JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

STATISTICAL APPROACHES FOR ACCURATE CEREBRAL STROKE PREDICTION IN IMBALANCED DATA ENVIRONMENTS

Prajna R

Department of PG Studies and Research in Statistics, Mangalore University, Karnataka, India

Abstract : This research focuses on predicting cerebral strokes within imbalanced data contexts, addressing the critical need for early detection through statistical methods. The study identifies stroke risk factors, develops and evaluates precision-oriented classification models (e.g., logistic regression, machine learning), and effectively manages data imbalances. Using the Kaggle Cerebral Stroke dataset with 12 attributes and imbalanced target variable, this investigation examines predictors like gender, age, hypertension, heart disease, marital status, work type, residence type, glucose level, BMI, and smoking status. Previous studies on stroke prediction using Naïve Bayes, decision trees, and neural networks are thoroughly reviewed. The research reveals key risk determinants and employs six data balancing techniques (ROSE, SMOTE, ADASYN, SVM-SMOTE, SMOTEEN, SMOTETOMEK), rigorously evaluating six classification models (Logistic regression, Decision Tree, Support Vector Machine, k-Nearest Neighbor, Random Forest, Naïve Bayes). Notably, combining ADASYN and KNN significantly enhances cerebral stroke prediction accuracy. This study advances early stroke prediction by leveraging advanced statistical techniques to mitigate imbalanced data challenges, holding potential to improve interventions and expedite timely medical responses.

Keywords: cerebral stroke, imbalanced data, statistical analysis, stroke risk factors, logistic regression, machine learning, data imbalances, Naïve Bayes, decision trees, neural networks, data balancing techniques, ROSE, SMOTE, ADASYN, SVM-SMOTE, accuracy.

I. INTRODUCTION

Cerebral stroke emerges as a critical medical condition, arising from disruptions in blood flow to specific brain regions, which deprive cells of vital nutrients and oxygen, ultimately leading to their demise. This demands immediate attention, with early detection and appropriate management crucial for minimizing damage to the affected brain area and mitigating complications.

Globally, cerebral stroke poses a significant threat to public health, marked by substantial morbidity, disability, and mortality. Disability-adjusted life years (DALYs) attributable to stroke rank second after ischemic heart disease, as highlighted by research like GBD1. The World Health Organization (WHO) reports fifteen million stroke cases annually worldwide, contributing to frequent fatalities. In the United States, stroke ranks as the sixth leading cause of death, accounting for around 11% of total fatalities, while in India, it stands as the fourth leading cause.

Cerebral strokes manifest as ischemic and hemorrhagic types. Ischemic strokes, more prevalent, result from blocked brain blood vessels due to clots or cholesterol buildup. Hemorrhagic strokes involve ruptured vessels, releasing blood into nearby tissues and increasing pressure on adjacent brain areas, exacerbating damage.

Clinical signs include weakness or numbness on one side of the face, impaired speech, visual disturbances, dizziness, balance issues, motor impairments, fainting, seizures, and sudden severe headaches.

Risk factors span modifiable elements like high blood pressure, heart disease, diabetes, smoking, oral contraceptive use, history of transient ischemic attacks, high red blood cell count, and elevated blood cholesterol. Unmodifiable factors comprise age, race, gender, prior stroke history, and hereditary predisposition.

Treatment strategies encompass clot-dissolving medications, therapies for brain swelling, neuroprotective drugs, life support, and surgical interventions like craniotomy. Swift clot-dissolving medication use within three hours of onset is pivotal. Comprehending cerebral stroke's intricacies, risk factors, presentations, and treatments is essential for healthcare professionals and the public to combat this condition effectively.

Furthermore, when considering binary classification for medical conditions like stroke, data imbalance often arises, where one class (e.g., stroke occurrence) is significantly rarer than the other (non-stroke). Addressing this imbalance is crucial to prevent the model from favoring the majority class and potentially missing critical cases. Techniques like oversampling, undersampling, and synthetic data generation can be employed to balance the data and improve the classifier's performance and generalization. These techniques ensure that both classes receive adequate representation, thus enhancing the model's accuracy in identifying stroke cases.

II. RELATED WORKS

In machine learning, data is crucial for training the model. In the real world, we constantly encounter the problem of imbalanced data. This section discusses the work completed towards the efficiency of some of the machine learning techniques while dealing with the different clinical datasets, as most of the clinical datasets are inherently imbalanced in nature. Various algorithms are designed to get rid of the consequences of imbalance. The very popular algorithms are studied and analyzed for the balancing of the datasets, and afterward, the different techniques of machine learning are employed to check their performances.

M. Mostafizur Rahman and D. N. Davis proposed a modified cluster-based under-sampling method for balancing the data, and a training set of good quality is generated for constructing classification models. SMOTE offers a new technique for oversampling. The blend of undersampling and SMOTE gives better performance than plain undersampling. SMOTE was applied on various datasets having variable imbalance degree and training datasets in different amounts, which provides a diverse test field.

Kumar et al. (2020) “Addressing Binary Classification over Class Imbalanced Clinical Datasets Using Computationally Intelligent Techniques” had applied of six classifiers, namely Decision Tree, k-Nearest Neighbour, Logistic regression, Artificial Neural Network, Support Vector Machine, and Gaussian Naïve Bayes, over five imbalanced clinical datasets, Breast Cancer Disease, Coronary Heart Disease, Indian Liver Patient, Pima Indians Diabetes Database, and Coronary Kidney Disease, with respect to seven different class balancing techniques, namely Under sampling, Random oversampling, SMOTE, ADASYN, SVM-SMOTE, SMOTEEN, and SMOTETOMEK . Result analysis demonstrates that SMOTEEN balancing method often performed better over all the other six data-balancing techniques with all six classifiers and for all five clinical datasets.

III. DESCRIPTION OF DATA-BALANCING ALGORITHMS

3.1 ROSE

In the realm of machine learning, "ROSE" stands for "Random Over-Sampling Examples." It's a technique used to counter class imbalance in datasets, a situation where one class has notably fewer samples than others. Such imbalances can hinder accurate learning, as algorithms lean towards the majority class, leading to suboptimal performance on minority classes.

ROSE focuses on the minority class and involves generating synthetic data. Here's how it works:

1. Identify the Minority Class: Pinpoint the class with fewer samples, which is considered the minority.

2. Select Samples: Randomly pick minority class samples for synthetic data generation. The number of synthetic samples depends on the desired balance or extent of imbalance correction.

3. Nearest Neighbors: For each chosen sample, find its K-nearest neighbors from existing minority samples, based on a distance metric like Euclidean distance. K is a parameter set beforehand.

4. Generate Synthetic Samples: Create synthetic samples by modifying feature values within the range defined by its K-nearest neighbors. This maintains the minority class's essence while introducing diversity.

5. Combine Data:Merge the original data with synthetic samples to form a balanced dataset.

6. Training:Train the model using this balanced dataset. The balanced data enables better learning from both classes, enhancing performance on the minority class during testing.

3.2 SMOTE: Synthetic Minority Oversampling Technique

SMOTE, which stands for Synthetic Minority Oversampling Technique, serves to balance class distribution by increasing minority class instances through replication. It works as follows: SMOTE generates new instances within the minority class by creating virtual training records through linear interpolation between existing instances. And it synthesizes synthetic training records by choosing k nearest neighbors for each minority class example. These records are added to reconstruct data, allowing various classification models to be applied.

Here's how the SMOTE algorithm operates:

1. Identify the k-nearest neighbors for each sample.

2. Randomly select samples from these neighbors.

3. Compute new samples as original samples + difference * random number (between 0 & 1).

4. Incorporate these new samples into the minority class, leading to the creation of a new dataset.

3.3 ADASYN: Adaptive Synthetic Data Generation

ADASYN (Adaptive Synthetic Data Generation) is an algorithm that tackles imbalanced data in machine learning. When one class has significantly fewer samples, model learning can suffer. ADASYN addresses this by generating synthetic data for the minority class, strategically focusing on "hard-to-learn" instances.

Here's how it works:

1. It calculates the minority-to-majority ratio. If this ratio is critically imbalanced, ADASYN is activated.

2. ADASYN determines the required synthetic data to achieve a desired balance post-generation.

3. The algorithm identifies instances in the minority class located within challenging majority-dominated neighborhoods.

4. Synthetic instances are created by blending attributes of instances from these challenging areas, boosting diversity within the minority class.

3.4 SVM-SMOTE (Support Vector Machine Synthetic Minority Over-sampling Technique)

SVM-SOMTE combines the SVM classifier with the SMOTE algorithm to address imbalanced datasets in machine learning. It involves two steps:

1. SVM Classification: The SVM classifier is used to identify the minority class data points and the decision boundary that separates them from the majority class.

2. SMOTE Generation: SMOTE then generates synthetic samples for the minority class along the decision boundary while considering the underlying distribution. This creates more balanced class proportions.

By integrating SVM and SMOTE, SVM-SMOTE improves the generalization of the classifier by increasing the diversity of the minority class while avoiding overfitting. The combination helps SVM to learn the decision boundary more accurately and increases the classifier's performance on imbalanced datasets.

3.5 SMOTE-ENN

SMOTE-ENN is a hybrid data sampling technique used in machine learning to address class imbalance in datasets. It combines two methods: SMOTE (Synthetic Minority Over-sampling Technique) and ENN (Edited Nearest Neighbors).

First, SMOTE generates synthetic examples of the minority class by interpolating between existing instances, thus expanding its representation. This reduces the class imbalance. Then, ENN is employed to clean the dataset by identifying and removing noisy samples, which enhances the quality of the dataset.

The combination of SMOTE and ENN aims to improve classification performance by simultaneously balancing class distribution and enhancing the data's overall quality.

3.6 SMOTE-Tomek

SMOTE-Tomek is a hybrid data sampling technique used in machine learning to address class imbalance. It combines two methods: Synthetic Minority Over-sampling Technique (SMOTE) and Tomek links. SMOTE enhances the minority class by creating synthetic instances, improving its representation. Tomek links, on the other hand, identify pairs of instances from different classes that are in close proximity and are considered noise or ambiguous data points.

The SMOTE-Tomek algorithm first applies Tomek links to remove borderline instances that might cause misclassification. Then, it applies SMOTE to the remaining minority class instances to generate synthetic examples. This process aims to create a balanced dataset while simultaneously reducing noisy data points. The outcome is a more evenly distributed dataset with enhanced class separation, ultimately leading to improved classification performance by reducing bias towards the majority class and removing noise from the data.

IV. DESCRIPTION OF CLASSIFICATION METHODS

An explanation in brief for every classification technique implemented in this study is given below so as to give the fundamental information regarding these classification methods:

4.1. Logistic Regression

Logistic regression is a statistical tool used to predict binary outcomes (e.g., yes/no) in research. Unlike linear regression, it models the probability of an event happening using the logistic function. This function transforms a linear combination of predictor variables into a probability between 0 and 1. The coefficients in the logistic regression equation are determined through methods like maximum likelihood estimation. This approach is valuable when investigating relationships between variables with dichotomous outcomes.

4.2 Decision Tree

Decision trees are vital tools in research for data analysis and prediction. They construct a tree-like structure by recursively dividing data into subsets, aiding decision-making. Internal nodes represent feature-based choices, and leaf nodes offer outcomes. Key features include splitting criteria selection, pruning for simplicity, and showcasing feature importance. Advantages encompass interpretability, nonlinear pattern recognition, and mixed data handling. Applications range from medical diagnoses to financial forecasting and ecological studies. Their transparency and versatility make decision trees a cornerstone in research analytics.

4.3 Support Vector Machines

Support Vector Machines (SVMs) are robust machine learning tools used for classification and regression tasks. SVMs work by identifying the best possible line or plane (hyperplane) that separates different classes in data space. Their effectiveness in various domains, such as computer vision and natural language processing, has contributed to their widespread adoption.

4.4 k-Nearest Neighbors (k-NN)

k-Nearest Neighbors (k-NN) classification assigns a class label to an unlabeled instance based on the classes of its nearest neighboring instances in the feature space. The class is determined by the majority vote among these neighbors. Distance metrics quantify similarity, and the parameter "k" defines the number of neighbors considered.

4.5 Random Forest

Random Forest classification is a powerful machine learning technique that assembles multiple decision tree models to make accurate predictions. Each tree is trained on different subsets of data and features, and their collective outputs determine the final classification. By mitigating overfitting and increasing robustness, Random Forest leverages the wisdom of the crowd to enhance accuracy and handle complex datasets, making it a popular choice for diverse classification tasks in various domains.

4.6 Gaussian Naïve Bayes

Gaussian Naïve Bayes classification is a machine learning approach that assumes features are independent and follow a Gaussian distribution per class. This simplifies probability calculations to determine the most likely class for new data based on Bayes' theorem, making it efficient and effective for classification tasks, particularly when features are approximately normally distributed.

V. PERFORMANCE METRICS OF CLASSIFIERS

The accuracy of the classifier on a given test set is the percentage of test tuples that are correctly classified by the classifier this can be represented in the table called confusion matrix.

True positives (TP): These refer to the positive tuples that were correctly labelled by the classifier. Let TP be the number of true positives.

True negatives (TN): These are the negative tuples that were correctly labelled by the classifier. Let TN be the number of true negatives.

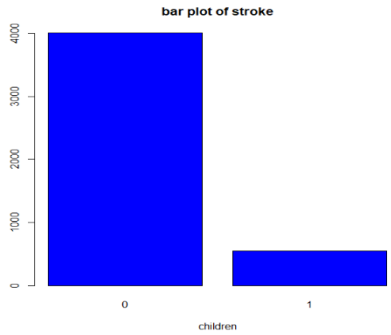
False positives (FP): These are the negative tuples that were incorrectly labelled as positive. Let FP be the number of false positives.

False negatives (FN): These are the positive tuples that were mislabelled as negative. Let FN be the number of false negatives.

$$ACCURACY = \frac{TP+TN}{TP+TN+FP+FN}, \qquad SENSITIVITY = \frac{TP}{TP+FN}, \qquad SPECIFICITY = \frac{TN}{TN+FP}$$

VI. RESULTS AND DISCUSSION

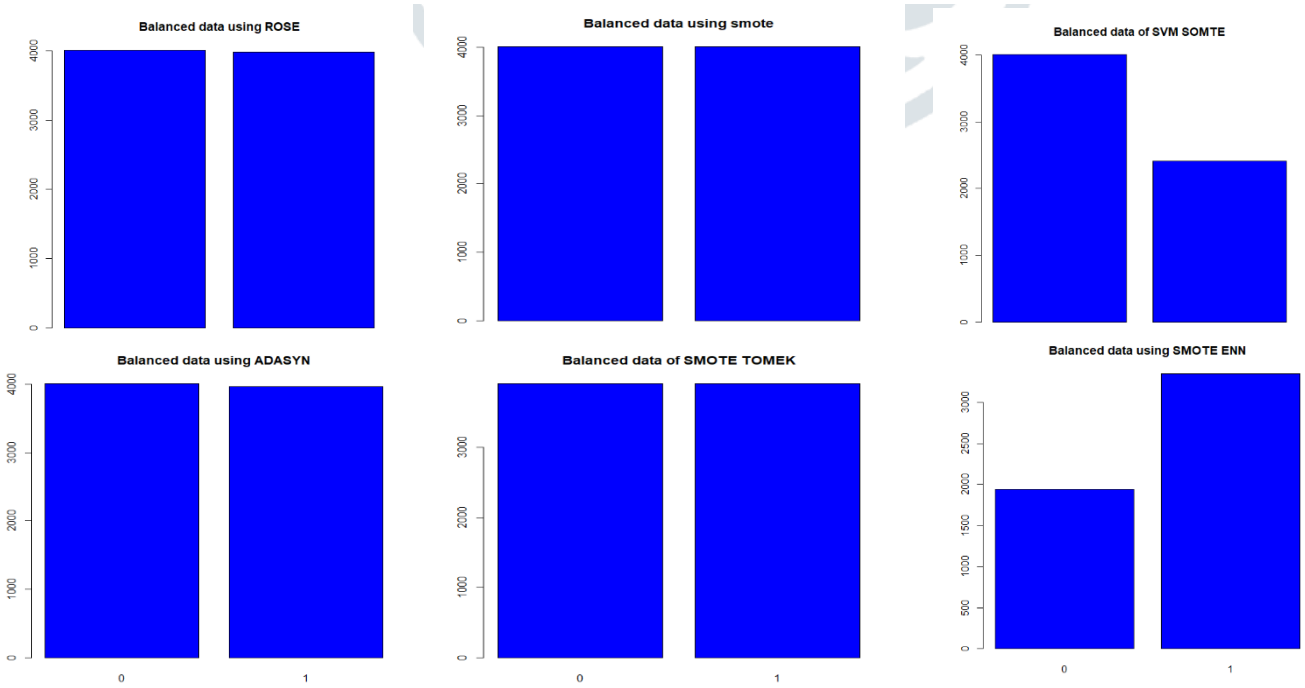
The experiments have been conducted for the review of six balancing techniques and six classification techniques over imbalanced Cerebral stroke dataset. To assess the results of classification, the evaluation has been performed on the basis of well-known performance measures, namely Accuracy, Sensitivity and Specificity.



Graphical representation of original data

The bar chart above illustrates the imbalance within the dependent variable. To enhance classification model performance for class prediction, it becomes necessary to employ balancing methods. As demonstrated in the subsequent bar chart, the use of these methods results in a more balanced representation of the data.

Bar plot for balanced data



CLASSIFICATION MODEL USING SMOTE METHOD

CONFUSION MATRIX

Logistic regression		
	predict	
actual	0	1
0	1062	121
1	490	730

Decision Tree		
	predict	
actual	0	1
0	1047	136
1	484	736

SVM		
	predict	
actual	0	1
0	1044	139
1	456	764

KNN		
	predict	
actual	0	1
0	885	298
1	330	890

Random Forest		
	predict	
actual	1	0
0	1035	148
1	445	775

Naïve-Bayes		
	predict	
actual	0	1
0	536	647
1	204	1016

OVERALL PERFORMANCE OF SMOTE

After fitting various models to this data, accuracy, sensitivity, and specificity are calculated and listed below.

Method	Accuracy	Sensitivity	Specificity
Logistic Regression	74.57345	89.7717	59.8361
Decision Tree	74.19892	88.5038	60.3279
SVM	75.23928	88.25021	62.62295
KNN	73.87	72.84	72.84
Random Forest	75.32251	63.52459	63.52459
Naïve-Bayes	64.59	61.09	61.09

CLASSIFICATION MODEL USING ADASYN METHOD

CONFUSION MATRIX

Logistic regression		
	predict	
actual	0	1
0	1060	140
1	479	710

Decision Tree		
	predict	
actual	0	1
0	1081	119
1	490	699

SVM		
	predict	
actual	0	1
0	1042	158
1	449	740

KNN		
	predict	
actual	0	1
0	860	340
1	334	855

Random Forest		
	predict	
actual	1	0
0	1045	155
1	453	736

Naïve-Bayes		
	predict	
actual	0	1
0	543	657
1	202	987

OVERALL PERFORMANCE OF ADASYN

After fitting various models to this data, accuracy, sensitivity, and specificity are calculated and listed below.

Method	**Accuracy**	**Sensitivity**	**Specificity**
Logistic Regression	74.50816	90.0833	58.7889
Decision Tree	74.08958	88.3333	59.71405
SVM	74.59188	86.8333	62.23717
KNN	71.79	71.67	71.91
Random Forest	74.55002	87.0833	61.90076
Naïve-Bayes	64.04	73.89	60.04

CLASSIFICATION MODEL USING ROSE

CONFUSION MATRIX

Logistic regression		
	predict	
actual	0	1
0	771	432
1	650	542

Decision Tree		
	predict	
actual	0	1
0	689	514
1	563	629

SVM		
	predict	
actual	0	1
0	695	508
1	388	804

KNN		
	predict	
actual	0	1
0	643	560
1	520	672

Random Forest		
	predict	
actual	1	0
0	641	562
1	520	672

Naïve-Bayes		
	predict	
actual	0	1
0	696	507
1	653	539

OVERALL PERFORMANCE OF ROSE

After fitting various models to this data, accuracy, sensitivity, and specificity are calculated and listed below.

Method	Accuracy	Sensitivity	Specificity
Logistic Regression	71.89979	62.62904	86.19958
Decision Tree	77.20251	68.93866	83.40643
SVM	76.36743	85.5891	69.4444
KNN	53.74	36.71	66.52
Random Forest	63.92	71.08	58.55
Naïve-Bayes	50.65	50.31	51.37

CLASSIFICATION MODEL USING SVM-SMOTE

CONFUSION MATRIX

Logistic regression		
	predict	
actual	0	1
0	1213	8
1	422	283

Decision Tree		
	predict	
actual	0	1
0	1148	73
1	361	344

SVM		
	predict	
actual	0	1
0	1166	55
1	372	333

KNN		
	predict	
actual	0	1
0	1040	181
1	287	418

Random Forest		
	predict	
actual	1	0
0	1211	10
1	414	291

Naïve-Bayes		
	predict	
actual	0	1
0	630	591
1	162	543

OVERALL PERFORMANCE OF SVM-SMOTE

After fitting various models to this data, accuracy, sensitivity, and specificity are calculated and listed below.

Method	Accuracy	Sensitivity	Specificity
Logistic Regression	77.67394	99.3448	40.14184
Decision Tree	77.46625	94.02129	48.79433
SVM	77.8297	95.4955	47.23404
KNN	75.7	78.37	69.78
Random Forest	77.98546	99.181	41.2766
Naïve-Bayes	60.9	79.55	47.88

CLASSIFICATION MODEL USING SMOTE-ENN

CONFUSION MATRIX

Logistic regression		
	predict	
actual	0	1
0	316	267
1	173	833

Decision Tree		
	predict	
actual	0	1
0	347	236
1	180	826

SVM		
	predict	
actual	0	1
0	393	190
1	217	789

KNN		
	predict	
actual	0	1
0	294	289
1	162	844

Random Forest		
	predict	
actual	1	0
0	362	221
1	172	834

Naïve-Bayes		
	predict	
actual	0	1
0	243	340
1	150	856

OVERALL PERFORMANCE OF SMOTE-ENN

After fitting various models to this data, accuracy, sensitivity, and specificity are calculated and listed below.

Method	Accuracy	Sensitivity	Specificity
Logistic Regression	64.62168	75.72727	72.30963
Decision Tree	69.51973	82.10736	72.82001
SVM	67.40995	78.42942	74.38641
KNN	64.47	74.49	71.62
Random Forest	60.89194	83.3002	75.07867
Naïve-Bayes	61.83	71.57	69.16

CLASSIFICATION MODEL USING SMOTE-TOMEK

CONFUSION MATRIX

Logistic regression		
	predict	
actual	0	1
0	1039	124
1	471	712

Decision Tree		
	predict	
actual	0	1
0	1025	135
1	492	691

SVM		
	predict	
actual	0	1
0	1023	137
1	441	742

KNN		
	predict	
actual	0	1
0	822	338
1	289	894

Random Forest		
	predict	
actual	1	0
0	1032	128
1	448	735

Naïve-Bayes		
	predict	
actual	0	1
0	471	689
1	186	997

OVERALL PERFORMANCE OF SMOTE-TOMEK

After fitting various models to this data, accuracy, sensitivity, and specificity are calculated and listed below.

Method	Accuracy	Sensitivity	Specificity
Logistic Regression	74.73325	89.5697	60.18597
Decision Tree	73.23944	88.36207	58.41082
SVM	75.33077	88.18966	62.72189
KNN	73.24	73.99	72.56
Random Forest	75.41613	88.96552	62.1318
Naïve-Bayes	62.65	71.69	59.13

Through rigorous evaluation of six balancing techniques and six classification methods on an imbalanced cerebral stroke dataset, this study sheds light on their combined impact. Balancing techniques such as SMOTE, ADASYN, ROSE, SVM SMOTE, SMOTE ENN, and SMOTE-TOMEK were

employed alongside classification methods including logistic regression, decision tree, SVM, KNN, random forest, and Naïve Bayes.

Key observations reveal that each approach has its merits and trade-offs. SMOTE-SVM yielded higher sensitivity at the cost of specificity, while ADASYN brought about a balanced trade-off between both metrics. ROSE exhibited varying performance, while methods like SMOTE ENN and SMOTE-TOMEK showcased balanced sensitivity and specificity.

VII. CONCLUSION

In this study, we aimed to address the challenge of class imbalance in a cerebral stroke dataset using six different balancing techniques and evaluated their performance using six classification methods. The overall performance of each combination of balancing technique and classification method was assessed in terms of accuracy, sensitivity, and specificity.

Among the balancing techniques, SMOTE, ADASYN, ROSE, SVM SMOTE, SMOTE ENN, and SMOTE-TOMEK were employed. For each technique, logistic regression, decision tree, SVM, KNN, random forest, and Naïve Bayes (NB) were chosen as the classification methods.

Results indicate that while the choice of balancing technique and classification method significantly influenced model performance, some consistent trends emerged. When utilizing the SMOTE balancing technique, SVM displayed higher sensitivity values, suggesting its effectiveness in correctly identifying positive instances. However, the trade-off was often lower specificity, implying a potential increase in false positives.

The ADASYN technique generally produced balanced sensitivity and specificity, enhancing classification across both classes. In contrast, the ROSE technique's performance varied across methods, underscoring its sensitivity to the choice of classification algorithm.

Balancing techniques like SMOTE ENN and SMOTE-TOMEK showcased improved balance between sensitivity and specificity, offering more reliable overall performance.

In conclusion, the choice of balancing technique and classification method should be made based on the specific requirements of the application. While no single approach outperformed others universally, this study provides insights into the trade-offs involved, aiding practitioners in making informed decisions for handling class imbalance in cerebral stroke prediction.

REFERENCES

1. Belarouci, S.; Chikh, M.A. Medical imbalanced data classification. *Adv. Sci. Technol. Eng. Syst.* 2017, *2*, 116–124.
2. Chan, P.K.; Stolfo, S.J. Toward Scalable Learning with Non-uniform Class and Cost Distributions: A Case Study in Credit Card Fraud Detection 1 Introduction. In Proceeding of the Fourth International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 27–31 August 1998.
3. Endo, A.; Shibata, T.; Tanaka, H. Comparison of Seven Algorithms to Predict Breast Cancer Survival (Contribution to 21 Century Intelligent Technologies and Bioinformatics). *Int. J. Biomed. Soft Comput. Hum. Sci. Off. J. Biomed. Fuzzy Syst. Assoc.* 2008, *13*, 11–16.
4. Han, J.; Kamber, M.; Pei, J. *Data Mining: Concepts and Techniques*; Elsevier: Amsterdam, The Netherlands, 2012.
5. Kubat, M.; Holte, R.C.; Matwin, S. Machine learning for the detection of oil spills in satellite radar images. *Mach. Learn.* 1998, *30*, 195–215.
6. Li, X.C.; Mao, W.J.; Zeng, D.; Su, P.; Wang, F.Y. Performance evaluation of machine learning methods in cultural modeling. *J. Comput. Sci. Technol.* 2009, *24*, 1010–1017.
7. Patel, H.; Rajput, D.S.; Reddy, G.T.; Iwendi, C.; Bashir, A.K.; Jo, O. A review on classification of imbalanced data for wireless sensor networks. *Int. J. Distrib. Sens. Netw.* 2020, *16*, 1–15.
8. Patel, H.; Rajput, D.S.; Stan, O.P.; Miclea, L.C. A New Fuzzy Adaptive Algorithm to Classify Imbalanced Data. *CMC-Comput. Mater. Contin.* 2022, *70*, 73–89.
9. Pavón, R.; Laza, R.; Reboiro-Jato, M.; Fdez-Riverola, F. Assessing the impact of class-imbalanced data for classifying relevant/irrelevant medline documents. In *Advances in Intelligent and Soft Computing*; Springer: Salamanca, Spain, 2011; Volume 93.