Stroke Risk Prediction: Comparing Different Sampling Algorithms

Qiuyang Yin[1], Xiaoyan Ye[2], Binhua Huang[3], Lei Qin[4], Xiaoying Ye[5], Jian Wang[6]
Department of Network Technology, Software Engineering Institute of Guangzhou, Guangzhou 310401, China[1,2,6]
Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China[3]
School of Computer Science, Universiti Sains Malaysia, Penang 11800, Malaysia[4]
School of Computer Science, Neusoft Institute Guangdong, Foshan 528225, China[5]

***Abstract*—Stroke is a serious disease that has a significant impact on the quality of life and safety of patients. Accurately predicting stroke risk is of great significance for preventing and treating stroke. In the past few years, machine learning methods have shown potential in predicting stroke risk. However, due to the imbalance of stroke data and the challenges of feature selection and model selection, stroke risk prediction still faces some difficulties.This article aims to compare the performance differences between different sampling algorithms and machine learning methods in stroke risk prediction. This study used the over-sampling algorithm (Random Over Sampling and SMOTE), the under-sampling algorithm (Random Under Sampling and ENN), and the hybrid sampling algorithm (SMOTE-ENN), and combined them with common machine learning methods such as K-Nearest Neighbors, Logistic Regression, Decision Tree and Support Vector Machine to build the prediction model.Through the analysis of experimental results, and found that the SMOTE combined with the LR model showed good performance in stroke risk prediction, with a high F1 score. In addition, this study found that the overall performance of the undersampling algorithm is better than that of the oversampling and hybrid sampling algorithms.These research results provide useful references for predicting stroke risk and provide a foundation for further research and application. Future research can continue to explore more sampling algorithms, machine learning methods, and feature engineering techniques to further improve the accuracy and interpretability of stroke risk prediction and promote its application in clinical practice.**

***Keywords*—Stroke prediction; data mining; machine learning; unbalanced data; sampling algorithms; classification algorithms**

I. Introduction

Stroke is a serious neurological disorder and its health burden is enormous worldwide. According to the World Health Organisation, millions of people die or become permanently disabled as a result of stroke each year [1]. Accurate prediction of the risk of stroke is therefore crucial for early intervention and treatment.

With the rapid development of machine learning techniques, the use of these techniques to predict stroke risk has become a hot topic of research. Machine learning models can predict the probability of stroke in individuals by learning and mining patterns and correlations in large amounts of patient data [2]. This provides clinicians with a new tool to aid decision-making and develop personalized treatment plans.

However, stroke risk prediction faces a number of challenges. Firstly, the mechanisms by which stroke events occur are complex and diverse, involving a variety of potential risk factors such as age, gender, hypertension, and diabetes [3]. Secondly, stroke data often suffer from a serious imbalance, i.e. there is a significant imbalance between the proportion of normal samples and stroke samples [4]. This data imbalance may result in the models having better predictive performance for most classes of samples, but poorer predictive performance for a few classes of samples (i.e. stroke samples). In addition, the generalisability and interpretability of the models are key issues in stroke risk prediction studies [5].

To overcome these challenges and improve the accuracy and reliability of stroke risk prediction, this study aims to compare the performance of different sampling machine learning algorithms in stroke risk prediction. The study will utilize various sampling algorithms, such as Random Over Sampling (ROS) and Synthetic Minority Over-sampling Technique (SMOTE), as well as undersampling algorithms like Random Under Sampling (RUS) and Edited Nearest Neighbors (ENN), along with the combination of SMOTE and Edited Nearest Neighbors (SMOTE-ENN), to address imbalanced data. Additionally, machine learning methods including K-Nearest Neighbors (KNN), Logistic Regression (LR), Decision Tree (DT), and Support Vector Machine (SVM) will be employed for prediction modeling [6]–[9]. Through the comparison of performance among various sampling algorithms and machine learning methods, the study aims to identify the optimal prediction model and enhance the accuracy and reliability of stroke risk prediction.

The paper is structured as follows: Section II reviews relevant previous work. Section III describes the experimental design and methods in detail. The Section IV presents the experimental results and analysis. Finally, Section V summarises the main findings of the paper and discusses directions for further research.

II. Related Works

The study of stroke risk prediction has attracted a great deal of interest and a lot of valuable work has emerged. This section provides an extensive review of relevant literature on stroke risk prediction, encompassing various methods and techniques employed in previous studies. Additionally, an in-depth analysis of the strengths and limitations of these approaches is presented.

Commonly used methods for stroke risk prediction in previous studies include traditional statistical models and machine learning models. Traditional statistical models such as

regression analysis, survival analysis, and decision trees are widely used for stroke risk prediction. These models can be based on large clinical datasets by building predictive models to identify patients' risk factors for stroke. Dennis et al. [10] used survival analysis to accurately estimate the long-term mortality risk of first-time stroke patients. Shao et al. [11] utilized the decision tree C4.5 algorithm to establish a stroke risk assessment model and identify the influences of various factors such as smoking, alcohol consumption, diet, sleep, and exercise on stroke risk. However, traditional statistical models have limitations in dealing with complex non-linear relationships and high-dimensional data.

In recent years, the development of machine learning techniques has opened up new possibilities for stroke risk prediction. Machine learning models can automatically learn patterns and correlations from data and are able to handle non-linear relationships and high-dimensional data. Common machine learning methods used in stroke risk prediction include SVM, DT and Random Forest (RF) and Artificial Neural Network (ANN) [12]–[15]. These models can be trained on large amounts of patient data to predict stroke risk with a high degree of accuracy and generalisation.

In previous research on stroke prediction using machine learning models, the focus has primarily been on the performance of machine learning models. Viswapriya et al. [12] proposed a hybrid model combining ANN and RF for stroke prediction, achieving a classification accuracy of 94%. Sailasya et al. [13] proposed the use of various machine learning algorithms to predict the risk of brain stroke, with Naïve Bayes (NB) achieving the highest accuracy of approximately 82%. Dritsas et al. [14] proposed a robust framework using machine learning models and a stacking method to accurately predict the long-term risk of stroke occurrence, achieving high performance with an AUC of 98.9% and an accuracy of 98%. Alageel et al. [15] proposed an analysis of factors enhancing stroke prediction using electronic health records, identifying age, average glucose level, heart disease, and hypertension as critical factors, and evaluating seven machine learning algorithms for stroke occurrence prediction with high accuracy and performance.

In addition to traditional statistical models and machine learning models, sampling algorithms are also widely used in stroke risk prediction. As stroke data usually exhibit a class imbalance problem, i.e. a significant imbalance between stroke and normal samples, sampling algorithms can balance the dataset by oversampling, undersampling or hybrid sampling. However, a critical aspect that has been overlooked in previous research is the comparison of different sampling algorithms for handling imbalanced datasets. Some researchers generally directly use the popular SMOTE algorithm to process imbalanced data [16]–[18], and , there are also some researchers simply use random sampling methods to compare with SMOTE algorithms [19].

In summary, stroke risk prediction is a challenging and important area of research. Traditional statistical and machine learning models provide powerful tools for stroke risk prediction, while sampling algorithms are able to handle unbalanced data sets. This study aims to further investigate the effectiveness of combining different sampling algorithms with machine learning models for stroke risk prediction. The findings of this research contribute to the advancement of methods and insights in the field of stroke risk prediction.

III. Methodology

This section is divided into four sections, including data collection, data pre-processing, machine learning Models, and evaluation metrics, and the proposed workflow is shown in Fig. 1.

A. Data Collection

The predicted stroke dataset in this study is from the Kaggle platform, which contains 5110 patient data [20]. It has 12 features, including seven categorical features, four quantitative features and a patient ID number. There is personal and health information about the patient, details of which are shown in the Table I.

TABLE I. Description of Stroke Dataset

Feature Name	Feature Description	Feature Type
Id	Patient unique id.	/
Gender	Male, Female, Other.	Quantitative
Age	Patient ages in years.	Quantitative
Hypertension	If the patient has hypertension, then 1 else 0.	Categorical
Heart disease	If the patient has heart disease, then 1 else 0.	Categorical
Ever married	No, Yes.	Categorical
Work type	Children, Govt job, Never worked, Private, Self-employed.	Categorical
Residence type	Rural, Urban.	Categorical
Avg glucose level	Average glucose level in blood.	Quantitative
BMI	Body Mass Index (BMI) is an indicator used to assess a person's weight status based on their weight and height.	Quantitative
Smoking status	Formerly smoked, Never Smoked, Smokes, Unknown.	Categorical
Stroke	If the patient has a stroke disease, then 1 else 0.	Categorical

The dataset has 4681 normal and 249 stroke patients, 95% of which are negative cases and only 5% are positive cases, a highly unbalanced dataset as shown in Fig. 2.

B. Data Pre-processing

1) Missing Value Handing: About 4% of the data have missing BMI values. To improve the robustness of the model, 0 is used for filling.

2) Meaningless Features Handing: In this study, the patient ID was considered irrelevant and subsequently excluded from the analysis.

3) Label Encoding: The values of some of the category features (Gender, Ever married, etc.) need to be converted to numerical values before being entered into the model.

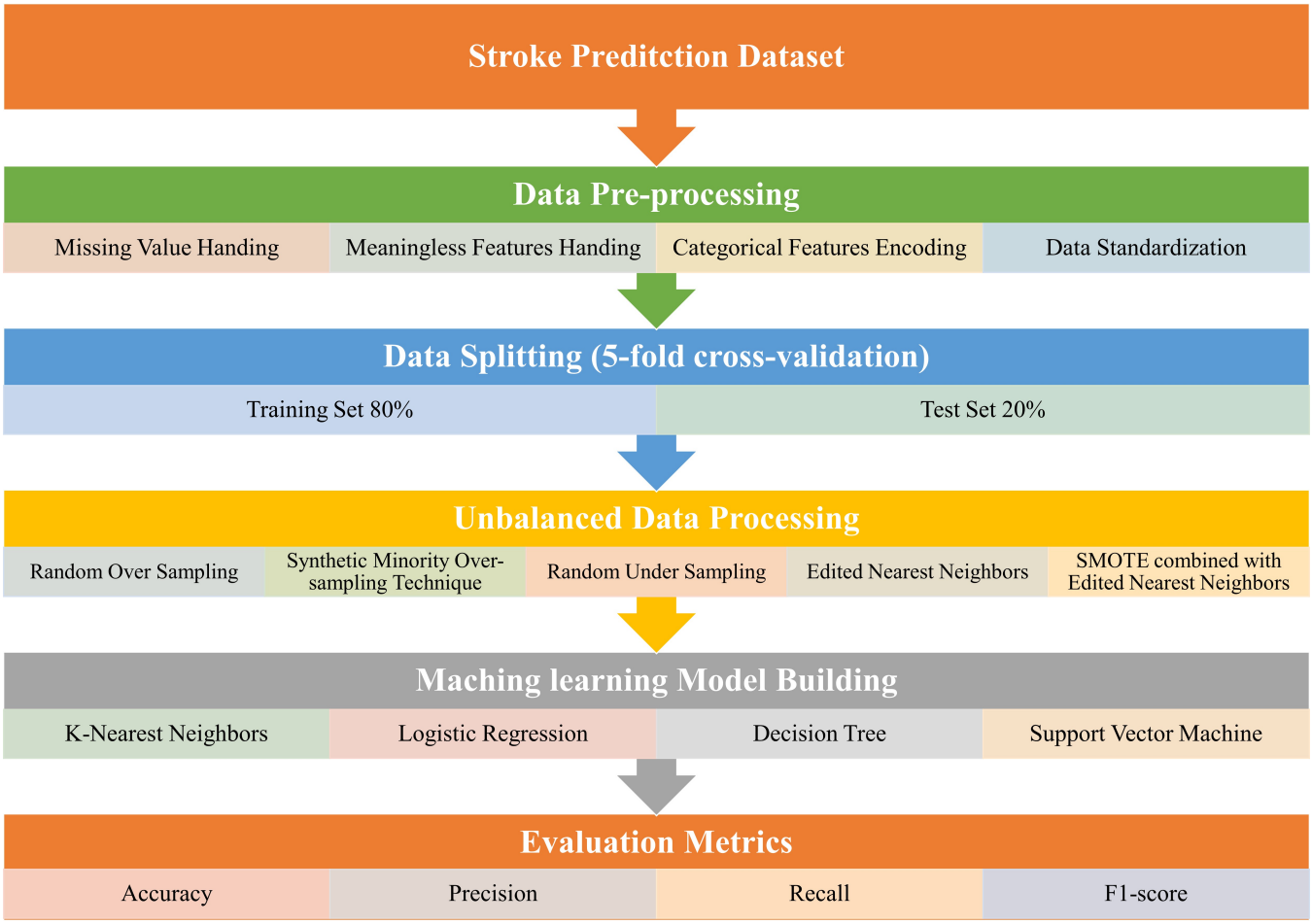

Fig. 1. Proposed workflow.

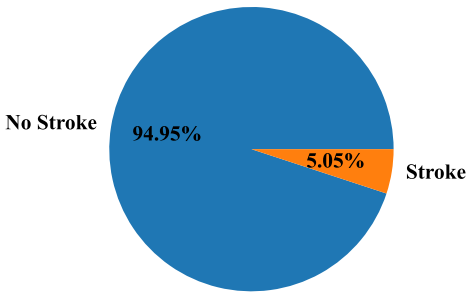

Fig. 2. No Stroke vs Stroke.

4) Data Standardization: The magnitude and unit of stroke patients' information data are different. For example, the value of avg glucose level is hundreds, while the value of BMI is only dozens. Data needs to be standardized before being input into the model. In this experiment, a separate standardization procedure was applied to the training set and test set after partitioning the dataset in order to avoid data leakage. Therefore, this study standardized the dataset, calculation as Equation (1):

$$x_{\text{standardized}} = \frac{x - \mu}{\sigma} \tag{1}$$

where $x_{\text{standardized}}$ represents the standardized value of x,μ is the mean of the data,σ is the standard deviation of the data, and x is the original data point.

C. Description of Sampling Algorithms

Stroke data typically suffer from category imbalance, i.e. a significant imbalance in the ratio between stroke and normal samples. To tackle the challenge of imbalanced data in stroke risk prediction, a range of sampling algorithms were utilized. This section provides an overview of the sampling algorithms employed, namely ROS, SMOTE, RUS, ENN, and SMOTE-ENN.

1) Random Over Sampler: ROS is a simple yet effective algorithm that randomly replicates minority class samples until the class distribution is balanced. It increases the number of stroke samples in the dataset, allowing the machine learning models to learn from more balanced data.

2) Synthetic Minority Over-sampling Technique: SMOTE is a widely used oversampling algorithm that generates synthetic minority class samples based on the characteristics of

existing minority samples. It creates synthetic samples by interpolating between randomly selected minority samples and their nearest neighbors.

3) Random Under Sampling: The RUS randomly selects a subset of majority class samples to match the number of minority class samples. It reduces the dominance of the majority class in the dataset, allowing the machine learning models to focus more on the minority class.

4) Edited Nearest Neighbors: ENN is an undersampling algorithm that removes misclassified majority class samples based on their nearest neighbors' class labels. It compares the class label of each majority sample with its k-nearest neighbors and removes the samples that are misclassified.

5) SMOTE-ENN: SMOTE-ENN first applies the SMOTE algorithm to generate synthetic samples and then applies the ENN algorithm to remove misclassified samples.

D. Description of Machine Learning Models

This section provides a detailed description of the machine learning methods utilized for stroke risk prediction in the study. The study utilized a selection of widely used machine learning algorithms, including KNN, LR, DT, and SVM. Each algorithm has its unique characteristics and advantages in handling different types of data and classification problems.

1) K-Nearest Neighbors: KNN is a non-parametric algorithm that classifies data points based on the majority class label of their k-nearest neighbors [6]. It can be used for stroke risk prediction by measuring the similarity between the input sample and other samples in the dataset. The common distance formula used in KNN algorithm is the Euclidean distance. For two sample points x and x_i, the Euclidean distance is calculated as Equation (2).

$$d(x, x_i) = \sqrt{\sum_{j=1}^{n}(x_j - x_{ij})^2} \tag{2}$$

where n is the feature dimension of the sample points, and x_j and x_{ij} denote the values of the sample x and x_i on the jth feature, respectively.

2) Logistic Regression: LR is a linear classification algorithm that models the probability of a sample belonging to a specific class. It can be used to predict the probability of stroke occurrence based on the input features [7]. The logistic regression model estimates the parameters of a logistic function using maximum likelihood estimation. The logistic function maps the input features to a probability value, and a threshold can be applied to classify the samples into different classes. The logistic regression model can be represented as Equation (3):

$$P(\text{Stroke} = 1|X) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_n X_n)}} \tag{3}$$

where $P(\text{Stroke} = 1|X)$ represents the probability of stroke occurrence given the input features X, and $\beta_0, \beta_1, \ldots, \beta_n$ are the coefficients estimated during model training.

3) Decision Tree: DT is a hierarchical tree-based algorithm that splits the feature space based on the values of different features [8]. It can be used to identify the important features and their thresholds that are associated with stroke risk. The decision tree creates a tree-like model where each internal node represents a decision based on a feature, and each leaf node represents a class label. The decision tree can be represented as a series of if-else statements, where each internal node represents a splitting condition based on a feature, and each leaf node represents a class label.

4) Support Vector Machine: SVM is a binary classification algorithm that aims to find an optimal hyperplane in the feature space that separates the data points of different classes with the maximum margin [9]. It can be used for stroke risk prediction by finding a decision boundary that distinguishes between samples with and without stroke. SVM can handle both linearly separable and non-linearly separable data by using different kernel functions to map the input features to a higher-dimensional space. The SVM classification function can be represented as Equation (4):

$$f(x) = \sum_{ix \in S} \alpha_i y_i K(x_i, x) + b \tag{4}$$

where x_i denotes the training patterns, $y_i \in \{+1, -1\}$ denotes the corresponding class labels and S denotes the set of Support Vectors [21].

These machine learning methods can effectively capture the underlying patterns and relationships in the data and make predictions about the stroke risk for individuals based on their input features.

E. Evaluation Metrics

To assess the effectiveness of the stroke risk prediction models, the evaluation of the confusion matrix and various evaluation metrics was performed. Stroke was considered the positive class, while no stroke was considered the negative class.

1) Confusion Matrix: The confusion matrix provides a tabular representation of the model's predictions, showing the counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). It helps in analyzing the model's performance in correctly classifying stroke and no stroke cases [13]. The confusion matrix is shown in Table II.

TABLE II. CONFUSION MATRIX

	Predicted Stroke	**Predicted No Stroke**
Actual Stroke	TP	FN
Actual No Stroke	FP	TN

The elements in the matrix have the following meanings:

TP: The number of samples that the model correctly predicted as strokes.

TN: The number of samples that the model correctly predicted a no stroke.

FP: The number of samples that the model incorrectly predicted as having a stroke.

FN: The number of samples that the model incorrectly predicted as no stroke.

2) Accuracy: Accuracy is the ratio of correctly predicted instances (both stroke and no stroke) to the total number of instances . It measures the overall correctness of the model's predictions, providing an indication of how well it classifies both positive and negative cases.It is calculated as Equation (5):

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

3) Precision: Precision quantifies the proportion of true positive predictions (correctly predicted strokes) out of the total predicted positive instances (predicted strokes). It measures the model's ability to accurately identify individuals at risk of stroke, minimizing false positive predictions.It is calculated as Equation (6):

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6}$$

4) Recall: Recall calculates the proportion of true positive predictions (correctly predicted strokes) out of the actual positive instances (actual strokes). It measures the model's ability to correctly identify individuals who have experienced a stroke, minimizing false negative predictions.It is calculated as Equation (7):

$$\text{Recall} = \frac{TP}{TP + FN} \tag{7}$$

5) F1-score: The F1-score is the harmonic mean of precision and recall, providing a balanced measure of the model's accuracy in predicting both stroke and no stroke cases. It considers both false positive and false negative predictions and provides an overall assessment of the model's performance.It is calculated as Equation (8):

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

By utilizing the confusion matrix and these evaluation metrics, a comprehensive evaluation of the stroke risk prediction models can be conducted to assess their performance in accurately identifying individuals at risk of stroke while minimizing false positive and false negative predictions.

IV. Result and Discussion

The performance of five different sampling algorithms was evaluated, including ROS, RUS, SMOTE, ENN, and SMOTE-ENN, in combination with four machine learning models: KNN, LR, DT and SVM. The evaluation metrics used were accuracy, precision, recall, and F1-score in Table III.

The accuracy comparison is shown in the Fig. 3 and the accuracy of the models is observed to decrease to some extent after applying the sampling algorithm. This may be because the sampling algorithm changes the distribution of the samples when processing unbalanced data, thus affecting the overall accuracy. Among all algorithms, the combination of ENN and LR achieved the highest accuracy (0.9454), while the combination of RUS and DT achieved the lowest accuracy (0.6945). Compared to other sampling algorithms, the model of ENN algorithm is higher, because the ENN algorithm improves accuracy by removing noise and redundant samples from the majority class, thereby preserving the distinctive features of the minority class samples.

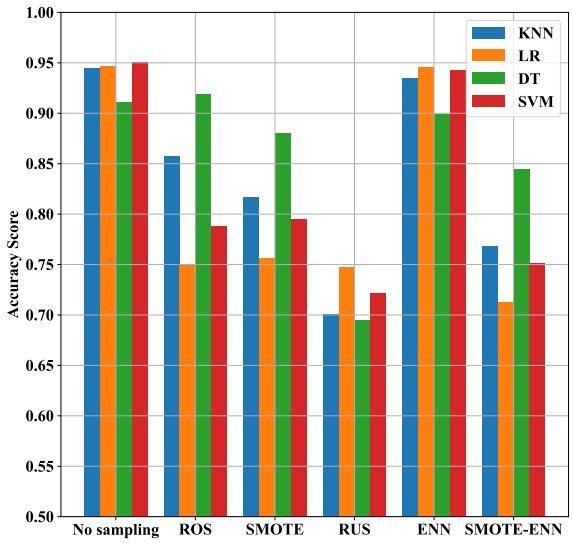

Fig. 3. Accuracy of each sampling algorithm combined with machine learning models.

The precision comparison is shown in Fig. 4, revealing that the combination of ENN and LR models achieved the highest precision rate of 0.2675, whereas the combination of RUS and DT algorithms obtained the lowest precision rate of 0.0967. And the precision of the model combined with ENN is significantly higher than the other sampling algorithms, indicating that the ENN algorithm is better able to identify the true positive samples and reduce the possibility of false positives.

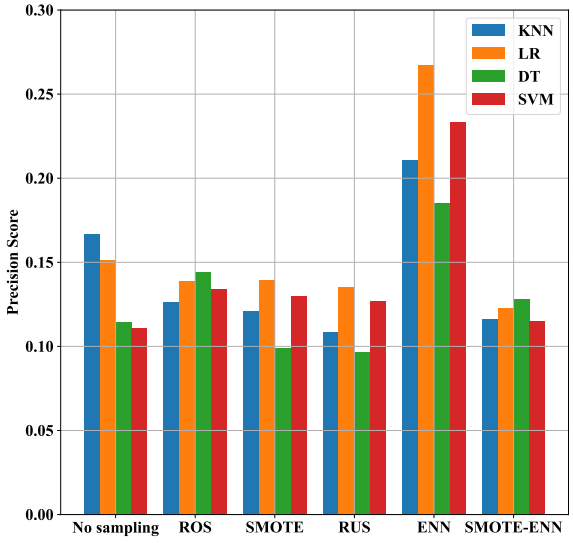

Fig. 4. Precision of each sampling algorithm combined with machine learning models.

The recall comparison is shown in Fig. 5, demonstrating a substantial improvement in the model's performance compared to unsampled data when utilizing the sampling algorithm. The

TABLE III. PERFORMANCE EVALUATION BY SAMPLING ALGORITHM AND MACHINE LEARNING MODEL

Sampling Algorithm[a]	Machine Learning Model[b]	Accuracy	Precision	Recall	F1-score
No Sampling	KNN	0.9442	0.1667	0.0189	0.0339
	LR	0.9464	0.1515	0.0255	0.0437
	DT	0.9105	0.1143	0.1176	0.1159
	SVM	**0.9508**	0.1111	0.0052	0.0099
ROS	KNN	0.8573	0.1262	0.3211	0.1803
	LR	0.7489	0.1390	0.7971	0.2363
	DT	0.9186	0.1442	0.1353	0.1393
	SVM	0.7881	0.1338	0.6081	0.2184
SMOTE	KNN	0.8168	0.1210	0.4328	0.1875
	LR	0.7556	0.1397	0.7772	**0.2368**
	DT	0.8804	0.0992	0.1817	0.1272
	SVM	0.7951	0.1297	0.5682	0.2108
RUS	KNN	0.7002	0.1088	0.7233	0.1889
	LR	0.7466	0.1353	0.7794	0.2305
	DT	0.6945	0.0967	0.6289	0.1673
	SVM	0.7211	0.1271	**0.8035**	0.2186
ENN	KNN	0.9341	0.2110	0.1240	0.1544
	LR	0.9454	**0.2675**	0.0542	0.0887
	DT	0.8990	0.1851	0.3071	0.2267
	SVM	0.9421	0.2331	0.0832	0.1192
SMOTE-ENN	KNN	0.7683	0.1160	0.5661	0.1920
	LR	0.7123	0.1226	0.8028	0.2117
	DT	0.8444	0.1281	0.3771	0.1912
	SVM	0.7509	0.1149	0.6102	0.1928

[a] ROS: Random Over Sampler; SMOTE: Synthetic Minority Over-sampling Technique; RUS: Random Under Sampling; ENN: Edited Nearest Neighbors; SMOTE-ENN:Synthetic Minority Over-sampling Technique combined with Edited Nearest Neighbors.
[b] KNN: K-Nearest Neighbors; LR: Logistic Regression; DT: Decision Tree; SVM: Support Vector Machine.

combination of RUS and SVM has the highest recall of 0.8035, while ENN+LR has the lowest recall of 0.0542. At this point, the model recall of the combination with ENN is significantly lower than the other sampling algorithms except for LR, and the model recall of the combination with RUS is overall higher than the other sampling algorithms, which indicates that the RUS algorithm can better identify the true positive samples and reduce the possibility of misclassification the possibility of misclassification.

The F1-score comparison is shown in Fig. 6. Under the metric, the model scores were all significantly higher after using the sampling algorithm than without the sampling algorithm. The combination of SMOTE and LR model obtained the highest score (0.2368), while the combination of ENN and LR model scored the lowest (0.0887). However, RUS performs better in combination with other algorithms, which indicates that the RUS is able to find a balance between accuracy and recall, resulting in a better overall model performance.

In a comprehensive comparison, the overall performance of the under sampling algorithm is better than that of the oversampling and hybrid sampling algorithms for the prediction stroke problem, and the models combined with the ENN algorithm generally perform better under the accuracy and precision metrics. The models combined with the RUS algorithm are generally better under the recll and F1-score metrics. Reviewing the findings depicted in Fig. 4, it becomes apparent that the utilization of the oversampling algorithm has resulted in only marginal improvements in model precision. Interestingly, in certain instances, the precision scores of certain models were lower when the oversampling algorithm was applied, compared to their performance without any sampling algorithm. Instead, the model precision scores of the ENN algorithm combined with the under sampling algorithm were significantly higher than the other algorithms. The observed phenomenon can be attributed to the limited number of positive samples, approximately 200 cases, and the significant variation in data distribution. The oversampling algorithm, in such cases,

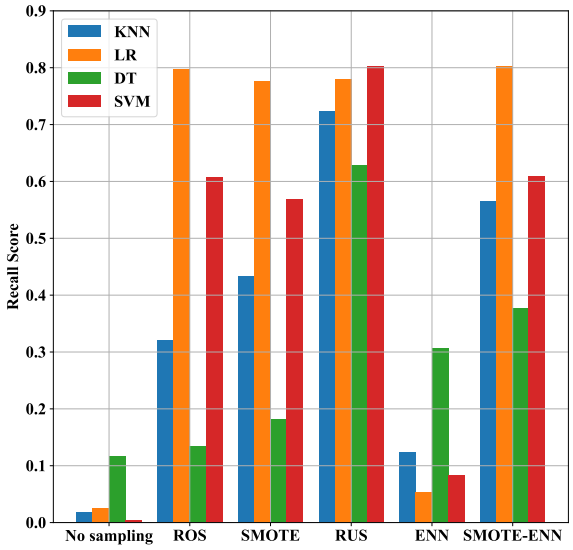

Fig. 5. Recall of each sampling algorithm combined with machine learning models.

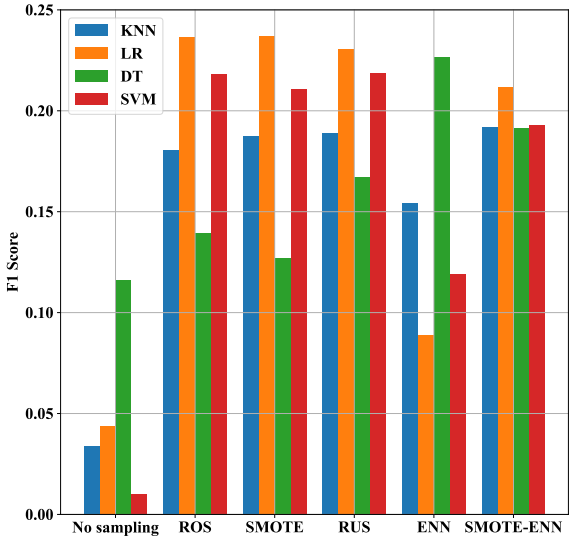

Fig. 6. F1-score of each sampling algorithm combined with machine learning models.

tends to generate duplicate, noisy, or unreliable samples.

V. Conclusion and Future Work

This study examines stroke risk prediction through a comparative analysis of various sampling algorithms and machine learning methods. The experimental results show that the use of an appropriate combination of sampling algorithms and machine learning methods can significantly improve the prediction performance in the stroke risk prediction task.

In the analysis conducted, the hybrid sampling algorithm (SMOTE+LR) combined with the KNN model demonstrated superior performance in stroke risk prediction, yielding a high F1 score. The combination of other sampling algorithms and machine learning methods also achieved some prediction performance, but was overall inferior to the combination of the SMOTE algorithm and the LR model.

Although the SMOTE+LR combination exhibited the highest F1 score, the analysis revealed that the overall performance of the oversampling algorithm outperformed the undersampling algorithm and the hybrid sampling algorithm when considering the performance of various sampling methods in conjunction with machine learning models. Selecting an appropriate combination of sampling algorithms and machine learning methods is pivotal in enhancing the accuracy of stroke risk prediction.

This research has achieved some beneficial results, there are still some limitations that need to be considered.

Regarding dataset selection, specific datasets were utilized for conducting the experiments, which may introduce domain-specific or sample distribution biases. To ensure the generalizability of the findings, future research should encompass a broader range of datasets for validation purposes.

In terms of model building, the study selected KNN, LR, DT and SVM as machine learning methods with oversampling, under sampling and hybrid sampling algorithms. However, there are other machine learning methods and sampling algorithms that can be tried, such as random forests, neural networks, and other variants of sampling methods. The comparison and exploration of these methods will contribute to a more comprehensive understanding of the problem of stroke risk prediction.

In future research, potential avenues for improvement can be explored in the following directions:

In terms of feature engineering, in stroke risk prediction, the selection and extraction of effective features are critical to prediction performance. Further research can explore better feature selection and feature engineering methods to improve prediction performance.

In terms of integrated learning, integrated learning methods can improve the accuracy and stability of stroke risk prediction by combining the prediction results of multiple models. Further research could try integrated learning methods and compare them with a single model.

In terms of interpretive analysis, the interpretation of stroke risk predictions is critical to clinical practice and decision support. Further research could explore how to interpret and explain the prediction results of models to increase their credibility and interpretability.

In conclusion, the results of this study provide a useful reference for stroke risk prediction and provide a basis for further research and application. Future research can continue to explore more sampling algorithms, machine learning methods and feature engineering techniques to further improve the accuracy and interpretability of stroke risk prediction and to promote its application in clinical practice.

Acknowledgment

The authors acknowledge support from Guangdong provincial innovation school project (Grant No.2022KTSCX172).

References

[1] S. Ramesh and K. Kosalram, "The burden of non-communicable diseases: A scoping review focus on the context of india," *Journal of Education and Health Promotion*, vol. 12, 2023.

[2] S. Dev, H. Wang, C. S. Nwosu, N. Jain, B. Veeravalli, and D. John, "A predictive analytics approach for stroke prediction using machine learning and neural networks," *Healthcare Analytics*, vol. 2, p. 100032, 2022.

[3] R. Alkahtani, "Molecular mechanisms underlying some major common risk factors of stroke," *Heliyon*, p. e10218, 2022.

[4] Y.-W. Chen, K.-c. Lin, Y.-c. Li, and C.-J. Lin, "Predicting patient-reported outcome of activities of daily living in stroke rehabilitation: a machine learning study," *Journal of NeuroEngineering and Rehabilitation*, vol. 20, no. 1, pp. 1–12, 2023.

[5] D. M. Oosterveer, H. Arwert, C. B. Terwee, J. W. Schoones, and T. P. V. Vlieland, "Measurement properties and interpretability of the promis item banks in stroke patients: a systematic review," *Quality of Life Research*, vol. 31, no. 12, pp. 3305–3315, 2022.

[6] X. Zhang, H. Xiao, R. Gao, H. Zhang, and Y. Wang, "K-nearest neighbors rule combining prototype selection and local feature weighting for classification," *Knowledge-Based Systems*, vol. 243, p. 108451, 2022.

[7] Y. Hu, Y. Fan, Y. Song, and M. Li, "A general robust low–rank multinomial logistic regression for corrupted matrix data classification," *Applied Intelligence*, pp. 1–17, 2023.

[8] P. Rani and R. Sharma, "Intelligent transportation system for internet of vehicles based vehicular networks for smart cities," *Computers and Electrical Engineering*, vol. 105, p. 108543, 2023.

[9] M. Tanveer, T. Rajani, R. Rastogi, Y.-H. Shao, and M. Ganaie, "Comprehensive review on twin support vector machines," *Annals of Operations Research*, pp. 1–46, 2022.

[10] M. S. Dennis, J. Burn, P. Sandercock, J. Bamford, D. Wade, and C. Warlow, "Long-term survival after first-ever stroke: the oxfordshire community stroke project." *Stroke*, vol. 24, no. 6, pp. 796–800, 1993.

[11] Z. Shao, Y. Xiang, Y. Zhu, A. Fan, and P. Zhang, "Influences of daily life habits on risk factors of stroke based on decision tree and correlation matrix," *Computational and Mathematical Methods in Medicine*, vol. 2020, 2020.

[12] S. Viswapriya and D. Rajeswari, "A systematic method of stroke prediction model based on big data and machine learning," in *2022 Smart Technologies, Communication and Robotics (STCR)*. IEEE, 2022, pp. 1–5.

[13] G. Sailasya and G. L. A. Kumari, "Analyzing the performance of stroke prediction using ml classification algorithms," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, 2021.

[14] E. Dritsas and M. Trigka, "Stroke risk prediction with machine learning techniques," *Sensors*, vol. 22, no. 13, p. 4670, 2022.

[15] N. Alageel, R. Alharbi, R. Alharbi, M. Alsayil, and L. A. Alharbi, "Using machine learning algorithm as a method for improving stroke prediction," vol. 14, no. 4.

[16] M. Ghosh *et al.*, "An enhanced stroke prediction scheme using smote and machine learning techniques," in *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. IEEE, 2021, pp. 1–6.

[17] Y. Ren, C. Wang, H. Wang, and Y. Xia, "Stroke prediction based on improved machine learning algorithm," in *International Symposium on Robotics, Artificial Intelligence, and Information Engineering (RAIIE 2022)*, vol. 12454. SPIE, 2022, pp. 496–504.

[18] M. Phankokkruad and S. Wacharawichanant, "Performance analysis and comparison of cerebral stroke prediction models on imbalanced datasets," in *2022 IEEE/ACIS 7th International Conference on Big Data, Cloud Computing, and Data Science (BCD)*. IEEE, 2022, pp. 161–165.

[19] Y. Wu and Y. Fang, "Stroke prediction with machine learning methods among older chinese," *International journal of environmental research and public health*, vol. 17, no. 6, p. 1828, 2020.

[20] Brain stroke prediction dataset. [Online]. Available: https://www.kaggle.com/datasets/zzettrkalpakbal/full-filled-brain-stroke-dataset

[21] S. Vishwanathan and M. N. Murty, "Ssvm: a simple svm algorithm," in *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No. 02CH37290)*, vol. 3. IEEE, 2002, pp. 2393–2398.