

## BUKTI KORESPONDENSI ARTIKEL JURNAL

|                     |   |                                                                                                                                              |
|---------------------|---|----------------------------------------------------------------------------------------------------------------------------------------------|
| Judul Artikel       | : | Development and Evaluation of Stroke Disease Classification Models: Classical Machine Learning, Deep Learning, and Explainable AI Approaches |
| Jurnal              | : | JUITA: Jurnal Informatika                                                                                                                    |
| Akreditasi / Scopus | : | Sinta 2                                                                                                                                      |
| Penulis             | : | Lianny Wydiastuty Kusuma, Andri Wijaya, Asahiro Nathanael Star Sitohang, Yo Ceng Giap                                                        |

| No | Perihal                                                      | Tanggal          |
|----|--------------------------------------------------------------|------------------|
| 1  | Bukti konfirmasi Submit artikel ke JUITA                     | 19 Agustus 2025  |
| 2  | Bukti konfirmasi revisi dengan status revisions are required | 15 Oktober 2025  |
| 3  | Bukti konfirmasi status accepted jurnal                      | 29 Oktober 2025  |
| 4  | Bukti informasi production                                   | 03 November 2025 |
| 5  | Publish jurnal                                               | 08 November 2025 |

Bukti Akreditasi:



Kementerian Pendidikan Tinggi,  
Sains, dan Teknologi

### SERTIFIKAT

Akreditasi Jurnal



TERAKREDITASI

No. SK : 10/C/C3/DT.05.00/2025

Tanggal : 21 Maret 2025

Direktur Penelitian, dan Pengabdian kepada Masyarakat  
dengan ini memberikan kepada

**JUITA : Jurnal Informatika**

EISSN : 25798901  
Publisher : Prodi Teknik Informatika Universitas Muhammadiyah Purwokerto

**Peringkat Akreditasi Jurnal Ilmiah Periode I 2025**

Akreditasi Ulang di Peringkat 2 mulai  
Volume 11 Nomor 2 Tahun 2023 sampai Volume 16 Nomor 1 Tahun 2028



I Ketut Adnyana  
NIP 196805151994031004

## Bukti Submit Jurnal

**JUITA: Jurnal Informatika**

Workflow Publication

Submission Review Copyediting Production

### Submission Files

| ID    | File Name                                                                       | Date            | Type         |
|-------|---------------------------------------------------------------------------------|-----------------|--------------|
| 79204 | Development and Evaluation of Stroke Disease Classification Models.docx         | August 19, 2025 | Article Text |
| 79205 | Development and Evaluation of Stroke Disease Classification Models Turnitin.pdf | August 19, 2025 | Other        |

[Download All Files](#)

---

**Submissions**

My Queue (1) Archives Help

### My Assigned

| ID    | Author        | Title                                                                                                   | Status      | Action      |
|-------|---------------|---------------------------------------------------------------------------------------------------------|-------------|-------------|
| 27868 | Kusuma et al. | Development and Evaluation of Stroke Disease Classification Models: Classical Machine Learning Approach | In Progress | Review View |

[JUITA] Editor Decision



Pengirim

Ardhista Shabrina Fitri <jurnalump01@gmail.com>

Penerima

Lianny Wydiastuty Kusuma <lianny.wydiastuty@ubd.ac.id>, Ceng Giap Yo <cenggiap@ubd.ac.id>, Andri Wijaya <andri.wijaya@ubd.ac.id>

Tanggal

2025-10-15 18:36

 C-Development and Evaluation of Stroke Disease Classification Models Revisi 1.docx (~664 KB)

Lianny Wydiastuty Kusuma, Ceng Giap Yo, Andri Wijaya:

We have reached a decision regarding your submission to JUITA: Jurnal Informatika, "Development and Evaluation of Stroke Disease Classification Models: Classical Machine Learning, Deep Learning, and Explainable AI Approaches".

Our decision is: revisions are required.

Thank you for revising the manuscript, but there are still some things that need attention.

Authors are required not only to revise the manuscript based on the reviewers' comments, but also to address the comments and suggestions provided by the editor.

JUITA is very strict in script writing. It's important to carefully review and follow the **author's guidelines and journal template** when revising your manuscript. As stated in the author guidelines, the maximum length of a manuscript is 10 pages, with a tolerance of up to 2 additional pages.

There is no need to include a separate "Discussion" subsection; instead, the discussion should be integrated within the "Results and Discussion" section to provide a holistic interpretation of the results and findings.

Employ grammar in accordance with academic requirements. Kindly ensure that all figures, tables, and equations are properly cited within the main text; that in-text citations conform to the IEEE referencing style—which excludes author names and publication years; and that the reference list is accurately and consistently formatted in accordance with academic conventions. Each author's name must be written using only the last name, followed by the initials of the given names, and all the author's names of the references are written (don't use "et al.") with the correct name writing order. The volume and issue, the page number, and the DOI are included in references originating from journal articles. It is highly recommended to cite papers published in JUITA.

Ensure that the plagiarism check has been carried out in accordance with JUITA's guidelines. Although the check have been completed, the responsibility for the originality and integrity of the manuscript with respect to plagiarism lies entirely with the author.

Thank you.  
JUITA Editor.

=====

## MATRIX OF REVISION NOTE

Paper ID : 27868  
 Title : Development and Evaluation of Stroke Disease Classification Models: Classical Machine Learning, Deep Learning, and Explainable AI Approaches  
 Authors : Lianny Wydiastuty Kusuma, Andri Wijaya, Asahiro Nathanael Star Sitohang, Yo Ceng Giap

| No. | Reviewer's & Editor's Comments                                                                                                                                                                                                                                                                                                                                                                  | Revision*                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
|-----|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1.  | <ul style="list-style-type: none"> <li>Clarify that the study uses secondary, public Kaggle data, not clinical hospital or population-based data.</li> <li>Avoid overstating clinical applicability; results show methodological potential, not direct clinical validation.</li> </ul>                                                                                                          | <p>Added explanation about:</p> <p>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</p> <p>These findings highlight the methodological potential of combining SMOTE with machine learning, deep learning, and XAI, but should not be interpreted as direct clinical validation.</p> <p>Introduction:<br/>       this study does not use clinical hospital data, but rather secondary, publicly available Kaggle data</p>                                                                                                                                                                                                                 |
| 2.  | <p>Data</p> <ul style="list-style-type: none"> <li>The main limitation is the dataset credibility: it is not a real-world clinical dataset, which affects generalizability.</li> <li>Authors should explicitly acknowledge this and discuss risks of bias in oversampled (SMOTE) data.</li> <li>Recommend adding a table of variables with clinical meaning to improve transparency.</li> </ul> | <p>Added explanation about:</p> <p>This study used the Stroke Prediction Dataset from Kaggle, which is a secondary, publicly available dataset compiled from synthetic and aggregated sources rather than real-world hospital or population-based records. While this dataset provides a useful 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.</p> <p>Another important limitation lies in the application of the Synthetic Minority Oversampling Technique (SMOTE). While SMOTE helps mitigate class imbalance and improves minority-class detection, it may introduce synthetic</p> |

|    |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
|----|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|    |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       | <p>patterns that do not fully reflect real clinical variation, potentially leading to bias or overfitting. Therefore, the observed performance gains should be viewed with caution until validated on authentic clinical data. To improve transparency, the main variables used in this study and their clinical meanings are presented in Table I. This mapping allows readers, particularly from the healthcare domain, to understand the relevance of each attribute to stroke risk prediction.</p> <p>Added table I to the method which explains the variable dataset and its explanation.</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
| 3. | <p>Methods &amp; Results</p> <ul style="list-style-type: none"> <li>• SMOTE improved minority class detection, but external validation is lacking.</li> <li>• Risk of overfitting is high since both ML and DL models show near-perfect metrics after oversampling on a single dataset.</li> <li>• Authors should tone down claims of "optimal" or "clinically applicable" without multi-source validation.</li> </ul>                                                                | <p>Added explanation about:</p> <p>After applying the SMOTE oversampling technique, the performance of both models improved substantially. The Logistic Regression model, which previously exhibited a large performance gap between classes, achieved an accuracy of 89% with more balanced precision (0.91) and recall (0.87) across both classes. This improvement suggests that a more even data distribution enables the model to recognize patterns from the minority class more effectively.</p> <p>Added explanation to the discussion about:</p> <p>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 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.</p> |
| 4. | <p>Discussion &amp; Conclusion</p> <ul style="list-style-type: none"> <li>• Strengthen discussion on dataset limitations: Kaggle data may not reflect actual stroke prevalence, population diversity, or real clinical workflows.</li> <li>• Reframe contributions as methodological (SMOTE + XAI pipeline) rather than as ready-to-use clinical tools.</li> <li>• Suggest future work with prospective or hospital-based data and external validation across populations.</li> </ul> | <p>Revisi discussion</p> <p>This study has significant limitations, even though the results are promising. The use of the Kaggle Stroke Prediction Dataset means that the results may not accurately represent stroke prevalence, 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. Also, the oversampling process with SMOTE, while it does help</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |

|    |                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
|----|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|    |                                                                                                                                                                                                                                                                                                                                                              | find minority classes better, could also create fake patterns that aren't really there in real patients. Consequently, the robust performance metrics observed should be regarded as an indication of methodological potential rather than direct clinical applicability.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
| 5. | References                                                                                                                                                                                                                                                                                                                                                   | Journal references from JUITA have been added and inappropriate references have been removed.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
| 6. | Presentation                                                                                                                                                                                                                                                                                                                                                 | Corrections have been made to the journal manuscript                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
| 7. | Language                                                                                                                                                                                                                                                                                                                                                     | Corrections have been made to the journal manuscript                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
| 8. | Abstract <ul style="list-style-type: none"> <li>• Add essential dataset details (total number of samples and stroke/non-stroke class distribution) to convey data scale.</li> <li>• Consider a brief mention of the key clinical implication—e.g., improved detection of rare stroke cases.</li> </ul>                                                       | <p>Kaggle’s Stroke Prediction Dataset (n = 5,110; 249 stroke cases, 4,861 non-stroke cases).</p> <p>These findings highlight the methodological potential of combining SMOTE with machine learning, deep learning, and XAI, but should not be interpreted as direct clinical validation.</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
| 9. | Introduction & Originality <ul style="list-style-type: none"> <li>• Recent literature (2023–2024) is cited, but the <i>state-of-the-art</i> (SOTA) contribution could be stated more explicitly.</li> <li>• Reference check: Citation [9] (EEG neurofeedback in alcohol dependence) appears unrelated; either justify its inclusion or remove it.</li> </ul> | <p>Remove Citation [9] and add an explanation about</p> <p>Another significant problem with using AI in healthcare is that it's challenging to comprehend how the models operate. Many high-performing models function as "black boxes," rendering them less reliable and complicating their application in clinical environments. In this research, we emphasize post-hoc explainability for tabular data by utilizing SHAP (global feature attribution) and LIME (local, case-level explanation) to clarify model behavior for domain experts. These choices align with current methodologies for structured clinical variables and avoid exaggerating the efficacy of image-based interpretability techniques, which fall beyond the scope of this study.</p> <p>add an Originality / SOTA paragraph at the end of the Introduction</p> <p>Recent studies utilizing ML or DL for stroke prediction predominantly (i) assess a singular family of models, (ii) present metrics on imbalanced data without explicitly quantifying the impact of data balancing, or (iii) offer limited interpretability for tabular predictors. We present a methodological pipeline that: (1)</p> |

|     |                                                                                                                                                                                                                                                                                                                                                  |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
|-----|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|     |                                                                                                                                                                                                                                                                                                                                                  | <p>evaluates both classical machine learning (Logistic Regression, Random Forest) and deep learning (Sequential, TabNet) on the same dataset; (2) isolates and quantifies the effect of SMOTE on minority-class detection using accuracy, precision, recall, F1, and PR-oriented metrics; (3) combines SHAP and LIME to provide complementary global and local explanations for structured clinical covariates; and (4) outlines dataset limitations and bias risks (Kaggle, non-clinical, oversampling artifacts) in advance to ensure transparent and reproducible reporting. We want to emphasize that these results demonstrate the feasibility of the methods, not that they are ready for use in a clinical setting. This indicates that further testing is necessary using future hospital data from multiple sources.</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| 10. | <ul style="list-style-type: none"> <li>Dataset description: Specify the total sample size, demographic composition, and exact class distribution (stroke vs. non-stroke) before and after SMOTE.</li> <li>SMOTE application: Quantify the number of synthetic minority samples generated and present the final balanced dataset size.</li> </ul> | <p>The dataset used in this study had a total initial sample of 5,110 data points, consisting of 249 data points labeled as stroke and 4,861 data points labeled as non-stroke. After pre-processing, 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 indicates that there is a much greater dominance of non-stroke data compared to stroke data, which has the potential to cause bias in the machine learning model that is built.</p> <p>The Synthetic Minority Oversampling Technique (SMOTE) method was used to fix this problem. This method increased the amount of data in the minority class (stroke) so that it was the same as the majority class. The results of using SMOTE show that the class distribution is balanced, with 4,700 stroke data and 4,700 non-stroke data. So, after resampling, the total sample size is 9,400 data points.</p> <p>Another significant limitation pertains to the implementation of the Synthetic Minority Oversampling Technique (SMOTE). SMOTE can help with class imbalance and make it easier to find minority classes, but it can also create synthetic patterns that don't fully show 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 shows the main variables used in this study and what</p> |

|     |                                                                                                                                                                                                               |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
|-----|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|     |                                                                                                                                                                                                               | <p>they mean in a clinical setting to make things more clear. This mapping helps readers, especially those in healthcare, see how each attribute is related to predicting stroke risk.</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
| 11. | <p>Pre-processing transparency: Report the number of records removed as duplicates, how missing values were imputed (mean, median, or other), and which features were dropped after correlation analysis.</p> | <p>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.</p> <p>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 <code>gender_Other</code> (-0.003010), <code>work_type_Never_worked</code> (-0.014149), and <code>work_type_children</code> (-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.</p> |
| 12. | <p>Experimental settings: Provide details on random seed, train-test split, and any cross-validation to allow replication.</p>                                                                                | <p>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 <code>train_test_split</code> 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% (<code>test_size=0.04</code>). At this point, a random seed of 12 (<code>random_state=12</code>) is used to ensure that the data remains the same every time the code is run.</p> <p>After dividing the data, normalization was performed using <code>StandardScaler</code> so that each feature had the same scale. The training data was then divided by Keras into training data and validation data through the <code>validation_split=0.4</code> parameter. This means that 40% of the training data was used as the validation set,</p>                                                                                                                                                                                                                                                                         |



|     |                                                                                                                                                                                                                  |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
|-----|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|     |                                                                                                                                                                                                                  | 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.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| 13. | Figure 1: Ensure the caption fully explains each pipeline stage so the figure can stand alone.                                                                                                                   | More detailed explanations have been provided.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
| 14. | Metrics emphasis: While accuracy is reported, the discussion should highlight recall and F1-score as more clinically meaningful for rare-event detection.                                                        | Revisions have been made to the journal manuscripts for both machine learning and deep learning.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
| 15. | Model comparison: Go beyond reporting numbers; discuss <i>why</i> TabNet outperformed other models (e.g., its ability to handle tabular data with attention mechanisms).                                         | It has been explained and depicted in the form of Model Comparison (Random Forest vs. TabNet)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
| 16. | XAI interpretation: SHAP and LIME results are described but could be linked more explicitly to medical knowledge, reinforcing that feature importance (age, smoking status, occupation) is clinically plausible. | 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. In the same way, the occupational variable could show stress from lifestyle choices or levels of physical activity, both of which are known to affect 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. |
| 17. | Limitations of interpretability: Briefly note that SHAP can be sensitive to feature correlations and that LIME explanations may vary with parameter settings                                                     | However, even though SHAP and LIME make models much more transparent, they aren't always easy to understand. SHAP values can be affected by how closely related features are. This means that if two or more clinical variables (like age and blood pressure) are very closely related, their individual contributions may not be distributed evenly, which could lead to misinterpretation of importance. LIME explanations also depend on a number of parameter settings, like 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.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
| 18. | <p>Discussion</p> <ul style="list-style-type: none"> <li>• Compares findings to recent studies effectively and acknowledges limitations (single dataset, no external validation).</li> <li>• Recommendation: Move the limitation statement to the beginning of the Discussion to foreground the issue of generalizability.</li> <li>• Consider adding a short paragraph on practical integration challenges (e.g., privacy, regulatory approval) for clinical deployment</li> </ul> | Added explanation for discussion                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
| 19. | <p>Conclusion</p> <p>Well written, contained in a single paragraph, and clearly summarizes contributions and future directions. Ensure no repetition of detailed results already covered in the Discussion.</p>                                                                                                                                                                                                                                                                     | <p>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.</p> |

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

Lianny Wydiastuty Kusuma<sup>1</sup>, Andri Wijaya<sup>2</sup>, Asahiro Nathanael Star Sitohang<sup>3</sup>, Yo Ceng Giap<sup>4</sup>

<sup>1,2,3,4</sup> *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 the World Stroke Organization [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. 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. Consequently, the results should be regarded as indicative of methodological feasibility rather than direct clinical applicability. Further validation on real-world clinical or population-based datasets is necessary to confirm their practical utility in healthcare settings.

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. Garg et al. [3] 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 Prasetyawati et al. [4] 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. Dai et al. [5] employed a Convolutional Neural Network (CNN) architecture for classifying CT images of stroke patients, achieving higher accuracy compared

to traditional methods. Chen et al. [6] 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 Synthetic Minority Over-sampling Technique (SMOTE) [7], [8] 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 [9].**

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 a difficult time 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 the predictions or diagnoses it did in the medical field. The Explainable AI (XAI) method addresses this issue by utilizing SHAP and LIME to provide explanations for the predictions [10]. **Another significant problem with using AI in healthcare is that it's challenging to comprehend how the models operate. Many high-performing models function as "black boxes," rendering them less reliable and complicating their application in clinical environments. In this research, we emphasize post-hoc explainability for tabular data by utilizing SHAP (global feature attribution) and LIME (local, case-level explanation) to clarify model behavior for domain experts. These choices align with current methodologies for structured clinical variables and avoid exaggerating the efficacy of image-based interpretability techniques, which fall beyond the scope of this study.**

Based on existing literature reviews, three main research gaps can be identified in the application of AI technology for stroke detection. First, the performance of classical machine learning is still limited when dealing with unbalanced stroke datasets, resulting in low model sensitivity to minority classes. Second, deep learning excels at analyzing medical images, but it still requires large, well-balanced datasets to function effectively. It involves data balancing strategies, such as SMOTE, to work optimally. Third, the application of XAI in stroke detection systems remains limited. The technology needs to be easy to understand so that people can trust it and use it more in medical settings. The technology needs to be easy to understand so that people can trust it and use it more in medical settings. These gaps demonstrate the pressing necessity for research that integrates classical machine learning and deep learning, employs data balancing techniques, and incorporates explainable artificial intelligence (XAI). Therefore, this study aims to provide a comprehensive solution to all three issues simultaneously. This effort is expected to promote more effective and trustworthy AI applications in the medical field, particularly in cases involving strokes.

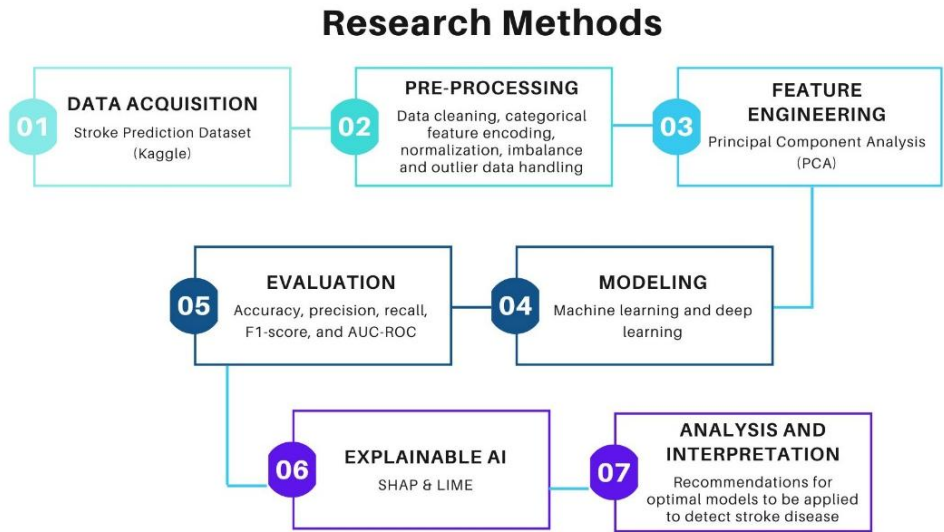
This study aims to evaluate the performance of classical machine learning and deep learning algorithms in classifying stroke diseases in unbalanced datasets by applying SMOTE to improve sensitivity to minority classes and integrating XAI methods (SHAP, LIME, and Grad-CAM) to improve prediction transparency. This study offers three primary contributions. First, comparing the performance of classical machine learning and deep learning algorithms on imbalanced stroke data will help us understand the pros and cons of each method. Second, SMOTE is a suitable method for balancing the data, making it less biased toward the majority class and facilitating the identification of stroke cases. Third, incorporating XAI into the stroke detection system enables medical staff to obtain predictions that are easily understandable and interpretable. This ensures that the model's results are both accurate and precise. With this approach, the resulting stroke detection system will exhibit high performance, be easily adoptable, and make a meaningful contribution to clinical practice. Ultimately, this research is expected to serve as a foundation for the development of more adaptive, transparent, and beneficial AI technology in the healthcare field.

**Recent studies utilizing machine learning or deep learning for stroke prediction predominantly (i) assess a singular family of models, (ii) present metrics on imbalanced data without explicitly quantifying the impact of data balancing, or (iii) offer limited interpretability for tabular predictors. We present a methodological pipeline that: (1) evaluates both classical machine learning (Logistic Regression, Random Forest) and deep learning (Sequential, TabNet) on the same dataset; (2) isolates and quantifies the effect of SMOTE on minority-class detection using accuracy, precision, recall, F1, and PR-oriented metrics; (3) combines SHAP and LIME to provide complementary global and local explanations for structured clinical covariates; and (4) outlines dataset limitations and bias risks (Kaggle, non-clinical, oversampling artifacts) in advance to ensure transparent and reproducible reporting. We want to emphasize that these**

results demonstrate the feasibility of the methods, not that they are ready for use in a clinical setting. This indicates that further testing is necessary using future hospital data from multiple sources.

## II. METHOD

This study utilizes secondary data obtained from the Stroke Prediction Dataset available on the Kaggle platform, which contains various demographic and clinical attributes relevant to stroke risk prediction. The research method is designed to develop a stroke detection system using a combination of machine learning, deep learning, feature engineering, data balancing techniques, and Explainable Artificial Intelligence (XAI) methods. The method flow consists of seven main stages, as shown in Fig. 1, namely: data acquisition, pre-processing, feature engineering, modeling, evaluation, XAI, and analysis and interpretation of results.



**Fig. 1 Research methods**

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. 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, as well as 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 [4], [11].

The dataset used in this study comprised a total of 5,110 initial data points, consisting of 249 data points labeled as stroke and 4,861 data points labeled as non-stroke. After pre-processing, 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 that there is a significant dominance of non-stroke data compared to stroke data, which has the potential to introduce bias into the machine learning model being built.

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.

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                                |

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 [12], and feature selection using a correlation matrix [13]. 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 [7], [14], [15].

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.

The next step is feature engineering using Principal Component Analysis (PCA). PCA is used to reduce the dimension of data, thereby reducing the risk of overfitting and speeding up the model training process [16]. By projecting data onto principal components, the information contained is preserved even though the number of variables is reduced. This method is beneficial for data with highly correlated variables. The application of PCA also helps improve model generalization [16], [17].

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) [3], [4]. The deep learning model used is a Convolutional Neural Network (CNN) adapted for tabular data [5], [6]. 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 [18].

The model was evaluated by measuring accuracy, precision, recall, F1-score, and AUC-ROC [19]. 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 [3], [5].

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) [10], [20]. SHAP is used to measure the contribution of each feature to the model's prediction results globally [21]. LIME is used to explain individual predictions locally, making it easier to interpret specific cases [21], [22]. The use of XAI is essential to increase medical professionals' trust in the results of prediction systems [10], [20].

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 [11], [14].

### 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.

TABLE II  
MACHINE LEARNING MODEL RESULTS

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

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.

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.

TABLE III  
DEEP LEARNING MODEL RESULTS

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

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.

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.

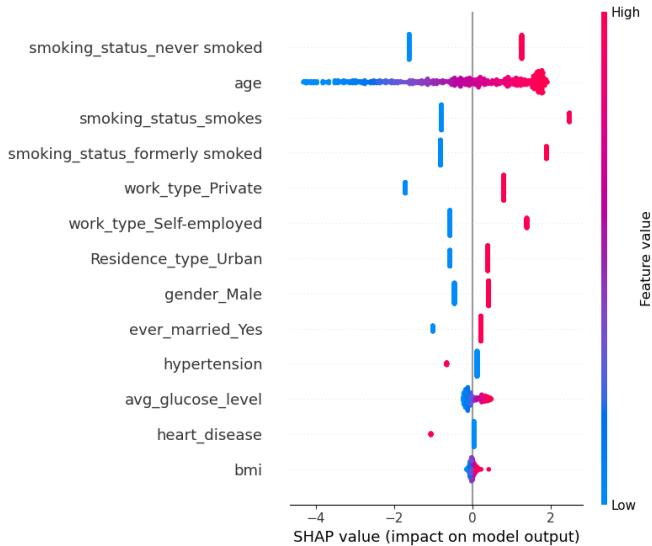


Fig. 2 Summary plot of SHAP results

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.

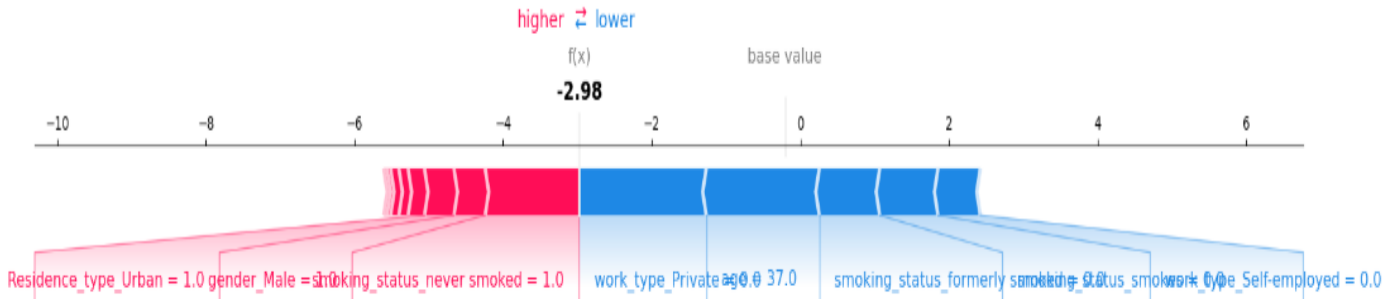


Fig. 3 Force plot of SHAP results

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.

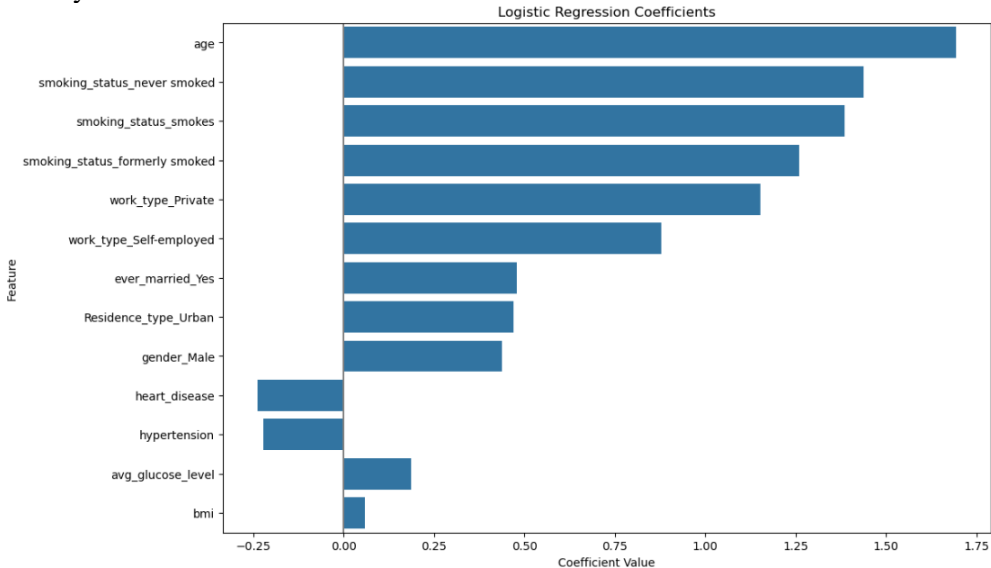


Fig. 4 Bar plot logistic regression coefficients

#### 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.

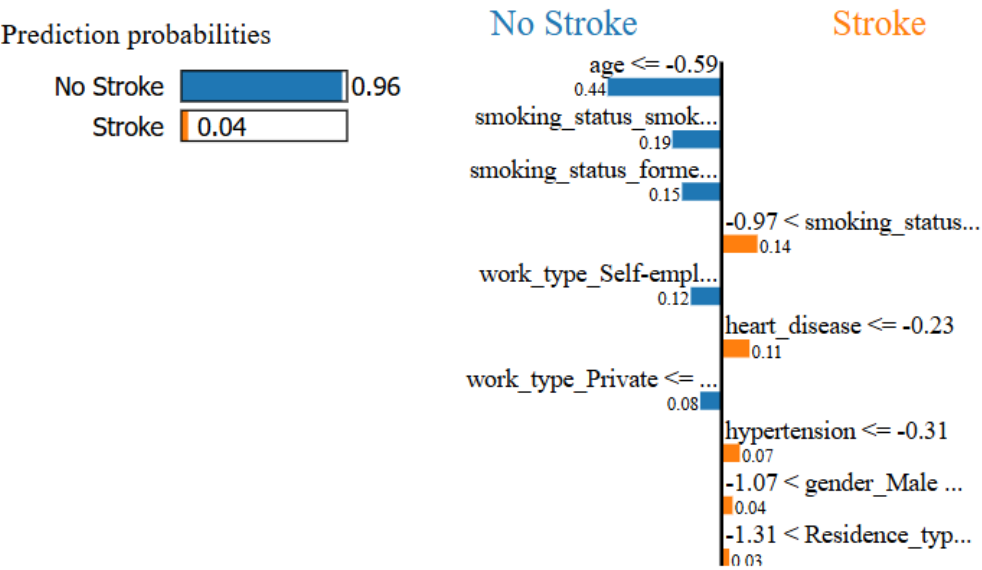


Fig. 5 Lime output

Based on the LIME visualization in Fig. 5, several factors negatively contribute to the predicted probability of stroke, thus strengthening the "No Stroke" classification. One dominant factor is a normalized age value  $\leq -0.59$ , indicating a relatively young age in the context of this dataset. Furthermore, certain smoking statuses—for example, never smoked or formerly smoked—also play a role in lowering stroke risk. Certain occupations also significantly reduce the risk of stroke, possibly due to lower levels of physical activity or exposure to environmental hazards. This combination of factors leads to a substantial decrease in the likelihood of stroke, thus shifting the model's prediction toward the "No Stroke" class. The clear visual and narrative relationships in Fig. 5 help readers understand how these factors work together.

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.

## E. Discussion

The results of this study show that the application of the Synthetic Minority Oversampling Technique (SMOTE) significantly improves the model's performance, both in machine learning approaches (Logistic Regression and Random Forest models) and deep learning (Sequential and TabNet Classifier models), in detecting stroke cases. This finding confirms the initial hypothesis that data imbalance is a factor that decreases model performance, and the oversampling technique can improve it [18], [23]. This finding is consistent with the theory outlined in the literature review, which suggests that balancing the class distribution can improve recall of the minority class without compromising overall accuracy. XAI analysis using SHAP provides global insights into the role of features such as age, smoking status, and occupation, which is in line with the stroke epidemiology literature [24]. Meanwhile, LIME provides local interpretations that show the variation in factor contributions to each prediction [22]. This combination of SMOTE and XAI supports the theoretical view that accuracy and interpretability need to go hand in hand to ensure the reliability of decision support systems in the healthcare sector [10].

Compared with previous studies, these results align with those of Chawla et al. and Fernández et al., who found that SMOTE improved minority class detection in the healthcare domain. However, the improvement was higher in this study, particularly in the post-SMOTE Random Forest model, possibly due to the combination of clinically

relevant features and optimal data pre-processing [25]. In contrast to the findings of Zhang et al., who reported that deep learning without data balancing was sufficiently accurate, this study demonstrates that data balancing is still necessary to enhance the detection of rare stroke cases. This difference can be attributed to variations in dataset characteristics, sample size, and the complexity of the features employed. Furthermore, the integration of SHAP and LIME provides additional value not always utilized in previous studies, providing a more comprehensive understanding of the prediction mechanisms [21], [22]. These methodological factors are the primary differences that may account for the variations in results between this study and previous studies.

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.

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 [24]. 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 [10]. 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 [25]. 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," 2023.
- [2] Centers for Disease Control and P. (CDC), "Stroke Facts," 2022.
- [3] N. Garg, H. Patel, and A. Sharma, "Predictive Modeling for Stroke Diagnosis Using Machine Learning Algorithms," *Sci Rep*, vol. 14, no. 1, pp. 1–10, 2024, doi: 10.1038/s41598-024-61665-4.

- [4] R. D. Prasetyawati, A. A. Nugroho, and R. Kurniawan, "Comparison of Machine Learning and Deep Learning Techniques for Stroke Prediction," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 2, pp. 45–53, 2024.
- [5] C. Dai, W. Zhou, Y. Liu, and others, "Deep learning for automatic detection of acute ischemic lesions on diffusion-weighted imaging," *Methods Inf Med*, vol. 62, no. 3, pp. 103–110, 2023, doi: 10.1016/j.zefq.2023.101739.
- [6] C. Chen, W. Zhao, J. Li, and others, "Stroke severity prediction using deep learning-based analysis of brain MRI," *Front Neurosci*, vol. 16, p. 945722, 2022, doi: 10.3389/fnins.2022.945722.
- [7] B. J. Erickson, P. Korfiatis, Z. Akkus, and T. L. Kline, "Artificial Intelligence and Machine Learning for Medical Imaging," *Radiographics*, vol. 41, no. 2, pp. 581–599, 2021, doi: 10.1148/rg.2021200169.
- [8] R. Sebastian and C. Juliane, "Comparison of Data Mining Classification Algorithms for Stroke Disease Prediction Using the SMOTE Upsampling Method," 2023. doi: <https://doi.org/10.30595/juita.v11i2.17348>.
- [9] W. P. Indahwati and F. M. Afendi, "Improving Stroke Detection with Hybrid Sampling and Cascade Generalization," May 2024. doi: <https://doi.org/10.30595/juita.v12i1.19386>.
- [10] G. Vilone and L. Longo, "Explainable Artificial Intelligence: A Systematic Review of Machine Learning Interpretability Methods," *Entropy*, vol. 23, no. 1, p. 18, 2021, doi: 10.3390/e23010018.
- [11] R. K. Yadav, S. Natarajan, and S. Radhakrishnan, "Machine Learning and Deep Learning Algorithms in Stroke Medicine," *Cureus*, vol. 16, no. 2, p. e38292, 2024, doi: 10.7759/cureus.38292.
- [12] E. Rahm and H. H. Do, "Data Cleaning: Problems and Current Approaches," 2000. [Online]. Available: <https://www.researchgate.net/publication/220282831>
- [13] P. Schober 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.
- [14] A. Hosny, C. Parmar, J. Quackenbush, L. H. Schwartz, and H. J. Aerts, "The Evolution of Artificial Intelligence in Medical Imaging," *Nat Biomed Eng*, vol. 7, pp. 995–1010, 2023, doi: 10.1038/s41551-023-01031-6.
- [15] M. Sulistiyono, Y. Pristyanto, S. Adi, and G. Gumelar, "SISTEMASI: Jurnal Sistem Informasi Implementasi Algoritma Synthetic Minority Over-Sampling Technique untuk Menangani Ketidakseimbangan Kelas pada Dataset Klasifikasi," 2021. [Online]. Available: <http://sistemasi.ftik.unisi.ac.id>
- [16] I. T. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," Apr. 13, 2016, *Royal Society of London*. doi: 10.1098/rsta.2015.0202.
- [17] R. Huang, L. Wang, J. Xu, and others, "Comparative Analysis of Machine Learning and Deep Learning Models in Clinical Stroke Prediction," *Sci Rep*, vol. 14, no. 1, p. 231, 2024, doi: 10.1038/s41598-024-82931-5.
- [18] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," 2002.
- [19] 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, Mar. 2015, doi: 10.1371/journal.pone.0118432.
- [20] E. Tjoa and C. Guan, "Transparency of Deep Neural Networks for Medical Image Analysis: A Review of Interpretability Methods," 2021.
- [21] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," Nov. 2017, [Online]. Available: <http://arxiv.org/abs/1705.07874>
- [22] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?' Explaining the predictions of any classifier," in *Proceedings of the ACM SIGKDD International Conference on Knowledge*

*Discovery and Data Mining*, Association for Computing Machinery, Aug. 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.

- [23] A. Fernández, S. García, M. Galar, R. Prati, B. Krawczyk, and F. Herrera, *Learning from Imbalanced Data Sets*. 2018. doi: 10.1007/978-3-319-98074-4.
- [24] V. L. Feigin *et al.*, “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.
- [25] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, “Machine learning applications in cancer prognosis and prediction,” 2015, *Elsevier B.V.* doi: 10.1016/j.csbj.2014.11.005.

[JUITA] Editor Decision



**Pengirim**

Mustafidah <jurnalump01@gmail.com>

**Penerima**

Lianny Wydiastuty Kusuma <lianny.wydiastuty@ubd.ac.id>, Ceng Giap Yo <cenggiap@ubd.ac.id>, Andri Wijaya <andri.wijaya@ubd.ac.id>

**Tanggal**

2025-10-29 08:00

Lianny Wydiastuty Kusuma, Ceng Giap Yo, Andri Wijaya:

We have reached a decision regarding your submission to JUITA: Jurnal Informatika, "Development and Evaluation of Stroke Disease Classification Models: Classical Machine Learning, Deep Learning, and Explainable AI Approaches".

Our decision is: Accept Submission.

Congratulations!

Conceptually, your manuscript has been accepted for publication in JUITA, Vol. 13 (3), November 2025.

Next, we will provide an invoice for the APC payment; afterward, we will edit and format the manuscript. There is a possibility that we will ask the author to make editorial corrections that are not substantive. In the last stage, we asked the author to do proofreading.

This acceptance notice also serves as the **Letter of Acceptance (LoA)**. Should you require a separate LoA file, please contact the technical team using the contact information provided on the journal's website.

Best Regards,

JUITA Editor

JUITA <https://jurnalnasional.ump.ac.id/index.php/JUITA>



[JUITA] Editor Decision



**Pengirim**

Zaeni <jurnalump01@gmail.com>

**Penerima**

Lianny Wydiastuty Kusuma <lianny.wydiastuty@ubd.ac.id>, Ceng Giap Yo <cenggiap@ubd.ac.id>, Andri Wijaya <andri.wijaya@ubd.ac.id>

**Tanggal**

2025-11-03 17:04

Lianny Wydiastuty Kusuma, Ceng Giap Yo, Andri Wijaya:

The editing of your submission, "Development and Evaluation of Stroke Disease Classification Models: Classical Machine Learning, Deep Learning, and Explainable AI Approaches," is complete. We are now sending it to production.

Submission URL: <http://jurnalnasional.ump.ac.id/index.php/JUITA/authorDashboard/submission/27868>