machine learning & knowledge extraction

MDPI

Article

Impact of Nature of Medical Data on Machine and Deep Learning for Imbalanced Datasets: Clinical Validity of SMOTE Is Questionable

Seifollah Gholampour

Department of Neurological Surgery, University of Chicago, Chicago, IL 60637, USA; seifgholampour@bsd.uchicago.edu

Abstract: Dataset imbalances pose a significant challenge to predictive modeling in both medical and financial domains, where conventional strategies, including resampling and algorithmic modifications, often fail to adequately address minority class underrepresentation. This study theoretically and practically investigates how the inherent nature of medical data affects the classification of minority classes. It employs ten machine and deep learning classifiers, ranging from ensemble learners to cost-sensitive algorithms, across comparably sized medical and financial datasets. Despite these efforts, none of the classifiers achieved effective classification of the minority class in the medical dataset, with sensitivity below 5.0% and area under the curve (AUC) below 57.0%. In contrast, the similar classifiers applied to the financial dataset demonstrated strong discriminative power, with overall accuracy exceeding 95.0%, sensitivity over 73.0%, and AUC above 96.0%. This disparity underscores the unpredictable variability inherent in the nature of medical data, as exemplified by the dispersed and homogeneous distribution of the minority class among other classes in principal component analysis (PCA) graphs. The application of the synthetic minority oversampling technique (SMOTE) introduced 62 synthetic patients based on merely 20 original cases, casting doubt on its clinical validity and the representation of real-world patient variability. Furthermore, post-SMOTE feature importance analysis, utilizing SHapley Additive exPlanations (SHAP) and tree-based methods, contradicted established cerebral stroke parameters, further questioning the clinical coherence of synthetic dataset augmentation. These findings call into question the clinical validity of the SMOTE technique and underscore the urgent need for advanced modeling techniques and algorithmic innovations for predicting minority-class outcomes in medical datasets without depending on resampling strategies. This approach underscores the importance of developing methods that are not only theoretically robust but also clinically relevant and applicable to real-world clinical scenarios. Consequently, this study underscores the importance of future research efforts to bridge the gap between theoretical advancements and the practical, clinical applications of models like SMOTE in healthcare.

Keywords: medical data; imbalanced dataset; minority class prediction; machine learning; deep learning; cost-sensitive learning; ensemble learning; SHapley Additive exPlanations (SHAP); feature importance; principal component analysis (PCA)

Citation: Gholampour, S. Impact of Nature of Medical Data on Machine and Deep Learning for Imbalanced Datasets: Clinical Validity of SMOTE Is Questionable. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 827–841. https://doi.org/10.3390/make6020039

Academic Editor: Dominik Heider

Received: 7 March 2024
Revised: 8 April 2024
Accepted: 10 April 2024
Published: 15 April 2024

1. Introduction

In the medical and healthcare domains, the rarity of certain conditions or diseases naturally results in fewer instances of positive cases compared to normal or negative cases, leading to an imbalanced dataset. Similarly, in the financial sector, instances of fraud or significant market shifts occur less frequently than regular transactions or stable market conditions. Hence, imbalances in datasets are more pronounced in medical science and financial science, posing a significant challenge in the process of classification and prediction [1–4]. In these datasets, the distribution of instances across certain classes is

disproportionately skewed, with some classes exhibiting significantly higher frequencies than others. Such imbalances present unique challenges in machine and deep learning, as standard algorithms optimized for balanced datasets may not perform effectively, often overlooking the minority class, which usually represents the most crucial information to be predicted [5,6]. Besides the hurdles in predicting minority classes in supervised learning scenarios, our recent study has brought to the forefront the intricate challenges associated with minority classes in unsupervised learning within the medical field [7]. Two approaches to tackling imbalanced datasets in classification include resampling data points and modifying the classification algorithm [8]. Each of these approaches aims to mitigate the skewness inherent in imbalanced datasets, enhancing the model's ability to accurately identify and classify minority-class instances. Hence, the former involves resampling the training data to balance classes (e.g., undersampling, oversampling, or hybrid methods) [9–13], while the latter involves altering the classifier's learning process, known as algorithmic modifications (e.g., cost-sensitive learning, thresholding, or ensemble methods) [14,15]. Haixiang et al. have demonstrated that in the domain of medical datasets, re-sampling-based ensemble classifiers are extensively utilized to tackle imbalances [4]. Conversely, within financial management datasets, feature engineering techniques are more commonly employed to address similar issues of imbalance [4]. However, the majority of previous studies preferred modifying and balancing classes using resampling techniques to address the classification challenges of imbalanced datasets: Beckmann et al. and Yu et al. utilized the undersampling technique [16,17]; Sáez et al. used SMOTE [18,19]; and Gong et al. and Alejo et al. employed a combination of oversampling and undersampling methods [20,21].

Birla et al. have utilized logistic regression and classification and regression trees (CART), alongside methods like undersampling, prior probabilities, loss matrix, and matrix weighing, to address the challenges of imbalanced data [22]. Generally, undersampling can lead to the discarding of valuable or meaningful information, particularly in the context of medical datasets. Viloria et al. demonstrated the effectiveness of oversampling, particularly the synthetic minority oversampling technique (SMOTE), for addressing class imbalances in gene expression medical datasets [23]. However, oversampling techniques may perform well on training data but less effectively on unseen data, potentially compromising the generalizability of findings [24,25]. Fernandez et al. acknowledged the existing challenges associated with SMOTE, emphasizing the importance of improving how small disjuncts, noise, data scarcity, overlap, dataset shift, and the curse of dimensionality are addressed [26]. Recently, Azhar et al. also highlighted that SMOTE's performance continues to be unsatisfactory, indicating its inadequacy in effectively managing data complexities and its potential to increase such complexities [27]. Bao et al. also revealed that noise samples could be involved in creating new samples, thereby compromising the integrity of these newly synthesized samples [28]. This, in turn, negatively affects the network's classification accuracy. To address this issue, they introduced the center point SMOTE and inner and outer SMOTE methods, aimed at enhancing the rationality of newly synthesized samples [28]. Guan et al. combined SMOTE with a data cleaning approach, specifically the weighted edited nearest neighbor rule, to tackle prediction challenges by leveraging SMOTE [29]. Raghuwanshi et al. proposed a SMOTE-based, class-specific kernelized extreme learning machine to mitigate performance fluctuations caused by the random initialization of weights between the input and hidden layers in a neural network [30]. Hosenie et al. applied level-wise data augmentation using methods such as SMOTE in combination with a hierarchical classification framework to address the classification problem of imbalanced datasets [31]. Some studies also tried to enhance the performance of oversampling methods, specifically by using techniques such as radial-based, density-based, and weighted oversampling, as well as kernel functions [9,32–35]. Other studies have also attempted to combine resampling the training data with algorithmic modifications to tackle the classification of imbalanced datasets [36–39]. On the other hand, studies such as the one by Vargas et al. have utilized hybrid sampling techniques (a combination of

under and oversampling) alongside cost-sensitive methods [40], demonstrating that this comprehensive approach, when coupled with ensemble learning and neural network methods, delivered the most effective performance. Some studies also strove to use ensemble learners and cost-sensitive classifiers to overcome the classification problem of imbalanced datasets [34,41–43].

One challenge that has been less explored in previous studies is the impact of dataset nature on prediction outcomes. The characteristics and inherent nature of data can significantly influence the behavior and effectiveness of machine and deep learning algorithms, particularly in practical applications dealing with imbalanced datasets. In medical datasets, data often involve complex relationships between variables, making it challenging to capture the underlying patterns of the minority class as well as to maintain the model's generalizability. On the other hand, financial data typically come in structured formats but can be highly volatile, requiring algorithms that are capable of adapting to sudden changes in data distribution. This study aims to compare the impact of different types natures of datasets, including medical and financial, on the classification of the minority class in multi-class classification settings. While previous studies have explored some limitations of resampling techniques, our study uniquely concentrates on the clinical usability of SMOTE, one of the most popular and widely used resampling methods, thereby bridging a gap in the current research landscape. Specifically, it examines the relevance and performance of SMOTE in clinical environments, assessing whether outcomes generated by SMOTE are not only theoretically valid but also practically meaningful and clinically actionable for healthcare professionals within real-world applications.

2. Materials and Methods

2.1. Datasets

Considering the significance of managing predictions in imbalanced data within the medical and financial sciences, we employed two comparably imbalanced datasets pertinent to cerebral stroke and bankruptcy prediction in firms. The first dataset comprises 1102 cases of individuals aged over 78 years, which was previously available on Kaggle.com (accessed on 11 December 2023) and utilized in a study by Liu et al. [44]. For the financial dataset, we compiled data from Sec.gov (accessed on 7 April 2024) on 1221 U.S. medical and healthcare stock market entities for the fiscal year 2022, applying the modified Altman Z-score method to assess bankruptcy risk [45]. Details of the features within these datasets are delineated in Table 1.

Table 1. The features used for machine and deep learning analysis of the cerebral stroke (1102 data) and financial datasets (1221 data). EBIT: Earnings before interest and taxes.

	Features of Cerebral Stroke Dataset	Mean	Variance	Features of Financial Health and Risk Dataset	Mean	Variance
1	Gender	---	---	Price-to-earnings ratio	−2.35	1.65
2	Age	80.9	0.7	Return on assets	−0.65	3.07
3	Hypertension	---	---	Return on investment	11.87	6.53
4	Heart disease	---	---	Quick ratio	3.46	0.57
5	Work	---	---	Receivable turnover ratio	−64.89	3.89
6	Resident	---	---	Enterprise value to sales	1890.83	3813.01
7	Average glucose level	122.5	3322.0	Enterprise value to EBIT ratio	2.00	1.06
8	Body mass index (BMI)	27.7	29.2	Equity ratio	6.37	6.13
9	Smoke	---	---	Cash return on assets	−0.42	4.27
10	---	---	---	Total liabilities to total assets ratio	0.57	2.22

In the medical dataset, three target classes were identified: 0, 1, and 2, with category 1 denoting cerebral stroke cases, constituting the minority class at 8.3% (Figure 1). Similarly, the financial dataset was categorized into three target classes: 0, 1, and 2, where category 1 represents firms under financial uncertainty (gray area), marked as the minority class at 5.2% (Figure 1), with Z-scores ranging from 1.1 to 2.6 [45]. It is important to note that

neither of the two datasets utilized in this study contained missing data, thus precluding the need for missing data techniques.

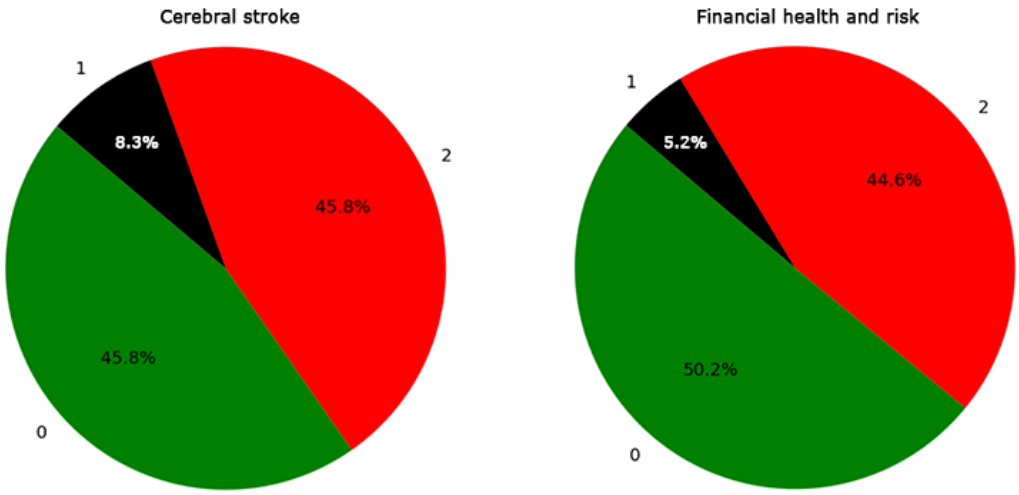

Figure 1. Distribution of target classes in cerebral stroke and financial datasets and the percentage of minority classes (class 1).

2.2. Machine and Deep Learning

To mitigate certain aspects of dataset imbalance, we initially employed strategies such as using the stratify option during data splitting and adopting stratified K-fold cross-validation. While these methods contribute valuable insights, they were found to be inadequate for our datasets, which exhibited severe imbalances. It should be noted that our initial strategy for addressing this issue was to avoid resampling the data to balance classes. Instead, we focused on employing various machine and deep learning algorithms, along with algorithmic modifications, to tackle the imbalance issue during prediction. Hence, we implemented ten algorithms, including random forest, adaptive boosting (AdaBoost), gradient boosting, extreme gradient boosting (XGBoost), categorical boosting (CatBoost), voting, balanced random forest, neural network (multilayer perceptron), stacking, and cost-sensitive algorithms (Table 2).

The algorithms' predictions were subsequently evaluated against the test outcomes using confusion matrices to assess accuracy, specificity, precision, recall, and F1 scores, alongside receiver operating characteristic (ROC) evaluation and learning curve analysis. It should be noted that two algorithms demonstrating optimal performance metrics were selected for integration into the voting and stacking methods. Additionally, we applied the cost-sensitive approach to all algorithms, adjusting the correct class weights based on the imbalance percentage of minority class 1. In conclusion, the algorithms that exhibited superior performance metrics were ultimately chosen as the definitive predictive method for each dataset.

3. Results

Three-dimensional principal component analysis (PCA) was conducted on the cerebral stroke and financial datasets, comprising data points and their corresponding features (Figure 2). The PCA process involves identifying the principal components that account for the maximum variance in the data. These components are linear combinations of the original variables and are orthogonal to each other, ensuring that the redundancy and correlation among the variables are minimized. The positioning of each data point reflects the synthesized features, thereby illuminating the variations in the distribution of data points across these two distinct datasets. Consequently, the graphical representations in Figure 2 adeptly preserve the inherent variance within each dataset. The PCA scatter plot for the cerebral stroke dataset exhibits a distribution of data points across all three principal

component axes, with a relatively even spread that suggests multidimensional variability within the dataset. In contrast, the financial dataset's PCA plot shows a high concentration of data points, especially along the first principal component axis.

Table 2. Brief description of various machine learning and deep learning algorithms used in the present study to address the issue of minority class prediction in medical and financial datasets.

Algorithm Name	Algorithm Description
Random Forest	An ensemble learning method that operates by constructing a multitude of decision trees at training time to output the mode of the classes of the individual trees.
AdaBoost	A boosting algorithm that combines multiple weak learners (a single tree) to create a strong learner. It sequentially adjusts the weights of incorrectly classified instances so they are correctly classified in subsequent rounds.
Gradient Boosting	An ensemble technique that builds models sequentially, with each new model correcting errors made by previous models. It minimizes loss via gradient descent, enhancing prediction accuracy for diverse datasets.
XGBoost	Stands for extreme gradient boosting, an optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable.
CatBoost	A gradient boosting algorithm on decision trees was developed to provide high performance with categorical data. It generally reduces overfitting and is effective across a broad range of datasets and problems.
Voting	An ensemble machine learning model that combines predictions from multiple models. In our project, it integrates the outputs of the two classifiers that had the best individual performance. Then, it predicts the class based on the majority vote for classification or the average of predicted probabilities.
Balanced Random Forest	A variant of the random forest algorithm that adjusts weights inversely proportional to class frequencies in the input data or specified by the user. This classifier should be generally useful for addressing imbalanced datasets.
Neural network (Multilayer perceptron)	A foundational deep learning algorithm, MLP models complex relationships between features and targets using a multilayer network of neurons. An MLP includes an input layer, several hidden layers, and an output layer. Connections between neurons across layers are adjusted during training via backpropagation—calculating loss function gradients—to minimize prediction errors.
Stacking	An ensemble learning technique that combines multiple classification models via a meta-classifier. In our project, it combines the outputs of the two classifiers that demonstrated the best individual performance. First, the base-level models are trained using a complete dataset. Then, the meta-model is trained on the outputs of the base-level models as features.
Cost-sensitive	Adjusts the classification algorithms to emphasize the minority class, aiming to improve prediction accuracy in imbalanced datasets by modifying error weights or altering the decision threshold.

Figure 3 displays a heatmap of the confusion matrix for all ten classifiers, fine-tuned after hyperparameter adjustments, providing a vivid visual representation of each model's performance across the various classes. The gradation of colors in the heatmap offers an intuitive guide to understanding classification accuracy and patterns of misclassification at a glance. Additionally, the figure includes the receiver operating characteristic (ROC) curves, which are crucial for our analysis. Given the imperative to minimize false negatives in predictions related to cerebral stroke and financial risk, the ROC curve emerges as a particularly relevant tool. The results reveal that, despite employing a variety of classifiers, including ensemble learners, applying cost-sensitive algorithms, and making algorithmic modifications, and despite achieving a high overall accuracy (greater than 90.5%), none of the classifiers succeeded in effectively classifying the minority class of the cerebral stroke dataset (class 1). The sensitivity for this class was less than 5.0% across all classifiers, and the area under the curve (AUC) was also below 57.0%. In contrast, for the financial dataset,

all classifiers demonstrated the capability to accurately classify the minority class, with the sensitivity for class 1 exceeding 73%, and the overall accuracy of all classifiers for this dataset was also above 95.0%. The AUC surpassing 96.0% suggests that the model has a high probability of correctly distinguishing between positive and negative cases, indicating strong discriminative power. The performance details of the top-performing classifiers (XGBoost and gradient boosting) for the financial dataset are provided in Figure 4. The overall accuracies for the XGBoost and gradient boosting classifiers were remarkably high, at 99.2% and 99.6%, respectively. Furthermore, the F1 scores and sensitivities, standing at 99.0% for XGBoost and a perfect 100.0% for gradient boosting, indicate that both classifiers achieved an exceptional balance between precision and recall. These metrics underscore the efficacy of the classifiers in handling the datasets, showcasing their robustness in achieving accurate predictions of the minority class as well. The detailed values for test, train, and cross-validation accuracy, as well as other metrics of these classifiers in Figure 4, showed there is no risk of overfitting, and the predicted results are totally generalizable based on the small final gap (<2.9%) in their learning curves. These findings affirm the consistency and robustness of the results obtained from these classifiers, showcasing the significant potential of these methodologies for predicting financial risk.

To address the prediction issue for the minority class of the cerebral stroke dataset, we employed the SMOTE technique to balance the classes, followed by reapplying all ten classifiers. The best performances were achieved by the XGBoost and CatBoost classifiers. The detailed metrics of their performance are illustrated in Figure 5. Implementing SMOTE effectively mitigated our concerns regarding the prediction of the minority class, as evidenced by the sensitivity for class 1 reaching 88.0% for both XGBoost and CatBoost classifiers with an overall accuracy of 88.1%. The F1 score of 88.0% for both classifiers demonstrates the model's effectiveness in capturing both true positives and true negatives, achieving a balance between its ability to identify relevant instances and avoid false positives. Additionally, the AUC for the minority class improved to 93%. The learning curve stability results were also satisfactory, with the gap between training and cross-validation accuracies on the right side being less than 4.9%.

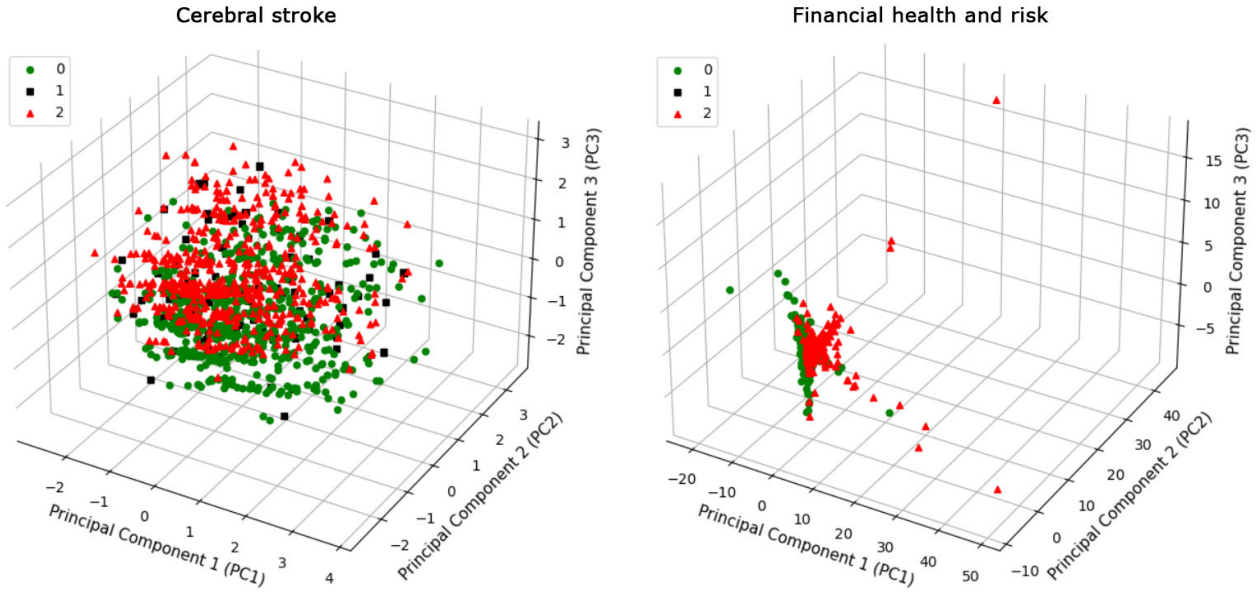

Figure 2. Three-dimensional principal component analysis (PCA) of cerebral stroke and financial datasets is visualized through the first three principal components. Class 1 in the financial dataset is obscured behind classes 0 and 2.

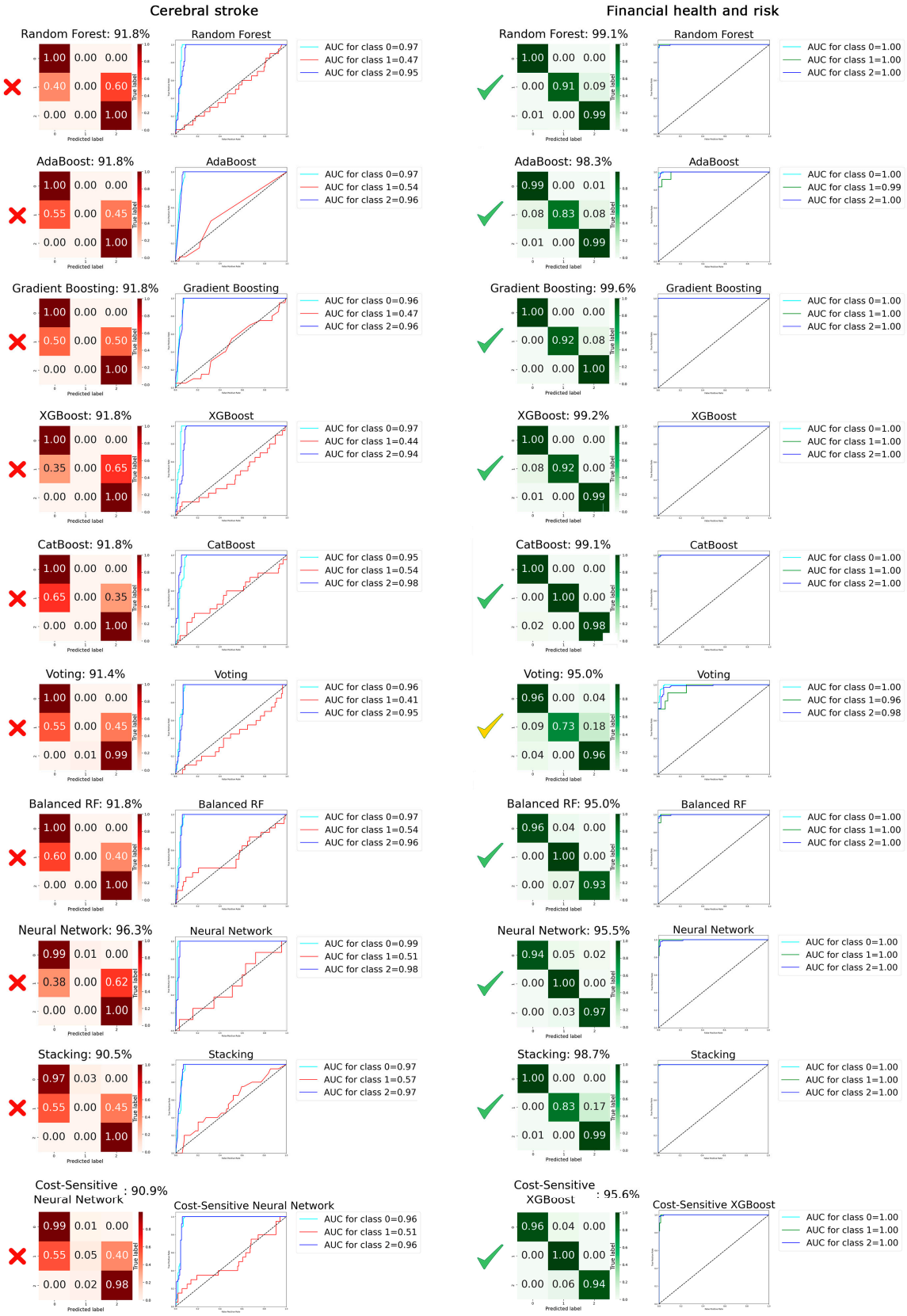

Figure 3. The heatmaps of the confusion matrices demonstrate the performance of the classification model in predicting the minority class for 10 classifiers. Additionally, the receiver operating characteristic (ROC) curves illustrate the true positive rate against the false positive rate at various threshold settings for the 10 classification models utilized.

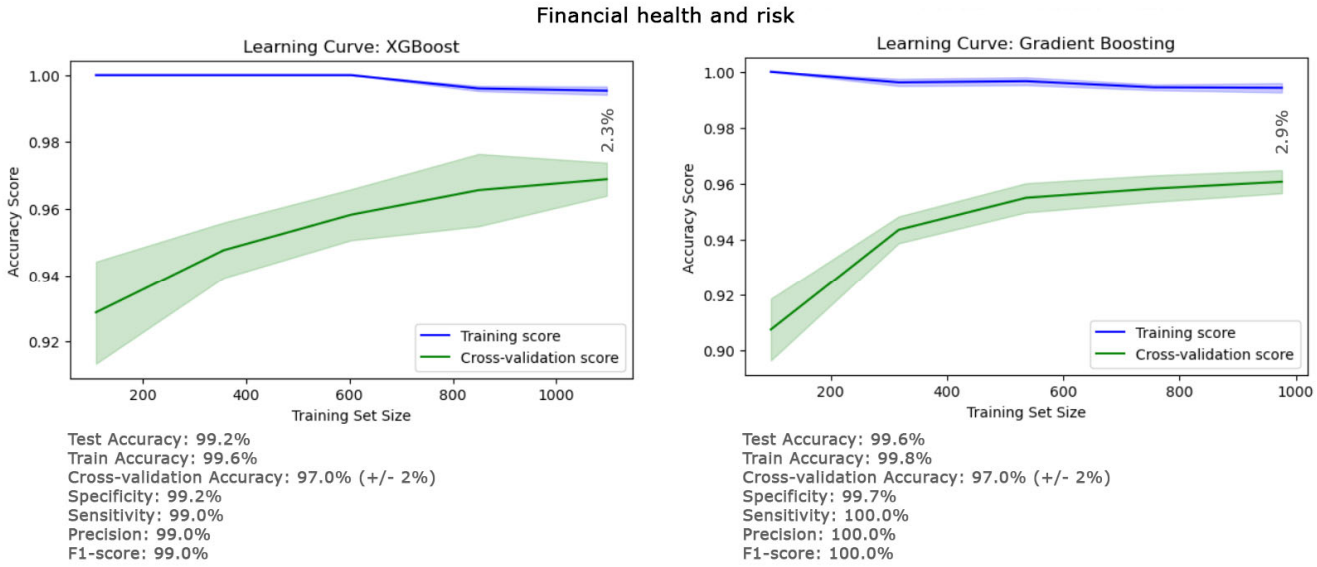

Figure 4. The performance of XGboost and gradient boosting as the best classifiers in predicting the minority class of the financial dataset. The generalization of the prediction results is compared using training and cross-validation accuracies in their learning curves.

4. Discussion

Imbalances in datasets are acutely observed in the fields of medical and financial sciences, presenting formidable challenges in prediction [1–4]. To address this pervasive issue, some strategies have been employed, including the adoption of resampling methods and modifications to classification algorithms. These approaches have been geared towards achieving a better balance within datasets, aiming to enhance the accuracy of predictive models when dealing with underrepresented minority classes. However, they often remained insufficient for effectively classifying the minority class due to a variety of complex factors [24,25,33]. Previous studies have paid less attention to the impact of a dataset's inherent nature on the prediction outcomes for minority classes. This study endeavors to shed light on this critical yet underexplored factor, examining how the intrinsic characteristics of medical data (specifically cerebral stroke in this study) affect the classification of minority classes. It also aims to ascertain whether the theoretical results derived from SMOTE are clinically valid and applicable for use by clinicians. The datasets used in this study were comparable in size, with the cerebral stroke dataset comprising 1102 cases and the financial dataset including 1221 firms. The cerebral stroke dataset featured nine attributes, while the financial dataset was characterized by a roughly similar number of features, ten. However, our analysis yielded a striking outcome: none of the 10 classifiers we employed managed to accurately classify the minority class within the cerebral stroke dataset, with the sensitivity for class 1 being less than 5.0% across all classifiers (Figure 3).

Notably, the proportion of the minority class in the cerebral stroke dataset was 8.3%, which was higher than the 5.2% observed in the financial dataset (Figure 1). Despite this, the same classifiers achieved perfect predictions for the minority class in the financial dataset (Figure 3). This significant discrepancy underscores the potential impact of the inherent differences in the nature of these datasets on classifier performance. It suggests that the characteristics, distribution, and complexities of the data itself may play a crucial role in influencing the effectiveness of classification algorithms, particularly in the context of minority class prediction. The PCA results from Figure 2 illustrate that the financial dataset exhibits concentrated data within each class, implying that classification might be more straightforward for these classes. In contrast, the cerebral stroke dataset does not show concentrated and tight clustering. Instead, it reveals a dispersed and homogeneous distribution of class 1 (the minority) among the other classes (0 and 2). This distribution

pattern significantly complicates the separation (classification) of this class from the others, a challenge that is substantiated by the performance of our 10 classifiers (Figure 3).

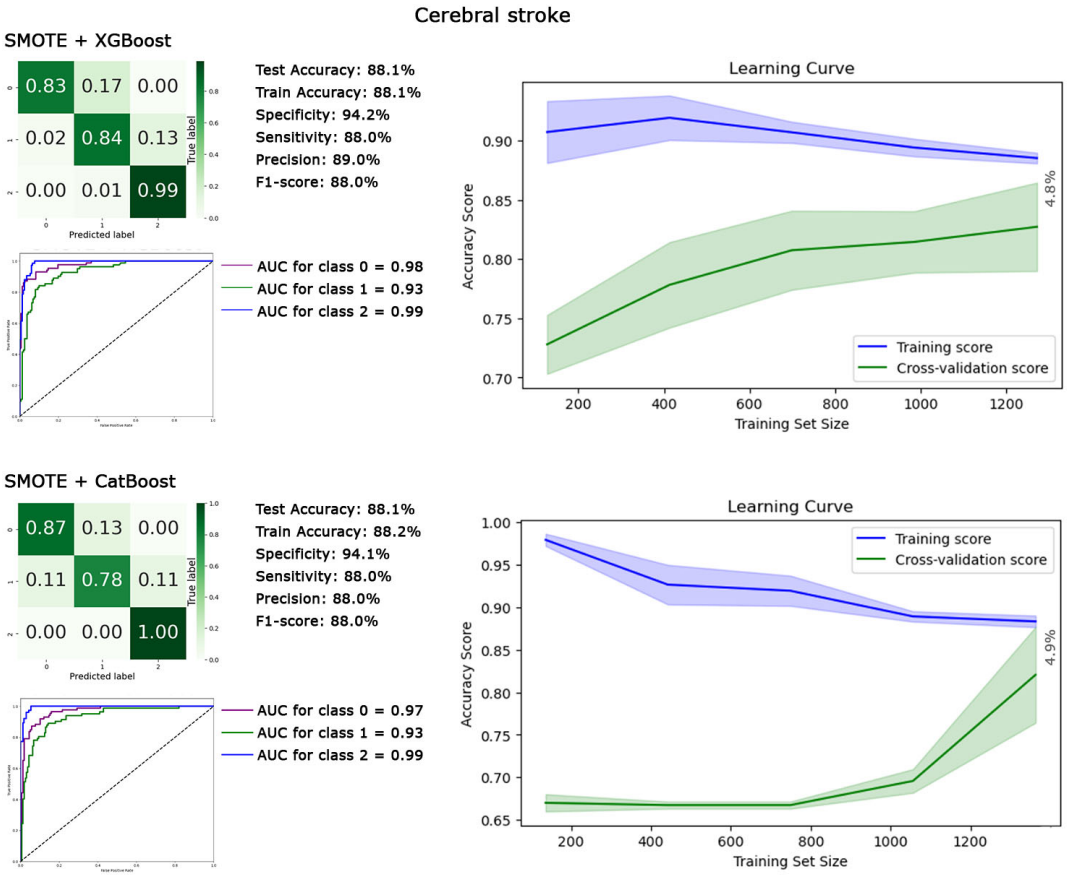

Figure 5. The heatmap of the confusion matrix, ROC curves, and performance metrics of XGBoost and CatBoost are analyzed as the top classifiers after applying the synthetic minority oversampling technique (SMOTE) to predict the minority class in the cerebral stroke dataset. The generalization of prediction results is compared using training and cross-validation accuracies depicted in their respective learning curves.

The distribution patterns observed in the PCA graphs for other medical datasets, focusing on hydrocephalus [7,46–49], cerebral aneurysms [50,51], orthopedic drilling [52], and Chiari malformation I [53,54], also bore striking similarities to those noted in this cerebral stroke dataset. In all instances, the data distribution across classes was dispersed, non-concentrated, and homogeneously distributed among the other classes. This phenomenon may be inherently tied to the nature of medical data. Unlike non-medical and biological datasets, where phenomena can often be explained by established, deterministic rules—such as the financial behaviors of firms being influenced by similar tax laws, insurance regulations, and management principles—the realm of medical science deals with human conditions that are far less predictable. The behavior of medical disorders does not always adhere to a set of fixed rules, contributing to the non-concentrated distribution of class data and making the task of classification notably more challenging. In contrast to the more deterministic nature of non-medical datasets, the medical data derived from human subjects encapsulates a broader spectrum of variability and unpredictability. This inherent unpredictability in medical conditions may account for the observed difficulties in classifying data into distinct classes, as the overlap between different conditions can blur the lines that typically define class boundaries. While increasing the number of discriminative features within a dataset could potentially mitigate these classification challenges, our analysis shows that the cerebral stroke and financial datasets were already comparable

in size and feature count. This suggests that simply augmenting the number of features or increasing the population may not be a sufficient solution to the problem of predicting minority classes in a medical context.

The results depicted in Figure 5 illustrate that the SMOTE apparently resolved the prediction issues related to the minority class in the medical dataset, achieving an overall accuracy of 88.1%, with the F1 score and sensitivity for the minority class (class 1) at 88.0%. However, this raises an important question regarding the underlying mechanism through which SMOTE facilitates the improvement of prediction outcomes for the minority class. As demonstrated in Figure 6, the comparison between the PCA representations and data distribution before and after applying SMOTE, specifically within the context of the XGBoost classifier, provides insight into this mechanism. By applying SMOTE, we observe a notable increase in the density of Class 1, achieved through the generation of synthetic data points (patients with cerebral stroke). This strategic augmentation directly tackles the class imbalance by enriching the dataset with more examples of the minority class, thereby creating a more balanced distribution of classes. The addition of these synthetic patients is designed to simulate the diversity and variability inherent in the real data, enhancing the dataset's representativeness. This change transcends the mere increase in the number of patients within the dataset. Rather, it involves a strategic enrichment and manipulation of the dataset, specifically designed to reduce the complexity and challenges associated with predicting the minority class.

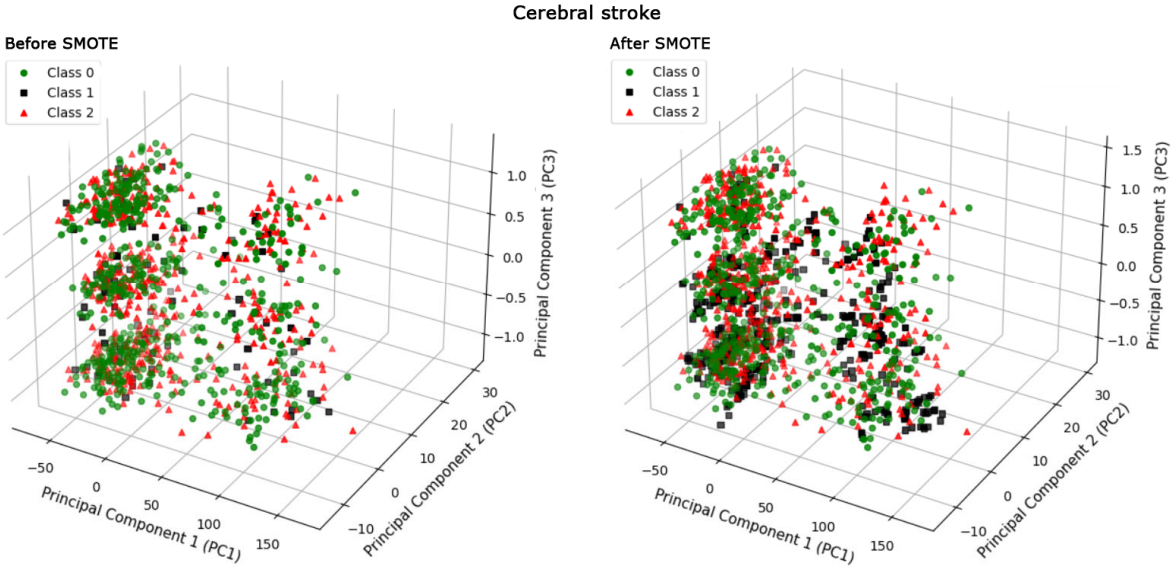

Figure 6. The comparison of three-dimensional principal component analysis (PCA) for the cerebral stroke dataset with the XGBoost classifier before and after applying the SMOTE technique reveals significant differences. The addition of new synthetic patients with cerebral stroke, depicted by additional black squares in the PCA graph after the synthetic minority oversampling technique (SMOTE), illustrates the expansion of the dataset aimed at addressing the prediction challenge associated with the minority class.

SMOTE works by identifying k nearest neighbors (kNN) within the minority class for each data point (patients). Then, it randomly selects a neighbor along the line segment, connecting the original patient to its neighbor. Finally, it creates new synthetic patients by interpolating (adding a weighted difference) between the original patient and the chosen neighbor. The initial number of patients in the test group in class 1 was 20. SMOTE generated 62 new synthetic patients in class 1, mimicking the distribution of the existing 20 patients to address the prediction problem of the minority class. The SMOTE strategy, unlike random oversampling, leads to a more dispersed distribution of minority-class samples, potentially reducing the risk of overfitting. However, SMOTE operates under

the assumption that the areas designated for generating new samples are optimal, an assumption that does not always hold true. The presence of dispersed minority outliers in many datasets can lead to overlaps between these designated areas and the existing distribution of the majority class. This overlap generates artificial samples that may not accurately represent the true data distribution. While expanding these areas beyond the reach of random oversampling could be beneficial, neighborhood-oriented methods risk extending into areas that are not representative of the minority class, inadvertently ignoring the distribution of the majority class [9]. This challenge is exacerbated in datasets with noise, especially when such noise impacts the labeling, increasing the likelihood of encountering independent outliers. Beinecke et al. also demonstrated that Gaussian noise up-sampling is more effective than SMOTE for classification purposes [55]. In response to the limitations of class balancing techniques like SMOTE, some studies advocate for an algorithm-level approach. For instance, Ganaie et al. demonstrated an improved classifier based on support vector machines (SVMs) using fuzzy least squares projection to address imbalances, showcasing enhanced prediction performance [56]. Additionally, boosting SVM has been proposed as a solution to this issue [8]. Efforts to refine the performance of SMOTE have explored methods such as radial-based oversampling, weighted SMOTE, kernel functions, or density-based solutions, aiming for more informative oversampling to tackle local instance-level challenges [9,32–34]. While these approaches can lead to increased class overlap, techniques like borderline-SMOTE, which focus on instances near the decision boundary, implicitly consider the position of the majority class. Nonetheless, the generation of new synthetic instances might not fully account for the majority class distribution, potentially leading to an altered decision boundary that does not optimally reflect the original data structure.

The ethical implications of manipulating clinical data by introducing synthetic patients to address dataset imbalances are profound. In clinical scenarios, the unique characteristics of even a single patient can pivot the management process of a disorder, often warranting a case report or series of articles due to the complexity of human subjects. The question arises: does the addition of 62 synthetic patients, based on merely 20 initial patients, hold clinical validity, and can it reflect real-world patient variability? Feature importance analysis using SHapley Additive exPlanations (SHAP) and tree-based methods post-SMOTE application further fuels this debate (Figure 7). These analyses revealed that "residence" and "work" emerged as the most significant contributors to the prediction of cerebral stroke, whereas 'hypertension' and 'heart disease' were deemed less important. This finding is in stark contrast to previous research that underscores the critical role of 'hypertension' and 'heart disease' as key risk factors in the incidence of cerebral stroke [57–63]. Additionally, the results from Figure 7 also indicate that glucose levels, which are widely recognized for their significance in stroke patients [64,65], were not significantly highlighted as important features. Such findings contradict clinical understanding. These contradictions and discrepancies between the predictive model's outcomes and well-established medical evidence prompt a reevaluation of the methodology and the potential biases introduced by the use of synthetic data in modeling complex health conditions. It raises concerns that the synthetic augmentation of datasets via SMOTE, while potentially enhancing classifier performance, may not yield clinically coherent results.

Our findings highlight a critical open question for future research: the urgent need to develop more sophisticated modeling techniques and algorithmic innovations specifically designed to navigate the complex landscape of medical data. These innovations aim to predict outcomes for minority classes without relying on traditional resampling techniques. By adopting such advanced approaches in future research, we can achieve classifications that are not only more accurate but also clinically meaningful, despite the significant dispersion and heterogeneity inherent in medical datasets. Furthermore, we advocate for the conduct of comparative studies across different data domains. Such research can deepen our understanding of how the characteristics and inherent nature of data influence model performance, particularly in addressing real-world problems. This comprehensive

approach will not only enhance the precision of predictive modeling in healthcare but also contribute to the broader field of machine learning, bridging the gap between theoretical research and practical application.

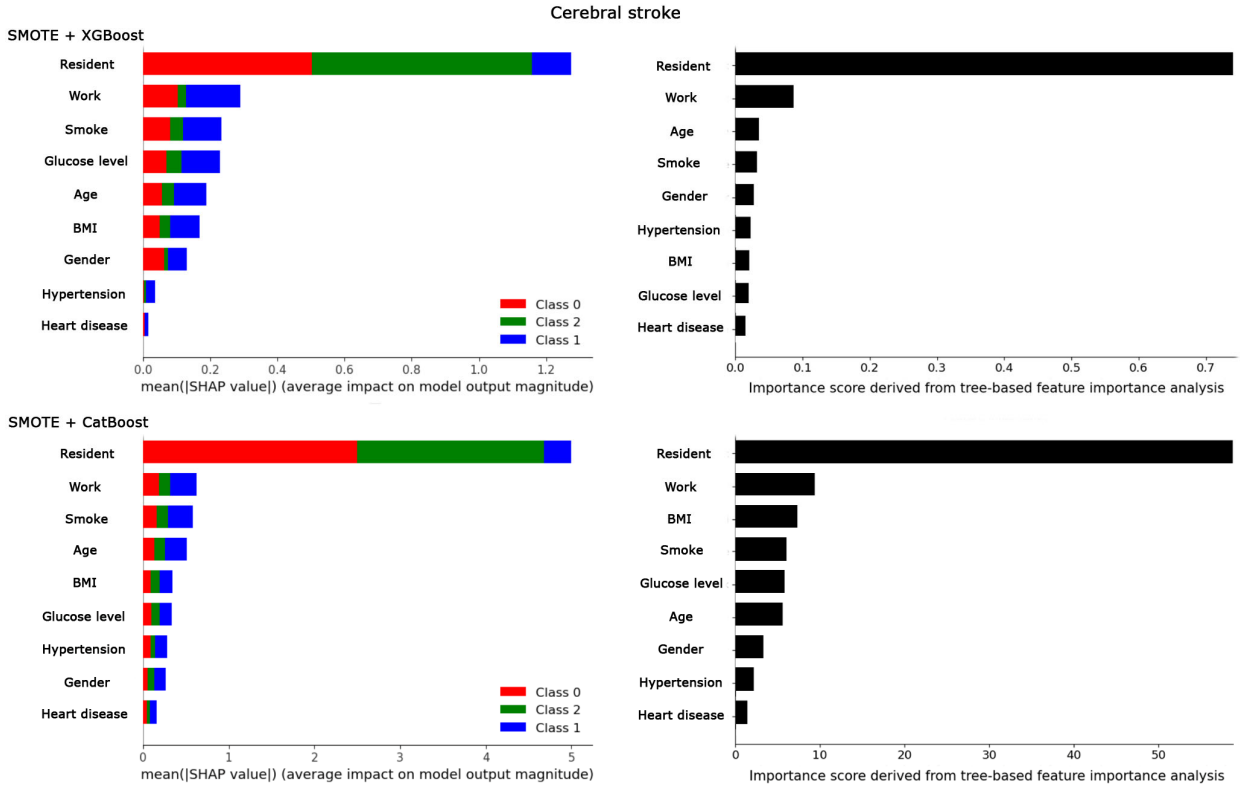

Figure 7. The comparison of feature importance analysis using two different methods, SHapley Additive exPlanations (SHAP) and tree-based methods, is conducted for the cerebral stroke dataset after applying SMOTE for the best classifiers, XGBoost and CatBoost.

5. Conclusions

Our study employed advanced machine learning and deep learning classifiers, along with algorithm modifications, to address one of the most prevalent challenges in medical datasets: the accurate classification of minority classes within imbalanced datasets. This effort underscored a persistent difficulty in classifying the minority class in medical datasets, a challenge starkly contrasted by the financial dataset, which exhibited robust discriminative power. By critically examining the clinical validity and real-world applicability of these models, our work highlights a significant gap between theoretical advancements and practical healthcare outcomes, particularly regarding the use of SMOTE in medical datasets.

1. Nature of Medical Data: The differential performance of machine and deep learning algorithms on medical versus financial datasets can be attributed to the unique characteristics and inherent nature of medical data. The dispersion and variability evident in PCA graphs pose substantial classification challenges, underscoring the necessity for models adept at navigating these complexities.

2. Implications of SMOTE: Introducing synthetic patients through SMOTE, aimed at balancing class representation, revealed contradictions in feature importance and raised concerns about the technique's clinical validity. These findings question SMOTE's ability to accurately reflect real-world variability, casting doubt on its efficacy in clinical settings.

3. Clinical Perspective: Our critical evaluation of SMOTE from a clinical viewpoint offers a distinctive perspective that enhances the ongoing methodological discourse. This scrutiny underscores the importance of ensuring that machine and deep learning applications in healthcare are both technically sound and clinically relevant.

4. Call for Sophisticated Modeling Techniques: Our findings not only unveil critical limitations in current predictive modeling approaches for medical data but also highlight the urgent need for more sophisticated modeling techniques. These techniques must be specifically designed to meet the unique challenges of medical datasets, including their inherent variability and complexity. Our study clearly demonstrates that future research should concentrate on devising innovative solutions that enhance prediction accuracy for minority classes without relying on resampling strategies. Such advancements are essential for achieving outcomes that are not only theoretically robust but also hold significant clinical value.

By bridging the gap between theoretical advancements in machine learning and deep learning and their practical application in medical science, our study lays the groundwork for future innovations. These innovations hold the potential to not only advance the frontiers of scientific knowledge but also significantly enhance clinical decision-making and treatment strategies. This dual emphasis highlights the significance of our findings and sets a clear direction for subsequent research efforts. The aim is to harness the full potential of machine and deep learning technologies in enhancing healthcare outcomes, particularly for imbalanced medical datasets.

Funding: This research received no external funding.

Data Availability Statement: The original financial dataset is available from the corresponding author. The cerebral stroke dataset is available at https://www.kaggle.com/datasets/shashwatwork/cerebral-stroke-predictionimbalaced-dataset (accessed on 11 December 2023).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Fotouhi, S.; Asadi, S.; Kattan, M.W. A comprehensive data level analysis for cancer diagnosis on imbalanced data. *J. Biomed. Inform.* **2019**, *90*, 103089. [CrossRef] [PubMed]
2. Li, Z.; Huang, M.; Liu, G.; Jiang, C. A hybrid method with dynamic weighted entropy for handling the problem of class imbalance with overlap in credit card fraud detection. *Expert Syst. Appl.* **2021**, *175*, 114750. [CrossRef]
3. Wu, X.; Meng, S. E-commerce customer churn prediction based on improved SMOTE and AdaBoost. In Proceedings of the 2016 13th International Conference on Service Systems and Service Management (ICSSSM), Kunming, China, 24–26 June 2016; pp. 1–5.
4. Guo, H.; Li, Y.; Jennifer, S.; Gu, M.; Huang, Y.; Gong, B. Learning from class-imbalanced data: Review of methods and applications. *Expert Syst. Appl.* **2017**, *73*, 220–239.
5. Japkowicz, N.; Stephen, S. The class imbalance problem: A systematic study. *Intell. Data Anal.* **2002**, *6*, 429–449. [CrossRef]
6. Ghosh, K.; Bellinger, C.; Corizzo, R.; Branco, P.; Krawczyk, B.; Japkowicz, N. The class imbalance problem in deep learning. *Mach. Learn.* **2022**, *111*, 1–57. [CrossRef]
7. Waterstraat, G.; Dehghan, A.; Gholampour, S. Optimization of Number and Range of Shunt Valve Performance Levels in Infant Hydrocephalus: A Machine Learning Analysis. *Front. Bioeng. Biotechnol.* **2024**, *12*, 1352490. [CrossRef] [PubMed]
8. Wang, B.X.; Japkowicz, N. Boosting support vector machines for imbalanced data sets. *Knowl. Inf. Syst.* **2010**, *25*, 1–20. [CrossRef]
9. Koziarski, M.; Krawczyk, B.; Woźniak, M. Radial-based oversampling for noisy imbalanced data classification. *Neurocomputing* **2019**, *343*, 19–33. [CrossRef]
10. Lin, C.; Tsai, C.-F.; Lin, W.-C. Towards hybrid over-and under-sampling combination methods for class imbalanced datasets: An experimental study. *Artif. Intell. Rev.* **2023**, *56*, 845–863. [CrossRef]
11. Vairetti, C.; Assadi, J.L.; Maldonado, S. Efficient hybrid oversampling and intelligent undersampling for imbalanced big data classification. *Expert Syst. Appl.* **2024**, *246*, 123149. [CrossRef]
12. Alamri, M.; Ykhlef, M. Hybrid Undersampling and Oversampling for Handling Imbalanced Credit Card Data. *IEEE Access* **2024**, *12*, 14050–14060. [CrossRef]
13. Liu, Y.; Zhu, L.; Ding, L.; Sui, H.; Shang, W. A hybrid sampling method for highly imbalanced and overlapped data classification with complex distribution. *Inf. Sci.* **2024**, *661*, 120117. [CrossRef]
14. Chawla, N.V.; Cieslak, D.A.; Hall, L.O.; Joshi, A. Automatically countering imbalance and its empirical relationship to cost. *Data Min. Knowl. Discov.* **2008**, *17*, 225–252. [CrossRef]
15. Ahmed, S.; Mahbub, A.; Rayhan, F.; Jani, R.; Shatabda, S.; Farid, D.M. Hybrid methods for class imbalance learning employing bagging with sampling techniques. In Proceedings of the 2017 2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS), Bengaluru, India, 21–23 December 2017; pp. 1–5.
16. Beckmann, M.; Ebecken, N.F.; Pires de Lima, B.S. A KNN undersampling approach for data balancing. *J. Intell. Learn. Syst. Appl.* **2015**, *7*, 104–116. [CrossRef]

17. Yu, H.; Ni, J.; Zhao, J. ACOSampling: An ant colony optimization-based undersampling method for classifying imbalanced DNA microarray data. *Neurocomputing* **2013**, *101*, 309–318. [CrossRef]
18. Sáez, J.A.; Krawczyk, B.; Woźniak, M. Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets. *Pattern Recognit.* **2016**, *57*, 164–178. [CrossRef]
19. Yun, Z.; Nan, M.; Da, R.; Bing, A. An effective over-sampling method for imbalanced data sets classification. *Chin. J. Electron.* **2011**, *20*, 489–494.
20. Gong, J.; Kim, H. RHSBoost: Improving classification performance in imbalance data. *Comput. Stat. Data Anal.* **2017**, *111*, 1–13. [CrossRef]
21. Alejo, R.; Valdovinos, R.M.; García, V.; Pacheco-Sanchez, J.H. A hybrid method to face class overlap and class imbalance on neural networks and multi-class scenarios. *Pattern Recognit. Lett.* **2013**, *34*, 380–388. [CrossRef]
22. Birla, S.; Kohli, K.; Dutta, A. Machine learning on imbalanced data in credit risk. In Proceedings of the 2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, BC, Canada, 13–15 October 2016; pp. 1–6.
23. Viloria, A.; Lezama, O.B.P.; Mercado-Caruzo, N. Unbalanced data processing using oversampling: Machine learning. *Procedia Comput. Sci.* **2020**, *175*, 108–113. [CrossRef]
24. Tarawneh, A.S.; Hassanat, A.B.; Altarawneh, G.A.; Almuhaimeed, A. Stop oversampling for class imbalance learning: A review. *IEEE Access* **2022**, *10*, 47643–47660. [CrossRef]
25. Kumari, A.; Behera, R.K.; Sahoo, K.S.; Nayyar, A.; Kumar Luhach, A.; Prakash Sahoo, S. Supervised link prediction using structured-based feature extraction in social network. *Concurr. Comput. Pract. Exp.* **2022**, *34*, e5839. [CrossRef]
26. Fernández, A.; Garcia, S.; Herrera, F.; Chawla, N.V. SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *J. Artif. Intell. Res.* **2018**, *61*, 863–905. [CrossRef]
27. Azhar, N.A.; Pozi, M.S.M.; Din, A.M.; Jatowt, A. An investigation of smote based methods for imbalanced datasets with data complexity analysis. *IEEE Trans. Knowl. Data Eng.* **2022**, *35*, 6651–6672. [CrossRef]
28. Bao, Y.; Yang, S. Two novel SMOTE methods for solving imbalanced classification problems. *IEEE Access* **2023**, *11*, 5816–5823. [CrossRef]
29. Guan, H.; Zhang, Y.; Xian, M.; Cheng, H.-D.; Tang, X. SMOTE-WENN: Solving class imbalance and small sample problems by oversampling and distance scaling. *Appl. Intell.* **2021**, *51*, 1394–1409. [CrossRef]
30. Raghuwanshi, B.S.; Shukla, S. Classifying imbalanced data using SMOTE based class-specific kernelized ELM. *Int. J. Mach. Learn. Cybern.* **2021**, *12*, 1255–1280. [CrossRef]
31. Hosenie, Z.; Lyon, R.; Stappers, B.; Mootoovaloo, A.; McBride, V. Imbalance learning for variable star classification. *Mon. Not. R. Astron. Soc.* **2020**, *493*, 6050–6059. [CrossRef]
32. Pérez-Ortiz, M.; Gutiérrez, P.A.; Tino, P.; Hervás-Martínez, C. Oversampling the minority class in the feature space. *IEEE Trans. Neural Netw. Learn. Syst.* **2015**, *27*, 1947–1961. [CrossRef] [PubMed]
33. Islam, S.; Sara, U.; Kawsar, A.; Rahman, A.; Kundu, D.; Dipta, D.D.; Karim, A.; Hasan, M. Sgbba: An efficient method for prediction system in machine learning using imbalance dataset. *Int. J. Adv. Sci. Comput. Appl.* **2021**, *12*, 430–441. [CrossRef]
34. Jeyalakshmi, K. Weighted Synthetic Minority Over-Sampling Technique (WSMOTE) Algorithm and Ensemble Classifier for Hepatocellular Carcinoma (HCC) In Liver Disease System. *Turk. J. Comput. Math. Educ. (TURCOMAT)* **2021**, *12*, 7473–7487.
35. Wang, C.; Deng, C.; Wang, S. Imbalance-XGBoost: Leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost. *Pattern Recognit. Lett.* **2020**, *136*, 190–197. [CrossRef]
36. Devi, D.; Biswas, S.K.; Purkayastha, B. Correlation-based oversampling aided cost sensitive ensemble learning technique for treatment of class imbalance. *J. Exp. Theor. Artif. Intell.* **2022**, *34*, 143–174. [CrossRef]
37. Abedin, M.Z.; Guotai, C.; Hajek, P.; Zhang, T. Combining weighted SMOTE with ensemble learning for the class-imbalanced prediction of small business credit risk. *Complex Intell. Syst.* **2023**, *9*, 3559–3579. [CrossRef]
38. Kaisar, S.; Chowdhury, A. Integrating oversampling and ensemble-based machine learning techniques for an imbalanced dataset in dyslexia screening tests. *ICT Express* **2022**, *8*, 563–568. [CrossRef]
39. Khuat, T.T.; Le, M.H. Evaluation of sampling-based ensembles of classifiers on imbalanced data for software defect prediction problems. *SN Comput. Sci.* **2020**, *1*, 108. [CrossRef]
40. Werner de Vargas, V.; Schneider Aranda, J.A.; dos Santos Costa, R.; da Silva Pereira, P.R.; Victória Barbosa, J.L. Imbalanced data preprocessing techniques for machine learning: A systematic mapping study. *Knowl. Inf. Syst.* **2023**, *65*, 31–57. [CrossRef] [PubMed]
41. Chamlal, H.; Kamel, H.; Ouaderhman, T. A hybrid multi-criteria meta-learner based classifier for imbalanced data. *Knowl. Based Syst.* **2024**, *285*, 111367. [CrossRef]
42. Chen, Z.; Duan, J.; Kang, L.; Qiu, G. Class-imbalanced deep learning via a class-balanced ensemble. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 5626–5640. [CrossRef] [PubMed]
43. Wang, B.; Pineau, J. Online bagging and boosting for imbalanced data streams. *IEEE Trans. Knowl. Data Eng.* **2016**, *28*, 3353–3366. [CrossRef]
44. Liu, T.; Fan, W.; Wu, C. A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset. *Artif. Intell. Med.* **2019**, *101*, 101723. [CrossRef] [PubMed]

45. El Khoury, R.; Al Beaïno, R. Classifying manufacturing firms in Lebanon: An application of Altman's model. *Procedia Soc. Behav. Sci.* **2014**, *109*, 11–18.
46. Gholampour, S.; Fatouraee, N.; Seddighi, A.; Seddighi, A. Numerical simulation of cerebrospinal fluid hydrodynamics in the healing process of hydrocephalus patients. *J. Appl. Mech. Tech. Phys.* **2017**, *58*, 386–391. [CrossRef]
47. Gholampour, S.; Fatouraee, N.; Seddighi, A.S.; Seddighi, A. Evaluating the effect of hydrocephalus cause on the manner of changes in the effective parameters and clinical symptoms of the disease. *J. Clin. Neurosci.* **2017**, *35*, 50–55. [CrossRef] [PubMed]
48. Gholampour, S. FSI simulation of CSF hydrodynamic changes in a large population of non-communicating hydrocephalus patients during treatment process with regard to their clinical symptoms. *PLoS ONE* **2018**, *13*, e0196216. [CrossRef] [PubMed]
49. Gholampour, S. Feasibility of assessing non-invasive intracranial compliance using FSI simulation-based and MR elastography-based brain stiffness. *Sci. Rep.* **2024**, *14*, 6493. [CrossRef] [PubMed]
50. Gholampour, S.; Mehrjoo, S. Effect of bifurcation in the hemodynamic changes and rupture risk of small intracranial aneurysm. *Neurosurg. Rev.* **2021**, *44*, 1703–1712. [CrossRef] [PubMed]
51. Hajirayat, K.; Gholampour, S.; Sharifi, I.; Bizari, D. Biomechanical simulation to compare the blood hemodynamics and cerebral aneurysm rupture risk in patients with different aneurysm necks. *J. Appl. Mech. Tech. Phys.* **2017**, *58*, 968–974. [CrossRef]
52. Gholampour, S.; Droessler, J.; Frim, D. The role of operating variables in improving the performance of skull base grinding. *Neurosurg. Rev.* **2022**, *45*, 2431–2440. [CrossRef]
53. Gholampour, S.; Gholampour, H. Correlation of a new hydrodynamic index with other effective indexes in Chiari I malformation patients with different associations. *Sci. Rep.* **2020**, *10*, 15907. [CrossRef] [PubMed]
54. Gholampour, S.; Taher, M. Relationship of morphologic changes in the brain and spinal cord and disease symptoms with cerebrospinal fluid hydrodynamic changes in patients with Chiari malformation type I. *World Neurosurg.* **2018**, *116*, e830–e839. [CrossRef] [PubMed]
55. Beinecke, J.; Heider, D. Gaussian noise up-sampling is better suited than SMOTE and ADASYN for clinical decision making. *BioData Mining* **2021**, *14*, 49. [CrossRef] [PubMed]
56. Ganaie, M.; Tanveer, M. Fuzzy least squares projection twin support vector machines for class imbalance learning. *Appl. Soft Comput.* **2021**, *113*, 107933. [CrossRef]
57. Boehme, A.K.; Esenwa, C.; Elkind, M.S. Stroke risk factors, genetics, and prevention. *Circ. Res.* **2017**, *120*, 472–495. [CrossRef] [PubMed]
58. Arboix, A. Cardiovascular risk factors for acute stroke: Risk profiles in the different subtypes of ischemic stroke. *World J. Clin. Cases WJCC* **2015**, *3*, 418. [CrossRef] [PubMed]
59. Webb, A.J.; Werring, D.J. New insights into cerebrovascular pathophysiology and hypertension. *Stroke* **2022**, *53*, 1054–1064. [CrossRef]
60. Phillips, S.J. Pathophysiology and management of hypertension in acute ischemic stroke. *Hypertension* **1994**, *23*, 131–136. [CrossRef] [PubMed]
61. Sidhu, N.S.; Kaur, S. Cerebrovascular Disease and Hypertension. In *Cerebrovascular Diseases-Elucidating Key Principles*; IntechOpen: London, UK, 2021.
62. Gorgui, J.; Gorshkov, M.; Khan, N.; Daskalopoulou, S.S. Hypertension as a risk factor for ischemic stroke in women. *Can. J. Cardiol.* **2014**, *30*, 774–782. [CrossRef] [PubMed]
63. Han, L.; Wu, Q.; Wang, C.; Hao, Y.; Zhao, J.; Zhang, L.; Fan, R.; Liu, Y.; Li, R.; Chen, Z. Homocysteine, ischemic stroke, and coronary heart disease in hypertensive patients: A population-based, prospective cohort study. *Stroke* **2015**, *46*, 1777–1786. [CrossRef] [PubMed]
64. Graor, R.A.; Hetzer, N.R. Current Concepts of Cerebrovascular Disease and Stroke. *Stroke* **1988**, *19*, 869–872.
65. Zhang, S.; Song, X.-Y.; Xia, C.-Y.; Ai, Q.-D.; Chen, J.; Chu, S.-F.; He, W.-B.; Chen, N.-H. Effects of cerebral glucose levels in infarct areas on stroke injury mediated by blood glucose changes. *RSC Adv.* **2016**, *6*, 93815–93825. [CrossRef]