https://doi.org/10.48047/AFJBS.7.4.2025.24-30

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper Open Access

Enhancing Stroke Prediction with Imputation and SMOTE

[a]Priyamvatha Krishnadas, [b]Thottassery Divya Anandan, [c]Varun S Kumar, [d]Ani R

[*] Amrita School of Computing, Amrita Vishwa Vidyapeetham, Amritapuri Campus

[a] am.en.u4cse22045@am.students.amrita.edu, [b] am.en.u4cse22058@am.students.amrita.edu, [c] am.en.u4cse22060@am.students.amrita.edu, [d]anir@am.amrita.edu

Volume 7, Issue 4,Apr 2025

Received: 15 Feb 2025

Accepted: 05 Mar 2025

Published: 05 Apr 2025

doi:10.48047/AFJBS.7.4.2025.24-30

Abstract— In today's fast-paced world, strokes are one of the five most serious health threats in terms of annual deaths and long-term disabilities. Early diagnosis and treatment of the disease plays a vital role in decreasing mortality rate and improving patient's wellbeing by preventing permanent damage to the brain. Developing computational methods using machine learning and deep learning models helps in accurately predicting chances of stroke and enhancing early detection. Several studies have explored the application of AI on publicly available datasets; however, a major drawback is the imbalance in the dataset, where minority class instances are significantly lower than the majority class. This study introduces a novel approach to overcoming this limitation by employing SMOTE and imputation to improve model accuracy. It evaluates six machine learning and deep learning models for the early prediction of stroke. The dataset used in this study is a publicly available imbalanced one containing information like population characteristics, health history and lifestyle habit. The data used was preprocessed and then split into train data and test data for training the Machine Learning models - RandomForest, XGBoost, and Stack Ensembles, and Deep Learning models such as Multilayer Perceptron, Long Short-Term Memory, and TabNet. The performance was analyzed and compared using various performance metrics like accuracy, precision, recall, F1-score, ROC Curve etc. with accuracy having the highest preference. Among the machine learning models, Stack Ensembles is the most accurate one, having an accuracy of 96.2%, whereas MLP outperformed the other deep learning models, with an accuracy of 97.12%. Although the dataset was imbalanced, the achieved accuracy surpassed that of existing studies, making this a novel finding. This research highlights the crucial role of imputation and the Synthetic Minority Over-sampling Technique (SMOTE) in improving model accuracy for stroke prediction.

Index Terms— Stroke Prediction, Machine Learning, Deep Learning, SMOTE

I. NTRODUCTION

Stroke, a neurological disorder, is caused due to the shortage in the supply of blood to the brain caused by blockages and is one of the most prominent reasons for death worldwide. Namely, there are 2 types of strokes - Ischemic stroke and Haemorrhagic stroke. An ischemic stroke occurs when the circulation of blood to the brain is reduced or blocked which causes the brain tissue to get insufficient oxygen and nutrients leading to the death of the tissues in minutes. Ischemic stroke reports about 87% of all strokes making it a major area of focus. A Haemorrhagic stroke happens when a blood vessel in the brain bursts, leading to bleeding in the brain. The leaked blood builds up, putting pressure on the brain cells, causing damage.

The urgent need for early prediction of stroke is underscored by the World Health Organization (WHO, 2023), which reports that 11.6% of deaths globally are caused by stroke. Additionally, 5 million stroke survivors live with long-term disabilities, with 15 million new cases reported annually [7][6]. Studies show that patients who have already suffered from stroke have higher or possible chances of repeated strokes than those patients who never had it [25]. One piece of the survey reveals that about 30% of stroke patients may suffer from another stroke within a time span of five years. This is because vulnerability is increased by tissue and blood vessel damage, and risk factors including heart disease, diabetes, and hypertension continue even after a stroke.

Early detection[30] and accurate risk assessment are two very important factors for timely interventions yet early diagnosis and prediction of stroke still remains as a challenge due to factors such as class imbalance, missing and noisy data, feature complexity and model generalizability. Existing researches include studying various machine learning [8] and deep learning models in this area of study yet they fail in bringing higher accuracy rates because of the above mentioned limitations. To address these challenges, this study integrates Synthetic Minority Oversampling Techniques(SMOTE) to address the issue of class imbalance and improve stroke detection sensitivity while mean imputation is used to handle missing values. Machine Learning algorithms like Random Forest, XGBoost, Stack Ensemble, and Deep Learning Algorithms like MLP, LSTM and TabNet are used to train a predictive model on a publicly available imbalanced dataset containing key patient information, including hypertension, BMI, heart disease, average glucose level, smoking status, history of stroke, and age.

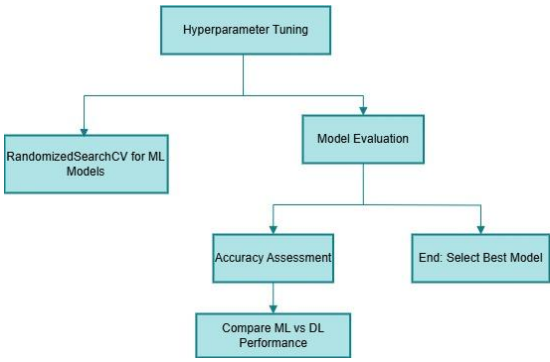
II. RELATED WORKS

The existing work in literature has explored different methods involved in predicting a stroke [3]. [7] Article also covers ML techniques for diagnosis of stroke and prediction of outcome. The paper aim in analyzing the risk factors for stroke prediction[12] by examining data from patients' Electronic Health Records (EHR). This will help identify the most critical factors for accurately predicting strokes. A case study conducted by Soumyabrata Dev provides valuable insights into these risk factors, building on previous work by Biswas et al. (2022) and Dritsas & Trigka (2022). Using a machine learning approach, we seek to enhance our understanding of stroke prediction. Example- Olamilekan Shobayao[5], the data of 5100 patients with stroke factors like hypertension, age, heart disease was collected to predict the stroke disease based on classification models like Logistic Regression, Decision Tree and Random Forest. The result of their experiment established that out of all classification models used to predict stroke, Random Forest yielded the highest accuracy. The study done by Yanqiu Ge[4][26] compares performance of Recurrent Neural Network (RNN) and Multi Layer Perceptron (MLP) in stroke prediction (Biswas et al., 2022; Dritsas & Trigka, 2022). Data of 13,930 patients were collected for the study. The MLP had three layers, and all layers except the input layer used a nonlinear activation function. Results showed that RNN performed better than MLP in predicting stroke for patients. Despite the advancements in stroke prediction using ML and DL models, existing studies suffer from class imbalance leading to poor sensitivity in identifying high-risk patients and lack a systematic way of evaluating the data preprocessing techniques like imputation and feature-selection. A study by Yagin et al. [31] explored the impact of handling class imbalance in stroke using SMOTE. Their work demonstrated that applying SMOTE to an imbalanced dataset improved their classification accuracy to 88.58% and reduced bias in model prediction. Using a Gradient Boosting Tree Classifier, they showed that SMOTE improved the sensitivity and specificity leading to more reliable stroke predictions. A recent study by Abiodun and Wreford[32] further explored the impact of SMOTE in stroke prediction by employing XGBoost and KNN in a stacked ensemble approach. Their result showed that applying SMOTE to balance the dataset improved model performance. The stacked model combining XGBoost and KNN achieved an accuracy of 97%. By tackling these research gaps, this study presents a robust and accurate result that aids in early diagnosis of stroke.

III. METHODOLOGIES

3.1 Data preprocessing and Label Encoding

The dataset used in this study was created by McKinsey and Company and obtained through Kaggle [1]. It consists of medical records from 5,110 individuals, aiming to predict stroke occurrence based on 12 key factors, including gender, age, hypertension, heart disease, marital status, work type, residence type, glucose levels, BMI, and smoking history.



The raw dataset initially contained 4,861 non-stroke cases and 249 stroke cases before applying any balancing techniques.

Preprocessing plays a crucial role in ensuring high-quality predictions, as missing values or noisy data can significantly impact model performance. To refine the dataset, we first performed data cleaning, which included removing duplicates, selecting relevant features, and categorizing data. Since the dataset contained categorical attributes such as gender, ever_married, residence type, glucose level category, and smoking status, these were transformed into numerical values using label encoding to facilitate effective model training and evaluation.

Additionally, we normalized the numerical features to scale their values between 0 and 1, followed by standardization to ensure a mean of 0 and a standard deviation of 1—a crucial step for maintaining consistency across different feature distributions [2]. The 'id' column was dropped as it held no predictive significance, and missing values in the 'bmi' feature were imputed using the mean value to maintain data integrity. Figure 3.1 illustrates the overall flow of our preprocessing methodology.

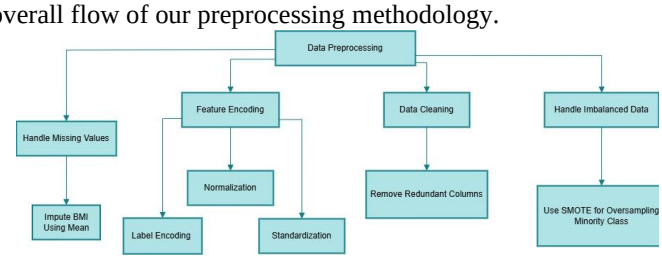


Figure 3.1 Data Preprocessing flowchart

3.2 Imbalanced data handling

In the proposed framework we have used SMOTE-Synthetic Minority Oversampling Technique to correct the imbalanced distribution among individuals in the target component 'stroke' which is the minority class. This target component was oversampled so as to have equal distribution of classes (Figure 3.2).

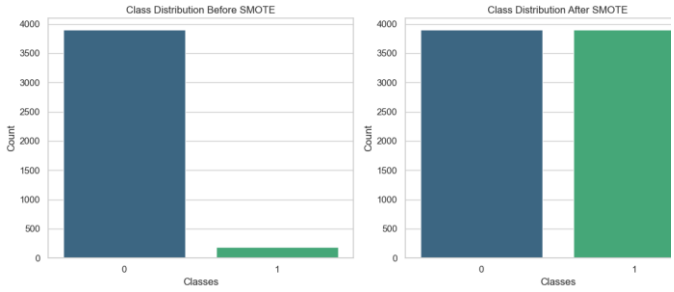


Figure 3.2 Class distribution after SMOTE

3.3 Hyperparameter tuning

Hyperparameter tuning is used to accelerate the performance and minimize errors of ML models. RandomizedSearchCV has been used along with cross-validation. In this way all combinations are checked and the values are passed in the dictionary and evaluated on the model for all combinations. This gives us the optimal accuracy for each of the combinations of hyperparameters [5]. A flowchart of how we used hyperparameter tuning for our study is shown below (see Figure 3.4).

3.4 Data-visualization

From the graph (Figure 4.4) we can infer that there is a higher count of female participants (0.0510%) having strokes than compared to male participants (0.0470%). The chart on the right in Figure 1 illustrates the distribution of strokes across various age groups, emphasizing that strokes are a significant issue among older adults, especially those aged 70–85 (0.1969%).

Figure 3.4 Hyperparameter tuning flowchart

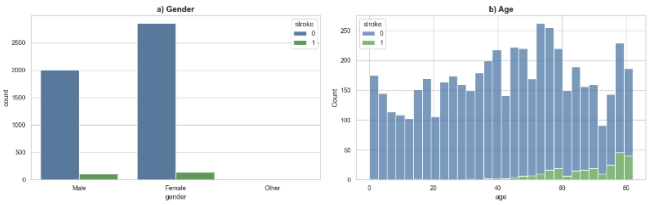


Figure 3.4: individual's distribution in each class with respect to gender

Figure 3.5 highlights how common heart disease is among people who have experienced a stroke. A majority of the participants do not have hypertension (0.0396%) but then those with hypertension (0.1325%) show a higher proportion of strokes. As in the case of heart disease we can infer that those with heart disease (0.1702%) has a higher proportion of strokes.

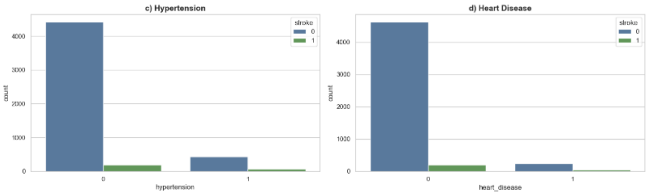


Figure 3.5: hypertension and heart disease distribution

Figure 3.6 illustrates the individual's background across two categories: their place of residence and their type of work. As we see, the percentage of stroke patients living in urban areas (0.0520%) is higher than that in rural areas (0.0453%). Also, participant's occupation is the majority of private(0.0793%) and nearly 40% of them had stroke.

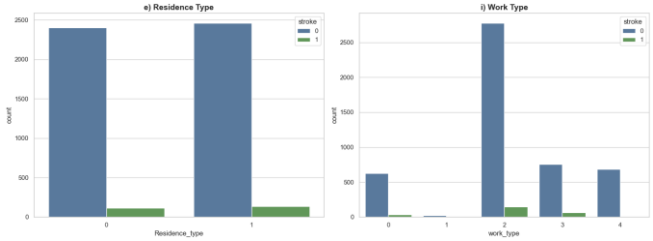


Figure 3.6: Residence type and work type distribution

Figure 3.7 depicts the individual's distribution among BMI where we see that there is a slight increase in the occurrence of stroke among the range 20-40 BMI. Towards the right of Figure 1, we're seeing the distribution of individuals on the basis of smoking habits. Participants who have never smoked (0.0470%) form the majority class but individuals with a history of smoking(0.0790%)or of those who have a habit of smoking even currently(0.0532%)have a notable proportion of

strokes.

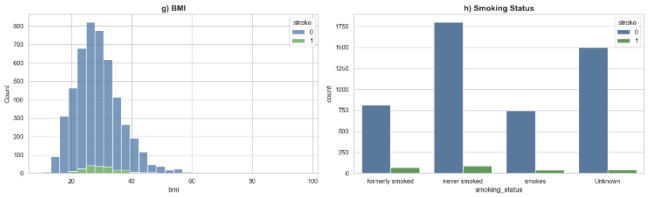


Figure 3.7:BMI and Smoking status distribution

Figure 3.8 displays how participants are distributed based on their average glucose levels and marital status.We can clearly infer that individuals with elevated glucose levels exhibit a higher incidence of stroke than those with lesser glucose levels. As in the case of ever_married we can notice that participants who have been married (0.0656%) form the larger group with a slightly higher proportion of strokes observed in this category compared to those never married (0.0165%).

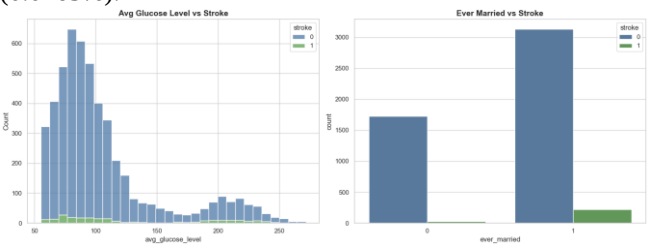


Figure 3.8: glucose level and married status distribution

3.5 Machine Learning classifiers

The performance is measured using 3 different classifiers, that is, Random Forest, XGBoost and Stack Ensembles (Biswas et al., 2022) which is highlighted in this session. The best model is determined by selecting the one with the highest accuracy among the three [2].

The methodology used in this study in order to improve the performance of all the machine learning and deep learning models is SMOTE (Synthetic Minority Over-sampling Technique) which is a data augmentation method used to handle data imbalance by generating synthetic samples of minority classes(stroke). SMOTE mainly improves minority class representation, enhances model accuracy and also prevents overfitting by identifying the k nearest neighbors of minority class. It then generates more synthetic examples by interpolating between the existing points. This generates more datapoints of the minority class making it a balanced dataset.

3.5.1 Random Forest

A Random Forest classifier works by combining multiple decision trees. Every individual tree is built using a unique subset of the data created through resampling. These trees independently perform classification or regression. The final prediction is determined by either averaging the results (for regression) or using majority voting (for classification).

3.5.2 XGBoost

XGBoost which is the abbreviation for Extreme Gradient Boosting" is a sophisticated gradient Boosting library used for efficient training of predictive models. XGBoost uses ensemble sequential learning to produce a stronger prediction by combining multiple weak models.

3.5.3 Stack Ensembles

Stacking is a powerful technique that combines the prediction of multiple base models to create more accurate final prediction. Each base model makes a prediction on the validation set which is further used as features for the meta model to predict the final output.

3.6 Deep Learning models

The performance is assessed using three models: Multi-Layer Perceptron (MLP), Long Short-Term Memory (LSTM), and TabNet, which are discussed in this section. The best model is determined based on the one that achieves the highest accuracy among the three.

3.6.1 Multi Layer Perceptron(MLP)

A Multi Layer Perceptron, commonly known as an MLP, is a feed-forward neural network consisting of an isolated layer of nodes that are interconnected by each other to ensure that the network can learn the complex relations of data. MLP has three main layers, the Input layer, hidden layer and output layer. The activation function is the function that allows the network to capture the non-linear relationships in the data.

3.6.2 TabNet

TabNet is a powerful Deep Learning model designed for tabular data. By using Sequential Attention Mechanism the model is able to handle complex relationships within the tabular data.The input data is first passed through feature transformer to extract relevant features.The attentive transformer further selects subsets of features at each decisive step and is masked to prevent information leak.The decoder processes the selected features and generates the final prediction

3.7 Evaluation Metrics

For the evaluation of the predictions made by the trained models, we need to consider various measures like Accuracy, Precision, Recall and F1-score. Recall is defined as the fraction of individuals who have a stroke and are correctly classified as positive considering the total number of participants who are positive. The above two Evaluation metrics are used to find the errors of a model while dealing with a dataset that is not balanced. Precision indicates how many people among the total of those who suffered from a stroke belong to the class 'stroke'. Meanwhile recall tells us the percentage of those who suffered from a stroke which I s predicted accurately. F-measure is a balance between precision and recall which sums up the predictive performance of a model. The formulas to calculate the same are given below:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + FP} \qquad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{Total}}$$

Note that TP stands for True Positive i.e. patients who were correctly identified as stroke patients, TN stands for True Negative i.e. patients who were rightly classified as patients without stroke, FP, that is, False Positive give us insights into the number of those patients who were wrongly identified to be having a stroke but are actually healthy patients, FN or False Negative i.e. patients who were wrongly identified as healthy patients but are actually stroke patients.

We are also using curves like AUC(Area under the curve) and ROC (Receiver Operating Characteristic) Curves as evaluation metrics where AUC provides gives numeric values in the range [0,1],that is, 1 for a perfect classifier and 0 for a classifier that classified all non strokes as strokes, implying how accurately the model is able to categorize between cases where as ROC Curves provide a visual insight of the performance of a classifier. These evaluation metrics are often considered while dealing with imbalanced datasets.

Another important evaluation metric used is logloss (logarithmic loss) also known as binary cross entropy loss which is mostly used for classification models which quantifies the difference in value between the predicted probabilities and the actual class labels.

IV. RESULTS AND DISCUSSION

With the help of visual representations, we can analyze and observe the dataset in data visualization. Correlation matrix calculates how two variables are related. Age is highly correlated with ever_married which is most positive relation (Figure 4.1).

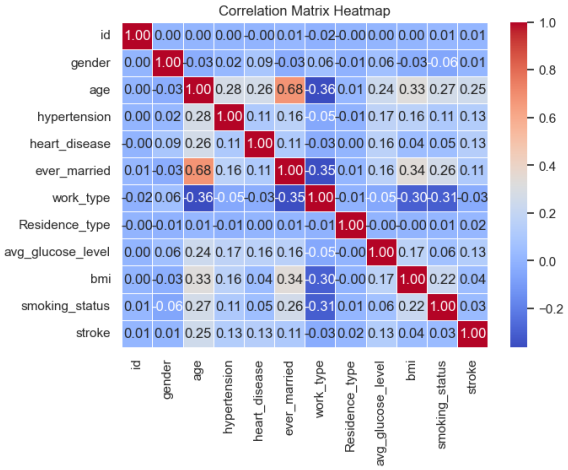


Figure 4.1 Correlation Matrix

4.1 Machine Learning Models

4.1.1 Random forest

The following is the confusion matrix of random forest-

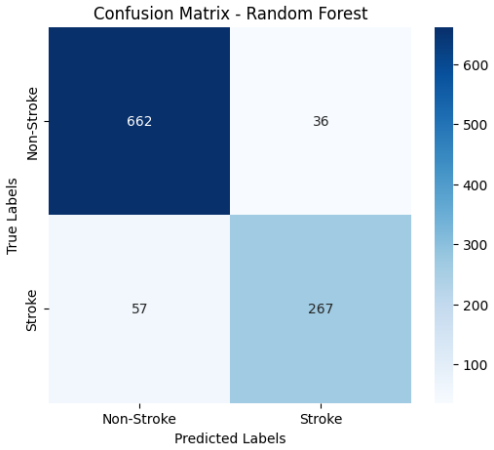


Figure 4.1.1

Accuracy: 91.0% -The model has a good accuracy
Precision: 88.0% - Nearly 88% cases were correct reducing false positives
Recall: 82.5% -The model correctly identifies 82.5% of the actual positive instances
F1 Score: 85.1% -Balance between recall and precision, ensuring the model is reliable
AUC: 0.935 -The model has a good discriminatory power.

4.2.2 XGBoost

The following is the confusion matrix of XGBoost-

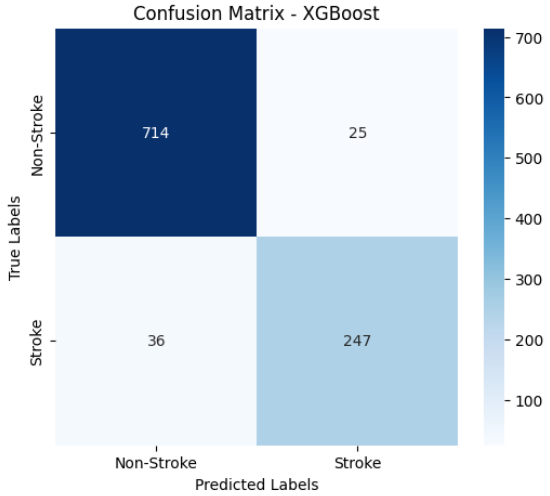


Figure 4.1.2: Confusion matrix of XGBoost model

Accuracy: 94.0% -The model has a good accuracy, higher than Random Forest
Precision: 90.8% -Higher precision means lower false positives
Recall: 87.3% -The model correctly identifies 87.3% of the actual positive instances
F1 Score: 89.0% -Strong balance of precision and recall
AUC: 0.960 -The model has a good discriminatory power, better than Random Forest

4.2.3 Stack Ensembles

Accuracy: 96.2% - Highest Accuracy compared to all other models

Precision: 92.5% - Out of the predicted strokes, 92.5% were correct.
Recall: 91.0% -The model correctly identifies 91% of the actual positive instance in the database
F1 Score: 91.7% - The Model is performing well
AUC: 0.980 - The model has the best discriminatory power compared to the other 2 model

Methods	Random Forest	XGBoost	Stack Ensembles
Precision	88.0%	90.8%	92.5%
Recall	82.5%	87.3%	91.0%
F-Measure	85.1%	89.0%	91.7%
Accuracy	91.0%	94.0%	96.2%
AUC	0.935	0.960	0.980

Table 1: Comparison of Machine Learning Models

Thus from all the evaluation metrics considered we can infer that among the three machine learning[28] models (Table 1), Stack Ensembles[27] is the most accurate model (accuracy - 96.2%) for Stroke Prediction as it combines the strengths of both Random Forest as well as XGBoost.

4.2 Deep Learning Models
4.2.1 Multilayer Perceptron (MLP)
Accuracy: 97.12% - The model is highly accurate
Recall: 85.64% -The model correctly identifies 85.64% of the actual positive instance in the database
Precision: 98.21% - Implies that there are few false positives
F1 Score: 91.42% - A balanced model with good balance between precision and recall
AUC: 0.9017 -The model has a good discriminatory power
Log Loss: 15.27 -The models predictions are almost accurate as it is indicated by low log loss

The following is the confusion matrix for MLP:
Confusion Matrix:

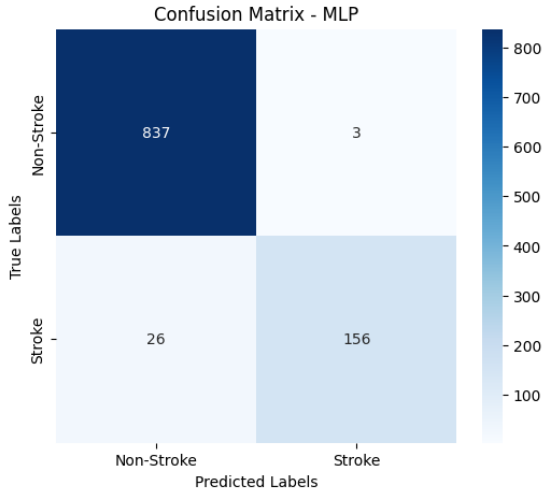


Figure 4.2.1: Confusion matrix of MLP model

4.2.2 Long Short Term Memory (LSTM)

The following is the confusion matrix of Long Short Term Memory

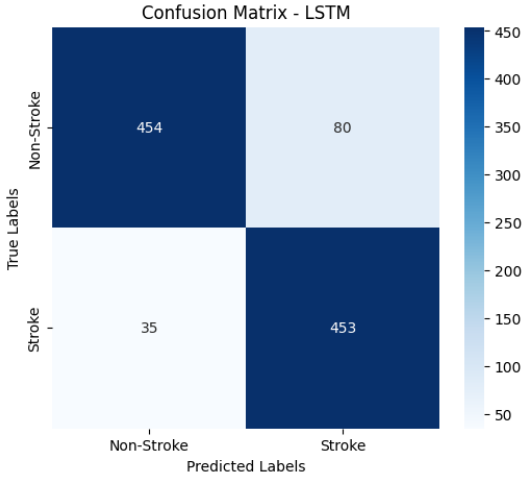


Figure 4.2.3: Confusion matrix of LSTM model

Accuracy: 88.74% - Less accuracy compared to the other MLP
Precision: 84.93% -There are False positives as the precision is not 1
Recall: 92.91% -The model correctly identifies 92.91% of the actual positive instance in the database
AUC: 0.9383 - The model has a good discriminatory power
F1 Score: 88.75% - The model maintains a descent balance between precision and recall
Log Loss: 34.79% -The models predictions are not as accurate as MLP

4.2.3 TabNet
The following is the confusion matrix of TabNet:

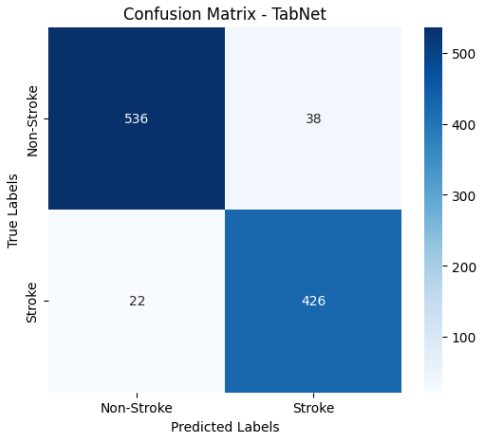


Figure 4.2.4: Confusion matrix of TabNet model

Accuracy: 94.18% -The model is reasonably accurate
Precision: 91.84% -There are False positives as the precision is not 1
Recall: 95.12% -The model correctly identifies 95.12% of the actual positive instance in the database
AUC : 0.7536 - Not a good discriminator
Log Loss: 79.83% - A higher log loss indicates less confidence in predictions
F1 Score: 93.47% - A good balance between precision and recall

Methods	MLP	LSTM	TabNet
Accuracy	97.12%	88.74%	94.18%
Log Loss	15.27%	34.79%	79.83%
Precision	98.21%	84.93%	91.84%
Recall	85.64%	92.91%	95.12%
AUC	0.9017	0.9382	0.7536
F1 Score	91.42%	88.75%	93.47%

Table 2: Comparison of Deep Learning Models

Among the three Deep Learning models (Table 2) considered, Multi Layer Perceptron (MLP) has the highest accuracy of 97.34% with a precision value 1. In contrast Long Short Term Memory (LSTM) has the least accuracy as LSTM is a more efficient choice for sequential data.

As we can see in Table 3, on training the machine learning models with the original dataset we get a lower accuracy as compared to the model being trained on the balanced dataset which is attained using SMOTE. Thus we can conclude that applying SMOTE on imbalanced dataset can help improve the efficiency of the model ,with Stack Ensembles being the most efficient model having the highest accuracy of 96.2%. Even in case of Deep Learning models, Figure 4, we can see a similar trend as balancing the data with SMOTE improved the accuracy of all the Deep Learning models ,with MLP being the most efficient model having an accuracy of 97.34%.

Machine Learning Methods	Without SMOTE	With SMOTE
Random Forest	88.6%	91.0%
XGBoost	92.1%	94.0%
Stack Ensembles	93.4%	96.2%

Table 3: Comparison of accuracy of Machine Learning Models without and with using SMOTE

Deep Learning Methods	Without SMOTE	With SMOTE
MLP	94.51	97.12%
LSTM	83.34	88.74%
TabNet	92.08	94.18%

Table 4: Comparison of accuracy of Deep Learning Models without and with using SMOTE

REFERENCES

1. Kaggle. (n.d.). Stroke prediction dataset. Retrieved from https://www.kaggle.com
2. Biswas, N., Uddin, K. M. M., Rikta, S. T., & Dey, S. K. (2022). A comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach. *Healthcare Analytics, 2*, 100032. https://doi.org/10.1016/j.health.2022.100032
3. Dritsas, E., & Trigka, M. (2022). Stroke risk prediction with machine learning techniques. *Sensors, 22(13)*, 4670. https://doi.org/10.3390/s22134670
4. Viswanatha, V., Ramachandra, A. C., Parameshachari, B. D., & Sharma, A. K. (2023). Brain stroke prediction using decision tree algorithm. *2023 IEEE 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 1–6. https://doi.org/10.1109/ICACCS55779.2023.10137451
5. Kaur, A., Kaushal, C., Hassan, M. M., & Aung, S. T. (2024). *Federated deep learning for healthcare: A practical guide with challenges and opportunities.* CRC Press. https://doi.org/10.1201/9781032694870
6. World Health Organization (WHO). (2023). The top 10 causes of death. Retrieved from https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death
7. PMC. (n.d.). Stroke-related research articles. Retrieved from https://pmc.ncbi.nlm.nih.gov
8. Nature. (n.d.). Machine learning and stroke detection research. Retrieved from https://www.nature.com
9. Dev, S., Wang, H., Nwosu, C. S., Jain, N., Veeravalli, B., & John, D. (2022). A predictive analytics approach for stroke prediction using machine learning and neural networks. *Healthcare Analytics, 2*, Article 100032. https://doi.org/10.1016/j.health.2022.100032
10. Chin, C. L., Lin, B. J., Wu, G. R., Weng, T. C., Yang, C. S., Su, R. C., & Pan, Y. J. (2017). An automated early ischemic stroke detection system using CNN deep learning algorithm. *Proceedings of the IEEE 8th International Conference on Awareness Science and Technology (iCAST)*, 8–10 November, Taichung, Taiwan.
11. Uppal, M., Gupta, D., Juneja, S., Gadekallu, T. R., El Bayoumy, I., Hussain, J., & Lee, S. W. (2023). Enhancing accuracy in brain stroke detection: Multi-layer perceptron with Adadelta, RMSProp, and AdaMax optimizers. *Frontiers in Bioengineering and Biotechnology, 11*, 1257591. https://doi.org/10.3389/fbioe.2023.1257591
12. Dristsas, E., & Trika, M. (2022). Stroke risk prediction with machine learning techniques. *Sensors, 22(13)*, 4670. https://doi.org/10.3390/s22134670
13. Shobayo, O., Zachariah, O., Odusami, M. O., & Ogunleye, B. (2023). Prediction of stroke disease with demographic and behavioural data using random forest algorithm. *Analytics, 2*, 604–617.
14. Ge, Y., Wang, Q., Wang, L., Wu, H., Peng, C., Wang, J., ... Yi, Y. (2022). Predicting post-stroke pneumonia using deep neural network approaches. *Elsevier, 1386*, 5056.
15. Vyas, S., Upadhyaya, A., Bhargava, D., & Kumar Shukla, V. (2023). *Edge-AI in healthcare: Trends and future perspectives.* CRC Press. https://doi.org/10.1201/9781003244592
16. Sharma, V. S., Mahajan, S., Nayyar, A., & Kant Pandit, A. (2024). *Deep learning in engineering, energy and finance: Principles and applications.* CRC Press. https://doi.org/10.1201/9781003564874
17. Pillai, A. S., & Menon, B. (2024). *Machine learning and deep learning in neuroimaging data analysis.* CRC Press. https://doi.org/10.1201/9781003264767
18. Gupta, A., Verma, H., Prasad, M., Singh Kirar, J., & Lin, C. (2023). *Computational intelligence aided systems for healthcare domain.* CRC Press. https://doi.org/10.1201/9781003368342
19. Al-Jumaily, A., Crippa, P., Mansour, A., & Turchetti, C. (2024). *Non-invasive health systems based on advanced biomedical signal and image processing.* CRC Press. https://doi.org/10.1201/9781003346678
20. Kumar Swarnkar, S., Sharma, A., Somasekar, J., & Bhushan, B. (2024). *Machine learning in multimedia:*

Unlocking the power of visual and auditory intelligence. CRC Press. https://doi.org/10.1201/9781003477280
21. Sharmila, V., Kannadhasan, S., Kannan, A. R., Sivakumar, P., & Vennila, V. (2024). *Challenges in information, communication and computing technology: Proceedings of the 2nd International Conference on Challenges in Information, Communication, and Computing Technology (ICCICCT 2024)*. CRC Press. https://doi.org/10.1201/9781003559085
22. Dev, S., Wang, H., Nwosu, C. S., Jain, N., Veeravalli, B., & John, D. (2022). A predictive analytics approach for stroke prediction using machine learning and neural networks. *Healthcare Analytics, 2*, Article 100032. https://doi.org/10.1016/j.health.2022.100032
23. Al-Jumaily, A., Crippa, P., Mansour, A., & Turchetti, C. (2024). *Non-invasive health systems based on advanced biomedical signal and image processing.* CRC Press. https://doi.org/10.1201/9781003346678
24. Uppal, M., Gupta, D., Juneja, S., Gadekallu, T. R., El Bayoumy, I., Hussain, J., & Lee, S. W. (2023). Enhancing accuracy in brain stroke detection: Multi-layer perceptron with Adadelta, RMSProp, and AdaMax optimizers. *Frontiers in Bioengineering and Biotechnology, 11*, 1257591. https://doi.org/10.3389/fbioe.2023.1257591
25. Akshara, A. Sajeev and A. T, "Cerebral Stroke Risk Assessment: A Comparative Study of Machine Learning Techniques," *2024 9th International Conference on Communication and Electronics Systems (ICCES)*, Coimbatore, India, 2024, pp. 2020-2026, doi: 10.1109/ICCES63552.2024.10859791.
26. J. K and N. K. Prakash, "Prediction of Brain Stroke using Machine Learning with Relief Algorithm," *2022 International Conference on Edge Computing and Applications (ICECAA)*, Tamilnadu, India, 2022, pp. 1255-1260, doi: 10.1109/ICECAA55415.2022.9936142.
27. S. C. Patra, B. U. Maheswari and P. B. Pati, "Forecasting Coronary Heart Disease Risk With a 2-Step Hybrid Ensemble Learning Method and Forward Feature Selection Algorithm," in *IEEE Access*, vol. 11, pp. 136758-136769, 2023, doi: 10.1109/ACCESS.2023.3338369.
28. J. G. Sukumar, M. S. Ram Reddy, N. Sambangi, S. Abhishek, and A. T, ”Enhancing salary projections: a supervised machine learning approach with flask deployment,” in 2023 5th International Conference on Inven tive Research in Computing Applications (ICIRCA), Coimbatore, India, 2023, pp. 693-700. doi: 10.1109/ICIRCA57980.2023.10220707.
29. A. Suresh, R. S. Abhishek, P. Akash, B. Anudev and K. H. Vishakh, ”Functional Electrical Stimulation with Machine Learning for Stroke Prediction and Rehabilitation,” 2024 International Conference on E-mobility, Power Control and Smart Systems (ICEMPS), Thiruvananthapuram, India, 2024, pp. 01-08, doi: 10.1109/ICEMPS60684.2024.10559321.
30. C. Dev, K. Kumar, A. Palathil, T. Anjali, and V. Panicker, ”Machine Learning Based Approach for Detection of Lung Cancer in DICOM CT Image,” in Ambient Communications and Computer Systems, Y. C. Hu, S. Tiwari, K. Mishra, and M. Trivedi, Eds. Singapore: Springer, 2019, vol. 904, pp. 157–166. doi: 10.1007/978-981-13-5934-7 15.
31. Yagin, Fatma Hilal, Ipek Balikci Cicek, and Zeynep Kucukakcali. "Classification of stroke with gradient boosting tree using smote-based oversampling method." *Medicine* 10.4 (2021): 1510-5.
32. Abiodun, Oladunjoye John, and Andrew Ishaku Wreford. "Stroke Prediction Using Smote for Data Balancing, XGBoost and KNN Ensemble Algorithms." *Journal of Applied Physical Science International* 15.1 (2023): 42-53.