Development and Evaluation of Stroke Disease Classification Models: Classical Machine Learning, Deep Learning, and Explainable AI Approaches

Lianny Wydiastuty Kusuma[1*], Andri Wijaya[2], Asahiro Nathanael Star Sitohang[3], Yo Ceng Giap[4]

[1,2,3,4] *Faculty of Science and Technology, Buddhi Dharma University*

*corr-author: lianny.wydiastuty@ubd.ac.id

Abstract - This study evaluates the impact of the Synthetic Minority Oversampling Technique (SMOTE) on improving machine learning and deep learning performance in stroke risk classification using secondary, publicly available data from Kaggle's Stroke Prediction Dataset (n = 5,110; 249 stroke cases, 4,861 non-stroke cases), for deep learning. Performance was measured using accuracy, precision, recall, and F1-score, while Explainable AI (XAI) methods (SHAP, LIME) were utilized for interpretability. The results show that applying SMOTE improves the model's sensitivity to the minority "Stroke" class, with Random Forest after SMOTE achieving 97% accuracy and a balanced precision–recall. These findings highlight the methodological potential of combining SMOTE with machine learning, deep learning, and XAI; however, they should not be interpreted as direct clinical validation. Future work with clinical and population-based datasets is necessary to assess the applicability in real-world healthcare settings.

Keywords: deep learning, machine learning, SMOTE, stroke, XAI

I. INTRODUCTION

Stroke is one of the leading causes of death and disability worldwide. According to [1], every year, more than 12 million people experience their first stroke, and approximately 6.5 million of them die as a result of this condition. In Indonesia, the prevalence of stroke continues to increase in line with changes in people's lifestyles, increased life expectancy, and the high prevalence of risk factors such as hypertension, diabetes, obesity, and smoking [2]. This condition not only increases the cost of medical care but also worsens patients' lives, incurring additional costs for families and the country [3]. Good stroke management includes stopping strokes from happening, finding them early, diagnosing them quickly, and giving them the proper treatment. This study utilizes secondary, publicly accessible Kaggle data rather than clinical hospital data, despite the capacity of modern data-driven methodologies, such as machine learning and deep learning, to enhance predictive modeling.

In healthcare, machine learning has made significant progress in a relatively short period, particularly in the areas of disease diagnosis and risk prediction. Machine learning can analyze complex medical data and identify patterns that are difficult for doctors to discern without specialized tools. Ref. [4] utilized Random Forest and Logistic Regression algorithms to predict stroke risk using clinical data, and the results show good generalization ability on balanced datasets. Another study by [5] conducted a comparative analysis of the performance of Support Vector Machine (SVM) and k-Nearest Neighbor (k-NN), demonstrating that SVM outperformed k-NN in managing multidimensional data. However, traditional machine learning techniques often encounter challenges with imbalanced datasets, characterized by a significantly higher number of non-stroke cases (the majority class) compared to stroke cases (the minority class). This imbalance causes the model to favor the majority class, which means it may not be susceptible to stroke cases and could result in incorrect predictions for patients who require medical attention.

In addition to classical machine learning methods, deep learning has become one of the most promising technologies, especially in medical image analysis. Deep learning has the advantage of automatically extracting complex features without requiring manual feature engineering processes. Ref. [6] employed a Convolutional Neural Network (CNN) architecture for classifying CT images of stroke patients, achieving higher accuracy compared to traditional methods. Ref. [7] utilized ResNet on diffusion MRI data to identify brain areas affected by stroke with higher precision. However, Deep learning performance is highly dependent on the availability of large and balanced datasets, whereas in the case of stroke, such datasets are relatively complex to obtain. This condition emphasizes the importance of applying data balancing techniques,

such as the Synthetic Minority Over-sampling Technique (SMOTE) [8,9], to reduce bias towards the majority class and improve model performance in stroke case detection. The use of SMOTE enables the replication of synthetic data in the minority class, thereby making the data distribution more proportional. Combined SMOTE–ENN with cascade ensemble learning, achieving higher balanced accuracy and sensitivity for stroke classification using IFLS5 data. This progression highlights the growing emphasis on hybrid sampling and multi-model integration in improving clinical prediction performance [10].

Another major problem with using AI technology in medicine is that it's hard to understand how the models work. Many deep learning algorithms, as well as a few advanced machine learning algorithms, operate as black boxes. This means that medical staff, who are the end-users, have difficulty understanding how the model makes decisions [8]. This raises doubts and resistance to the adoption of AI technology for clinical decision-making, especially in critical cases such as stroke. Healthcare professionals need clear and complete explanations of why the model made its predictions or diagnoses in the medical field. The Explainable AI (XAI) method addresses this issue by utilizing SHAP and LIME to provide explanations for the predictions [11]. The growing use of artificial intelligence in healthcare introduces a critical challenge: many high-performing models operate as “black boxes”, making their internal decision-making difficult to interpret and thus less suitable for clinical use. As emphasized by Sadeghi, explainable artificial intelligence (XAI) methods such as SHAP and LIME have become essential tools to improve model transparency, enable clinical interpretability, and foster trust among medical practitioners. In this study, these post-hoc explanation techniques are applied to tabular clinical data to clarify feature contributions and support domain experts in understanding model behavior [12].

Three significant research gaps exist in the application of AI for stroke detection, according to earlier studies. First, low sensitivity to minority (stroke) cases is a result of classical machine learning's difficulties with unbalanced datasets. Second, while deep learning works well with medical images, it needs large, balanced datasets, which is why data balancing techniques like SMOTE are useful. Third, because clinical trust in Explainable AI (XAI) depends on its interpretability, its application in stroke detection is still restricted.

These gaps indicate that to improve performance and interpretability, integrated research combining deep learning and classical methods, data balancing strategies, and XAI is required. By comparing deep learning (Sequential, TabNet) and classical (Logistic Regression, Random Forest) models on the same dataset, using SMOTE to enhance minority-class detection, and integrating SHAP and LIME for clear explanations, this study seeks to address all three problems at once.

The study makes three contributions: (1) evaluating the effects of SMOTE on stroke detection using precision, recall, F1-score, and PR metrics; (2) comparing classical and deep learning models on imbalanced data; and (3) improving model transparency through local and global XAI interpretations. The findings highlight the need for future validation using multi-source hospital data to ensure robustness and real-world applicability, as they are meant to show methodological feasibility rather than clinical readiness.

II. METHOD

This study utilized the Stroke Prediction Dataset from Kaggle, a secondary, publicly available dataset compiled from synthetic and aggregated sources, rather than real-world hospital or population-based records. This study used publicly available, anonymized data and did not involve human participants; thus, ethical clearance was not required. While this dataset provides a valuable benchmark for testing methodological approaches, its credibility as a clinical dataset is limited. Consequently, the results of this study should be interpreted as methodological demonstrations rather than clinically validated findings, and the generalizability to real patient populations remains uncertain. This dataset includes demographic, clinical, and patient health history variables, along with binary labels indicating whether the patient has ever had a stroke. The selection of this dataset is based on the completeness of attributes relevant to predictive analysis. The available data provides an opportunity to test various machine learning and deep learning models in detecting stroke risk. Additionally, this dataset has been used in previous studies, enabling objective comparison of results [5, 13].

The dataset used in this study comprised a total of 5,110 initial data points, consisting of 249 labeled as stroke and 4,861 labeled as non-stroke. After preprocessing, the number of samples was reduced to 4,909 data points, with a class distribution of 209 data points labeled as stroke and 4,700 data points labeled as non-stroke. This imbalance in the number of samples between classes suggests a significant dominance of non-stroke data over stroke data, which could introduce bias into the machine learning model being built.

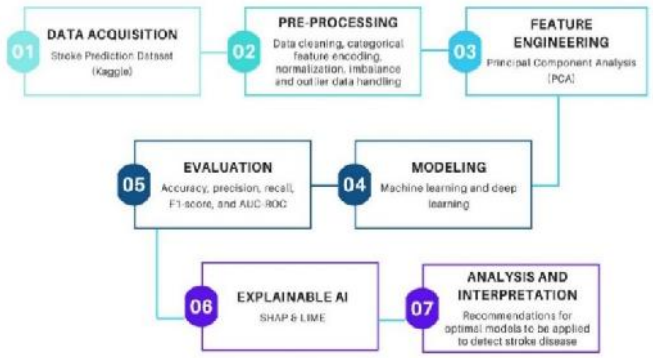

Fig. 1 Research methods

The Synthetic Minority Oversampling Technique (SMOTE) method was employed to address this issue. This method increased the amount of data in the minority class (stroke) to match that of the majority class. The results of using SMOTE indicate that the class distribution is balanced, with 4,700 stroke data points and an equal number of non-stroke data points. So, after resampling, the total sample size is 9,400 data points.

Another significant limitation pertains to the implementation of the Synthetic Minority Oversampling Technique (SMOTE). SMOTE can help address class imbalance and facilitate the identification of minority classes. Still, it can also create synthetic patterns that don't fully represent real clinical variation, which could lead to bias or overfitting. Consequently, the noted performance improvements must be regarded with caution until corroborated by genuine clinical data. Table I presents the main variables used in this study, along with their clinical significance, to clarify their meaning. This mapping helps readers, especially those in healthcare, see how each attribute is related to predicting stroke risk.

Pre-processing is performed to enhance data quality before entering the modeling stage. The steps taken include removing duplicate data (using the 'drop_duplicates' method), handling missing values through imputation methods [14] and feature selection using a correlation matrix [15]. This selection only keeps features that are strongly related to the target variable. Data balancing has not been performed yet because the SMOTE method will be used independently to compare model performance. This method facilitates an examination of how SMOTE impacts the final model results [9, 16, 17].

During the data pre-processing stage, several necessary steps were taken to ensure the quality of the dataset before it was used in the modeling process. First, data with missing values were removed entirely. The choice was made because the sample size was sufficiently large, comprising approximately 5,000 data points. Removing approximately 200 data points with missing values didn't significantly alter the overall representativeness of the dataset. This method was chosen over imputation methods, such as the mean or median, because it was believed to preserve the original data better.

Then, a correlation analysis was conducted to identify features that weren't particularly useful for the target variable. The correlation matrix showed that some features had very low negative correlation values. These were gender_Other (-0.003010), work_type_Never_worked (-0.014149), and work_type_children (-0.080971). After that, these features were removed from the dataset because they didn't significantly contribute to the classification process. The model should work better if these features are removed, as it will only use variables that are highly relevant to the prediction.

TABLE I
DATASET VARIABLES AND CLINICAL DESCRIPTIONS

Variable	Type	Clinical Meaning
age	Numerical	Patient’s age (older age is a major stroke risk factor)
gender	Categorical	Biological sex (males and females may have different stroke risk profiles)
hypertension	Binary	History of high blood pressure (major risk factor for stroke)
heart_disease	Binary	History of heart disease (linked with higher stroke risk)
ever_married	Categorical	Marital status (proxy for lifestyle/social determinants)
work_type	Categorical	Type of employment (affects lifestyle, stress, and health risk)
residence_type	Categorical	Urban vs rural living (proxy for environmental and lifestyle differences)
avg_glucose_level	Numerical	Blood glucose concentration (indicator of diabetes risk, linked to stroke)
bmi	Numerical	Body Mass Index (an indicator of obesity, a key risk factor)
smoking_status	Categorical	Smoking habits (well-established stroke risk factor)
stroke (target label)	Binary	Outcome variable: 1 = stroke, 0 = no stroke

In the experimental setup, multiple measures were implemented to ensure the reproducibility of the research process and achieve uniform outcomes. First, SMOTE was used to oversample the data, providing an even class distribution. Next, the train_test_split function is used to split the resampled dataset into a training set and a testing set. The training set comprises 96% of the data, and the testing set shall consist of 4% (test_size=0.04). At this point, a random seed of 12 (random_state=12) is used to ensure that the data remains the same every time the code is run.

After dividing the data, normalization was performed using StandardScaler so that each feature had the same scale. The training data was then divided by Keras into training data and validation data through the validation_split=0.4 parameter. This means that 40% of the training data was used as the validation set, allowing the model to be evaluated during training. With these settings, the experiment can be replicated in the same way, as all critical parameters, such as random seed, train–test split proportion, and validation, have been described transparently.

There are two primary approaches to modeling data: classical machine learning and deep learning. Some of the classical machine learning models that were used are Random Forest, Logistic Regression, Support Vector Machine (SVM), and k-Nearest Neighbor (k-NN) [4-5]. The deep learning model used is a Convolutional Neural Network (CNN) adapted for tabular data [6-7]. Each model was tested in two scenarios: one without SMOTE and the other with SMOTE. The SMOTE technique was used to address class imbalance by synthesizing minority data [20].

The model was evaluated by measuring accuracy, precision, recall, F1-score, and AUC-ROC [21]. We chose these metrics because they provide a comprehensive view of how well the model identifies minority classes. Precision and recall evaluate how well you can make correct optimistic predictions and how effectively you can identify positive cases. The F1-score provides a harmonic mean of these two metrics, while the AUC-ROC indicates how well the model distinguishes between positive and negative classes. This evaluation approach ensures that the best model is selected objectively [4, 6].

The explainable AI (XAI) stage is carried out to provide interpretations of the predictions generated by the model. Two methods used are SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) [11, 22]. SHAP is used to measure the contribution of each feature to the model's prediction results globally [23]. LIME is used to explain individual predictions locally, making it easier to interpret specific cases [23-24]. The use of XAI is essential to increase medical professionals' trust in the results of prediction systems [11, 22].

The final stage is the analysis and interpretation of results. At this stage, the model evaluation results are compared between scenarios with and without SMOTE. Essential factors affecting predictions are analyzed based on SHAP and LIME outputs. The best model is selected based on a combination of performance and interpretability. The resulting model recommendations are expected to be implemented in medical decision support systems for early detection of stroke risk [13, 16].

III. RESULT AND DISCUSSION

This study uses two main approaches, namely machine learning and deep learning, to compare the performance of stroke case detection in balanced and unbalanced datasets. In the machine learning approach, the algorithms used are Logistic Regression and Random Forest, which are evaluated on the No Stroke and Stroke classes. Meanwhile, in the deep learning approach, two different architectures are used: the Sequential model and the TabNet Classifier. Both methods are evaluated in two scenarios—before and after the implementation of the SMOTE technique—to assess the influence of data balancing on the detection of minority classes.

A. Machine Learning

This study used two machine learning algorithms: Random Forest and Logistic Regression. We used precision, recall, F1-score, and accuracy metrics to compare the performance of both models for the No Stroke and Stroke classes. The following table provides a summary of the evaluation results, facilitating a comprehensive analysis and performance comparison.

The results of evaluating the machine learning models without SMOTE indicate that the majority class (No Stroke) and the minority class (Stroke) do not perform equally well. The Logistic Regression model achieved 73% accuracy, but the Stroke class had a very low precision value of 0.07, despite a relatively high recall value of 0.83. This means that the model often guessed the Stroke class, but it did so with a high number of incorrect predictions (false positives), which lowered its F1-Score to only 0.13. The F1-Score, which is the harmonic mean of precision and recall, decreases significantly when recall is high but precision is low. This illustrates the significant disparity between the two metrics.

TABLE II
MACHINE LEARNING MODEL RESULTS

Model	SMOTE	Class	Precision	Recall	F1-Score	Support	Accuracy
Logistic Regression	No	*No Stroke*	0.99	0.73	0.84	245	0.73
		Stroke	0.07	0.83	0.13	6	
Random Forest	No	*No Stroke*	0.98	1.00	0.99	246	0.98
		Stroke	0.00	0.00	0.00	5	
Logistic Regression	Yes	*No Stroke*	0.91	0.87	0.89	236	0.89
		Stroke	0.88	0.91	0.90	244	
Random Forest	Yes	*No Stroke*	0.96	0.98	0.97	233	0.97
		Stroke	0.98	0.96	0.97	247	

The Random Forest model, on the other hand, achieved an overall accuracy of 98%. Still, it completely failed to identify any Stroke cases, as indicated by a recall value of 0.00 and an F1-Score of 0.00. This means that the model was very biased toward the majority class (No Stroke), only predicting it and completely ignoring the minority class. These results confirm the initial hypothesis that imbalanced data directly hinders the model's ability to identify minority classes, especially in medical datasets where positive (disease) cases are infrequent.

After applying the SMOTE oversampling method, both models showed significant improvements. The Logistic Regression model, which was used to demonstrate a substantial difference in performance between classes, achieved an accuracy of 89% with more even precision (0.91) and recall (0.87) across both classes. This improvement enables the model to identify patterns in the minority class more easily when the data is more evenly distributed.

After applying SMOTE, the Random Forest model also showed significant improvement, with an accuracy of 97% and both precision and recall equal to 0.97 for the Stroke class. As a result, its F1-Score also reached 0.97, which means that SMOTE improved the model's sensitivity (recall) without hurting its accuracy, resulting in a balanced predictive performance. However, these results should be interpreted with caution, as the evaluation was conducted on a singular secondary dataset without external validation. Such high performance could indicate that the model is overfitting to the oversampled data, meaning it learns specific patterns rather than generalizing from them.

In general, these results indicate that when there is an imbalance in the data, models tend to overlook minority classes, which are crucial to identify in medical prediction tasks. The use of SMOTE altered the model's learning process, enhancing its ability to identify minority classes while maintaining the accuracy of majority classes. The results also show that recall and F1-Score are essential measures of how well a model works with unbalanced medical data. Recall ensures the model identifies as many actual positive cases as possible, while the F1-Score ensures this improvement doesn't come at the cost of too many false positives.

Data balancing is essential for medical classification tasks. With a well-balanced dataset, both linear models, such as Logistic Regression, and tree-based models, such as Random Forest, can yield consistent and easy-to-understand results. This success demonstrates that SMOTE and other resampling methods can effectively handle imbalanced datasets in medical diagnostics. This opens the door for their use in disease prediction systems.

B. Deep Learning

This study utilized two architectures within the deep learning framework: the Sequential model and the TabNet Classifier. We used precision, recall, F1-score, and accuracy metrics to compare the performance of both models for the No Stroke and Stroke classes. The following table summarizes the test results to facilitate easier analysis and comparison of performance across different data conditions.

The evaluation results of the deep learning models without SMOTE application exhibited a similar issue identified in the machine learning models—specifically, a significant bias towards the No Stroke class. The Sequential model without SMOTE achieved a high overall accuracy of 96%, but it failed to identify any Stroke cases, resulting in precision and recall values of 0.00 for the minority class. This result shows that, although the accuracy may seem high, the model cannot reliably determine which class is most important for clinical purposes. For medical classification, this kind of performance is misleading because high accuracy can mean that the model is good at predicting the majority class.

TABLE III
DEEP LEARNING MODEL RESULTS

Model	SMOTE	Class	Precision	Recall	F1-Score	Support	Accuracy
Sequential	No	*No Stroke*	0.96	1.00	0.98	2965	0.96
		Stroke	0.00	0.00	0.00	128	
TabNet Classifier	No	*No Stroke*	0.97	0.81	0.88	4089	0.80
		Stroke	0.08	0.38	0.14	182	
Sequential	Yes	*No Stroke*	0.92	0.90	0.91	190	0.91
		Stroke	0.90	0.92	0.91	186	
TabNet Classifier	Yes	*No Stroke*	0.97	0.94	0.95	94	0.95
		Stroke	0.94	0.97	0.95	94	

The TabNet Classifier model without SMOTE performed slightly better, achieving a recall of 0.38 for the Stroke class. However, its accuracy remained very low at 0.08, indicating a high number of false positives. As a result, the F1-score for this class dropped to 0.14, indicating a clear imbalance between precision and recall. This relationship shows that the F1-score, which is the harmonic mean of precision and recall, will only be high when both metrics are equal. In this case, even though the recall improved slightly, the very low precision prevented the F1-score from increasing significantly. This suggests that data imbalance hinders the model's ability to generalize patterns associated with minority groups accurately.

Both models demonstrated significant improvements in performance after applying the SMOTE oversampling method. The Sequential model's accuracy went up to 91%, and the precision (0.90) and recall (0.92) were well-balanced across both classes. This balance demonstrates that SMOTE enabled the model to learn essential patterns from the minority class without compromising its ability to identify the majority class.

After SMOTE was used, the TabNet Classifier did even better, getting 95% accuracy and a precision and recall of 0.95 for the Stroke class. This balance between precision and recall resulted in a consistently high F1-score (0.95), indicating that the model maintained a high percentage of correct optimistic predictions while minimizing false classifications. The close match between recall and F1-score also indicates that the model not only identified more true positive cases but also maintained the reliability of its predictions, a crucial requirement for medical decision-support systems.

When comparing the effects of SMOTE on deep learning models to those on machine learning models, the impact on deep learning models is even more consistent. After data balancing, both Sequential and TabNet showed balanced performance. However, TabNet stood out for keeping high accuracy with a smaller dataset. This finding suggests that TabNet is more efficient in terms of computing power while maintaining high detection quality, making it a suitable choice for real-world applications with limited computing resources.

In general, these results demonstrate the importance of preprocessing and balancing data when working with medical datasets that aren't balanced. The increase in recall directly raised the F1-score, showing the connection between sensitivity (recall) and model reliability (F1-score) in health-related predictions. From a clinical perspective, enhancing recall ensures the accurate identification of a greater number of at-risk patients, while sustaining a high F1-score guarantees that these identifications are significant and not overwhelmed by false positives.

Using SMOTE made deep learning models for stroke prediction much more sensitive and overall stronger. These results demonstrate that balanced data distributions enable deep learning models to be both statistically and clinically relevant. This opens the door to more widespread use of SMOTE-based methods in other areas of disease diagnosis where identifying minority cases early is crucial.

The comparative analysis between traditional machine learning models (Logistic Regression, Random Forest) and deep learning models (Sequential, TabNet) reveals that they perform differently due to their distinct architectural designs. Without SMOTE, all of the models were very biased toward the No Stroke class. This is a common problem when training on datasets that aren't evenly distributed. Linear models, including Logistic Regression and Sequential, were unable to capture intricate nonlinear feature relationships, leading to inadequate Stroke detection despite a comparatively high overall accuracy.

On the other hand, Random Forest, which employs an ensemble approach, proved more robust for tabular medical data because it can model nonlinear patterns

using multiple decision trees. However, it still failed to address minority cases (recall = 0.00), even without balancing. After using SMOTE, the precision, recall, and F1-score all increased significantly. This shows that ensemble models work well with balanced data.

TabNet Classifier consistently outperformed other deep learning models, getting balanced precision and recall after SMOTE. TabNet uses sequential attention mechanisms to focus on the most important clinical features (like age, hypertension, and BMI), which makes it easier to understand and works better than regular neural networks. Its hybrid design, which combines tree-like feature selection with neural attention, enables it to learn effectively from sparse, diverse tabular data and maintain high accuracy even with smaller datasets.

In general, both Random Forest and TabNet achieved the most from SMOTE, as it utilized feature diversity to enhance discriminative power without overfitting. Random Forest is better suited for modeling nonlinear relationships, while TabNet excels at adapting to and explaining complex medical data. These results demonstrate that selecting a model and balancing the data are closely linked, and that attention-based architectures, such as TabNet, are effective for both prediction and interpretation in medical prediction tasks where the data is unbalanced.

C. Xai SHAP

The SHAP (SHapley Additive Explanations) analysis results in Fig. 2 show that age, smoking status, and occupation are the three most important factors that affect stroke risk prediction. Age is strongly linked to a higher risk of stroke, which means that older people are more likely to have one. On the other hand, not smoking lowers the risk of having a stroke by a lot. Occupation also plays a significant role, likely due to factors such as stress levels, physical activity levels, or exposure to certain environmental risks. This analysis corroborates prior medical research indicating that demographic and lifestyle factors are significant predictors of stroke.

Fig. 3 shows a SHAP waterfall plot visualization that provides a detailed view of how each feature affects a stroke prediction case. In this instance, variables such as residing in an urban environment, being male, and having never smoked reduce the likelihood of stroke. But older age and some jobs are essential factors that make a stroke more likely. This information is crucial because it can help identify risk factors that can be modified, such as making lifestyle changes or adjusting the work environment. The waterfall plot makes it easier for both patients and healthcare professionals to understand how decisions are made.

Additionally, Fig. 4 shows a graph of the average feature contributions made by SHAP, which were then compared to the coefficients in the Logistic Regression model. The results of this comparison demonstrate a high level of consistency, with age and lifestyle factors, such as smoking habits, remaining the most significant in both analyses. This consistency makes the results more reliable and reduces the likelihood of bias that could arise from using only one model. These findings have pragmatic implications for stroke prevention strategies, especially regarding smoking cessation and healthcare for older people. The agreement of the results from SHAP and Logistic Regression also makes the prediction system more reliable.

D. XAI LIME

The Local Interpretable Model-Agnostic Explanations (LIME) method was used in this study to give in-depth explanations of each prediction made by the Random Forest model. This method emphasizes local interpretation, which helps you understand better why a specific prediction was made. In the case examined, the model had a 96% chance of predicting the “No Stroke” class. In contrast, the probability for the “Stroke” class was only 4% in Fig. 5. This significant difference in probability indicates that the model is highly confident in its prediction of no stroke. LIME helps to sort out the factors that affect these probabilities, separating the beneficial and detrimental effects on the outcome.

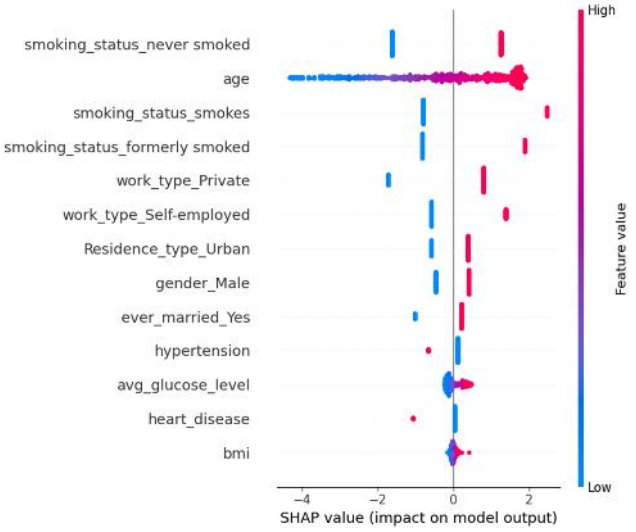

Fig. 2 Summary plot of SHAP results

Fig. 3 Force plot of SHAP results

Although the majority of factors in Fig. 5 show a negative contribution to stroke risk, some variables make a positive, though not dominant, contribution. These factors include a history of heart disease, hypertension, male gender, and urban residence. These variables increase the probability of stroke by a small amount, but are still relevant for risk interpretation. The presence of these positive factors suggests that even if a patient is predicted to be "No Stroke," there are certain conditions that require further attention. This LIME analysis complements the global findings from SHAP, providing a more contextualized, personalized picture for each individual. Integrating the results from Fig. 5 with the SHAP visualization offers a more comprehensive approach, increasing the transparency and accountability of the machine learning-based prediction system.

The SHAP and LIME explanations successfully pinpointed significant predictive attributes, including age, smoking status, and occupation, aligning with recognized medical risk factors for stroke. However, the interpretation can be enhanced by more clearly linking these model-derived feature importances to clinical reasoning. For instance, SHAP values that show a strong positive effect of advanced age and smoking status on stroke risk are in line with epidemiological studies that show that both factors make the cerebrovascular system more vulnerable. Similarly, the occupational variable may indicate stress resulting from lifestyle choices or levels of physical activity, both of which are known to impact heart health. Focusing on these clinical correlations strengthens the idea that the model's feature importance is not only statistically sound but also clinically reasonable and understandable. This makes it more trustworthy for medical decision support.

However, even though SHAP and LIME make models much more transparent, they aren't always easy to understand. SHAP values can be influenced by the degree of relatedness between features. This means that if two or more clinical variables (such as age and blood pressure) are closely related, their individual contributions may not be distributed evenly, which could lead to a misinterpretation of their importance. LIME explanations also depend on several parameter settings, such as the number of perturbation samples or the width of the kernel. Changing these parameters could lead to slightly different local explanations, which could make the interpretation less stable and less reproducible. Consequently, the insights offered by SHAP and LIME should be viewed as auxiliary diagnostic explanations rather than definitive causal evidence. Recognizing these limitations is essential to guarantee that explainability techniques are interpreted judiciously and in conjunction with clinical expertise, preserving both statistical validity and medical relevance in decision support applications.

In line with Kourou observations regarding the significance of data heterogeneity and feature selection in medical prediction, this study achieved even greater improvement—particularly in the post-SMOTE Random Forest model—than previous works, which emphasized the importance of data quality and feature relevance in improving healthcare classification models. This likely resulted from the integration of clinically relevant attributes and optimized data preprocessing [27]. Overall, the results confirm that SMOTE improves minority class detection across both machine learning and deep learning models. However, the absence of external validation across independent datasets limits the generalizability of these findings. Moreover, the combination of oversampling and a relatively small dataset size increases the risk of overfitting, which may artificially inflate performance metrics. To address these issues, future work should include validation on multi-source or hospital-based clinical datasets to ensure robustness, reliability, and potential clinical relevance.

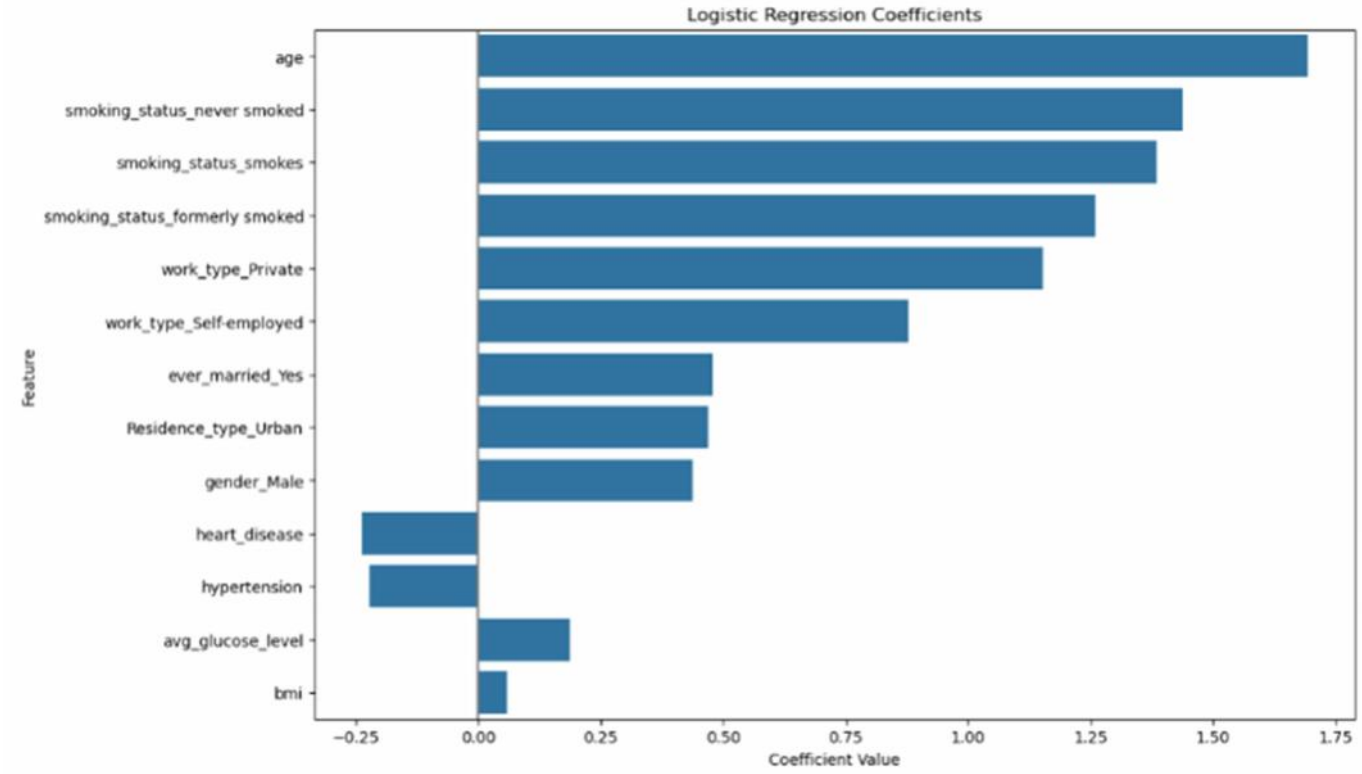

Fig. 4 Bar plot logistic regression coefficients

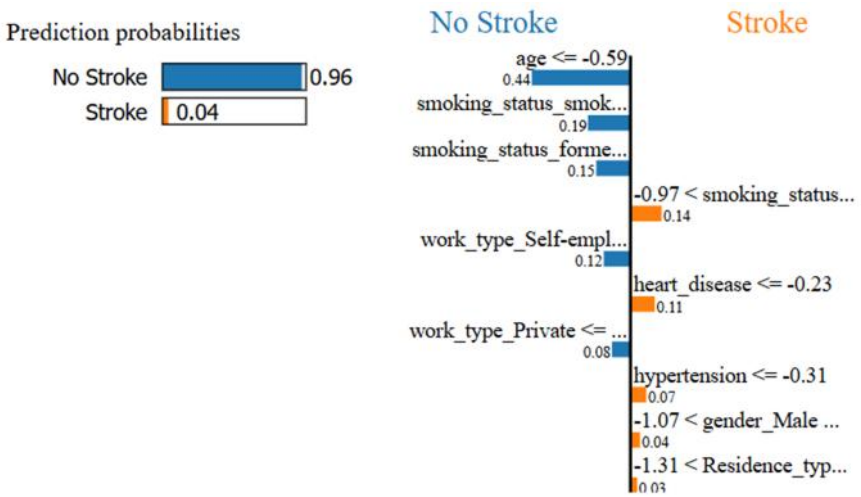

Fig. 5 Lime output

The practical implication of this research is the need to apply data balancing methods and model interpretability to medical decision support systems, especially for low-incidence diseases such as stroke [26]. The research results can be used by healthcare institutions to develop screening algorithms that are more sensitive to high-risk cases, while maintaining transparency in clinical decision-making [11]. Theoretically, this research contributes to the development of a hybrid approach that combines data balancing and explainable AI as a strategy to improve accuracy and user trust. However, this research has limitations, including the use of a single dataset and the failure to test the model on real-time data or across different populations [27]. For future research, it is recommended to conduct external validation using multi-source data, explore more adaptive balancing techniques, and compare the results with other interpretability methods, such as counterfactual explanations. This will allow continued research to

strengthen the practical application and academic relevance of these findings in the future.

This study has significant limitations, despite its promising results. The use of the Kaggle Stroke Prediction Dataset means that the results may not accurately reflect the prevalence of strokes, population diversity, or clinical workflows in the real world. It is a synthetic and aggregated dataset, so it doesn't have the depth and variety of data from hospitals or populations. Additionally, the oversampling process with SMOTE, while it does help identify minority classes more effectively, can also create artificial patterns that aren't present in real patients. Consequently, the robust performance metrics observed should be regarded as an indication of methodological potential rather than direct clinical applicability.

IV. CONCLUSION

This research demonstrates that integrating data balancing methodologies, such as SMOTE, with interpretable machine learning and deep learning models can significantly enhance the identification of minority classes in medical prediction tasks, particularly in stroke risk evaluation. By using explainable AI methods like SHAP and LIME, the models not only made accurate predictions but also provided clinically useful information that aligned with known risk factors. The findings underscore the significance of explainability and data preprocessing in guaranteeing the reliability and credibility of AI-driven medical decision support systems. However, since this study utilized a singular dataset without external validation, subsequent research should concentrate on cross-institutional datasets, prospective data collection, and real-world implementation to assess generalizability and practical applicability. Expanding this framework to additional disease areas and incorporating it with secure, privacy-compliant clinical infrastructures will promote the utilization of explainable and ethically robust AI in healthcare.

REFERENCES

[1] World Stroke Organization, “Impact of Stroke,” https://www.world-stroke.org/world-stroke-day-campaign/about-stroke/impact-of-stroke.

[2] U.S. Centers for Disease Control and Prevention (CDC), “Stroke Facts,” https://www.cdc.gov/stroke/data-research/facts-stats/index.html.

[3] S. Strilciuc, “The economic burden of stroke: a systematic review of cost of illness studies,” *J Med Life*, vol. 14, no. 5, pp. 606–619, Jan. 2021, doi: 10.25122/jml-2021-0361.

[4] A. Hassan, S. Gulzar Ahmad, E. Ullah Munir, I. Ali Khan, and N. Ramzan, “Predictive modeling and identification of key risk factors for stroke using machine learning,” *Sci Rep*, vol. 14, no. 1, p. 11498, May 2024, doi: 10.1038/s41598-024-61665-4.

[5] S. Yakut and N. Barişçi, “Comparison of Machine Learning and Deep Learning Techniques for Stroke Prediction,” vol. 17, no. 1, pp. 11–27, 2025, doi: 10.29137/ijerad.1432162.

[6] P. L. Chiang, “Deep Learning-Based Automatic Detection of ASPECTS in Acute Ischemic Stroke: Improving Stroke Assessment on CT Scans,” *J Clin Med*, vol. 11, no. 17, Sep. 2022, doi: 10.3390/jcm11175159.

[7] H. Yu, “Prognosis of ischemic stroke predicted by machine learning based on multi-modal MRI radiomics,” *Front Psychiatry*, vol. 13, Jan. 2023, doi: 10.3389/fpsyt.2022.1105496.

[8] R. Sebastian and C. Juliane, “Comparison of Data Mining Classification Algorithms for Stroke Disease Prediction Using the SMOTE Upsampling Method,” vol. 11, no. 2, pp. 311–321, 2023, doi: https://doi.org/10.30595/juita.v11i2.17348.

[9] A. Barragán-Montero, “Artificial intelligence and machine learning for medical imaging: A technology review,” *Physica Medica*, vol. 83, pp. 242–256, Mar. 2021, doi: 10.1016/j.ejmp.2021.04.016.

[10] W. P. Indahwati and F. M. Afendi, “Improving Stroke Detection with Hybrid Sampling and Cascade Generalization,” vol. 12, no. 1, pp. 9–18, May 2024, doi: https://doi.org/10.30595/juita.v12i1.19386.

[11] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, “Explainable AI: A Review of Machine Learning Interpretability Methods,” *Entropy*, vol. 23, no. 1, p. 18, Dec. 2020, doi: 10.3390/e23010018.

[12] Z. Sadeghi, “A review of Explainable Artificial Intelligence in healthcare,” *Computers and Electrical Engineering*, vol. 118, p. 109370, Aug. 2024, doi: 10.1016/j.compeleceng.2024.109370.

[13] M. Issaiy, D. Zarei, S. Kolahi, and D. S. Liebeskind, “Machine learning and deep learning algorithms in stroke medicine: a systematic review of hemorrhagic transformation prediction models,” Jan. 01, 2025, *Springer Science and Business Media Deutschland GmbH*. doi: 10.1007/s00415-024-12810-6.

[14] E. Rahm and H. H. Do, “Data Cleaning: Problems and Current Approaches,” 2000, [Online]. Available: https://www.researchgate.net/publication/220282831

[15] P. Schober, C. Boer, and L. A. Schwarte, “Correlation Coefficients: Appropriate Use and Interpretation,” *Anesth Analg*, vol. 126, no. 5, pp. 1763–1768, May 2018, doi: 10.1213/ANE.0000000000002864.

[16] M. Avanzo, J. Stancanello, G. Pirrone, A. Drigo, and A. Retico, “The Evolution of Artificial Intelligence in

Medical Imaging: From Computer Science to Machine and Deep Learning," Oct. 2024, doi: 10.20944/preprints202410.0025.v1.

[17] M. Sulistiyono, Y. Pristyanto, S. Adi, and G. Gumelar, "Implementasi Algoritma Synthetic Minority Over-Sampling Technique untuk Menangani Ketidakseimbangan Kelas pada Dataset Klasifikasi," *SISTEMASI*, vol. 10, pp. 445–459, May 2021, [Online]. Available: http://sistemasi.ftik.unisi.ac.id

[18] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *The Royal Society*, vol. 374, no. 2065, p. 20150202, Apr. 2016, doi: 10.1098/rsta.2015.0202.

[19] K. Moulaei, L. Afshari, R. Moulaei, B. Sabet, S. M. Mousavi, and M. R. Afrash, "Explainable artificial intelligence for stroke prediction through comparison of deep learning and machine learning models," *Sci Rep*, vol. 14, no. 1, p. 31392, Dec. 2024, doi: 10.1038/s41598-024-82931-5.

[20] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," 2002.

[21] T. Saito and M. Rehmsmeier, "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets," *PLoS One*, vol. 10, no. 3, p. e0118432, Mar. 2015, doi: 10.1371/journal.pone.0118432.

[22] Z. Salahuddin, H. C. Woodruff, A. Chatterjee, and P. Lambin, "Transparency of deep neural networks for medical image analysis: A review of interpretability methods," *Comput Biol Med*, vol. 140, p. 105111, Jan. 2022, doi: 10.1016/j.compbiomed.2021.105111.

[23] S. M. Lundberg, P. G. Allen, and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 4768–4777, 2017, [Online]. Available: https://github.com/slundberg/shap

[24] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?,'" in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, Aug. 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.

[25] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning from Imbalanced Data Sets*. Cham: Springer International Publishing, 2018. doi: 10.1007/978-3-319-98074-4.

[26] V. L. Feigin, "Global, regional, and national burden of stroke and its risk factors, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021," *Lancet Neurol*, vol. 23, no. 10, pp. 973–1003, Oct. 2024, doi: 10.1016/S1474-4422(24)00369-7.

[27] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Comput Struct Biotechnol J*, vol. 13, pp. 8–17, 2015, doi: 10.1016/j.csbj.2014.11.005.