An Enhanced Stroke Prediction Scheme Using SMOTE and Machine Learning Techniques

Ferdib-Al-Islam[1] and Mounita Ghosh[2]
[1]*Department of Computer Science and Engineering, Northern University of Business and Technology Khulna*
[2]*Department of Biomedical Engineering, Khulna University of Engineering & Technology*
Khulna, Bangladesh

ferdib.bsmrstu@gmail.com and mounitamouni22@gmail.com

2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT) | 978-1-7281-8595-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCCNT51525.2021.9579648

***Abstract*—Stroke is the second driving reason for death worldwide, answerable for around 11% of all out passings. Throughout the long term, scientists are attempting to connect up various elements to the onset of stroke. Early awareness of various threat issues of stroke can limit the chance of stroke. The prediction of stroke is essential to counter health damage or passing. In this research, machine learning has been utilized to predict stroke inpatients. A popular oversampling method called SMOTE with several machine learning classifiers (Logistic Regression, Random Forest, and XGBoost) has been applied to the dataset to predict the consequence. The random forest model achieved better performance among these algorithms with 99.07% of accuracy, 99.0% of precision and recall. The feature importance scores have been shown to understand the feature's impact on the model development. The proposed model outperformed the existing works with higher accuracy.**

Keywords—Stroke, Exploratory Data Analysis, SMOTE, Machine Learning, Feature Importance.

I. INTRODUCTION

A stroke is known as a cerebrovascular injury which occurs when a portion of the brain's blood flow is cut off. Stroke is one of the supreme risky conditions for individuals above the oldness of 65. It injures the brain in the same way as a "heart attack" damages the heart and is the 2nd leading cause of demise in both developed and developing countries. When a stroke condition happens, it not only charges a lot of currency on clinic treatment and causes permanent injury, but it can also lead to death. A stroke kills someone every 4 minutes, but up to 80% of strokes can be avoided if we can recognize or forecast the onset of stroke in its primary stages. A stroke happens when the blood flow to the brain is cut off or decreased. A stroke denies a person's brain oxygen and nutrients, which can result in brain cell death. Frequent studies have been conducted to compare the success of predictive data mining systems to forecast different diseases [1].

In recent years, cerebral stroke has emerged as a major global public health concern. The optimal solution to this problem is to avoid it from happening in the first place by monitoring associated metabolic variables. However, until clinical signs are irregular, it is difficult for medical staff to determine if additional procedures are required for a prospective patient solely dependent on monitoring [2]. Most healthcare professionals use the word "stroke" to describe damage to the brain and spinal cord caused by irregularities in blood flow. Stroke's definition is projected through various perspectives; nevertheless, stroke elicits an overt emotional reaction on a global scale. 100 billion neurons and glia, rolled into more than three pounds of tissue, which incorporates each memory, encrypts, and retains it in a network. Every person's breath and expression are supported by brain activity. For more than five decades, the percentage of people who die as a result of a stroke has been ten times higher in developed nations [3].

Precisely the prediction of stroke results from a collection of predictive variables can help to distinguish high-danger patients and direct cure methods, resulting in lower morbidity. Several models are useful for identifying and validating predictive variables. Machine learning algorithms, on the other hand, provide an alternative, especially for large-scale multi-institutional data, with the added benefit of quickly integrating recently accessible data to increase prediction efficiency [4]. Long-term outcome estimation in stroke patients can be helpful in cure decisions as well as in controlling prognostic expectations. For this reason, several prognostic scoring systems have been developed. The application of the technique in the medical field has generated impressive findings in light of recent developments in machine learning. In certain cases, machine learning algorithms have proved to be better at describing the dynamic and uncertain essence of human physiology [5].

In this study, exploratory data analysis has been conducted on the dataset, and various machine learning algorithms (Logistic Regression, Random Forest, and XGBoost) have been applied to determine whether the patient will be affected by stroke or not. The feature importance scores have been demonstrated to realize the effect of features in the development of the machine learning model.

The remaining parts of the paper have been organized as follows - the "Related Work" presents the latest researches in predicting and classifying stroke with machine learning and other approaches. The explanation of the implementation of this research has been explained in the "Methodology" in several subsections. The result of this study has been represented in "Result and Discussion". The "Conclusion" expresses the conclusion of the paper.

II. RELATED WORK

Singh et. al. [1] compared various approaches for stroke prediction with their methodology on the "Cardiovascular Health Study (CHS)" dataset. To construct a classification model, the decision tree was applied for feature collection, the PCA algorithm was utilized for dimension reduction as well as the backpropagation neural network algorithm was used. The analysis had the best prognostic model for stroke disease with 97.7% of accuracy after studying and

combining classification efficiencies with various approaches and heterogeneity models. A hybrid machine learning (ML) approach was exhibited by Liu et al. [2] for predicting cerebral stroke for clinical prognosis based on insufficient physiological evidence and class imbalance measures. The procedure included two steps which were evolved with the whole process. To begin, random forest regression was used to assign missing values before the classification. Secondly, on an imbalanced dataset, an automatic hyperparameter optimization (AutoHPO) built on a deep neural network (DNN) was utilized to predict stroke. The medical dataset holds 43400 records of possible patients, including 783 stroke cases. The prediction strategy had a false negative rate of just 19.1%, which had decreased by an average of 51.5% as compared to other conventional approaches. The suggested solution predicted a false positive rate, accuracy, as well as sensitivity of 33.1%, 71.6%, as well as 67.4%, respectively. Govindarajan et al. [3] proposed an idea to extract patients' signs from case sheets and train the machine with the resulting data. The case sheets of 507 patients were obtained. Artificial neural networks trained with a stochastic gradient descent algorithm achieved 95% of classification accuracy and a 14.69 standard deviation.

Asadi et al. [4] executed a systematic analysis of a prospectively compiled record of acute ischaemic stroke treated with the endovascular intervention was carried out. SPSS, MATLAB, and Rapidminer were used to perform conventional statistics and ANN analysis. Support vector machine algorithms were used to create a supervised system proficient in classifying these predictors as potentially positive or bad outcomes. Using randomly divided data, these algorithms were learned, validated, and evaluated. They included 107 patients who had endovascular treatment for acute anterior circulation ischaemic stroke. 66 men were present, with an average oldness of 65.3. The models incorporated all accessible socioeconomic, practical, and clinical considerations. The neural network's final confusion matrix showed an average congruency of 80% between the goal and output groups, with favorable receiving operative characteristics. However, after optimization, the help vector machine performed somewhat better, with a root mean squared error of 2.064. Heo et al. [5] proposed a retrospective study that recruited patients with acute ischemic stroke from a prospective cohort. At 3 months, a favorable outcome was described as an adjusted Rankin Scale score of 0, 1, or 2. They developed and compared the predictability among DNN, random forest, and logistic regression. Letham et al. [6] introduced Bayesian Rule which listed a generative model that generated a posterior distribution over potential decision lists. The study demonstrated that Bayesian Rule Lists had a high level of predictive accuracy. On par with the latest top machine learning prediction algorithms, the approach was inspired by recent advances in precision medicine and it was able to be used to create highly precise and interpretable patient ($CHADS_2$ score) scoring systems. Emon et al. [7] had classified stroke patients by applying several machine learning techniques as Logistics Regression, SGD, Decision Tree, AdaBoost, KNN, Gradient Boosting, and XGBoost. In that work, the highest accuracy of 97% had been obtained by the weighted voting classifier.

After reviewing the mentioned works, it has been found that these can be enhanced in different performance metrics. The feature importance score calculation was also missing in the previous works. The proposed work in this paper intends to fill these gaps.

III. METHODOLOGY

The graphical illustration of the proposed scheme has been represented in Fig. 1, which contains several data-preprocessing steps, exploratory data analysis, data splitting, model training, and testing phase, and model performance evaluation steps. The methodology of this work has been divided into the subsequent parts –

i. Data Description and Pre-processing
ii. Exploratory Data Analysis
iii. SMOTE for Handing Imbalanced Data
iv. Machine Learning for Classification

A. Data Description and Preprocessing

The dataset that has been utilized in this study was an open-source dataset, available in Kaggle [8]. Emon et al. [7] collected this dataset from a medical hospital in Bangladesh. The dataset contained 5110 instances of 12 columns – "id", "gender", "age", "hypertension", "heart_disease", "ever_married", "work_type", "Residence_type", "avg_glucose_level", "bmi", "smoking_status", and "stroke". The description of each column of the dataset has been depicted in Table 1.

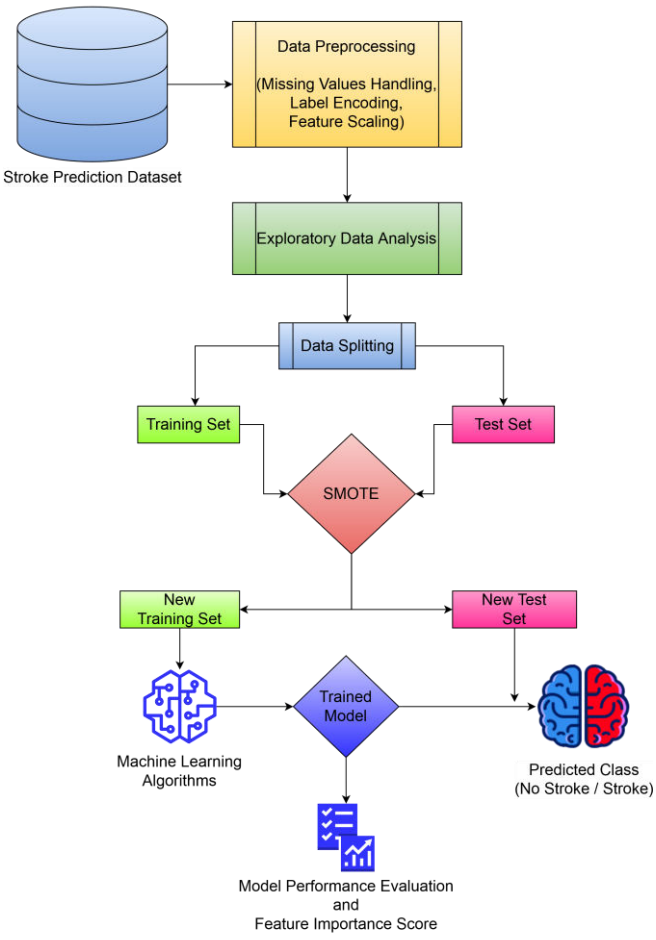

Fig. 1. The architecture of the proposed system.

TABLE I. DATASET DESCRIPTION

Column	Description
id	Patient unique identifier
gender	"Male", "Female" or "Other"
age	The age of the patient
hypertension	1 = The patient has hypertension 0 = The patient doesn't have hypertension
heart_disease	1 = The patient has heart disease(s) 0 = The patient doesn't have any heart disease
ever_married	Marital status of the patient - "Yes" or "No"
work_type	"children", "Govt_job", "Never_worked", "Private" or "Self-employed"
Residence_type	"Urban" or "Rural"
avg_glucose_level	The average level of glucose in the blood
bmi	"Body Mass Index" of the patient
smoking_status	"formerly smoked", "never smoked", "smokes" or "Unknown"
stroke	The target variable. 1 = The patient had a stroke 0 = The patient didn't have a stroke

There were missing values in the dataset. There are several missing value handling techniques are available. In this research, the "Mean Value Imputation" technique has been applied using the "Simple Imputer" module of Scikit-learn. However, the column "id" had no significance in this study, so it has been eliminated from the dataset while doing the preprocessing tasks. Numerical data is required for almost every machine learning algorithm. This means, when the dataset contains categorical data, it should be initially encoded into a numerical presentation. Label encoding is one of the widely used concepts of converting categorical labels into numeric forms. It has been done for the categorical features. Feature scaling is an approach to scale all the features to the same scale. In this research, min-max scaling or normalization has been performed. It is a scaling principle where the values are rescaled in the range of 0 and 1. The formula for computing normalization has been given in (1):

$$F' = \frac{F - F_{\min}}{F_{\max} - F_{\min}} \quad (1)$$

where F_{max} and F_{min} are the top and the bottom values of the feature correspondingly.

B. *Exploratory Data Analysis*

Exploratory data analysis is an important assessment intended to reveal the hidden structure of a data index and to find patterns, trends, and associations that are not promptly clear [9]. The histogram is applied to show the pattern and distribution of the categorical variables, which are represented in Fig. 2 – Fig. 6. In Fig. 2, the histogram of the target variable "stroke" represents the class imbalance, where "No Stroke" is the majority and "Stroke" is the minority class. From Fig. 3, it is can be understood that the female patient suffers stroke more than male and "other" gender. Fig. 4 represents that among the work type, people who are engaged in private jobs face stroke more than others. People of the urban areas suffer from stroke more than rural areas, according to the representation of Fig. 5. The relation between the smoking status and stoke has been illustrated in Fig. 6. People who never did smoking are mostly in the "No Stroke" class. The correlation coefficient is needed to find out the correlation between the two features/variables. Correlation has control over feature significance. As two features/variables are related, a variation in one will create variation in another. So there is no reason to keep each of them.

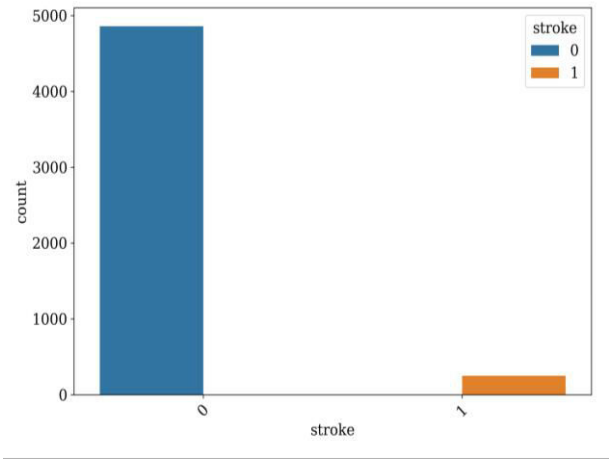

Fig. 2. Target variable's class distribution.

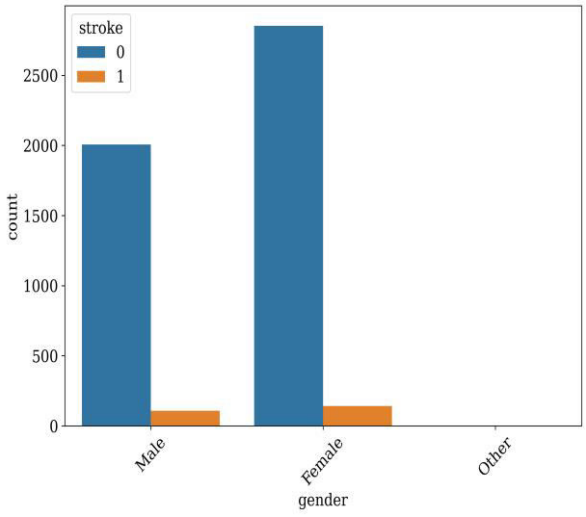

Fig. 3. "gender" distribution according to target variable "stroke".

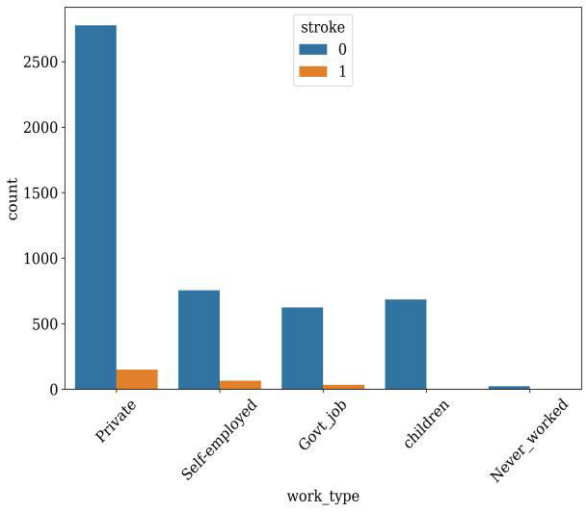

Fig. 4. "work_type" distribution according to "stroke".

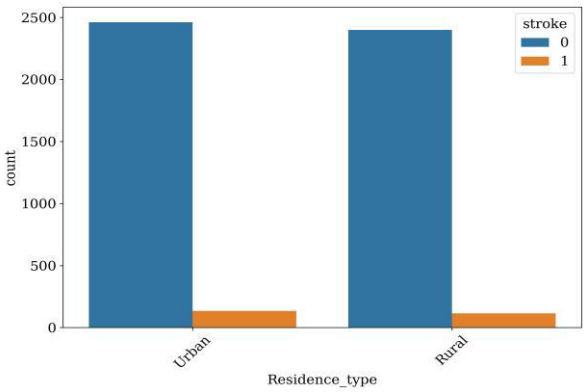

Fig. 5. "Residence_type" distribution according to "stroke".

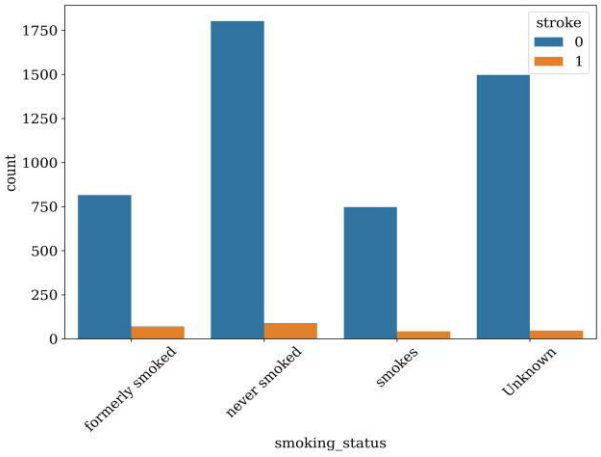

Fig. 6. "smoking_status" distribution according to "stroke".

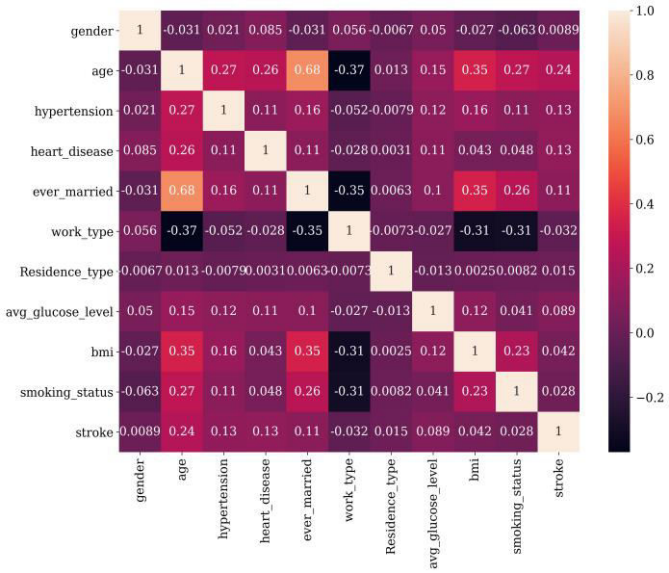

Fig. 7. Correlation among the variables.

TABLE II. SMOTE IMPACT ON DATASET

Before Using SMOTE		After Using SMOTE	
No. of Training Data	*No. of Test Data*	*No. of Training Data*	*No. of Test Data*
4088	1022	7782	1940

TABLE III. LOGISTIC REGRESSION CLASSIFIER'S PARAMETERS

Parameter Name	Chosen Value
"C"	0.01
"penalty"	"l2"
"solver"	"lbfgs"

Fig. 7 demonstrates the correlation among the variables. There is no strong correlation among the input variables and no feature is strongly correlated with the target variable. So, no variable was eliminated in this study.

C. *SMOTE for Handing Imbalanced Data*

An imbalanced classification concern is an illustration of a classification concern where the distribution of models across the known classes is biased or slanted. This consequence in models that have low performance, obviously for the minority class. This is a concern as, the minority class has more significance, and hence the concern is more sensitive to classification errors for the minority class than the majority class. The dataset which has been utilized in this research was imbalanced. There are several strategies to counter the imbalanced class issue in the dataset. In this work, "Synthetic Minority Oversampling Technique" was used [10]. SMOTE is an oversampling strategy where the synthetic data are produced for the minority class [11]. This calculation assists in overcoming the over-fitting issue presented by random oversampling. It centers on the feature space to create new occurrences with the assistance of interjection between the positive occasions that lie together. Firstly, the absolute no. of oversampling perceptions, N is set up. Normally, it is chosen with the end goal that the binary class distribution is 1:1. In any case, that could be tuned down dependent on need. At that point, the sequence begins by first choosing a positive class occurrence at arbitrary. Next, the KNN's for that occurrence is acquired. Finally, N of these K occasions is picked to add new synthetic occurrences. To do that, utilizing any distance metric the distinction in distance between the component vector and its neighbors is determined. Presently, this difference is increased by any arbitrary incentive in (0, 1] and is added to the prior feature vector.

Due to the imbalanced distribution of data in the target variable in the dataset, SMOTE has been utilized to eliminate this issue. The modifications in the dataset after utilizing SMOTE have been described in Table 2. Before applying SMOTE, the ratio of the dataset was not balanced, there were a majority and minority in classes. But, after applying SMOTE, the ratio had become equal.

D. *Machine Learning for Classification*

It has been mentioned earlier that, in this work, logistic regression, random forest, and XGBoost algorithm have been used for accomplishing the classification task. By applying the concept of the "percentage split" technique with the help of "train_test_split" in Scikit-learn, the training set, and test set have been prepared. 80% of data has been used in training and the remaining 20% of data has been used in the test. "Grid Search CV" algorithm has been utilized to obtain the finest parameters of the classifiers.

a) Logistic Regression Classifier: Logistic regression is one of the most basic and generally utilized machine learning algorithms [12]. Logistic regression isn't a

regression method yet a probabilistic classification algorithm. The motivation in Logistic regression is to solve the problem as a summed up linear regression algorithm as in (2):

$$\hat{y} = \beta_0 + \beta_1 x_1 + \ldots\ldots + \beta_n x_n \quad (2)$$

where ŷ = predicted esteem, x = independent features, and the β = learning coefficients.

The selected parameters' values for the logistic regression classifier have been listed in Table 3.

b) Random Forest Classifier: Random forest is a supervised learning technique. The "forest" it assembles, is a collection of decision trees, normally trained with the "bagging" principle [13]. The overall motivation of the "bagging" principle is that a combination of learning models builds the general consequence. Random forest is a compliant, ordinary to utilize machine learning strategy that yields, even without hyper-parameter tuning, an inconceivable consequence more often than not. Accordingly, in a random forest, just an irregular subset of the features is thought about by the calculation for parting a node. Even it can be made trees more arbitrary by also developing thresholds for each feature instead of searching for the most ideal thresholds (like a general decision tree does). RF algorithm works in 4 steps:

1. Selecting arbitrary samples from the specified dataset.
2. Constructing a decision tree for each instance and acquire an anticipated result from every decision tree.
3. Performing a vote in approval of each estimated outcome.
4. Selecting the outcome with the most votes is the final estimation.

The selected parameters' values for the random forest classifier have been listed in Table 4.

c) XGBoost Classifier: Among the gradient boosting (ensemble) methods in tree-based machine learning algorithms, Extreme Gradient Boosting (XGBoost) is one of the mainstream algorithms that possesses improved and quick execution [14]. In the collection of ensemble learning methods, XGBoost represents the boosting method set. A set of classifiers which are the combination of several models that are used for delivering superior classification performance is the concept of ensemble learning. According to the boosting method, the classification errors that have been done in the prior models, are endeavored to be fixed by the subsequent models by summing up extra weights to it. Gradient boosting methods use optimized loss function where the other boosting techniques the weights of the misclassification are larger. XGBoost algorithm is the advancement of gradient boosting methods that have regularization factors.

The objective method is an aggregate of a particular loss assessed classifications and a whole of regularization term for all classifiers. Mathematically, the formula of calculating the objective method is in (3) –

$$obj(\theta) = \sum_{i}^{n} l(y_i - \hat{y}_i) + \sum_{k=1}^{n} \Omega(f_k) \quad (3)$$

The selected parameters' values for the XGBoost classifier have been listed in Table 5.

IV. Result and Discussion

It was mentioned previously, the logistic regression, random forest, and the XGBoost algorithm have been implemented for classifying stroke and non-stroke patients. The implemented system performance has been evaluated based on accuracy, precision, and recall using the subsequent principles in (4), (5), and (6) respectively.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (4)$$

$$\text{Pr}ecision = \frac{TP}{TP + FP} \quad (5)$$

$$\text{Re}call = \frac{TP}{TP + FN} \quad (6)$$

A significant variation in the model's performance has been observed before and after the utilization of SMOTE in the dataset. The detailed description of each model has been represented in Table 6. It can be seen clearly that all models performed well after applying SMOTE, and among the models, the Random Forest model achieved better performances in terms of accuracy, precision, and recall. In Fig. 8, the confusion matrix of the Random Forest model has been illustrated. It represents that, the RF model misclassified (No Stroke vs. Stroke) only 18 instances out of the 1940 instances of the test set.

TABLE IV. Random Forest Classifier's Parameters

Parameter Name	Chosen Value
"n_estimators"	100
"random_state"	5
"max_depth"	30

TABLE V. XGBoost Classifier's Parameters

Parameter Name	Chosen Value
"colsample_bytree"	0.5
"learning_rate"	0.1
"max_depth"	200
"alpha"	10
"n_estimators"	1000
"objective"	"binary:logistic"
"booster"	"gbtree"

TABLE VI. Model Performance

Model	Before SMOTE			After SMOTE		
	Acc. (%)	*Prec. (%)*	*Rec. (%)*	*Acc. (%)*	*Prec. (%)*	*Rec. (%)*
Logistic Regression	95.69	95.69	91.87	97.94	97.5	98.0
Random Forest	96.1	94.74	95.32	99.07	99.0	99.0
XGBoost	95.89	94.73	92.82	98.35	98.5	98.5

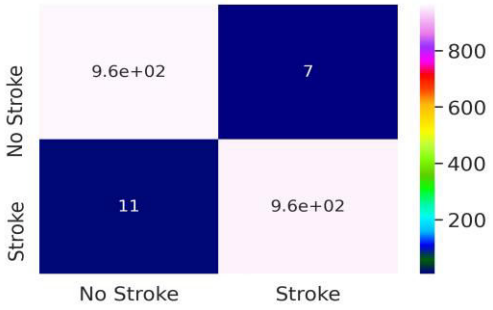

Fig. 8. Confusion matrix of the Random Forest model.

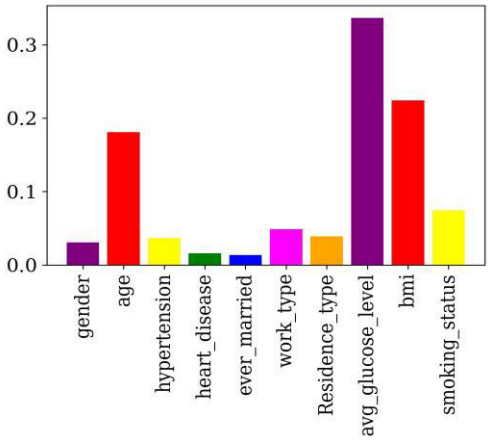

Fig. 9. Feature importance scores from the Random Forest model.

TABLE VII. PROPOSED MODEL COMPARISON

Author	Method Used	Accuracy (%)	Feature Importance Calculation
Singh et. al. [1]	DT + PCA + NN	97.7	No
Liu et al. [2]	DNN	71.6	No
Govindarajan et al. [3]	ANN + SGD	95	No
Emon et al. [7]	Weighted Voting	97	No
Proposed Work	SMOTE + RF	99.07	Yes

The feature importance score from the Random Forest model has been computed using permutation feature importance [15]. Fig. 9 represents the corresponding feature importance scores of the RF model. The feature named "avg_glucose_level", "bmi", and "age" ranked in top-3 among the features. In Table 7, the proposed model has been compared with the previous works and it can be seen that the proposed method achieved significantly better results than the existing works, as well as the feature importance score has been shown also in this research.

V. CONCLUSION

Stroke is a worldwide public health problem, a primary cause of adult disability, and a major cause of death. There is an evidence deficit in developed countries about the public health impact of stroke. Early stroke diagnosis is critical for prompt prevention and recovery. According to research, measurements derived from different risk factors provide useful knowledge for the prediction of stroke. Machine learning algorithms have been proposed as useful decision-making methods in the medical field. This work aims to create a machine learning-based system to predict stroke in people who have stroke symptoms or risk factors with high performance. In this study, a method using oversampling and machine learning has been proposed. Exploratory data analysis and several machine learning algorithms (Logistic Regression, Random Forest, and XGBoost) have been utilized for stroke prediction. SMOTE with Random Forest classifier had shown a dominant performance and outperformed existing works. This research will impact the early prediction of stroke in patients.

REFERENCES

[1] M. S. Singh and P. Choudhary, "Stroke prediction using artificial intelligence," 2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON), 2017, pp. 158-161.

[2] T. Liu, W. Fan and C. Wu, "A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset", *Artificial Intelligence in Medicine*, vol. 101, p. 101723, 2019.

[3] P. Govindarajan, R. Soundarapandian, A. Gandomi, R. Patan, P. Jayaraman and R. Manikandan, "Classification of stroke disease using machine learning algorithms", *Neural Computing and Applications*, vol. 32, no. 3, pp. 817-828, 2019.

[4] H. Asadi, R. Dowling, B. Yan and P. Mitchell, "Machine Learning for Outcome Prediction of Acute Ischemic Stroke Post Intra-Arterial Therapy", *PLoS ONE*, vol. 9, no. 2, p. e88225, 2014.

[5] J. Heo, J. Yoon, H. Park, Y. Kim, H. Nam and J. Heo, "Machine Learning–Based Model for Prediction of Outcomes in Acute Stroke", *Stroke*, vol. 50, no. 5, pp. 1263-1265, 2019.

[6] B. Letham, C. Rudin, T. McCormick and D. Madigan, "Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model", *The Annals of Applied Statistics*, vol. 9, no. 3, 2015.

[7] M. U. Emon, M. S. Keya, T. I. Meghla, M. M. Rahman, M. S. A. Mamun and M. S. Kaiser, "Performance Analysis of Machine Learning Approaches in Stroke Prediction," 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020, pp. 1464-1469.

[8] "Stroke Prediction Dataset", Kaggle.com, 2021. [Online]. Available: https://www.kaggle.com/fedesoriano/stroke-prediction-dataset. [Accessed: 20- Apr- 2021]

[9] S. Morgenthaler, "Exploratory data analysis", *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 1, no. 1, pp. 33-44, 2009.

[10] N. Chawla, K. Bowyer, L. Hall and W. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique", *Journal of Artificial Intelligence Research*, vol. 16, pp. 321-357, 2002.

[11] Ferdib-Al-Islam, L. Akter and M. M. Islam, "Hepatocellular Carcinoma Patient's Survival Prediction Using Oversampling and Machine Learning Techniques," 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), 2021, pp. 445-450.

[12] S. Nusinovici, Y. Tham, M. Chak Yan, D. Wei Ting, J. Li, C. Sabanayagam, T. Wong and C. Cheng, "Logistic regression was as good as machine learning for predicting major chronic diseases", Journal of Clinical Epidemiology, vol. 122, pp. 56-69, 2020.

[13] A. Mosavi, F. Sajedi Hosseini, B. Choubin, M. Goodarzi, A. Dineva and E. Rafiei Sardooi, "Ensemble Boosting and Bagging Based Machine Learning Models for Groundwater Potential Prediction", *Water Resources Management*, vol. 35, no. 1, pp. 23-37, 2020.

[14] L. Akter and Ferdib-Al-Islam, "Dementia Identification for Diagnosing Alzheimer's Disease using XGBoost Algorithm," 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD), 2021, pp. 205-209.

[15] J. Gómez-Ramírez, M. Ávila-Villanueva and M. Fernández-Blázquez, "Selecting the most important self-assessed features for predicting conversion to mild cognitive impairment with random forest and permutation-based methods", *Scientific Reports*, vol. 10, no. 1, 2020.