Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

Machine Learning for Stroke Prediction: Evaluating the Effectiveness of Data Balancing Approaches

Muhamad Indra [-1], Ilham Maulana[-2], Siti Ernawati [-3],

[1,2]Ilmu Komputer / Fakultas Teknologi Informasi
Universitas Nusa Mandiri
[1]indra.orion@gmail.com, [2]k4ilham@gmail.com,

[3]Sistem Informasi / Fakultas Teknologi Informasi
Universitas Nusa Mandiri
[3]siti.ste@nusamandiri.ac.id

Abstract

Stroke occurs due to disrupted blood flow to the brain, either from a blood clot (ischemic) or a ruptured blood vessel (hemorrhagic), leading to brain tissue damage and neurological dysfunction. It remains a leading cause of death and disability worldwide, making early prediction crucial for timely intervention. This study evaluates the impact of data balancing techniques on stroke prediction performance across different machine learning models. Random Forest (RF) consistently achieves the highest accuracy (98%) but struggles with precision and recall variations depending on the balancing method. Decision Tree (DT) and K-Nearest Neighbors (KNN) benefit most from SMOTE and SMOTETomek, improving their F1-scores (11.21% and 9.18%), indicating better balance between precision and recall. Random Under Sampling enhances recall across all models but reduces precision, leading to lower overall predictive reliability. SMOTE and SMOTETomek emerge as the most effective balancing techniques, particularly for DT and KNN, while RF remains the most accurate but requires further optimization to improve precision and recall balance.

Keywords: Stroke Prediction; Data Balancing Technique; Artificial Intelegence; Classification; Imbalanced Data.

Abstrak

Stroke terjadi akibat gangguan aliran darah ke otak, baik karena adanya gumpalan darah (iskemik) maupun pecahnya pembuluh darah (hemoragik), yang menyebabkan kerusakan jaringan otak dan disfungsi neurologis. Penyakit ini tetap menjadi salah satu penyebab utama kematian dan kecacatan di seluruh dunia, sehingga prediksi dini sangat penting untuk intervensi yang tepat waktu. Penelitian ini mengevaluasi dampak teknik penyeimbangan data terhadap kinerja prediksi stroke menggunakan berbagai model machine learning. Hasil menunjukkan bahwa Random Forest (RF) secara konsisten mencapai akurasi tertinggi (98%), namun mengalami variasi dalam presisi dan recall, tergantung pada metode penyeimbangan yang digunakan. Decision Tree (DT) dan K-Nearest Neighbors (KNN) menunjukkan peningkatan kinerja dengan SMOTE dan SMOTETomek, yang meningkatkan F1-score masing-masing menjadi 11,21% dan 9,18%, menunjukkan keseimbangan yang lebih baik antara presisi dan recall. Random Under Sampling meningkatkan recall pada semua model, tetapi mengurangi presisi, sehingga menurunkan keandalan prediksi secara keseluruhan. Secara keseluruhan, SMOTE dan SMOTETomek merupakan teknik penyeimbangan data yang paling efektif, terutama untuk DT dan KNN, sedangkan RF tetap menjadi model paling akurat namun memerlukan optimasi lebih lanjut untuk meningkatkan keseimbangan antara presisi dan recall.

Kata kunci: Prediksi Stroke; Teknik Penyeimbangan Data; Kecerdasan Buatan; Klasifikasi; Data Tidak Seimbang.

INTRODUCTION

Stroke is a neurological disorder caused by disrupted blood flow to the brain due to a blood clot (ischemic) or a ruptured blood vessel (hemorrhagic)(Auer & Sommer, 2021; Niles et al., 2024). This interruption deprives the brain of oxygen, resulting in brain cell death. Ischemic stroke stems from arterial blockages, while hemorrhagic stroke arises from ruptured vessels, often associated with hypertension, vessel wall weakness, or blood thinner use(Musmar et al., 2022). As a sudden and progressive condition, stroke is a leading global cause of death and long-term disability(Avan & Hachinski, 2021; W. Li et al., 2020), presenting symptoms such as muscle weakness, facial or limb paralysis, speech difficulties, changes in consciousness, and vision issues(Al Hashmi et al., 2022). Also referred to as cerebrovascular accident (CVA), it is the third leading cause of death worldwide, following heart disease and cancer, predominantly affecting the elderly(Woodward, 2019). Hypertension is a primary risk factor for ischemic stroke, while younger populations may experience stroke due to clotting disorders, carotid dissection, or drug abuse (Murphy & Werring, 2020). Treatment options include medication, surgery, and rehabilitation therapy, tailored to the type of stroke(S. Chen et al., 2021; Duncan et al., 2021). Advances in Artificial Intelligence now facilitate early stroke risk prediction, enabling timely interventions to reduce its impact(Zeng et al., 2020) . Artificial Intelligence (AI) technology offers significant potential for early prediction of Cerebrovascular Accident (CVA) risk. By mimicking human intellectual abilities, AI enables more intelligent and flexible interactions, supported by technology acceptance theory(Elfa & Dawood, 2023; Sohn & Kwon, 2020). The development of explainable AI further enhances trust in human-computer interactions(Shneiderman, 2020). AI algorithms, including Machine Learning and Deep Learning, are increasingly used in medical predictions, particularly for stroke, leveraging advancements in processing capabilities(Bohr & Memarzadeh, 2020). Early detection of stroke risk factors facilitates timely interventions to mitigate its effects. Research by Gangavarapu Sailasya and Gorli L. Aruna Kumari demonstrates the effectiveness of Machine Learning algorithms for stroke prediction, identifying Naïve Bayes as the best-performing algorithm with an accuracy of approximately 82% (Sailasya & Kumari, 2021).

Research by Ivan G. Ivanov, Yordan Kumchev, and Vincent James Hooper focuses on improving stroke prediction by addressing data imbalance and algorithmic bias. Their study developed a Machine Learning model utilizing Support Vector Machine (SVM) with an accuracy of 98% and recall of 97%, emphasizing the importance of data quality and preprocessing(Ivanov et al., 2023). Similarly, Hatice Nizam-Ozogur and Zeynep Orman tackled class imbalance in healthcare datasets using GASMOTEPSO_ENN, a hybrid method combining SMOTE, ENN, GA, and PSO. Applied to datasets for chronic kidney disease (CKD), stroke prediction (CSP), and PIMA Indian diabetes (PID), this method achieved high performance, with MCC values of 1.00 for Logistic Regression (CKD), 0.94 for XGBoost (CSP), and 0.87 for SVM (PID)(Nizam-Ozogur & Orman, 2024).

Based on previous research, this study compares the performance of Decision Tree, Random Forest, K-NN algorithms in predicting stroke on the Cerebral Stroke Prediction-Imbalanced Dataset. Data balancing techniques such as Resampling Technique and Hybrid Sampling Technique are used to address class imbalance in the dataset, ensuring good model performance and effective prediction of all classes. This study also focuses on data cleaning and preprocessing to produce high-quality data. The determination of the algorithm with the best accuracy, recall, precision, and f-measure is one of the main outcomes of this research.

RESEARCH METHODS

This study aims to enhance early stroke detection and mitigate its impact through comprehensive data analysis. The research methodology involves systematic steps to identify risk factors and predict stroke occurrence. The key stages include Data Collection, Pre-Processing, Data Balancing using Resampling and Hybrid Sampling Techniques, Classification with Machine Learning and Deep Learning, Feature Evaluation, and Results and Discussion. These steps are organized and depicted in the research flowchart, providing a clear overview of the process. Each stage is presented in Figure 1.

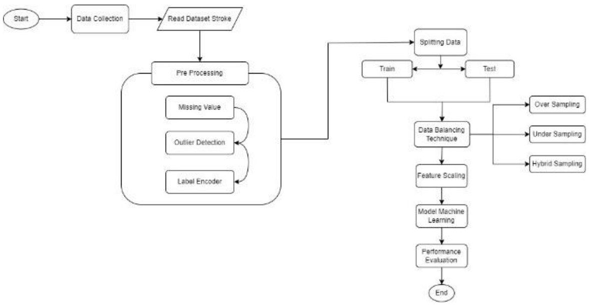

Figure 1. Research Method

Types of research

This study uses historical data from stroke patients, such as patient demographics, medical history, and treatment outcomes. The study intends to use and analyse data mining algorithms to improve stroke diagnosis, treatment decision-making, and patient outcome prediction. This study aims to improve clinical decision-making, optimise treatment protocols, and refine risk assessment models for stroke patients by analysing patterns in previous instances.

Research Target / Subject

The subject of this research is Stroke Identification on Unbalanced Datasets derived from publicly obtained datasets. This research will study how unbalanced data affects prediction accuracy and how to obtain a balanced dataset for the best classification results.

Data Collection

Data collection involves gathering and measuring information from various sources to accurately understand a specific subject (Ganesha & Aithal, 2022). The dataset used in this study was obtained from a repository on the Kaggle platform, a site offering resources and competitions for data science and coding, facilitating data analysis and processing for technology applications and scientific research. The dataset was downloaded from the following link: Kaggle Dataset - Cerebral Stroke Prediction.

Pre-Processing

Data pre-processing is the process of identifying and correcting (or removing) corrupted or inaccurate records in a dataset(Jassim & Abdulwahid, 2021). Data preprocessing is typically performed by removing irrelevant data(Xiao et al., 2022). Additionally, the data is transformed into a format that is easier for the system to understand, making this process crucial for simplifying the next steps, namely data analysis. Several pre-processing techniques performed in this study include: Handling Missing Data, handling categorical data (label encoding), feature scaling, and handling outliers.

Missing Value

The data collected still contains some empty fields in several functions, so this data is considered to have missing values. Therefore, a method is needed to normalize this data(Jäger et al., 2021). Missing values occur because information about the object is unavailable, hard to find, or does not exist (Johnson et al., 2021). Some approaches to handle this issue include replacing the missing data with the average value from the available historical data or deleting the entire row(Khattab et al., 2023).

Category Data (Lable encoding)

Classification involves grouping features into distinct categories. In this study, the data is initially in string format, which is converted into nominal values for processing. For instance, class A might be converted to class H, class B to class P, and so forth. This study uses label encoding to handle categorical data by transforming it into numerical form. Label encoding assigns numerical labels to categorical text values, converting all textual features into corresponding numbers for seamless data processing(Dahouda & Joe, 2021).

Outlier Detection

Anomaly detection, a crucial aspect of data processing, identifies data points or objects that differ significantly from the rest of the dataset (Bergmann et al., 2021). Outliers can heavily influence the mean, increase the standard deviation, and drastically alter the overall data distribution. Ideally, in the absence of outliers, the data should follow a normal distribution. However, the presence of outliers may distort this, resulting in a left-skewed (negative skewness) or right-skewed (positive skewness) distribution(Jones, 2019).

Data Balancing Technique

In machine learning, the issue of imbalanced data occurs when the number of observations in one class is significantly lower than in the other, which can lead to bias towards the majority class and affect classification performance (Fornacon-Wood et al., 2020). To address class imbalance, data balancing techniques are crucial, especially in the healthcare field, where class distribution can impact model performance.

Various approaches have been proposed to improve classification accuracy(Ferdinandy et al., 2020).

Oversampling Technique

To address the class imbalance issue, where one class significantly outweighs the other, a data augmentation method known as oversampling aims to rebalance the training data distribution by increasing the number of examples in the underrepresented class(Khan et al., 2024). Several Oversampling Techniques that are commonly used include:

Random Over Sampling

In class imbalance scenarios, the Random Over Sampling method is employed to address the issue. This method involves randomly duplicating samples from the minority class to balance the class distribution with the majority class (Hasanin et al., 2019). The duplication can be performed with or without replacement, ensuring the minority class sample size matches or closely approximates that of the majority class. This approach balances the dataset while preserving the original data structure, as no modifications are made to the existing data(Tran et al., 2021).

SMOTE (Synthetic Minority Over-sampling Technique)

The Synthetic Minority Over-sampling Technique (SMOTE) is used to address class imbalance by generating synthetic samples through interpolation, which is different from random oversampling that simply duplicates minority samples. SMOTE reduces the risk of overfitting by producing more varied data, helping the model identify new patterns(Rendón et al., 2020; Yan et al., 2019).

Undersampling Technique

The undersampling method is used to address class imbalance in datasets. The main goal of this method is to create a more balanced class distribution so that machine learning models can learn patterns from both classes fairly. This method is used to reduce the number of samples from the majority class until it is close to the number of samples from the minority class(Dablain et al., 2023).

Random Under Sampling

Undersampling is a technique that involves randomly sampling from the majority class and adding it to the minority class to create a new training dataset(Fujiwara et al., 2020). To achieve a more balanced proportion between the majority and minority classes, this method reduces the number of samples from the majority class through random sampling.

Cluster Centroids

Cluster Centroids Undersampling is a data balancing method that falls under the category of undersampling. However, this method differs from random undersampling. It reduces the majority class data by using centroids, or central points, of several clusters formed through clustering algorithms like K-Means(Zhang et al., 2019).

Hybrid Sampling Technique

A larger majority class compared to the minority class can lead to bias in the predictive model when handling imbalanced data. SMOTE-Tomek is an excellent combined technique to address this issue. It integrates two popular methods: SMOTE for augmenting the minority class data and Tomek Links for cleaning the data from noise and redundancy(Hairani et al., 2023).

Data splitting

After the data is processed into a usable form, data splitting—training and testing—is performed in machine learning. The training data is used to train the classification model, while the testing data is used to evaluate it. This data splitting, known as the hold-out method, divides the dataset into proportions such as 80% for training and 20% for testing (J. Li et al., 2024).

Transformasi Data

During the data processing stage, feature scaling plays a crucial role in standardizing independent variables or the range of features in the data(Wang et al., 2021). In feature scaling, we can use the normalization technique. This technique normalizes the feature column data within the range of [0,1] using Z-score Normalization (Standardization). This method standardizes features so they have a mean of 0 and a standard deviation of 1. Essentially, this method uses a Z-transformation by subtracting each observed value of a variable by its mean and then dividing the result by the standard deviation of the variable(Peng et al., 2019). The formula for standard transformation can be seen in Equation (1).

$$Z = \frac{(x - \mu)}{\sigma} \quad (1)$$

Explanation:

- z = normalized data value.
- x = original data value.
- μ = mean of the data.
- σ = standard deviation of the data.

Model Machine Learning

In this study, after the data transformation process and separating the data into test and train datasets, the next step is classification using Machine Learning algorithms, including Decision Tree, K-Nearest Neighbors, and Random Forest. These algorithms are employed for stroke prediction classification.

Decision Tree

Decision Tree is a machine learning technique for regression and classification that splits data into subsets based on attribute-based decision rules(Charbuty & Abdulazeez, 2021). The tree consists of a root node (the initial feature for splitting), internal nodes (tests on attributes), and leaves (prediction outcomes). The model is built recursively until conditions such as uniform labels at a node or the exhaustion of features for splitting are met. Decision Trees excel at producing models that are easy to understand and interpret(ÇETİNKAYA & HORASAN, 2021).

K-Nearest Neighbors

The K-Nearest Neighbors (KNN) algorithm is a non-parametric method for regression and classification, which operates by identifying the k nearest neighbors using distance metrics such as Euclidean or Manhattan(Kiyak & Ghasemkhani, 2023). Classification is based on the majority class of the neighbors, while regression uses the average target value. KNN is suitable for various datasets but has drawbacks such as high computational cost for large datasets and sensitivity to the choice of k and feature scaling. Feature normalization and selecting an appropriate k are essential for optimal results(Uddin et al., 2022).

Random Forest

Random Forest is an ensemble-based machine learning algorithm that enhances prediction accuracy by combining multiple decision trees(J. Chen et al., 2024). The algorithm employs the bagging method to train each tree on random subsets of data and features, thereby reducing the risk of overfitting. The final prediction is obtained through majority voting for classification or averaging for regression. Random Forest excels in handling numerous features, missing data, and outliers, and it provides feature importance for further data analysis(Yadav & Pal, 2020).

Evaluating

The final stage of this study involves comparing the six algorithms based on accuracy, recall, precision, and f-measure derived from the data balancing techniques used. This evaluation process aims to assess how changes in data balancing techniques affect the model's ability to predict stroke, especially on datasets with class imbalance. The evaluation of the stroke prediction model's accuracy is done using the Classification Report and Confusion Matrix.

Classification Report

A Classification Report is an evaluation tool in machine learning that provides a detailed overview of a classification model's performance. This report shows the results of prediction analysis, such as accuracy, precision, recall, F1-score, and support. This evaluation is important for understanding how well the model predicts each category, especially on imbalanced datasets (Noaman et al., 2024).

Accuracy is the most basic and commonly used evaluation metric to assess the performance of a classification model. It measures the proportion of correct predictions compared to the total number of samples in the dataset (Vujović, 2021).

$$Accuracy = \frac{True\ Positives\ (TP) + True\ Negatives\ (TN)}{TP+TN+FP+\ FN}\ x\ 100 \qquad (2)$$

Precision is a measure of the number of correct positive predictions made by the model. A high precision indicates that most of the positive predictions made by the model are correct. It is calculated by dividing the number of true positive predictions (TP) by the total number of positive predictions made by the system(Vujović, 2021).

$$Precision = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) +\ False\ Positives\ (FP)}\ x\ 100 \qquad (3)$$

Recall, also known as sensitivity, is a measure of the number of actual positive data points that are correctly identified by the model. A high recall indicates that the model successfully captures most of the positive data. It is calculated by dividing the total number of actual positive data points and the true positive data points.

$$Precision = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Positives\ (FP)}\ x\ 100 \qquad (4)$$

The F1 Score, which is the harmonic mean of recall and precision, provides a balanced view between the two metrics. It is useful when there is a need to minimize both False Positives and False Negatives in a balanced way(Vujović, 2021).

$$F1 - score = 2\ x \frac{Precision\ x\ Recall}{Precision + Recall} \qquad (5)$$

The support represents the actual number of samples for each class, which is an important metric for understanding the data distribution(Vujović, 2021). It helps in evaluating how well the model performs for each class, especially in imbalanced datasets.

Confusion Matrix

A confusion matrix is a crucial evaluation tool for machine learning classification that provides detailed information about the model's performance. It consists of four key elements—True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN)—that show the relationship between the model's predictions and the actual labels (Krstinić et al., 2020). The confusion matrix table is presented in Figure 2 below.

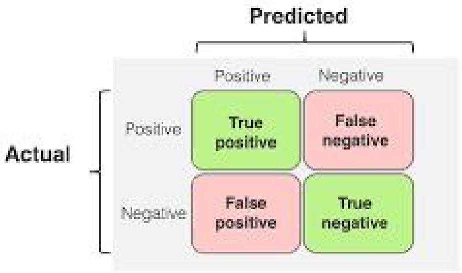

Figure 2. Confusion Matrix

RESULTS AND DISCUSSION

Using an imbalanced dataset for stroke prediction, with class 0 (No Stroke) making up about 98% of the total data, while class 1 (Stroke) accounts for only about 2% of the total data, as shown in Figure 3.

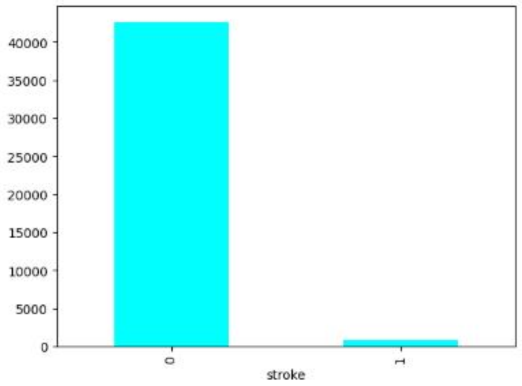

Figure 3. Unbalanced target class distribution

Data imbalance can lead to a model disproportionately predicting the majority class (No Stroke) while neglecting the minority class (Stroke), reducing its effectiveness in detecting rare diseases. To mitigate this, various data balancing techniques were implemented, including Random Over Sampling, SMOTE, Random Under Sampling, Cluster-Based Sampling, and Hybrid Sampling, specifically SMOTETOMEK. Random Over Sampling and SMOTE were employed to increase the sample size of the minority class, balancing it with the majority class. The outcomes of these techniques are illustrated in Figures 4 and 5.

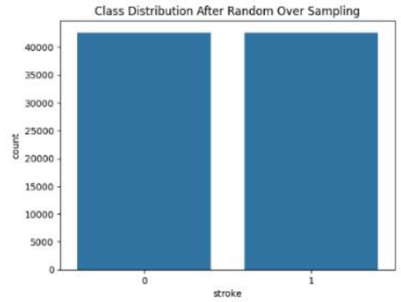

Figure 4. Target class distribution after Random Over Sampling

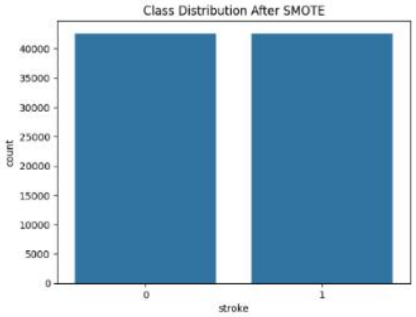

Figure 5. Target class distribution after SMOTE

Unlike Random Over Sampling and SMOTE, the Random Under Sampling and Cluster Based

Sampling techniques are used to reduce the number of samples in the majority class to balance it with the minority class. The results of the Random Under Sampling and Cluster Based Sampling techniques can be seen in Figure 6 and Figure 7.

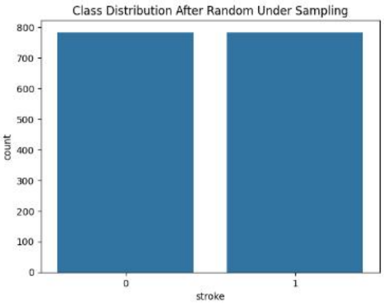

Figure 6. Target class distribution after Random Under Sampling

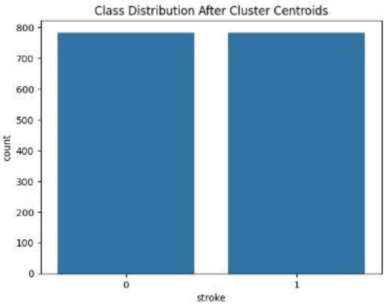

Figure 7. Target class distribution after Cluster Based Sampling

In addition to using Over Sampling and Under Sampling methods, the issue of class imbalance in the stroke dataset is addressed by implementing a combination of two data balancing techniques, namely SMOTE and TOMEK Links. SMOTE is used to generate synthetic samples for the minority class, while TOMEK Links is used to identify and reduce the noise that arises from oversampling. The results of the combined SMOTE-TOMEK technique can be seen in Figure 8.

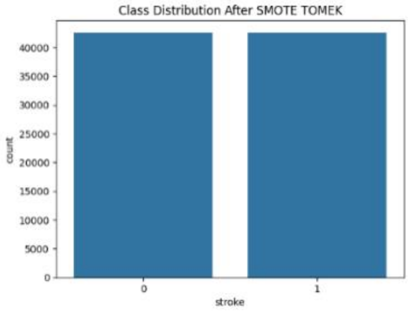

Figure 8. Target class distribution after SMOTETOMEK

Table 1 shows the results reveal that data balancing strategies have a considerable impact on stroke prediction accuracy across multiple machine learning models. Random Forest (RF) regularly outperforms other approaches, with 98% accuracy except for SMOTE and SMOTETomek (93%), demonstrating its ability to handle imbalanced data. K-Nearest Neighbours (KNN) exhibits a significant decline in accuracy with Random Under Sampling (70%), implying that lowering data hurts its performance. Meanwhile, Decision Tree (DT) benefits the most from SMOTE and SMOTETomek (81%), demonstrating that oversampling increases predictive power.

Oversampling approaches (SMOTE, SMOTETomek, and Random Over Sampling) improve model performance, particularly for DT and KNN. Random Under Sampling, on the other hand, produces the lowest accuracy across all models, especially for KNN and RF. Cluster-Centroid retains high accuracy for KNN and RF (98%), but has no meaningful impact on DT.

Table 1. Result of Accuracy Each Algorithm

Algoritma	DT	KNN	RF
Dataset without Balancing	75%	98%	98%
Random Over Sampling	74%	93%	98%
SMOTE	81%	86%	93%
Random Under Sampling	71%	70%	72%
Cluster-Centroid	75 %	98%	98%
SMOTETomek	81%	86%	93%

Table 2 compares the precision of data balancing approaches with stroke prediction accuracy across multiple machine learning models. Random Forest (RF) has the highest precision (16.67%) when compared to Cluster-Based and the original dataset, showing that these methods retain important predictive features. However, oversampling techniques such as SMOTE (6.28%) and SMOTETomek (5.99%) reduce RF precision, indicating potential overfitting or misclassification issues.

Decision Tree (DT) had the highest precision with SMOTE and SMOTETomek (6.15%), demonstrating that these techniques improve minority class recognition. Meanwhile, K-Nearest Neighbours (KNN) performs poorly with 0% accuracy in the original dataset and Cluster-Based techniques, but marginally better with Random Over Sampling (5.75%) and SMOTE (5.21%).

Table 2. Result of Precision Each Algorithm

Algoritma	DT	KNN	RF
Dataset without Balancing	4.85%	0%	16.67%
Random Over Sampling	5.35%	5.75%	8.70%
SMOTE	6.15%	5.21%	6.28%
Random Under Sampling	4.84%	4.61%	4.95%
Cluster-Based	4.85%	0%	16.67%
SMOTETomek	6.15%	5.12%	5.99%

Table 3 shows the comparison results of recall that Random Forest (RF) has the highest recall (80.89%) with the original dataset, but it is significantly lower with oversampling techniques such as Random Over Sampling (2.48%) and Cluster-Centroid (1.27%), indicating a problem with classification. Decision Tree (DT) performed best with Random Under Sampling (80.89%), while SMOTE and SMOTETomek (63.98%) reduced recall, indicating that oversampling can result in noise. K-Nearest Neighbours (KNN) has a low recall on the original dataset (1,91%), but it improves with SMOTE and SMOTETomek (38.51%) and reaches 79.62% with Random Under Sampling.

Table 3. Result of Recall Each Algorithm

Algoritma	DT	KNN	RF
Dataset without Balancing	70.06%	1.91%	80.89%
Random Over Sampling	76.40%	17.39%	2.48%
SMOTE	63.98%	38.51%	19.25%
Random Under Sampling	80.89 %	79.62 %	78.98 %
Cluster-Centroid	70.06%	0%	1.27%
SMOTETomek	63.98%	38.51%	20.50%

The F1-score results in table 4 show that SMOTE and SMOTETomek produce the highest scores for Decision Tree (DT) (11.21%), K-Nearest Neighbours (KNN) (9.18% and 9.04%), and Random Forest (RF) (9.47% and 9.27%, respectively). This shows that these oversampling algorithms successfully balance precision and recall, resulting in improved overall performance. In comparison, KNN struggles with the original dataset and Cluster-Centroid (0%), illustrating its difficulty dealing with imbalanced data. RF performs slightly better with SMOTE (9.47%) than with the original dataset (2.37%), implying that oversampling improves predictive balance.

Table 4. Result of F1-Score Each Algorithm

Algoritma	DT	KNN	RF
Dataset without Balancing	9.07%	0%	2.37%
Random Over Sampling	9.99%	8.64%	3.86%
SMOTE	11.21%	9.18%	9.47%
Random Under Sampling	9.13%	8.71%	9.31%
Cluster-Centroid	9.07%	0%	2.37%
SMOTETomek	11.21%	9.04%	9.27%

The following figure depicts the evaluation results of the stroke prediction model using the Confusion Matrix:

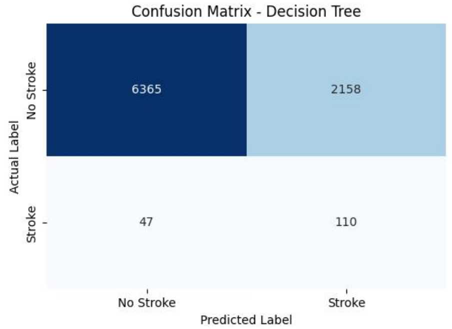

Figure 9. Confusion Matrix of Decision Tree Algorithm Without Balancing Technique

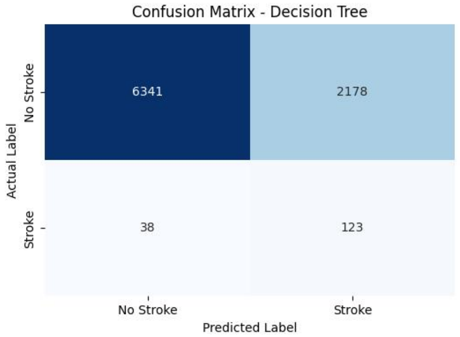

Figure 10. Confusion Matrix of Decision Tree Algorithm With Balancing Technique Random Over Sampling

Figure 9 displays the confusion matrix for the Decision Tree model before data balancing, whereas Figure 10 shows its performance after random oversampling. Prior to balancing, the model accurately identified 6,365 non-stroke cases while misclassifying 2,158 as strokes, resulting in a significant proportion of false positives. In stroke instances, 47 were mistakenly labelled as non-stroke, while only 110 were correctly identified. This demonstrates that the model struggled to detect actual strokes, most likely because of the skewed sample.

Figure 10 shows that the model's stroke detection improves marginally when Random Over

Sampling is used. The number of successfully diagnosed stroke patients increased from 110 to 123, while false negatives (missed stroke cases) decreased from 47 to 38, demonstrating improved sensitivity. However, false positives increased slightly, from 2,158 to 2,178, indicating that oversampling may have resulted in noise. Overall, Random Over Sampling improved stroke identification while only slightly affecting non-stroke misclassification, making it a viable technique for dealing with class imbalance in stroke prediction.

CONCLUSIONS AND SUGGESTIONS

Conclusion

The study's findings demonstrate that Random Forest (RF) outperforms most approaches in terms of accuracy (98%), showing great overall performance. However, Decision Tree (DT) improves the most from SMOTE and SMOTETomek, improving its accuracy to 81%, whilst K-Nearest Neighbours (KNN) suffers greatly with Random Under Sampling (70%), demonstrating its sensitivity to data reduction. In terms of accuracy, RF outperforms the original dataset (16.67%), but its precision declines with balancing strategies, implying a trade-off between recall and false positive reduction. SMOTE provides the maximum precision for DT and KNN (6.15% and 5.21%, respectively), demonstrating that oversampling improves minority class recognition.

On the original dataset, RF has the highest recall (80.89%), whereas DT and KNN improve with Random Under Sampling (80.89% and 79.62%), demonstrating that undersampling improves stroke detection. However, this is at the expense of precision. F1-scores are highest for all models that use SMOTE and SMOTETomek, demonstrating that these oversampling strategies provide the optimum balance of precision and recall. Overall, SMOTE and SMOTETomek increase DT and KNN performance, whereas RF remains the most accurate, but suffers from accuracy and recall trade-offs when balancing approaches are used.

Suggestion

In future research, it is recommended to improve stroke prediction performance, Random Forest (RF) should be used when prioritizing overall accuracy, but additional tuning is needed to balance precision and recall. Decision Tree (DT) and K-Nearest Neighbors (KNN) benefit the most from SMOTE and SMOTETomek, making these oversampling techniques preferable when aiming for better F1-scores. However, Random Under Sampling enhances recall for all models but lowers precision, suggesting that a hybrid approach combining oversampling and undersampling (e.g., SMOTETomek) could optimize results. Further improvements can be achieved by experimenting with hyperparameter tuning, ensemble learning, and feature selection to enhance model stability and predictive power. Additionally, alternative data-balancing techniques, such as ADASYN or cost-sensitive learning, should be explored to reduce false positives while maintaining high recall for stroke detection.

REFERENCES

Al Hashmi, A. M., Shuaib, A., Imam, Y., Amr, D., Humaidan, H., Al Nidawi, F., Sarhan, A., Mustafa, W., Khalefa, W., Ramadan, I., Usman, F. S., Hokmabadi, E. S., Ghorbani, M., Nassir, T., Aladham, F., Salmeen, A., Kikano, R., Muda, S., Jose, S., … Mansour, O. Y. (2022). Stroke services in the Middle East and adjacent region: A survey of 34 hospital-based stroke services. *Frontiers in Neurology*, *13*. https://doi.org/10.3389/fneur.2022.1016376

Auer, R. N., & Sommer, C. J. (2021). Histopathology of Brain Tissue Response to Stroke and Injury. *Stroke: Pathophysiology, Diagnosis, and Management*, *November*, 0–8. https://doi.org/10.1016/B978-0-323-69424-7.00004-1

Avan, A., & Hachinski, V. (2021). Stroke and dementia, leading causes of neurological disability and death, potential for prevention. *Alzheimer's and Dementia*, *17*(6), 1072–1076. https://doi.org/10.1002/alz.12340

Bergmann, P., Batzner, K., Fauser, M., Sattlegger, D., & Steger, C. (2021). The MVTec Anomaly Detection Dataset: A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection. *International Journal of Computer Vision*, *129*(4), 1038–1059. https://doi.org/10.1007/s11263-020-01400-4

Bohr, A., & Memarzadeh, K. (2020). The rise of artificial intelligence in healthcare applications. In *Artificial Intelligence in Healthcare* (Issue January). https://doi.org/10.1016/B978-0-12-818438-7.00002-2

ÇETİNKAYA, Z., & HORASAN, F. (2021). Decision Trees in Large Data Sets. *Uluslararası Muhendislik Arastirma ve Gelistirme Dergisi*, *13*(1), 140–151. https://doi.org/10.29137/umagd.763490

Charbuty, B., & Abdulazeez, A. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends*, *2*(01), 20–28. https://doi.org/10.38094/jastt20165

Chen, J., Du, H., Mao, F., Huang, Z., Chen, C., Hu, M., & Li, X. (2024). Improving forest age prediction performance using ensemble learning algorithms base on satellite remote sensing data. *Ecological Indicators*, *166*(April), 112327. https://doi.org/10.1016/j.ecolind.2024.112327

Chen, S., Shao, L., & Ma, L. (2021). Cerebral Edema Formation After Stroke: Emphasis on Blood–Brain Barrier and the Lymphatic Drainage System of the Brain. *Frontiers in Cellular Neuroscience*, *15*(August), 1–17. https://doi.org/10.3389/fncel.2021.716825

Dablain, D., Krawczyk, B., & Chawla, N. V. (2023). DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(9), 6390–6404. https://doi.org/10.1109/TNNLS.2021.3136503

Dahouda, M. K., & Joe, I. (2021). A Deep-Learned Embedding Technique for Categorical Features Encoding. *IEEE Access*, *9*, 114381–114391. https://doi.org/10.1109/ACCESS.2021.3104357

Duncan, P. W., Bushnell, C., Sissine, M., Coleman, S., Lutz, B. J., Johnson, A. M., Radman, M., Pvru Bettger, J., Zorowitz, R. D., & Stein, J. (2021). Comprehensive Stroke Care and Outcomes: Time for a Paradigm Shift. *Stroke*, *52*(1), 385–393. https://doi.org/10.1161/STROKEAHA.120.029678

Elfa, M. A. A., & Dawood, M. E. T. (2023). Using Arti fi cial Intelligence for Enhancing. *Journal of Art, Design & Music*, *2*(2).

Ferdinandy, B., Gerencsér, L., Corrieri, L., Perez, P., Újváry, D., Csizmadia, G., & Miklósi, Á. (2020). Challenges of machine learning model validation using correlated behaviour data: Evaluation of cross-validation strategies and accuracy measures. *PLoS ONE*, *15*(7), 1–14. https://doi.org/10.1371/journal.pone.0236092

Fornacon-Wood, I., Mistry, H., Ackermann, C. J., Blackhall, F., McPartlin, A., Faivre-Finn, C., Price, G. J., & O'Connor, J. P. B. (2020). Reliability and prognostic value of radiomic features are highly dependent on choice of feature extraction platform. *European Radiology*, *30*(11), 6241–6250. https://doi.org/10.1007/s00330-020-06957-9

Fujiwara, K., Huang, Y., Hori, K., Nishioji, K., Kobayashi, M., Kamaguchi, M., & Kano, M. (2020). Over- and Under-sampling Approach for Extremely Imbalanced and Small Minority Data Problem in Health Record Analysis. *Frontiers in Public Health*, *8*(May), 1–15. https://doi.org/10.3389/fpubh.2020.00178

Ganesha, H. R., & Aithal, P. S. (2022). How to Choose an Appropriate Research Data Collection Method and Method Choice Among Various Research Data Collection Methods and Method Choices During Ph.D. Program in India? *International Journal of Management, Technology, and Social Sciences*, *7*(2), 455–489. https://doi.org/10.47992/ijmts.2581.6012.0233

Hairani, H., Anggrawan, A., & Priyanto, D. (2023). Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link. *International Journal on Informatics Visualization*, *7*(1), 258–264. https://doi.org/10.30630/joiv.7.1.1069

Hasanin, T., Khoshgoftaar, T. M., Leevy, J. L., & Bauder, R. A. (2019). Severely imbalanced Big Data challenges: investigating data sampling approaches. *Journal of Big Data*, *6*(1). https://doi.org/10.1186/s40537-019-0274-4

Ivanov, I. G., Kumchev, Y., & Hooper, V. J. (2023). An Optimization Precise Model of Stroke Data to Improve Stroke Prediction. *Algorithms*, *16*(9), 1–16. https://doi.org/10.3390/a16090417

Jäger, S., Allhorn, A., & Bießmann, F. (2021). A Benchmark for Data Imputation Methods. *Frontiers in Big Data*, *4*(July), 1–16. https://doi.org/10.3389/fdata.2021.693674

Jassim, M. A., & Abdulwahid, S. N. (2021). Data Mining preparation: Process, Techniques and Major Issues in Data Analysis. *IOP Conference Series: Materials Science and Engineering*, *1090*(1), 012053. https://doi.org/10.1088/1757-899x/1090/1/012053

Johnson, T. F., Isaac, N. J. B., Paviolo, A., & González-Suárez, M. (2021). Handling missing values in trait data. *Global Ecology and Biogeography*, *30*(1), 51–62. https://doi.org/10.1111/geb.13185

Jones, P. R. (2019). A note on detecting statistical outliers in psychophysical data. *Attention, Perception, and Psychophysics*, *81*(5), 1189–1196. https://doi.org/10.3758/s13414-019-01726-3

Khan, Z., Ali, A., & Aldahmani, S. (2024). Feature Selection via Robust Weighted Score for High Dimensional Binary Class-Imbalanced Gene Expression Data. *Heliyon*, *10*(19), e38547. https://doi.org/10.1016/j.heliyon.2024.e38547

Khattab, A. A. R., Elshennawy, N. M., & Fahmy, M. (2023). GMA: Gap Imputing Algorithm for time series missing values. *Journal of Electrical Systems and Information Technology*, *10*(1), 1–20. https://doi.org/10.1186/s43067-023-00094-1

Kiyak, E. O., & Ghasemkhani, B. (2023). High-Level K-Nearest Neighbors (HLKNN): A Supervised. *Electronics*, *12*, 1–20.

Krstinić, D., Braović, M., Šerić, L., & Božić-Štulić, D. (2020). *Multi-label Classifier Performance Evaluation with Confusion Matrix*. 01–14. https://doi.org/10.5121/csit.2020.100801

Li, J., Othman, M. S., Chen, H., & Yusuf, L. M. (2024). Optimizing IoT intrusion detection system: feature selection versus feature extraction in machine learning. *Journal of Big Data*, *11*(1). https://doi.org/10.1186/s40537-024-00892-y

Li, W., Yue, T., & Liu, Y. (2020). New understanding of the pathogenesis and treatment of stroke-related sarcopenia. *Biomedicine and Pharmacotherapy*, *131*(September), 110721. https://doi.org/10.1016/j.biopha.2020.110721

Murphy, S. J., & Werring, D. J. (2020). Stroke: causes and clinical features. *Medicine (United Kingdom)*, *48*(9), 561–566. https://doi.org/10.1016/j.mpmed.2020.06.002

Musmar, B., Adeeb, N., Ansari, J., Sharma, P., & Cuellar, H. H. (2022). Endovascular Management of Hemorrhagic Stroke. *Biomedicines*, *10*(1). https://doi.org/10.3390/biomedicines10010100

Niles, J., Bhasin, G., & Ganti, L. (2024). Large right middle cerebral artery stroke with hemorrhagic transformation. *International Journal of Emergency Medicine*, *17*(1). https://doi.org/10.1186/s12245-024-00739-6

Nizam-Ozogur, H., & Orman, Z. (2024). A heuristic-based hybrid sampling method using a combination of SMOTE and ENN for imbalanced health data. *Expert Systems*, *41*(8), 1–22. https://doi.org/10.1111/exsy.13596

Noaman, A. Y., Gad-Elrab, A. A. A., & Baabdullah, A. M. (2024). Towards Scientists and Researchers Classification Model (SRCM)-based machine learning and data mining methods: An ISM-MICMAC approach. *Journal of Innovation and Knowledge*, *9*(3), 100516. https://doi.org/10.1016/j.jik.2024.100516

Peng, M., Zhang, Q., Xing, X., Gui, T., Huang, X., Jiang, Y. G., Ding, K., & Chen, Z. (2019). Trainable undersampling for class-imbalance learning. *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, 4707–4714. https://doi.org/10.1609/aaai.v33i01.33014707

Rendón, E., Alejo, R., Castorena, C., Isidro-Ortega, F. J., & Granda-Gutiérrez, E. E. (2020). Data sampling methods to dealwith the big data multi-class imbalance problem. *Applied Sciences (Switzerland)*, *10*(4). https://doi.org/10.3390/app10041276

Sailasya, G., & Kumari, G. L. A. (2021). Analyzing the Performance of Stroke Prediction using ML Classification Algorithms. *International Journal of Advanced Computer Science and Applications*, *12*(6), 539–545. https://doi.org/10.14569/IJACSA.2021.0120662

Shneiderman, B. (2020). Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Transactions on Interactive Intelligent Systems*, *10*(4). https://doi.org/10.1145/3419764

Sohn, K., & Kwon, O. (2020). Technology acceptance theories and factors influencing artificial Intelligence-based intelligent products. *Telematics and Informatics*, *47*, 101324. https://doi.org/10.1016/j.tele.2019.101324

Tran, N., Chen, H., Jiang, J., Bhuyan, J., & Ding, J. (2021). Effect of class imbalance on the performance of machine learning-based network intrusion detection. *International Journal of Performability Engineering*, *17*(9), 741–755. https://doi.org/10.23940/ijpe.21.09.p1.741755

Uddin, S., Haque, I., Lu, H., Moni, M. A., & Gide, E. (2022). Comparative performance analysis of

K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. *Scientific Reports*, *12*(1), 1–11. https://doi.org/10.1038/s41598-022-10358-x

Vujović, Ž. (2021). Classification Model Evaluation Metrics. *International Journal of Advanced Computer Science and Applications*, *12*(6), 599–606. https://doi.org/10.14569/IJACSA.2021.0120670

Wang, S., Dai, Y., Shen, J., & Xuan, J. (2021). Research on expansion and classification of imbalanced data based on SMOTE algorithm. *Scientific Reports*, *11*(1), 1–11. https://doi.org/10.1038/s41598-021-03430-5

Woodward, M. (2019). Cardiovascular disease and the female disadvantage. *International Journal of Environmental Research and Public Health*, *16*(7). https://doi.org/10.3390/ijerph16071165

Xiao, H. H., Yang, W. K., Hu, J., Zhang, Y. P., Jing, L. J., & Chen, Z. Y. (2022). Significance and methodology: Preprocessing the big data for machine learning on TBM performance. *Underground Space (China)*, *7*(4), 680–701. https://doi.org/10.1016/j.undsp.2021.12.003

Yadav, D. C., & Pal, S. (2020). Prediction of heart disease using feature selection and random forest ensemble method. *International Journal of Pharmaceutical Research*, *12*(4), 56–66. https://doi.org/10.31838/ijpr/2020.12.04.013

Yan, Y., Tan, M., Xu, Y., Cao, J., Ng, M., Min, H., & Wu, Q. (2019). Oversampling for imbalanced data via optimal transport. *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, 5605–5612. https://doi.org/10.1609/aaai.v33i01.33015605

Zeng, Z., Chen, P. J., & Lew, A. A. (2020). From high-touch to high-tech: COVID-19 drives robotics adoption. *Tourism Geographies*, *22*(3), 724–734. https://doi.org/10.1080/14616688.2020.1762118

Zhang, J., Chen, L., & Abid, F. (2019). Prediction of Breast Cancer from Imbalance Respect Using Cluster-Based Undersampling Method. *Journal of Healthcare Engineering*, *2019*. https://doi.org/10.1155/2019/7294582