# Mediguide - Brain stroke detection on imbalanced dataset

***Kumar Gaurav(mailto:kumargaurav.3815@gmail.com), Sparsh Singh(mailto:sparshsingh980@gmail.com), Yash Kumar(mailto:yash_kumar.ug21@nsut.ac.in),***
***Mentor- Dr. Gaurav Singal***
***Computer Science and Engineering, Netaji Subhas University of Technology***

***Abstract***-Cerebral stroke is a medical condition in which sudden disruption of blood flow to the brain causes damage to brain tissue. This happens due to either of two reasons blockage in the blood vessel or the rupture of the blood vessels. This results in a lack of blood flow that prevents oxygen and nutrients from reaching brain cells and eventually causes death.
However, by early detecting brain stroke we can significantly improve the outcomes and reduce long-term consequences as early detection of brain stroke leads to minimizing brain damage, prevents disability and impairment and helps in identifying the cause earlier leads in improving the patient's conditions and will increase the chances of successful recovery by taking correct treatment on time. The dataset of brain stroke is highly imbalanced in terms of a minority class of stroke in a dataset. It contains 12 features out of which 43,400 samples only have 783 strokes which are only 1.804% of the total patients and detecting the stroke correctly in this highly imbalanced dataset is a major challenge as the basic machine learning algorithms fail in detecting the minority class which
leads to a wrong and bad prediction. This research paper aims to predict brain stroke with a highly imbalanced dataset with very high accuracy as imbalanced datasets are a major and common problem in the medical field and that makes the correct prediction difficult. In our research, we are solving this problem by giving the maximum and correct accuracy out of all algorithms currently present. First, we balance the dataset using different machine learning techniques like Oversampling minority class, Under sampling majority class, Synthetic Minority Oversampling Technique (SMOTE) and its variants like(SMOTE-TOMEK),(SMOTE-ENN),(SVM-SMOTE),(Borderline SMOTE) and our novel model (SMOTEN) out of all these algorithms we found that balancing data with (SMOTEN) gives maximum accuracy on our baseline model which is Artificial neural networks (ANN) and after that, we achieve maximum accuracy of 99% and (AUC of 0.99) using ensemble learning voting classifier on the best-performing models random forest classifier(99%) and k nearest neighbors(KNN)(98%) out of all tested algorithms which include ANN(92%), logistic regression(81%), Adaboost(89%), Naive bias(60%), support vector machine (SVM)(91%).

## INTRODUCTION

The medical data classification faces the[3] imbalanced count of data samples, here at least one class forms only a very small minority of the data, but it is a drawback of most of the machine learning algorithms. The class labels are the reason that medical datasets are mostly imbalanced. The existing classification techniques generally perform badly on minority classes due to an imbalance in the dataset.

Brain stroke is a medical condition in which sudden disruption of blood flow to the brain causes leading damage to brain tissue that results in a lack of blood flow which prevents oxygen and nutrients from reaching brain cells and eventually causes death. Stroke is the most general neurological reason for death and inability[1] worldwide. Approximately 16 million people suffer from stroke in the world. Stroke is the world's second-leading cause of mortality; as a result, it requires prompt[2] treatment to avoid brain damage. Lowering the death rates can be achieved by having early brain stroke detection which can prevent or lessen the severity of stroke. It is a promising method to use machine learning to identify risk variables.

The dataset of our problem which is brain stroke detection is highly imbalanced as out of 43,400

samples only 783 people have stroke which made it difficult to predict as the machine learning algorithms struggle with the imbalanced datasets as it fails to predict the minority class accurately.

This research paper aims to find an effective solution for handling imbalanced datasets, the first step of the solution is to find a way to balance the dataset in the best possible way and for that, we have to test it with a baseline algorithm and by using the accuracy of the baseline algorithm we can find the perfect balanced dataset and once we have the balanced dataset then we can apply various machine learning and deep learning algorithms to get the best accuracy. In our proposed model we take an Artificial neural network(ANN) as the baseline algorithm and for balancing the dataset we use different techniques such as Oversampling minority class, Under sampling majority class, Synthetic Minority Oversampling Technique (SMOTE) and its variants like(SMOTE-TOMEK),(SMOTE-ENN),(SVM-SMOTE),(Borderline SMOTE) and our novel model (SMOTEN) and out of all these we found that which is a novel model (SMOTEN) is the best for balancing and outperforms all other balancing techniques mentioned and for the accuracy we have ensemble learning voting classifier which gives maximum accuracy of 99% and AUC score of 0.99 which is performed on the best-performing models random forest classifier(99%) and k nearest neighbors(KNN)(98%) out of all tested algorithms which includes ANN(92%), logistic regression(81%) , AdaBoost(89%), Naive bias(60%), support vector machine (SVM)(91%).
The remaining paper is organized as follows:
We presented a literature review in Section 2. In section 3 we briefly represented the dataset and methods. In section 4 we interpreted the performance metrics for each experiment. The conclusion is given in Section 5.

**LITERATURE WORK**

Brain Stroke has been a major leading cause of death all over the world. Brain Stroke is a disability that remains for a long time. The medical datasets are mostly imbalanced in their class labels. Medical datasets, however, are frequently unbalanced in their class label, with the tendency to poorly predict minority classes by machine learning algorithms. A higher amount is being yielded due to early brain stroke prediction that is profitable for the initiating time. Various lifestyle decisions by people particularly in the current scenario by
evolving elements such as high blood sugar, heart disease, obesity, diabetes, and hypertension are the major causes of Brain stroke. V. Krishna, J. Sasi Kiran[4] In this research study has used various machine learning (ML) algorithms like K nearest neighbor, logistic regression, random forest (RF) classifier and SVC. One of the given algorithms with high accuracy is used to design a model in this research work to predict strokes for new inputs.

S. Vasa, P. Borugadda and A. Koyyada in[5] research to develop the optimal model to predict brain stroke using Machine Learning Algorithms (MLA's), namely Logistic Regression (LR), Decision Tree Classifier (DTC), Random Forest Classifier (RFC), Naive Bayes Classifier(NBC), XGBoost Classifier(XGB), Support Vector Machine (SVC) and KNN Classifier (KNN). Apply Grid Search CV (CV=5) along with the above algorithms with hyperparameter on the given dataset. Due to the given unbalanced dataset, while training the models, a few difficulties were addressed like underfitting, null values in the dataset and boosting the performance of the models by balancing the data using data sampling methods such as SMOTE. In the 7 models given XGB is the optimal model with 96.34% accuracy.
study by S. U. Sabha and A. Assad [6] aims to evaluate the effectiveness of five oversampling strategies that are intended to address data imbalance, namely random oversampling, SMOTE, borderline SMOTE, ADASYN, and Deep SMOTE. On the basis of evaluation metrics, a comparative analysis is carried out and the effectiveness of each strategy is examined. On small and imbalanced results Deep SMOTE outperformed other oversampling techniques while demonstrating experimental results.

The proposed soft voting model[7] improved the accuracy and robustness of the final prediction compared to a single classifier. The limitation of stroke type classification study can lead to the reduction of healthcare costs and appropriate use of resources. To resolve this problem, the future introduced a swarm intelligence-based optimization to improve classification accuracy. The accuracy of the proposed model is 96.88%.
Low precision, recall, and F1 scores but high accuracy is often obtained due to an imbalance of the real-world datasets.
The paper[8] aims to solve these issues by using an oversampling technique called Synthetic Minority Over-sampling Technique (SMOTE). A real-world imbalanced dataset called the Lending Club is obtained for training the models. The dataset comprises 151 features among which 38 features are categorical and the remaining 113 features are numerical. It has a total of 22,60,701 instances. To improve accuracy, precision, recall, and F1-score SMOTE is applied to the dataset to make it balanced. Four supervised classifier algorithms i.e. Logistic Regression, XGBoost, Decision Tree and Random Forest were used to predict brain stroke. Among them, random forest produced the highest accuracy with an accuracy of 87% with f1-score, recall and precision of 0.89,0.97 and 0.81 respectively.

**METHODOLOGY**

Data Cleaning and Preprocessing- The first step involves data cleaning and preprocessing which includes handling of null values and converting categorical columns to numerical columns for better analysis and performance of the model. In our dataset the columns containing null values are smoking status and BMI in which the number of instances of null values in smoking status
is very high so we need to drop it and for BMI the null values instances are low so we can handle it by imputing null values by median and also drop the id column as we don't need it in our further analysis of the data.
Data Analysis-In this step we analyze the data according to which we get the following results out of 43,400 instances only 783 have stroke and 42617 instances did not have stroke. the median of the BMI is 27.7 ,4061 instances have hypertension out of them only 200 people have stroke, 2062 instances have heart disease out of them 177 have stroke and out of 25665 females and 17724 males 431 female have stroke and 352 males have stroke and then we do the correlation analysis to analyze the correlation among the features.

Handling Data Imbalance-
Creating a baseline model using Artificial Neural Networks(ANN) which mimics the behavior of biogical neural network present in human brain. ANN has an input layer which takes data from outside as input this data then passes through hidden layers that transforms the input into data valuable to output layer then finally output layer provides output in the form of response to the input provided, In our project we first train ANN with imbalanced data after training we found that it completely fails in detecting stroke due to the highly imbalance nature of the data then we balance the data using different techniques and again train the balance data using ANN and the technique that gives us maximum accuracy on balanced dataset we will select the technique and move further.

Different Techniques for Balancing the Data-
1)Under sampling- In this technique we balance the dataset by randomly under sampling the majority class to balance with the minority class this improves the performance of ANN in predicting stroke but gives accuracy of only 78% and f1 score for stroke is .79.
2)Oversampling-In this technique we balance the dataset by oversampling the majority class to balance the ratio between majority class and minority class this improves the performance of ANN in detecting stroke and gives accuracy of 81% and f1 score for stroke is 0.83 which is better than under sampling.

3) SMOTE- It stands for synthetic minority oversampling technique it first selects a minority class instance at random and then finds its k nearest minority class neighbors. The synthetic instance is then created by choosing one of the k nearest neighbors b at random and connecting a

to b to form a line segment in the feature space the synthetic instances are generated as a combination of a and b instances. This gives an accuracy of 83% and f1 score of 0.84 which is better than oversampling.

4)SMOTE Tomek-This is a variation smote that combines the Smote ability to generate synthetic data for minority class and Tomek Links ability to remove the data that are identified as Tomek links from majority class (that is, samples of data from the majority class that is closest with the minority class data).It gives an accuracy of 83% and f1 score of .84 which is same as SMOTE.

5)SMOTE-ENN- ENN stands for edited nearest neighbor. This is a variation smote that combines the SMOTE ability to generate synthetic examples for minority class and ENN ability to delete some observations from both classes that are identified

as having different class between the observation's class and its K-nearest neighbor majority class. It gives an accuracy of 87% and f1 score of .88 which is best till now .

6)Adasyn-[9] The essential idea of ADASYN is to use a weighted distribution for different minority class examples according to their level of difficulty in learning, where more synthetic data is generated for minority class samples which are not as easy to learn as compared to minority samples. Thus, ADASYN improves learning in 2 ways with respect to data distribution (1) adaptively shifting the classification decision boundary toward the difficult examples and (2) reducing the bias introduced by the class imbalance. This gives an accuracy of 83% and f1 score of .84 which is the same as smote.

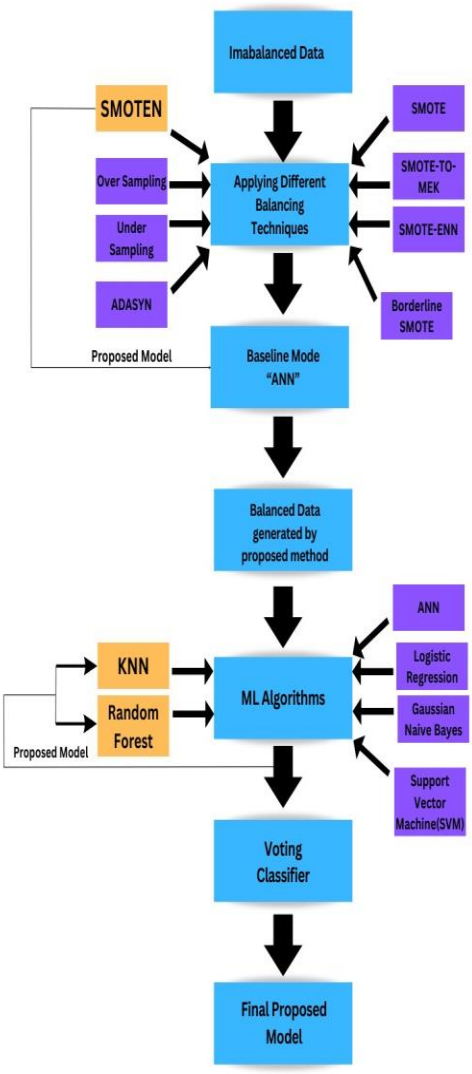
7)SVM SMOTE-[10] SVM-SMOTE is an over-sampling technique that is used to investigate how well it handles the trade-off. Its ancestor is SMOTE which balances class distribution by synthetically generating new minority class instances from given existing minority class instances along directions towards their neighbor instances. SVM-SMOTE focuses on generating new minority class instances near borderlines with SVM so as to help establish boundaries between classes. This gives an accuracy of 89% and f1 score of .82 which is highest accuracy till now.

8)BORDERLINESMOTE-[11] Borderline SMOTE[12] is an improved oversampling algorithm based on SMOTE, which uses only a few class samples on the border to combine new samples, thus improving the sample category distribution. Borderline SMOTE is divided into the following 3 categories: Noise, Danger and Safe. At last, a few number Danger samples were oversampled. This gives an accuracy of 91% and f1 score of .91 which is the highest accuracy till now.

9)SMOTEN- It stands[13] for Synthetic Minority Over-sampling Technique for Nominal It expects that the data to resample are only made of categorical features. This gives an accuracy of 93% and f1 score of .93 which is highest accuracy till now.

So, after all these techniques, we found that SMOTEN gives us the best-balanced data so we finalize this technique and then apply the other classification algorithms such as Random Forest Classifier, K nearest Neighbor, Logistic regression, Support vector machine, AdaBoost, Naive Bias and we found that out of all models the KNN and random forest gives us the best result so we
use ensemble learning voting classifier[14] on them which gives us an accuracy of 99% and f1 score of .99 which is our proposed model for maximum accuracy

FLOWCHART REPRESENTING THE WORKING OF THE MODEL-:




**RESULT ANALYSIS-**

In medical diagnosis, false positives are generally considered to be less concerning than false negatives. A false negative result occurs when the patient is actually diagnosed with a disease but the model predicts that the patient does not have that disease. The patient's situation can get worse due to delayed treatment because of it.

ROC curve, F1 score, recall, precision and accuracy are the commonly used evaluation metrics in binary classification problems. Different model comparisons are made by these metrics and suggest the best one for a given problem. Evaluation metrics are important because they help us to determine whether the model is meeting our expectations and how well the model is performing.

Precision: In terms of all positive predictions (TP + FP), precision is the percentage of true positive predictions (TP) among all positive predictions. Precision measures how often the model correctly identifies positive instances. High precision refers to the model making very few false positive predictions. Precision = True Positives / (True Positives + False Positives)

Recall: It is used to count accurate positive predictions(TP) percentage of all the instances of positive data (TP + false negative predictions (FN)). In other words, recall is used to measure how often the model correctly identifies actual positive instances. High recall refers to the model's ability to correctly identify most of the actual positive cases.
Recall = True Positives / (True Positives + False Negatives)

F1-score: The F1-score is the harmonic mean of precision and recall, giving equal weight to both metrics. It provides a single value that summarizes the model's performance, with 1 being the best possible value and 0 being the worst possible value,
F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

Accuracy: Accuracy measures the proportion of the accurate predictions among all the given predictions made by the model.

AUC-ROC-AUC ROC measures the performance of a binary classification model by plotting the true positive rate against the false positive rate at various classification thresholds. The higher the AUC ROC value, the better the model's discrimination.

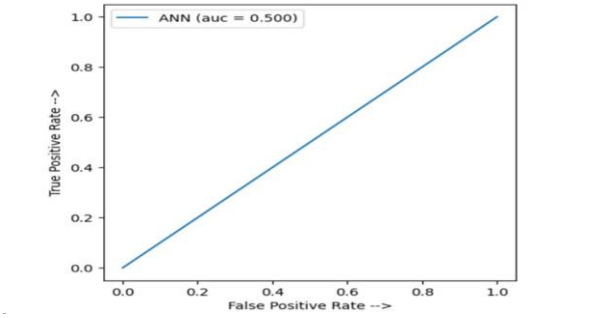
BASELINE MODEL WITHOUT SAMPLING-

ANN is our baseline model that mimics the behavior of biogical neural network present in human brain. ANN has an input layer which takes data from outside as input this data then passes through hidden layers that transforms the input into data valuable to output layer then finally output layer provides output in the form of response to the input provided. when ANN is fed with imbalanced data it gives precision and recall for stroke 0.00 and AUC score 0f .50 which indicates that it does not determine stroke correctly.

```
Epoch 48/50
1085/1085 [==============================] - 1s 807us/step - loss: 0.0728 - accuracy: 0.9820
Epoch 49/50
1085/1085 [==============================] - 1s 800us/step - loss: 0.0727 - accuracy: 0.9820
Epoch 50/50
1085/1085 [==============================] - 1s 896us/step - loss: 0.0726 - accuracy: 0.9820
272/272 [==============================] - 0s 828us/step - loss: 0.0764 - accuracy: 0.9819
[0.07636862993240356, 0.9819124341011047]
272/272 [==============================] - 0s 701us/step
Classification Report:
              precision    recall  f1-score   support

           0       0.98      1.00      0.99      8523
           1       0.00      0.00      0.00       157

    accuracy                           0.98      8680
   macro avg       0.49      0.50      0.50      8680
weighted avg       0.96      0.98      0.97      8680
```

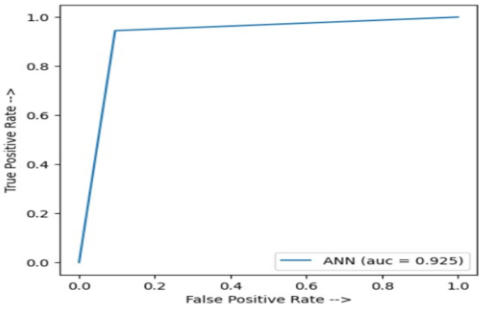



ANN when fed with the data balanced with our best proposed method gives accuracy of 93% precision of .90 recall of .96 ,f1 score of 0.96 and AUC=0.925 which is very good as compared to without sampling baseline model.

```
2131/2131 [==============================] - 2s 859us/step - loss: 0.2117 - accuracy: 0.9261
Epoch 49/50
2131/2131 [==============================] - 2s 910us/step - loss: 0.2112 - accuracy: 0.9260
Epoch 50/50
2131/2131 [==============================] - 2s 858us/step - loss: 0.2110 - accuracy: 0.9264
533/533 [==============================] - 0s 710us/step - loss: 0.2082 - accuracy: 0.9297
[0.20817238092422485, 0.9297236800193787]
533/533 [==============================] - 0s 631us/step
Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.90      0.93      8523
           1       0.91      0.96      0.93      8524

    accuracy                           0.93     17047
   macro avg       0.93      0.93      0.93     17047
weighted avg       0.93      0.93      0.93     17047
```

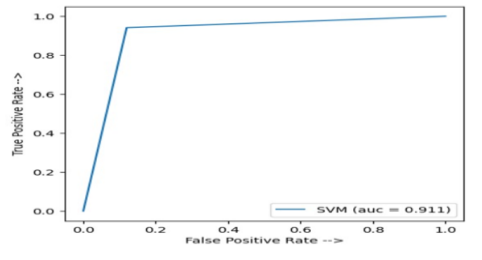



Support Vector Machine- Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification and regression. The main objective of the SVM algorithm is to find the optimal hyperplane that can separate the data points into different classes in the feature space in an N-dimensional space. [17] The hyperplane tries to maximize the margin between the closest points of different classes. When we fed the balanced data to SVM it gave an accuracy of 91% precision of .89 recall of .94,f1 score of 0.91 and AUC=0.911.

```
              precision    recall  f1-score   support

           0       0.94      0.88      0.91      8523
           1       0.89      0.94      0.91      8524

    accuracy                           0.91     17047
   macro avg       0.91      0.91      0.91     17047
weighted avg       0.91      0.91      0.91     17047
```

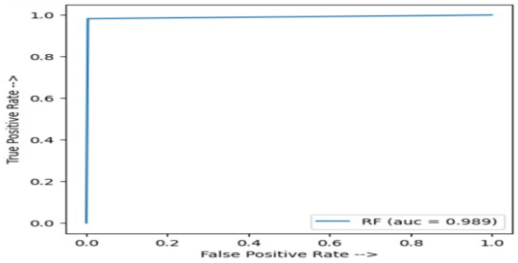



Random Forest Classifier-Random Forest[15] is an ensemble of decision trees; it creates multiple decision trees and combines their predictions to make the final decision. The main idea is that multiple trees[16] can work together to overcome overfitting and other limitations of only one decision tree, by reducing the variance and increasing the robustness of the model. Random Forest is a versatile and. powerful machine learning that can handle both non-linear and

linear relationships, missing values, and large datasets.

When we fed the balanced data to random forest it gives accuracy of 99% precision of 1 recall of .98 ,f1 score of 0.99 and AUC=0.989.

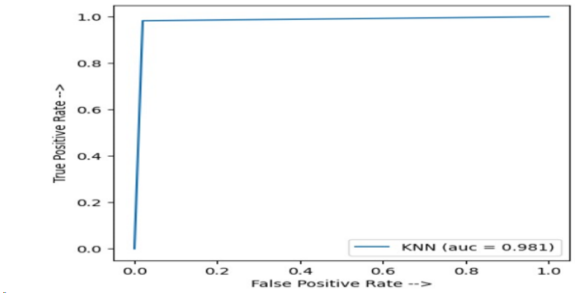
| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.98 | 1.00 | 0.99 | 8523 |
| 1 | 1.00 | 0.98 | 0.99 | 8524 |
| accuracy | | | 0.99 | 17047 |
| macro avg | 0.99 | 0.99 | 0.99 | 17047 |
| weighted avg | 0.99 | 0.99 | 0.99 | 17047 |




K-Nearest Neighbors (KNN): [18]KNN is an instance-based, non-parametric learning algorithm. The class of the K nearest neighbors in the feature space is used to determine new data points. The k hyperparameter's value is chosen by the user. The larger the value of K the smoother the decision boundaries, while smaller values of K can capture more local details in the data. It can handle imbalanced data by adjusting the number of nearest neighbors, which controls the decision.

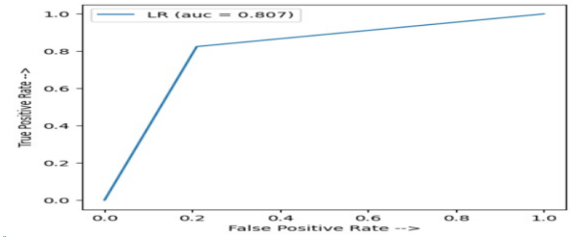
When we fed the balanced data to the random forest it gave an accuracy of 98% precision of 0.98 recall of .98,f1 score of 0.98 and AUC=0.981.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.98 | 0.98 | 0.98 | 8523 |
| 1 | 0.98 | 0.98 | 0.98 | 8524 |
| accuracy | | | 0.98 | 17047 |
| macro avg | 0.98 | 0.98 | 0.98 | 17047 |
| weighted avg | 0.98 | 0.98 | 0.98 | 17047 |



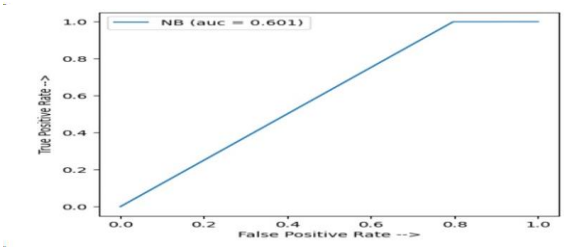

Logistic regression-[16] It is a supervised machine learning algorithm. This algorithm is mainly used for classification tasks where our main aim is to predict the probability that a given instance belongs to a given class. It is referred to as regression as its input is the linear regression function's output and probability for the given class is estimated by a sigmoid function. When we fed the balanced data to Logistic Regression it gave an accuracy of 81% precision of 0.80 recall of .82,f1 score of 0.81 and AUC=0.807.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.82 | 0.79 | 0.80 | 8523 |
| 1 | 0.80 | 0.82 | 0.81 | 8524 |
| accuracy | | | 0.81 | 17047 |
| macro avg | 0.81 | 0.81 | 0.81 | 17047 |
| weighted avg | 0.81 | 0.81 | 0.81 | 17047 |




Gaussian Naive Bayes: Gaussian Naive Bayes is a probabilistic-based algorithm[19] that is commonly used for classification tasks. It is a variant of the Naive Bayes algorithm that makes the assumption that the features are normally distributed, and it is called "naive" because it assumes that the features are independent of each other. It can be used for the classification of the imbalanced dataset, but the degree of imbalance in the data may affect its performance. When we fed the balanced data to Gaussian Naive Bayes it gave an accuracy of 60% precision of 0.56 recall of 1 ,f1 score of 0.71 and AUC=0.601.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 1.00 | 0.20 | 0.34 | 8523 |
| 1 | 0.56 | 1.00 | 0.71 | 8524 |
| accuracy | | | 0.60 | 17047 |
| macro avg | 0.78 | 0.60 | 0.53 | 17047 |
| weighted avg | 0.78 | 0.60 | 0.53 | 17047 |




Voting classifier (proposed model)-The voting classifier is an ensemble learning method that combines the predictions of multiple base models to make a final prediction. It is a type of ensemble learning technique in which improvement in the entire system is done by combining multiple models(classifiers). It works by aggregating the predictions of multiple base classifiers and combining them into a single prediction. The major reason for the usage of a Voting Classifier is to improve the overall accuracy and robustness of the model. In our model we combine random forest and KNN the two best performing models to achieve best accuracy out of all algorithms.

when we fed the balanced data to Logistic Regression it gave an accuracy of 99% precision of 1 recall of .98,f1 score of 0.99 and AUC=0.990.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.98 | 1.00 | 0.99 | 8523 |
| 1 | 1.00 | 0.98 | 0.99 | 8524 |
| accuracy | | | 0.99 | 17047 |
| macro avg | 0.99 | 0.99 | 0.99 | 17047 |
| weighted avg | 0.99 | 0.99 | 0.99 | 17047 |

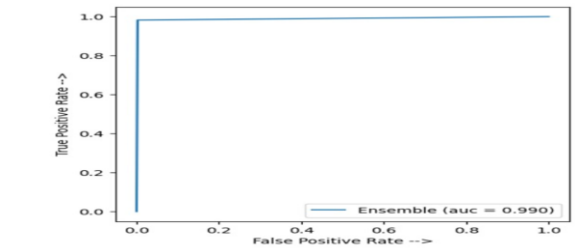



TABLE COMPARING DIFFERENT SAMPLING TECHNIQUES

| Balancing techniques | F1 Score | Accuracy | Macro Average | Dataset After Balancing |
|---|---|---|---|---|
| Under Sampling | 0.78 | 0.77 | 0.78 | 0-783<br>1-783 |
| Over Sampling | 0.83 | 0.82 | 0.83 | 0-42617<br>1-42617 |
| SMOTE | 0.83 | 0.82 | 0.83 | 0-42617<br>1-42617 |
| SMOTE TO MEK | 0.83 | 0.82 | 0.83 | 0-42312<br>1-42312 |
| SMOTE ENN | 0.89 | 0.87 | 0.88 | 0-36999<br>1-40302 |
| ADASYN | 0.84 | 0.83 | 0.84 | 0-42617<br>1-42832 |
| SVM SMOTE | 0.85 | 0.90 | 0.89 | 0-42617<br>1-27370 |
| Borderline SMOTE | 0.92 | 0.92 | 0.92 | 0-42617<br>1-42617 |
| SMOTEN | 0.93 | 0.92 | 0.93 | 0-42617<br>1-42617 |

TABLE COMPARING DIFFERENT ALGORITHMS

| Algorithms | Precision | Recall | F1 Score | Accuracy | AUC Score |
|---|---|---|---|---|---|
| ANN (Imbalanced) | 0.00 | 0.00 | 0.00 | 0.98 | 0.500 |
| ANN (Balanced) | 0.90 | 0.96 | 0.93 | 0.93 | 0.931 |
| SVM | 0.89 | 0.94 | 0.91 | 0.91 | 0.911 |
| Random Forest | 1.00 | 0.98 | 0.99 | 0.99 | 0.989 |
| KNN | 0.98 | 0.98 | 0.98 | 0.98 | 0.981 |
| Logistic Regression | 0.80 | 0.82 | 0.81 | 0.81 | 0.807 |
| Naive Bayes | 0.56 | 1.00 | 0.71 | 0.60 | 0.601 |
| Voting Classifier **(Proposed Model)** | 1.00 | 0.98 | 0.99 | 0.99 | 0.990 |

From the results obtained we can observe that SMOTEN is the best technique for sampling and voting classifier with random forest and KNN is giving the highest accuracy

**CONCLUSION**

In a world that is heavily dependent on data the problem of class Imbalance is one of the major problem especially in the field of medicines as it is difficult to deal with Imbalanced data using machine learning algorithms so by this research work the main aim is to tackle the challenge of handling Imbalanced datasets and provide a

useful and best solution to the problem so we propose a model that balance the Imbalanced dataset using SMOTEN technique which is a variation of SMOTE then after that we apply ensemble learning voting classifier on our two best performing algorithms that are Random Forest Classifier and KNN and this gives us the result with maximum accuracy of 99% with AUC score of 0.990.

REFERENCES


**[1]**Ozaltin O, Coskun O, Yeniay O, Subasi A. A Deep Learning Approach for Detecting Stroke from Brain CT Images Using OzNet. Bioengineering (Basel). 2022 Dec 8;9(12):783. doi: 10.3390/bioengineering9120783. PMID: 36550989; PMCID: PMC9774129.

**[2]**B. Akter, A. Rajbongshi, S. Sazzad, R. Shakil, J. Biswas and U. Sara, "A Machine Learning Approach to Detect the Brain Stroke Disease," 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2022, pp. 897-901, doi: 10.1109/ICSSIT53264.2022.9716345.

**[3]**Suresh T, Brijet Z, Subha TD. Imbalanced medical disease dataset classification using enhanced generative adversarial network. Comput Methods Biomech Biomed Engin. 2023 Oct-Dec;26(14):1702-1718. doi: 10.1080/10255842.2022.2134729. Epub 2022 Nov 2. PMID: 36322625.

**[4]**V. Krishna, J. Sasi Kiran, P. Prasada Rao, G. Charles Babu and G. John Babu, "Early Detection of Brain Stroke using Machine Learning Techniques," 2021 2nd International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2021, pp. 1489-1495, doi: 10.1109/ICOSEC51865.2021.9591840.

**[5]**S. Vasa, P. Borugadda and A. Koyyada, "A Machine Learning Model to Predict a Diagnosis of Brain Stroke," 2023 International Conference on Inventive Computation Technologies (ICICT), Lalitpur, Nepal, 2023, pp. 180-185, doi: 10.1109/ICICT57646.2023.1013419.

**[6]**S. U. Sabha, A. Assad, N. M. U. Din and M. R. Bhat, "Comparative Analysis of Oversampling Techniques on Small and Imbalanced Datasets Using Deep Learning," 2023 3rd International conference on Artificial Intelligence and Signal Processing (AISP), VIJAYAWADA, India, 2023, pp. 1-5, doi: 10.1109/AISP57993.2023.10134981.

**[7]**A. Srinivas, Joseph Prakash Mosiganti,A brain stroke detection model using soft voting based ensemble machine learning classifier,Measurement: Sensors Volume 29,2023,100871,ISSN 2665-9174,https://doi.org/10.1016/j.measen.2023.100871.

**[8]**M. Karim, M. F. Samad and F. Muntasir, "Improving Performance Factors of an Imbalanced Credit Risk Dataset Using SMOTE," 2022 4th International Conference on Electrical, Computer & Telecommunication Engineering (ICECTE), Rajshahi, Bangladesh, 2022, pp. 1-4, doi: 10.1109/ICECTE57896.2022.10114486.

**[9]**Haibo He, Yang Bai, E. A. Garcia and Shutao Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, 2008, pp. 1322-1328, doi: 10.1109/IJCNN.2008.4633969.

**[10]**L. Demidova and I. Klyueva, "SVM classification: Optimization with the SMOTE algorithm for the class imbalance problem," 2017 6th Mediterranean Conference on Embedded Computing (MECO), Bar, Montenegro, 2017, pp. 1-4, doi: 10.1109/MECO.2017.7977136.

**[11]**I. Dey and V. Pratap, "A Comparative Study of SMOTE, Borderline-SMOTE, and ADASYN Oversampling Techniques using Different Classifiers," 2023 3rd International Conference on Smart Data Intelligence (ICSMDI), Trichy, India, 2023, pp. 294-302, doi: 10.1109/ICSMDI57622.2023.00060.

**[12]**S. Rani, T. Ahmad and S. Masood, "Handling Class Imbalance Problem using Oversampling Techniques for Breast Cancer Prediction," 2023 International Conference on Recent Advances in Electrical, Electronics & Digital Healthcare Technologies (REEDCON), New Delhi, India, 2023, pp. 693-698, doi: 10.1109/REEDCON57544.2023.10150702.

**[13]**Q. P. Lau, W. Hsu, M. L. Lee, Y. Mao and L. Chen, "Prediction of Cerebral Aneurysm Rupture," 19th IEEE International Conference on Tools with Artificial Intelligence(ICTAI 2007), Patras, Greece, 2007, pp. 350-357, doi: 10.1109/ICTAI.2007.98.

**[14]**Uma R. Salunkhe, Suresh N. Mali,Classifier Ensemble Design for Imbalanced Data Classification: A Hybrid Approach, Procedia Computer Science,Volume 85,2016,Pages 725-732,ISSN 1877-0509,https://doi.org/10.1016/j.procs.2016.05.259.

**[15]**A. S. More and D. P. Rana, "Review of random forest classification techniques to resolve data imbalance," 2017 1st International Conference on Intelligent Systems and Information Management (ICISIM), Aurangabad, India, 2017, pp. 72-78, doi: 10.1109/ICISIM.2017.8122151.

**[16]**H. Luo, X. Pan, Q. Wang, S. Ye and Y. Qian, "Logistic Regression and Random Forest for Effective Imbalanced Classification," 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), Milwaukee, WI, USA, 2019, pp. 916-917, doi: 10.1109/COMPSAC.2019.00139.

**[17]**Sundar R., Punniyamoorthy M.,Performance enhanced Boosted SVM for Imbalanced datasets,Applied Soft Computing,Volume 83,2019,105601,ISSN 1568-4946,https://doi.org/10.1016/j.asoc.2019.105601.

**[18]**S. Liu, P. Zhu and S. Qin, "An Improved Weighted KNN Algorithm for Imbalanced Data Classification," 2018 IEEE 4th International Conference on Computer and Communications (ICCC), Chengdu, China, 2018, pp. 1814-1819, doi: 10.1109/CompComm.2018.8780580.

**[19]T**aeheung Kim, Jong-Seok Lee,Maximizing AUC to learn weighted naive Bayes for imbalanced data classification,Expert Systems with Applications,Volume 217,2023,119564,ISSN.