maximum average glucose level is 271.74, and the minimum is 55.12.
10. **BMI:**[29] This indicates body mass index.
11. **smoking status:**[30] This has four categories formerly smoked 17.31 %, "never smoked 37%", smokes 15.44 % or Unknown 30.21 %
12. **stroke:** This indicates whether the subject suffered a stroke.4.87 % have had a stroke.

3.1.2 Data Pre-processing

Prior to applying machine learning algorithms, the dataset undergoes preprocessing steps. These steps involve handling missing values, feature scaling, categorical variable encoding, and any other necessary transformations required to prepare the data for further analysis.

1. **Removal of missing values :** Quality of predictions may suffer if raw data is of poor quality, for example, if there are missing values or if the data is particularly noisy. To make data more suitable for analysis, pre-treatment steps, including removing duplicate values, choosing appropriate features, and discretizing the data, are required. As a result of completing pre-processing on the data, it has been discovered that the dataset with "BMI" contains null values; hence, it is essential to fill in missing or null values with a suitable value. In this, the BMI NULL values are filled with the mean of the BMI values
2. **Outlier detection and elimination:** Finding dataset observations that deviate significantly from the norm along with additional values is known as outlier identification. The accuracy of the model is improved by removing outliers. This study used box plots to identify outliers. There are outliers in the age, average blood glucose, and BMI characteristics. Outliers in these columns were eliminated.
3. **One hot encoding** is implemented to convert feature categories such as married, residence type, gender, work type, and smoking status to numerical data. The numerical features age, avg glucose level, and BMI are standardized using a standard scalar.
4. **Data Balancing:** The brain stroke dataset demonstrates the peculiarity of having a number of samples of a specific class that are under-represented in comparison to samples from other classes. This "class imbalance problem" is the challenge of gaining an understanding of a concept from a group that contains a limited number of examples.[31,32] In this case, the SMOTE method was used to increase equity in the dataset[19] and SMOTE ENN[20] method. To equalize their numbers with those of the dominant class, additional individuals of the minority category were generated.
   - **SMOTE:** It is a machine-learning strategy for coping with the difficulties of processing sparse data and enhancing the quality of the data by generating simulated information from the actual data. The benefit of SMOTE is that it provides synthetic data points which are slightly off compared to the originals rather than du-

plicate data points. It interpolates between minority class samples to create new ones. Consequently, over-fitting is prevented, and minority-class decision boundaries expand into majority-class space. Minority class means tend to approach majority class means. The amount and distribution of synthetic samples may change the balanced data set's mean. SMOTE effectively increases the variability within the minority class. Increased variability can raise the minority class's standard deviation in the balanced dataset.

– **SMOTE ENN:** In order to apply the Edited Nearest Neighbour approach, we must first determine the K-nearest neighboring observation, after which we must determine whether the neighbor's majority class corresponds to the class of the current observation. For a problem with two classes, the approach can be defined as follows: for each sample E_i, the set's three closest neighbors are found in the training set. Up till the issue is resolved, this process is repeated. If E_i is a member of the majority category and the designation that its three closest neighbors have given it is in conflict with the class that Ei was initially assigned to, then E_i will be eliminated. In the event that E_i is a member of a minority group and its three immediate neighbors incorrectly categorize E_i, then the immediate neighbors who are members of the majority group will be eliminated. It is clear that the age distribution of the stroke patient population is tilted to the left. The majority of the patients have ages ranging from sixty to eighty-two. There are also some young patients, including children, who are suffering from strokes. The age range of most patients at the hospital are male category is an age in span of 55 to 82 years old. The majority of female patients range in age from 48 to 82 years old.

3.1.3 Exploratory Data Analysis

Exploratory data analysis comes before creating an algorithm for machine learning since it helps to spot data issues and gives a broad overview of the entire data collection. As a result of our analysis, we can choose one feature from the categorical feature classification along with a feature from the numerical feature classification.

1. **Category I (Numerical Data Analysis: Average glucose level)**
   In the analysis of exploratory data analysis, we observed patients having a stroke with high blood glucose levels are more prone to get a stroke than those whose glucose levels are within the normal range. Data show that 11.3% of people who have elevated increased glucose levels are at an increased risk for stroke. For those with healthy blood sugar levels, just 3.6% suffered a stroke. High blood sugar levels triple a person's risk of having a stroke. In conclusion, those of us with high glucose levels are more likely to have a stroke. The possibility of a stroke be reduced in those with elevated glucose levels by increasing their amount of physical activity and improving their diet. But the data is imbalanced and hence the observation states that stroke patients with high glucose have fewer chances of stroke i.e. 11.3% as shown in Figure 1.[?]
   The need to balance the data for accurate prediction and analysis to reduce the false prediction, SMOTE was imposed on the imbalanced data of the average glucose level and we observed that accuracy has been increased to 23% as shown in the Figure 2.
2. **Category II (Categorical Data Analysis: Hyper-Tension and Heart Diseases)**
   Consequently, we can go on to the next finding in the data set, it is estimated that 20.3% of people with high blood pressure(hypertension)and cardiovascular disease have had a stroke, and 70.7% are stable also merely 3.4% of people who do not have hypertension as well as heart disease have never suffered a stroke. This means that the data set is unbalanced, as shown in fig3, which makes accurate classification

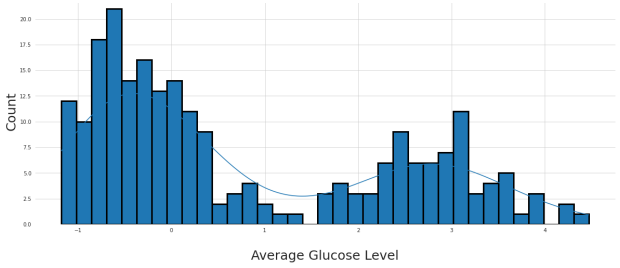

Fig. 1 The distribution of stroke patients' average glucose level in the imbalanced dataset.

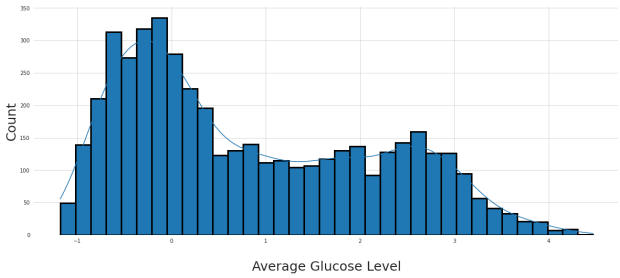

Fig. 2 Stroke patient's average glucose level after balancing dataset using SMOTE

by predictive modelling challenging. This inequity needed to be fixed before any modelling could be done. To improve the accuracy of predicting stroke for a person

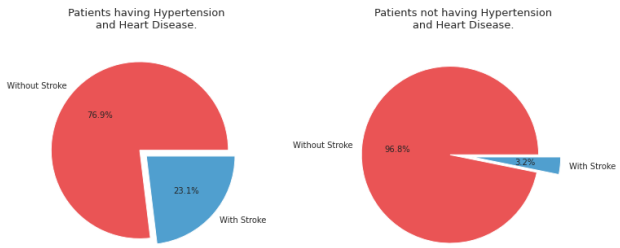

Fig. 3 Patients' distribution per hypertension and heart disease in the imbalanced dataset

having hypertension and heart disease, the data is imposed with SMOTE to balance the data, and the improved results are observed as shown in the pie chart fig4, which indicates that the patients with stroke are 46.7% which is higher than that of imbalanced data. To conclude, the results of the experiment applying SMOTE on imbalanced data have given improved results, and data was balanced with higher accuracy of stroke prediction, to precise this result of SMOTE, another experiment of applying SMOTE-ENN gave more improved balanced data and higher accuracy after applying SMOTE-ENN as shown in fig5, with an accuracy of stroke prediction in a person having hypertension and heart diseases is increased by 16.2% as compared to the accuracy of stroke prediction in a person having hypertension and heart diseases by applying only SMOTE.

Thus exploratory data analysis concludes with the observation that SMOTE showed better results on numerical data of average glucose levels in stroke persons, but the

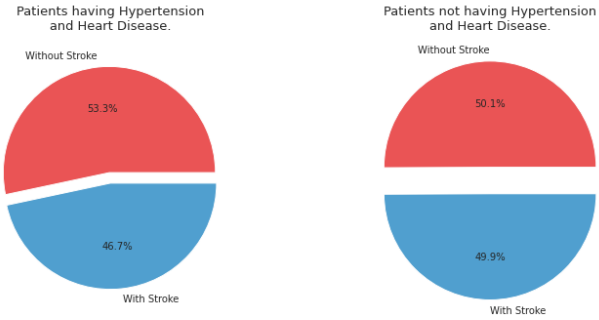

Fig. 4 Distribution of per hypertension for Patients and heart disease in balanced dataset using SMOTE

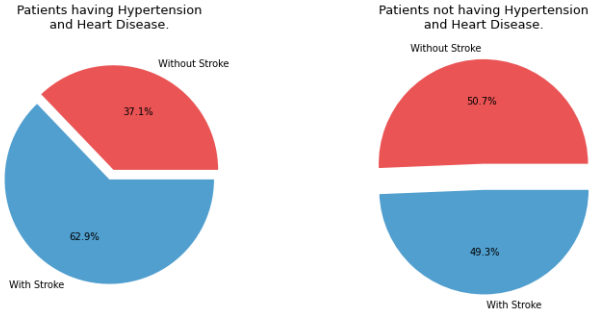

Fig. 5 Distribution of per hypertension for Patients and heart disease in balanced dataset using SMOTE ENN

results were improved more after applying SMOTE-ENN on categorical data like hypertension and heart diseases.

3.2 **Methods**

3.2.1 K Nearest Neighbors (KNN)

A machine learning method called K-nearest neighbours (KNN) is employed for regression and classification that is both easy to implement and very effective. It works as follows:

1. The feature vectors and labels from the training dataset are stored by KNN.
2. When a new data point is encountered during testing, KNN uses a distance metric like Euclidean distance to determine how far away it is from each of the training data points.
3. Using the determined distances, KNN chooses the nearest neighbours, where k is a positive number supplied by the user.
4. KNN uses a majority vote from the class labels of its k nearest neighbours to determine what label to give a new data point for a brain stroke.
5. The most numerous category is a proxy for its expected label.

Mathematically, given a new data point "x", the prediction function of KNN can be defined as:

For classification:

$$\hat{y} = \arg\max_{c} \sum_{i=1}^{k} I(y_i = c)$$

where $\hat{y}$ the anticipated value, y_i are target values of "k" nearest neighbors.

3.2.2 Logistic Regression (LR)

The maximum likelihood ratio determines the statistical significance of variables when creating the logistic regression equation. While computing the conditional likelihood P(Y=1/X), where X= (X1, X2, X3, ... , XN) is a representation of the n risk factors for brain stroke.[33] The logistic regression model can be written in Eq.1.

$$log\frac{P}{1-P} = \beta 0 + \beta 1X1 + \beta 2X2 + ------ + \beta nXn\,. \tag{1}$$

3.2.3 Support Vector Machine (SVM)

SVM works by separating data with hyperplanes and extending this to nonlinear boundaries. The performance and efficiency of SVM depend heavily on the kernel function. To achieve the highest performance, choosing the right kernel type is crucial. Different kernels were used in this study, including linear, polynomial, and Gaussian.[34] The linear, polynomial, and Gaussian kernel equations are represented in Eq.2, Eq.3, and Eq.4, respectively. Suppose x_i, and x_j are the two variables the kernel equations are written as

$$Linear\,kernel : K(x_i, x_j) = x_i \cdot x_j\,. \tag{2}$$

$$Polynomial\,kernel : K(x_i, x_j) = (x_i \cdot x_j + 1)^d\,. \tag{3}$$

$$Gaussian\,kernel : K(x_i, x_j) = exp(-\gamma \,||\, x_i - x_j \,||^2)\,. \tag{4}$$

Whereas d is the degree of the polynomial.

3.2.4 Decision Tree (DT)

Determination trees such as decision trees are used in machine learning for problems involving regression and classification.[35] It uses data to learn basic decision criteria, which it then incorporates into a model. A decision tree, in mathematical terms, is a function that converts a feature vector x into a target variable y:

For stroke classification:

$$f(\mathbf{x}) = \begin{cases} y, & \text{if leaf node is reached} \\ f_{\text{left}}(\mathbf{x}), & \text{if } \mathbf{x} \text{ satisfies the test at the internal node and goes left} \\ f_{\text{right}}(\mathbf{x}), & \text{if } \mathbf{x} \text{ satisfies the test at the internal node and goes right} \end{cases}$$

Decision trees learn optimal decision rules by data separated iterative based on qualities that increase information gain or minimize impurity. They are interpretable, handle categorical and numerical features, and capture non-linear relationships.

3.2.5 Random Forest (RF)

A sort of ensemble machine learning system, the Random Forest method creates predictions by aggregating the results of numerous decision trees.[36] A Random Forest is a form of artificial neural network that builds several decision trees using the bootstrap aggregating (as well as bagging) technique. Both the training data and the input characteristics are randomly chosen for each tree as they are used to educate them. The ultimate prediction is arrived at by adding up all of the individual predictions provided by the trees. The prediction of a Random Forest for stroke classification can be represented as:

$$f(\mathbf{x}) = \frac{1}{T}\sum_{t=1}^{T} f_t(\mathbf{x})$$

where $f_t(\mathbf{x})$ provides the prediction of the tth decision tree, $\mathbf{x}$ is the input feature vector, and T is the total number of decision trees in the Random Forest.

Random Forests reduce overfitting, handle high-dimensional data, and capture complex relationships. However, they can be computationally expensive and require hyperparameter tuning.

3.2.6 EXtreme gradient boosting tree (XGB)

XGBoost stands for An application of gradient-boosted decision-making trees is extreme gradient boosting. This algorithm is a popular ML algorithm, and it is widely used for supervised learning tasks such as regression, classification, and ranking. This is intended to push the limits of the calculation to the furthest reaches of machines based on the standards of the gradient boosting system. Decision trees are generated sequentially in this algorithm.[37] The HyperOpt technique is used to perform hyperparameter tuning for the XGB Classifier in our study. The general tree ensemble model in this algorithm is denoted in Eq.5

$$b_i = \phi(a_i) = \sum_{s=1}^{s} m_s(a_i), m_s \epsilon M\,. \tag{5}$$

3.2.7 Stacking

Stacking is a type of ensemble technique that makes predictions by combining the results of numerous models. It involves training base models, generating predictions, creating meta-features, training a meta-model, and making final predictions. The ensemble prediction is represented mathematically as:

$$P(\mathbf{x}) = g(\mathbf{h}_1(\mathbf{x}), \mathbf{h}_2(\mathbf{x}), \ldots, \mathbf{h}_n(\mathbf{x}))$$

where $P(\mathbf{x})$ is the final prediction, $\mathbf{x}$ is the input feature vector, $\mathbf{h}_i(\mathbf{x})$ represents the prediction of the ith base model, and $g(\cdot)$ is the meta-model that combines the base model predictions.

Stacking improves prediction accuracy by leveraging the strengths of different models. It requires careful selection and tuning to avoid overfitting and can handle complex relationships in the data.

3.2.8 Artificial Neural Network (ANN)

ANNs are a form of computational model that mimics some of the physical traits of the human brain and is based on current neurobiological research. An ANN uses training and learning techniques to calculate the discrepancy between each neuron's actual and predicted output. Each connection's weight is adjusted from the output layer through the hidden layer, followed by finally to the input layer to reduce error.[38] In this way, input pattern recognition increases in accuracy, which can be used to predict its probability. Following is the architecture in Table 1 used for ANN in this paper:

"

Table 1 Neural Network Architecture of proposed stroke model

Layer (type)	**Output Shape**	**Param #**
dense_7 (Dense)	(None, 16)	320
dense_8 (Dense)	(None, 32)	544
dropout_5 (Dropout)	(None, 32)	0
batch_normalization_5 (BatchNormalization)	(None, 32)	128
dense_9 (Dense)	(None, 64)	2112
dropout_6 (Dropout)	(None, 64)	0
batch_normalization_6 (BatchNormalization)	(None, 64)	256
dense_10 (Dense)	(None, 64)	4160
dropout_7 (Dropout)	(None, 64)	0
batch_normalization_7 (BatchNormalization)	(None, 64)	256
dense_11 (Dense)	(None, 64)	4160
dropout_8 (Dropout)	(None, 64)	0
batch_normalization_8 (BatchNormalization)	(None, 64)	256
dense_12 (Dense)	(None, 64)	4160
dropout_9 (Dropout)	(None, 64)	0
batch_normalization_9 (BatchNormalization)	(None, 64)	256
dense_13 (Dense)	(None, 1)	65
Total params		16,673
Trainable params		16,097
Non-trainable params		576

" **Parameters:** Learning Rate = 0.001, activation function= Relu, Loss Function = binary_crossentropy, Optimizer = Adam, Epochs = 100

4 Proposed brain stroke survival prediction model with Hyper-parameter tuning

The proposed approach includes stages to process a dataset, balance the dataset, and apply machine learning methods, and deep learning techniques: ann batch normalization. This is as shown in Figure 6. The model performance is assessed with the help of execution as follows: Data setup is the first step. Then data is preprocessed to update missing values, and remove outliers, then all the features are converted to numerical data using one-hot encoding. The given imbalanced data is balanced using SMOTE and SMOTE ENN techniques explained as follows:

Algorithm 1 SMOTE Algorithm

Input : Dataset D with minority class
Output: Balanced dataset D'

while *balance not achieved* **do**
  Select a random sample x from the minority class in D
  Find "K nearest neighbors" of x based on distance metric
  foreach *neighbor* n *in* K **do**
    1. Calculate vector v between x and n
    2. Create random number r between {0, 1}
    3. Multiply vector v by r to obtain new vector v'
    4. Add v' to x to create synthetic sample s
    5. Add s to the balanced dataset D'
  end
end

Algorithm 2 SMOTE-ENN Algorithm

Input : "Imbalanced dataset D"
Output: "
"Balanced dataset D'

1. Determine K as the number of nearest neighbours (if not determined, set $K = 3$)
2. **repeat**
   Find the K-nearest neighbor(s) for each observation in D based on distance metric **foreach** *observation* x *in* D **do**
     Find the majority class from the "K-nearest neighbors of x" **if** *class of* x *and majority class differ* **then**
       Remove x and its K-nearest neighbor(s) from D
     end
   end
   until *desired class proportion fulfilled*;
3. Apply SMOTE to the remaining minority class samples in D Create a balanced dataset D' by combining the SMOTE-generated synthetic samples and the remaining minority class samples

The ML models KNN, LR, SVM, and DT models with hyperparameter tuning are trained on both SMOTE-Balanced and SMOTE ENN-Balanced datasets. A neural network is trained where higher training and less overfitting are achieved by using batch normalisation and dropout techniques. Early stopping is used to avoid overfitting and determine the optimum number of epochs. Use measures like accuracy, precision, recall, and F1-score to assess how well each algorithm performs on the testing dataset. The overall efficiency of the model random forest, XG boosting, and stacking techniques are used.

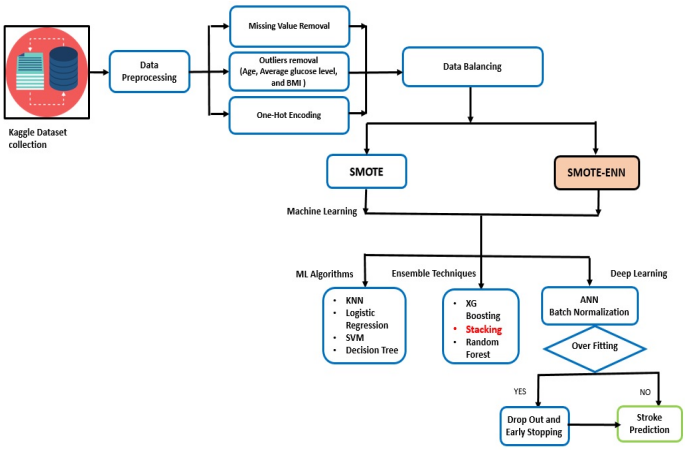

Fig. 6 Workflow of Proposed Model

The dataset is subjected to a preprocessing step that uses SMOTE-ENN to address the class imbalance and lower the overall noise level. After the dataset has been preprocessed, it is utilised to train four different types of base learners: DT, LR, RF, and KNN. Following the training of the base learners, the predictions generated by each are stacked and integrated using a Support Vector Machine (SVM) in the role of the meta-learner. The enhanced stacking strategy aims to increase predictive performance and effectively address imbalanced categorization issues. This is shown in algorithm 3 and Figure 7.

Algorithm 3 Proposed Stacking Model of SMOTE-ENN

Data: Training data with labels, Test data
Result: Ensemble stacking predictions on the test set

1. Split the training data into 80% training set and 20% meta training set
2. Train base learners LR, KNN, DT, RF on the training set
3. **foreach** *base learner* **do**
   | Make predictions on the test set to get $predictions_{\text{test}}$ using the trained model
   end
4. Create a new dataset called meta training set with n rows and 4 columns
   foreach *data point in the training set* **do**
   | $meta\_row \leftarrow$ concatenate $predictions_{\text{LR}}$, $predictions_{\text{KNN}}$, $predictions_{\text{DT}}$, $predictions_{\text{RF}}$, and the true label
   | Add $meta\_row$ to the meta training set
   end
5. Train the SVM meta learner on the meta training set **foreach** *base learner* **do**
   | Make predictions on the test set to get $predictions_{\text{base}}$ using the trained model
   end
6. Create a new dataset called meta test set with m rows and 4 columns
7. **foreach** *data point in the test set* **do**
   | $meta\_row \leftarrow$ concatenate $predictions_{\text{LR}}$, $predictions_{\text{KNN}}$, $predictions_{\text{DT}}$, $predictions_{\text{RF}}$
   | Add $meta\_row$ to the meta test set
   end
8. Make final predictions on the meta-test set using the trained SVM meta learner

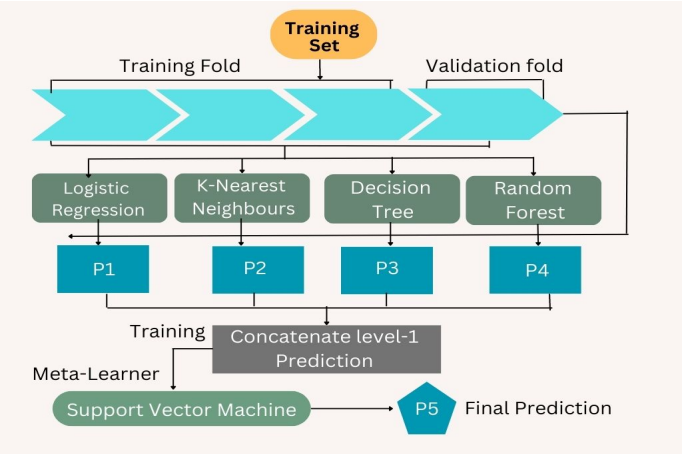

Fig. 7 Workflow of Proposed Model

5 Experimentation, Performance metrics, and Results

The following section provides information about the experimental setup required, different metrics used to evaluate the performance, and the results of the models.

5.1 Experimental Setup

In the experimental setup section, the ML model's performance is evaluated using Python and Google Collaboratory[39] environment also referred to as "Google Colab". It is a research initiative aimed at prototyping machine learning models on powerful hardware alternatives such as GPUs. It offers an environment for interactive development based on Jupyter Notebooks that do not require a server. Similar to the other tools in the G Suite, using Google Colab is completely free. It offers models for data preparation, categorization, segmentation, predictions, and visualization. Experiments were conducted on a computer with the following features: Intel(R) Core(TM) i5-8250U CPU @ 1.60GHz 1.80 GHz 16 GB Memory, 64-bit operating system, x64-based processor.

5.2 Performance Metrics

Several measures are used to evaluate the model's accuracy, and most of these are defined entirely by the confusion matrix values. Throughout the process of evaluating the Machine Learning models that were taken into consideration, numerous performance measures were obtained. In this analysis, we will focus on the terminology that is most frequently found in the pertinent research.[40] The phrase "recall," which also goes by the names "true positive rate" and "sensitivity," refers to the percentage of patients who had sepsis and were correctly identified as having the condition. A True Positive (TP) result indicates that there is value in both the actual class and the class that was predicted. The potential for positive value was accurately foreseen. True Negatives, also abbreviated as TN, are shorthand for real-class projected negative values. When the desired class and the observed class are different, this can lead to both false positives and false negatives. The term "false positive" (FP) refers to a situation in which the projected class is accurate, but the actual class is inaccurate. The occurrence of a false negative (FN) takes place when the actual class does not match the expected class.

Accuracy: is the number of correct predictions from the total positive and negative classes. The formula for accuracy is represented in Eq.6.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} * 100\,. \tag{6}$$

Precision: indicates out of all the positive predicted classes how many are actually positive. The formula for precision is represented in Eq.7.

$$Precision = \frac{TP}{TP + FP} * 100\,. \tag{7}$$

Recall: represents the ratio of correctly predicted positive observations to all observations in the actual class. The formula for recall is represented in Eq.8.

$$Recall = \frac{TP}{TP + FN} * 100\,. \tag{8}$$

F1-score: is calculated by using precision and recall. These metrics can be represented by using the following formula in Eq.9.

$$F1 - Score = 2 * \frac{(Precision * Recall)}{(Precision + Recall)}\,. \tag{9}$$

5.3 **Results**

In the proposed model, the results of the balanced dataset using SMOTE and SMOTE ENN are shown in Table 2 and Figure 8. When the data is normalized using SMOTE and SMOTE ENN, it has been discovered that SMOTE ENN performs better in terms of accuracy. In conclusion, the stacking technique continues to be the method with the best performance and is the primary recommendation that emerges from our research.

Model	SMOTE	SMOTE ENN
KNN	0.9178	0.9781
LR	0.8624	0.8997
SVM	0.9265	0.9756
DT	0.8756	0.9331
RF	0.9457	0.9817
XG	0.8821	0.9125
ANN	0.9376	0.9705
Stacking	0.9539	0.9884

Table 2 Machine Learning Models Accuracy Evaluation Using Smote and SMOTE ENN for Stroke survival prediction

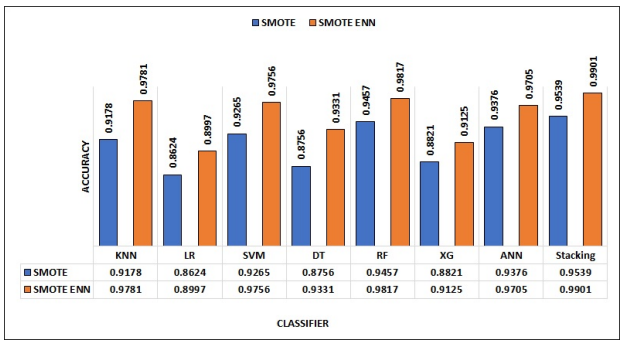

Fig. 8 Comparison of Accuracy of the proposed model using SMOTE and SMOTE ENN

6 Result Analysis and Discussion

The conclusions of the current research project are contrasted with those of a previously published study in[15] and[17] and used the same dataset.[18] All of the suggested models used SMOTE and SMOTE ENN techniques for data balancing. The models that used SMOTE ENN substantially outperform their performance, as measured by F-measure, precision, recall, and accuracy. In summary, the stacking technique continues to be the method with the best performance and is the primary recommendation that emerges from our research.

Model	Sailasya, Gangavarapu[15]	Elias Dritsas[17] et. al.	Proposed Model
KNN	0.8	0.81	0.98
LR	0.78	0.79	0.90
SVM	0.8	Not implemented	0.98
DT	0.66	0.91	0.93
RF	0.73	0.97	0.98
XG	Not implemented	Not implemented	0.91
ANN	Not implemented	0.92	0.97
Stacking	**Not implemented**	**0.98**	**0.99**

Table 3 Comparison of SMOTE ENN model's accuracy with existing models

The Table 3 and Figure 9, shows the comparison of proposed SMOTE ENN models with Sailasya, Gangavarapu et.al.[15] Elias Dritsas et.al.[17] Based on the accuracy scores in the Table 3, it has been observed that the proposed model achieved an accuracy score of 0.98 for KNN, SVM, and RF, which is higher than the existing models by,[15,17]. However, the proposed model achieved accuracy for ANN as 97% and Stacking as 99% which is 5% and 1% higher than the existing model by Elias Dritsas et. al.[17]

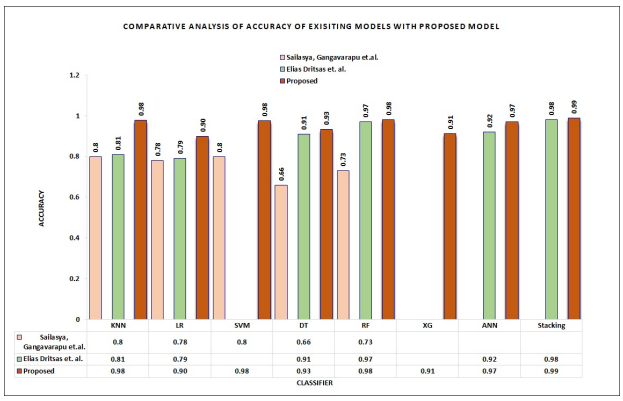

Fig. 9 Comparison of SMOTE ENN model's accuracy with existing models

Model	Sailasya, Gangavarapu[15]	Elias Dritsas et. al.[17]	Proposed model
KNN	0.77	0.92	0.98
LR	0.78	0.79	0.9
SVM	0.79	Not implemented	0.98
DT	0.78	0.91	0.94
RF	0.72	0.97	0.98
XG	Not implemented	0.79	0.92
Stacking	**Not implemented**	**0.97**	**0.99**

Table 4 Comparison of SMOTE ENN model's Precision with existing models

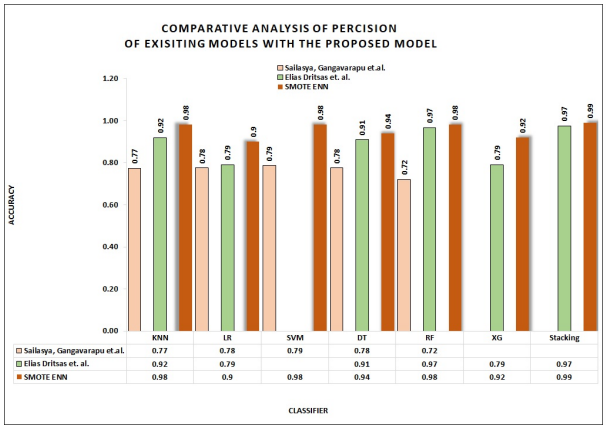

Fig. 10 Comparison of SMOTE ENN model's Precision with existing models

As shown in Table 4 and Figure 10, KNN, LR, SVM, DT, and RF models have shown a significant average increase of 18.6% as compared with the existing model.[15] On the other hand, the KNN, LR, DT, RF, XGBoost, and stacking models have shown a significant average increase of 5.14% as compared with the existing model.[17] This indicates that the model has identified true positives more accurately when the dataset is balanced using SMOTE ENN technique.

Model	Sailasya, Gangavarapu	Elias Dritsas et. al.	Proposed model
KNN	0.804	0.915	0.98
LR	0.776	0.791	0.9
SVM	0.818	Not implemented	0.98
DT	0.776	0.909	0.93
RF	0.727	0.966	0.98
XG	Not implemented	0.881	0.91
Stacking	**Not implemented**	**0.974**	**0.99**

Table 6 Comparison of SMOTE ENN model's F1-Score with existing models

Model	Sailasya, Gangavarapu[15]	Elias Dritsas et. al.[17]	Proposed model
KNN	0.84	0.92	0.98
LR	0.78	0.79	0.9
SVM	0.84	Not implemented	0.98
DT	0.78	0.91	0.93
RF	0.74	0.97	0.98
XG	Not implemented	0.82	0.91
Stacking	**Not implemented**	**0.97**	**0.99**

Table 5 Comparison of SMOTE ENN model's Recall with existing models

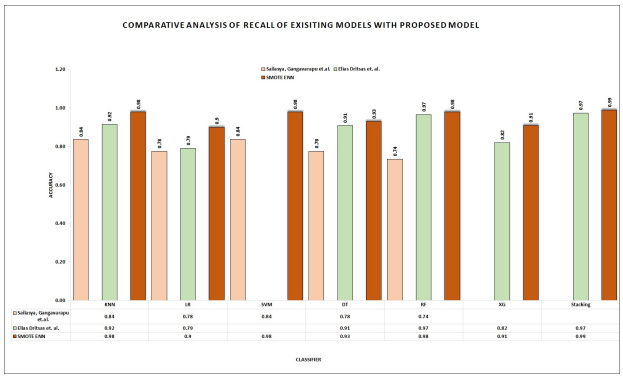

Fig. 11 Comparison of SMOTE ENN model's Recall with existing models

Moreover, comparing Table 5 and Figure 11, it is observed that the effect of the SMOTE-ENN technique on recall is also varied across different machine learning models. As per the observation, the KNN and SVM models have shown a significant increase of 14%. Similarly, LR, DT, and RF have shown increased recall of 12%, 18%, and 24% as compared with.[15] Moreover, the average increase of 5.16% is observed as compared to[17] for KNN, LR, DT, RF, XGBoost, and stacking. Because of this, the frequency of erroneous negative predictions in the model for predicting strokes was reduced, and a significant number of individuals with a higher likelihood of having a stroke were accurately recognized.

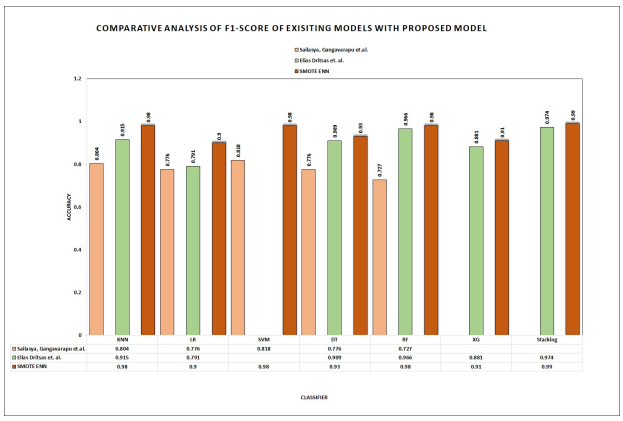

Fig. 12 Comparison of SMOTE ENN model's F1-Score with existing models

The overall performance of a machine learning model by taking into account both false positives and false negatives. In the Table 6 and Figure 12 and autore f Figure F1-score comparison, it was observed that the effect of the SMOTE-ENN technique on the F1 score varies across different machine learning models. has The proposed model have shown a significant increase of 17.4% in the F1-score measure for KNN, LR, SVM, DT, and RF as compared with Sailasya 2021-analysis. Moreover, models KNN, LR, DT, RF, XGBoost, and stacking have shown a 4.16% increase in the F1-score measure as compared with Dritsas 2022 stroke. This suggests that the model is already good at identifying both true positives and true negatives in the original dataset. There is the field of forecasting strokes, a high F1-score suggests that the model correctly identifies a greater part of patients who have a chance to suffer a stroke, while avoiding false positive results. The use of a freely accessible dataset is a limitation of this proposed work. In contrast to statistics from a hospital or institute, these are limited in scope and form. The latter may provide more informative data models with more features that capture a comprehensive health profile for the participants, however, due to privacy concerns, accessing this data is frequently time-consuming and challenging.

7 Conclusions

To reduce the risk of unanticipated issues, a stroke must be averted as a possible threat to human life. Clinical providers, medical experts, and decision-makers can now take advantage of the established models to find the most relevant has (or, as an alternative, risk factors) for the possibility of a stroke and to assess the chance or risk related to this occurrence as a consequence of the fast advancement of AI and ML. However, due to the intricacy of the issue, it is very challenging to develop methods for the prediction with the detection of stroke, which relies on machine learning as well as information mining. In this regard, machine learning can aid in the early diagnosis of stroke and assist in reducing its severe aftereffects. Numerous variables may have an impact on the outcomes produced by the learning systems in use today. Class imbalance, where examples from one class considerably exceed instances from another class in the training information, has been connected to one of these causes. The problem of the class imbalance within the Kaggle dataset was addressed in this paper by doing a thorough experimental evaluation utilizing SMOTE and SMOTE ENN. The recall and precision (F1-Score) and accuracy metrics used in the performance assessment of the classifiers are fundamentally appropriate for interpreting the results of the models and show the models' capacity to

categorize data. It has been found that the accuracy of various machine learning models is affected differently by the SMOTE as well as SMOTE-ENN approaches. The accuracy of the KNN model significantly increases from 0.8 to 0.9781. Similarly, the DT model exhibits a notable improvement in accuracy from 0.66 to 0.9331. The prediction accuracy of LR, on the other hand, is mostly unaffected by the SMOTE and SMOTE-ENN methods, indicating that the model is already effective at predicting the right class in the initial dataset. The SVM model also shows consistently high accuracy even without using the SMOTE-ENN SMOTE- technique, with an accuracy of 0.9265, also using SMOTE- ENN achieved a proposed model has achieved an accuracy of 97% with the SVM classifier. The RF model shows a moderate increase in accuracy using SMOTE, while it has been increased using SMOTE-ENN to 98.17%. The newly applied XGBoost model has shown an accuracy of 88% with SMOTE and 91.25% with the SMOTE-ENN technique. This suggests that the model has become better at predicting the correct class after balancing the dataset. The ANN and stacking models also show significant increases in accuracy with the use of SMOTE and SMOTE-ENN techniques, indicating that both models benefit from dataset balancing. Additionally, they show the models' accuracy and their potential for prediction in relation to the type of stroke. SMOTE ENN stacking classification performs better than the alternative methods, with a precision of 99%, recall of 99%, F1 measure of 99%, and accuracy of 99.01%. In order to identify the most precise method to estimate stroke based on various variables that reflect the participants' profiles, this study examines the effectiveness of various ML algorithms. This study specifically examines how effectively these algorithms can forecast strokes.

The application of deep learning strategies will be the focus of the next phase of this research project, which aims to improve the ML framework. The last step will involve gathering image data from brain CT scans and evaluating deep learning models' propensity to forecast the development of strokes. It is challenging but potentially advantageous to forecast strokes with precise accuracy.

8 Declarations

Conflict of Interest The authors declare no competing interests

Acknowledgments The authors are thankful to the management of the Indian Institute of Information Technology, Nagpur, and Visvesvaraya National Institute of Technology, Nagpur and Symbiosis Institute of Technology, Nagpur India.

Data availability statement All data that support the findings of this study are included in the article.

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

Authors Contribution

- **Snehal Shinde:** Conceived and designed the study, collected and analyzed the data, and wrote the manuscript.
- **Manish P Kurhekar:** Assisted in designing the study, conducted experiments, and contributed to data analysis and interpretation. Also provided critical revisions to the manuscript.
- **Monali Gulhane:** Contributed to the study design, provided statistical expertise, and assisted in data analysis and interpretation. Also played a major role in writing and editing the manuscript.

– **Nileshchandra Pikle:** Provided technical expertise, conducted experiments, and contributed to data analysis and interpretation. Also participated in drafting and revising the manuscript.

References

1. Eng H Lo, Turgay Dalkara, and Michael A Moskowitz. Mechanisms, challenges and opportunities in stroke. *Nature reviews neuroscience*, 4(5):399–414, 2003.
2. Dariush Mozaffarian, Emelia J Benjamin, Alan S Go, Donna K Arnett, Michael J Blaha, Mary Cushman, Sarah De Ferranti, Jean-Pierre Després, Heather J Fullerton, Virginia J Howard, et al. Heart disease and stroke statistics—2015 update: a report from the american heart association. *Circulation*, 131(4):e29–e322, 2015.
3. Jiaquan Q Xu, Sherry L Murphy, Kenneth D Kochanek, and Elizabeth Arias. Mortality in the united states, 2015. nchs data brief, no 267. hyattsville, md: Us department of health and human services, cdc. *National Center for Health Statistics*, 2016.
4. George Howard and David C Goff. Population shifts and the future of stroke: forecasts of the future burden of stroke. *Annals of the New York Academy of Sciences*, 1268(1):14–20, 2012.
5. U. Wilensky. NetLogo NetLogo 6.0.3 User Manual (2016). [Online]. Available: https://www.world-stroke.org/world-stroke-day-campaign/why-stroke-matters/learnabout-stroke.
6. Andrea Biasiucci, Robert Leeb, Iñaki Iturrate, Serafeim Perdikis, Abdul Al-Khodairy, Tiffany Corbet, Armin Schnider, T Schmidlin, Huaijian Zhang, Manuela Bassolino, et al. Brain-actuated functional electrical stimulation elicits lasting arm motor recovery after stroke. *Nature communications*, 9(1):1–13, 2018.
7. Jeyaraj D Pandian, Seana L Gall, Mahesh P Kate, Gisele S Silva, Rufus O Akinyemi, Bruce I Ovbiagele, Pablo M Lavados, Dorcas BC Gandhi, and Amanda G Thrift. Prevention of stroke: a global perspective. *The Lancet*, 392(10154):1269–1278, 2018.
8. Asha Latha Thandu, Vijaya Saradhi Thommandru, and Pradeepini Gera. Data science in healthcare monitoring under covid-19 detection by extended hybrid leader-based compressed neural network. *New Generation Computing*, pages 1–28, 2023.
9. Nidhi Agarwal, Sachi Nandan Mohanty, Shweta Sankhwar, and Jatindra Kumar Dash. A novel model to predict the effects of enhanced students' computer interaction on their health in covid-19 pandemics. *New Generation Computing*, pages 1–34, 2023.
10. Aditya Gupta and Amritpal Singh. An intelligent healthcare cyber physical framework for encephalitis diagnosis based on information fusion and soft-computing techniques. *New Generation Computing*, 40(4):1093–1123, 2022.
11. Talha Karadeniz, Gül Tokdemir, and Hadi Hakan Maraş. Ensemble methods for heart disease prediction. *New Generation Computing*, 39(3-4):569–581, 2021.
12. Nonita Sharma, Jaiditya Dev, Monika Mangla, Vaishali Mehta Wadhwa, Sachi Nandan Mohanty, and Deepti Kakkar. A heterogeneous ensemble forecasting model for disease prediction. *New Generation Computing*, pages 1–15, 2021.
13. Tasfia Ismail Shoily, Tajul Islam, Sumaiya Jannat, Sharmin Akter Tanna, Taslima Mostafa Alif, and Romana Rahman Ema. Detection of stroke disease using machine learning algorithms. In *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–6. IEEE, 2019.
14. Xuemeng Li, Di Bian, Jinghui Yu, Mei Li, and Dongsheng Zhao. Using machine learning models to improve stroke risk level classification methods of china national stroke screening. *BMC medical informatics and decision making*, 19:1–7, 2019.
15. Gangavarapu Sailasya and Gorli L Aruna Kumari. Analyzing the performance of stroke prediction using ml classification algorithms. *International Journal of Advanced Computer Science and Applications*, 12(6), 2021.
16. Tahia Tazin, Md Nur Alam, Nahian Nakiba Dola, Mohammad Sajibul Bari, Sami Bourouis, Mohammad Monirujjaman Khan, et al. Stroke disease detection and prediction using robust learning approaches. *Journal of healthcare engineering*, 2021, 2021.
17. Elias Dritsas and Maria Trigka. Stroke risk prediction with machine learning techniques. *Sensors*, 22(13):4670, 2022.
18. Stroke Prediction Dataset. Available online:, howpublished = https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset, note = Accessed: 18 March 2023.
19. Sebastián Maldonado, Julio López, and Carla Vairetti. An alternative smote oversampling strategy for high-dimensional datasets. *Applied Soft Computing*, 76:380–389, 2019.
20. Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, 6(1):20–29, 2004.
21. Sheng-Feng Sung, Ling-Chien Hung, and Ya-Han Hu. Developing a stroke alert trigger for clinical decision support at emergency triage using machine learning. *International Journal of Medical Informatics*, 152:104505, 2021.

22. Kathryn M Rexrode, Tracy E Madsen, Amy YX Yu, Cheryl Carcel, Judith H Lichtman, and Eliza C Miller. The impact of sex and gender on stroke. *Circulation research*, 130(4):512–528, 2022.
23. Yuda Turana, Jeslyn Tengkawan, Yook Chin Chia, Michael Nathaniel, Ji-Guang Wang, Apichard Sukonthasarn, Chen-Huan Chen, Huynh Van Minh, Peera Buranakitjaroen, Jinho Shin, et al. Hypertension and stroke in asia: A comprehensive review from hope asia. *The Journal of Clinical Hypertension*, 23(3):513–521, 2021.
24. Connie W Tsao, Aaron W Aday, Zaid I Almarzooq, Cheryl AM Anderson, Pankaj Arora, Christy L Avery, Carissa M Baker-Smith, Andrea Z Beaton, Amelia K Boehme, Alfred E Buxton, et al. Heart disease and stroke statistics—2023 update: a report from the american heart association. *Circulation*, 147(8):e93–e621, 2023.
25. KK Andersen and TS Olsen. Stroke case-fatality and marital status. *Acta Neurologica Scandinavica*, 138(4):377–383, 2018.
26. Juliet Addo, Luis Ayerbe, Keerthi M Mohan, Siobhan Crichton, Anita Sheldenkar, Ruoling Chen, Charles DA Wolfe, and Christopher McKevitt. Socioeconomic status and stroke: an updated review. *Stroke*, 43(4):1186–1191, 2012.
27. George Howard. Rural-urban differences in stroke risk. *Preventive Medicine*, 152:106661, 2021.
28. Lei Zhao, Shuangling Xiu, Lina Sun, Zhijing Mu, and Junling Fu. A study of the relationship between blood glucose and serum insulin in acute cerebrovascular disease. *Evidence-Based Complementary and Alternative Medicine*, 2022, 2022.
29. Salah Elsayed and Muath Othman. The effect of body mass index (bmi) on the mortality among patients with stroke. *Eur. J. Mol. Clin. Med*, 8:181–187, 2021.
30. Biqi Pan, Xiao Jin, Liu Jun, Shaohong Qiu, Qiuping Zheng, and Mingwo Pan. The relationship between smoking and stroke: a meta-analysis. *Medicine*, 98(12), 2019.
31. Guillaume Lemaître, Fernando Nogueira, and Christos K Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research*, 18(1):559–563, 2017.
32. Dennis L Wilson. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics*, (3):408–421, 1972.
33. Simon Nusinovici, Yih Chung Tham, Marco Yu Chak Yan, Daniel Shu Wei Ting, Jialiang Li, Charumathi Sabanayagam, Tien Yin Wong, and Ching-Yu Cheng. Logistic regression was as good as machine learning for predicting major chronic diseases. *Journal of clinical epidemiology*, 122:56–69, 2020.
34. William S Noble. What is a support vector machine? *Nature biotechnology*, 24(12):1565–1567, 2006.
35. Mohmad Badr Al Snousy, Hesham Mohamed El-Deeb, Khaled Badran, and Ibrahim Ali Al Khlil. Suite of decision tree-based classification algorithms on cancer gene expression data. *Egyptian Informatics Journal*, 12(2):73–82, 2011.
36. Yanjun Qi. Random forest for bioinformatics. *Ensemble machine learning: Methods and applications*, pages 307–323, 2012.
37. Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, et al. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4, 2015.
38. Anil K Jain, Jianchang Mao, and K Moidin Mohiuddin. Artificial neural networks: A tutorial. *Computer*, 29(3):31–44, 1996.
39. Ekaba Bisong et al. *Building machine learning and deep learning models on Google cloud platform*. Springer, 2019.
40. Mohammad Hossin and Md Nasir Sulaiman. A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, 5(2):1, 2015.