Analyzing risk factors and handling imbalanced data for predicting stroke risk using machine learning

Adiwijaya [a,1,*], Nur Ghaniaviyanto Ramadhan [a,2]

[a] School of Computing, Telkom University, Bandung, Jawa Barat, Indonesia,
[1] adiwijaya@telkomuniversity.ac.id; [2] nuruer@telkomuniversity.ac.id
* corresponding author

ARTICLE INFO

Article history
Received July 7, 2024
Revised January 30, 2025
Accepted February 6, 2025
Available online February 28, 2025

Keywords
General checkup data
Machine learning
Stroke prediction
ADASYN
Random forest

ABSTRACT

Stroke is a serious medical condition resulting from disturbances in blood flow to the brain, signaling a chronic health issue that requires an immediate response. Principal risk factors increasing the likelihood of stroke include the presence of pre-existing conditions such as Diabetes Mellitus (DM), hypertension, and high cholesterol levels. Effective preventive measures are crucial to minimize stroke risk, and using predictive methods based on data analysis from the clinical examination dataset over the last three years (2019-2021), known as the general checkup (GCU) dataset, presents an innovative approach. This study aims to predict an individual's stroke risk for the following year. In this context, the study also addresses the preprocessing stage of the GCU dataset, which includes solutions for missing values by substituting them with the statistical mean, label encoding, feature correlation analysis using entropy values, and addressing data imbalance with the Adaptive Synthetic (ADASYN) technique. To evaluate their predictive performance, the research involves comparisons among various machine learning models. The outcome of the experiment shows that the Random Forest model is the best model, with 98.7% accuracy and 63.9% F1-Score. This research highlights the importance of preemptive measures against stroke by utilizing predictive techniques on clinical data, with the Random Forest model proving most effective in forecasting stroke probability.

1. Introduction

Stroke prevalence in Indonesia has been increasing, as reported by the Basic Health Research of the Republic of Indonesia, which is conducted every five years [1], [2]. Between 2013 and 2018, stroke cases rose from 7% to 10.9% of the total population. Major risk factors, including Diabetes Mellitus (DM), hypertension, and high cholesterol, further contribute to this growing health concern. The 3.9% increase underscores the need for early detection and preventive strategies to mitigate stroke risks.

Machine learning (ML) models have been explored for stroke prediction based on clinical examination data. Prior research has evaluated various approaches: Emon, *et al.* [3] examined machine learning systems to choose the optimum model of stroke prediction. Sailasya, *et al.* [4] implemented DT and RF to predict the stroke of Kaggle dataset. Meanwhile, Kaur *et al.* [5] implemented some of deep learning-based models to estimate stroke using electroencephalogram (EEG) data.

Alshammari *et al.* [6] developed a national stroke data management system using Deepnet and Decision Tree (DT) but faced challenges in generalizability due to limited dataset representation. Dritsas *et al.* [7] used Naïve Bayes for long-term stroke risk prediction, but reliance on Kaggle data reduced its applicability to broader populations. Kokkotis *et al.* [8] tackled data imbalance with random under-sampling and Multi-Layer Perceptron (MLP). Santos *et al.* [9] applied Artificial Immune Systems and DT, managing imbalance through One-Sided Selection.

Thanka *et al.* [10] improved stroke prediction using an Artificial Neural Network (ANN) combined with SMOTE and Random Undersampling but risked overfitting due to limited data. Dahiya, *et al.* [11] and Abdullahi *et al.* [12] applied boosting-based models using SMOTE-ENN to handle imbalance. Meanwhile, Wu *et al.* [13] predicted stroke in the elderly in China but did not compare it with other techniques like ADASYN. Rahman *et al.* [14] analyzed feature correlations for early stroke prediction but lacked discussion on handling missing data, noise, and input errors.

Despite these advancements, challenges remain. Many studies rely on non-representative datasets, such as those from Kaggle, which is limiting their generalizability. Additionally, while performance metrics (AUC, accuracy, precision, recall) are commonly used for evaluation, fewer studies assess the clinical relevance of predictive features [7]. Complex deep learning models risk overfitting when trained on limited data [10]. Moreover, existing research often lacks discussions on handling missing values, noise, and input errors in medical datasets [14].

This study aims to address these gaps by developing a novel modified Random Forest (RF) model for annual stroke risk prediction based on comprehensive general checkup (GCU) data collected between 2019 and 2021. The key contributions of this study include:

- Real-world GCU Clinical Data Analysis: Unlike previous studies that rely on non-representative datasets, we use real-world GCU data, providing a more accurate and clinically relevant prediction.
- Enhanced Problem-Solving Methods: We systematically handle missing values, evaluate feature correlations, and apply robust oversampling techniques to address data imbalance, ensuring reliable model performance.
- Modified RF: Instead of standard RF, we introduce modifications to enhance predictive accuracy and reduce overfitting, making it more effective for annual stroke risk prediction. Overcoming this gap, our research enhances the robustness and applicability of stroke prediction, offering a more practical approach to early risk assessment and prevention strategies.

2. Method

2.1. Related Works

Stroke is a chronic disease caused by multiple risk factors. Therefore, several studies have focused on analyzing these risk factors [7], [15]. Previous research has identified several risk factors associated with stroke, including diabetes [7], high blood pressure, and unhealthy lifestyle patterns [15]. Analyzing stroke risk factors is crucial because it allows for the identification of people at high risk of stroke, thereby enabling early prevention and intervention measures. By understanding factors such as smoking and unhealthy lifestyle habits, healthcare professionals are expected to develop more effective strategies to reduce the incidence of stroke. This knowledge can also assist in educating the public about the importance of maintaining heart and vascular health through positive lifestyle changes.

Emon *et al.* [3] carried out a comprehensive analytical approach among ten basic machine-learning models for stroke prediction using a dataset from Bangladesh. This study replaced missing values with the mean/median, followed by data normalization and feature correlation analysis. The findings highlighted a significant correlation between age and stroke risk, with weighted voting yielding the best predictions. Sailasya *et al.* [4] implemented six machine-learning models for stroke prediction using a public dataset from Kaggle. This research included filling in null values with the mean, converting string data to integers through label encoding, and applying undersampling techniques to address imbalanced data. The Naïve Bayes model achieved the highest prediction accuracy of 82%. However, the imbalance technique used is undersampling, which removes majority class data. This poses a risk because, in medical data, every data point is highly valuable.

Santos, *et al.* [9] employed the DT to forecast for stroke using an imbalanced data from the study by Liu *et al.* [16]. This research eliminated features with missing values and applied the One-Sided Selection (OSS) undersampling technique, resulting in a 70% accuracy. However, the limitation of this study is that it only uses a DT developed with Genetic Programming (GP) without comparing it to other models such as RF, XGBoost, or Logistic Regression (LR). Additionally, it removes data with missing values but does not discuss the imputation approaches that could be used. Abdullahi *et al.* [12] analyzed the use of boosting-based models in stroke prediction with a dataset from Kaggle featuring twelve attributes. The preprocessing handling included null value imputation using K-Nearest Neighbors, converting categorical features to numerical, discretization, outlier removal, and applying SMOTE-ENN for imbalanced data. However, the study focuses only on achieving high accuracy without discussing the relevance of relationships between clinical features in the dataset. Additionally, the dataset used is not a primary but a public dataset from Kaggle.

Rahman *et al.* [14] evaluated the efficiency of several machine learning models in predicting stroke using a dataset sourced from Kaggle. Their analysis revealed significant correlations between stroke occurrences and factors such as age, heart disease, and hypertension. Additionally, they implemented the Random Oversampling technique to address imbalanced data issues, with the Light Gradient Boosting Machine model achieving the highest accuracy. Govindarajan *et al.* [17] classified stroke predictions with a dataset from Sugam Multispecialty Hospital, India, and conducted a series of preprocessing steps. However, in the study, the dataset used consists of only 507 patients, which is not representative enough and does not include handling for imbalanced data. Fang *et al.* [18] applied a feature selection strategy for stroke prognosis, using a dataset from the International Stroke Trial (IST) website, utilizing Pearson Correlation Analysis and Recursive Feature Elimination with Cross-Validation. Tazin *et al.* [19] performed stroke detection and prediction using Logistic Regression, RF, and Voting techniques, addressing data preprocessing that includes filling null values with the mean, applying label encoding, and employing the SMOTE for imbalanced data. Li *et al.* [15] enhanced the stroke risk screening level with data from the China National Stroke in 2017, addressing imbalanced data with SMOTE. However, the study does not analyze the correlation between clinical features in the dataset. Fernandez *et al.* [20] researched clinical factors in stroke patients to predict mortality using data from a European Tertiary Hospital, with preprocessing that involves feature correlation analysis and managing imbalanced data using Random Undersampling (RUS).

These studies collectively highlight the significance of addressing imbalanced data and the critical role of feature selection, and the effectivity of different machine learning models in medical applications. The chart for this exercise is illustrated in Fig. 1.

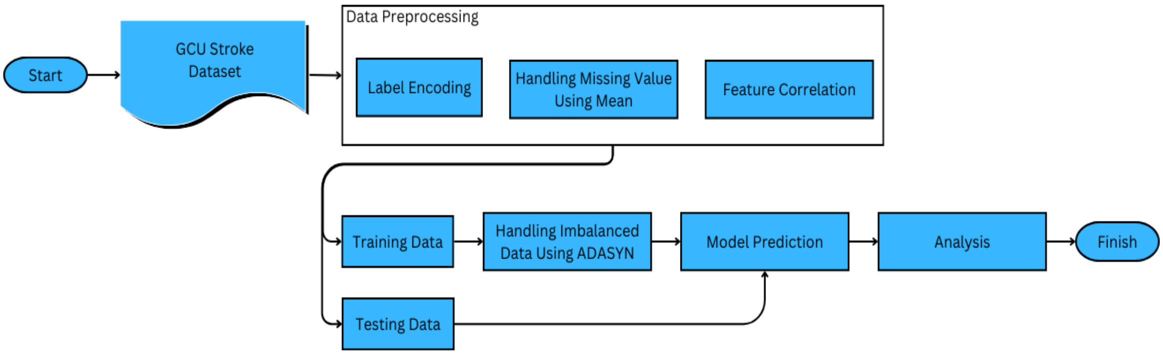

Fig. 1. Proposed Flowchart

Fig. 1 highlights key steps such as data preprocessing (handling missing values, feature correlation analysis, and balancing data with ADASYN), followed by training and evaluation using various machine learning models, and finally, performance comparison using a confusion matrix. Fig. 4 presents the effect of ADASYN implementation on model performance, showing improvements in all confusion matrix for the minority class (stroke cases). Table 2 gives a detailed evaluation of the predicted outcomes. In Section

3 of the paper, we discuss the importance of feature selection and ranking using entropy values (Fig. 3), explaining how key features such as sports history, blood sugar levels, and BMI contribute significantly to stroke prediction. We have elaborated on the effectiveness of RF in making predictions based on feature importance analysis.

2.2. Dataset

This research utilizes the General Checkup (GCU) dataset, which was comprehensively collected in a clinical setting over three consecutive years, from 2019 to 2021 [21]. This dataset consists of 28 of these variables with 1333 rows of data. Detailed characteristics of these variables can be seen at Table 1.

Table 1. Summary of the Dataset

Attributes	Null Value	Max.	Min.	Avg.	Med.	StdDev	Var.
Age	0	52	26	48.26	49	3.49	1.22
bloodpressure_sistole	14	190	90	124.44	120	14.95	2.23
bloodpressure_diastole	14	116	60	81.65	80	9.5	9.03
bloodpressure_Mean Arterial	14	136.66	73.33	95.91	93.33	10.7	1.14
BMI	14	59.72	15.35	27.54	27.34	4.41	1.94
stomach_circumference	14	143	60	90	90	9.95	9.91
bloodsugar_fast_glucose	14	390	57	113	100	44.85	2.01
bloodsugar_fast_glucose_2hour_pp	16	538	51	127.81	108	61.15	3.74
cholesterol	14	493	69	210.24	209	39.90	1.59
cholesterol_LDL	14	312	13	133.7	133	33.16	1.09
cholesterol_HDL	14	112	1.06	45.15	44	9.85	9.71
trigliserida	14	1174	28	140.26	118	96.97	9.40
cholesterol_CVD_riskratio	14	11.7	0.22	3.06	3.025	0.934	8.74
score stress	1203	43	1	13.13	11	10.25	1.05
kidney_creatinine	14	18.4	0.41	1.11	1.1	0.68	4.73
kidney_EGFR	16	247.6	3.8	82.08	79	18.23	3.32
kidney_ureum	14	156	8	24.37	23	10.64	1.13
bloodsugar_HbA1c	1331	231	221.25	224.5	224.5	9.19	8.45
hematology_leko	14	14700	3100	7266.3	7100	1750.74	3.06
hematology_mcv	14	99	27.8	84.74	85.5	5.80	3.36
hematology_mch	14	37.2	17.3	28.66	29	2.24	5.02
hematology_mchc	14	38.7	13.2	33.77	33.7	1.37	1.87

The summary of categorical data in this study is as follows: The gender variable has no missing data, with 245 females and 1,088 males. The history_smoke variable has 14 missing values, with 1,007 individuals who do not smoke, 100 light smokers (1-5 cigarettes per day), 105 moderate smokers (6-10 cigarettes per day), 80 heavy smokers (11-15 cigarettes per day), and 27 individuals who smoke more than 15 cigarettes per day. For the history_sport variable, there are 92 missing values, with 684 individuals who do not exercise, 281 individuals who exercise less than 3 times a week, and 276 individuals who exercise more than 3 times a week. Meanwhile, the category_stress variable has 1,203 missing values, with 37 individuals classified as having low stress, 56 with medium stress, and 37 with high stress. The urine_protein variable has 18 missing values, with 1,199 individuals showing no protein in their urine, 80 with a small amount (+), 28 with a moderate amount (++), 6 with a high amount (+++), and 3 with a very high amount (++++). The urine_glucose variable has 19 missing values, with 1,222 individuals showing no glucose reduction in their urine, 42 with a small reduction (+), 28 with a moderate reduction (++), 19 with a high reduction (+++), and 5 with a very high reduction (++++). Finally, the class variable has no missing data, with eight individuals classified as positive and 1,325 as negative.

In the General Checkup (GCU) dataset under study, the age of individuals undergoing examination ranges from 26 to 52 years, with a dominant age group between 46 and 52 years. Furthermore, the majority of the examination subjects are male. This is consistent with findings that most positive stroke cases are also identified in males, particularly in the age group of 45 to 50 years. Significantly, some individuals with a positive stroke have shown blood sugar_fast_glucose values reaching 303, indicating the condition of Diabetes Mellitus as a pre-existing disease.

The GCU dataset also includes variables related to physical activity, including sports history and smoking history. The sports history variable is categorized based on the frequency of sports activities per week, while the smoking history variable is categorized based on the number of cigarettes consumed daily. Analyzing these variables is important, considering that a healthy lifestyle can reduce the risk of stroke.

Moreover, the dataset includes variables that describe an individual's stress condition, namely stress score and category, filled out by the hospital medical team. However, it was found that many stress-related data are not filled in, possibly because respondents are reluctant to undergo testing to assess their stress levels.

2.3. Label Encoding

In this phase, values are encoded on the General Checkup (GCU) dataset, which contains string-formatted data [22]. This process entails converting string-type data into an integer or numeric format [4], resulting in a dataset with a numeric representation. There are five columns in this research dataset that are converted through the encoding process, namely: gender (gender), smoking history (history_smoke), sports history (history_sport), protein in urine (urine_protein), and protein glucose (protein_glucose).

2.4. Handling Missing Value

Resolving missing values in a dataset is an important step in analyzing data [23]. This research applies a statistical technique, mean imputation, to replace null values in the GCU dataset. Zero-value distribution within the dataset is displayed in Tables I and II. In this research, several features underwent column deletion (drop column) due to a significant number of null values, including blood sugar level HbA1c (bloodsugar_hbA1c), stress category (category_stress), and stress score (score_stress). Handling missing values is essential, particularly in disease prediction research [24].

2.5. Feature Correlation

Identifying the correlation among features within the dataset is an important step in disease prediction [25]. Feature correlation analysis can enhance prediction accuracy and disease classification [26]. This research initiates the identification of feature correlations by using a heatmap [27], [28]. Subsequently, this study applies entropy values in the RF to determine the importance of ranking features in the dataset. The equation for the entropy value is presented in the equation (1) [29].

$$\Sigma^{Ci} P_i log_2 P_i \tag{1}$$

Here C represents the count of both positive and negative classes, and P denotes the sample proportion for class i (stroke and non-stroke).

2.6. Handling Imbalanced Classes

The stroke data in this study is categorized as highly imbalanced, with a total of 1333 data entries, of which only 8 entries are classified as positive stroke cases and 1325 entries as negative. To assess the level of data imbalance, the Imbalance Ratio (IR) is used, calculated using the following equation [29].

$$IR = Instance\ Minority/Instance\ Majority \tag{2}$$

where the minority instance represents the number of entries in the minority class, while the majority instance is the number of entries in the majority class. Data is considered imbalanced when the IR value approaches 0 and less imbalanced as the IR value approaches 1. For the GCU stroke dataset analyzed, the IR value was found to be 0.00603, indicating a significant level of imbalance.

Given the low IR value, handling imbalanced data in this research becomes crucial. The Adaptive Synthetic (ADASYN) oversampling technique is chosen for its ability to avoid overfitting and its suitability for high-dimensional data [30]–[33]. ADASYN addresses the weaknesses of Random Oversampling, which often leads to overfitting by merely duplicating samples from the minority class [34]–[36]. Additionally, ADASYN (Adaptive Synthetic Sampling) generates synthetic data that is more focused on regions with a more complex minority class distribution [37], [38], unlike SMOTE, which only creates new samples using a simple interpolation approach [39]. ADASYN operates by calculating Euclidean distance, determining neighborhood, calculating the number of synthetic samples created, and performing interpolation [40], as explained in equations (3) through (7).

$$G = (m_j - m_i)\beta \tag{3}$$

where m_j total number of majority data and m_i total number of minority data. Thus, G total number of minorities data that will be produced. The factor β represents the ratio between the desired minority data compared to the majority data. $A\,\beta$ value of 1 indicates a highly balanced outcome after the application of ADASYN. The subsequent process involves finding the nearest K value for each minority sample and calculating the r_i value (4). This step signifies that each minority sample will be assigned with a divergent environment.

$$r_1 = instance\ majority\ /\ k \tag{4}$$

The value of r_i indicates the level of majority class dominance within a specific environment. Environments with higher ri values contain a larger proportion of majority class samples, making them more challenging to learn from. Therefore, the r_i values must be normalized so that the total sum of r_i values equals 1 (5).

$$r_i^{\wedge} = r_i / \sum r_i \tag{5}$$

Calculating the quantity of synthesized examples to be made for each environment is achieved through step (6). A higher value of r_i in environments dominated by the majority class results in generating more synthetic minority class examples for that environment. Consequently, this approach lends an adaptive characteristic to the ADASYN algorithm, ensuring that the augmentation of synthetic samples is focused on areas where the minority class is underrepresented.

$$G_i = G \times r_i \tag{6}$$

To generate G_i data in each environment, the first step is to select a minority example, named xi, from that environment. Subsequently, another minority example, named XZ_i, is randomly selected from the same environment. Thus, a fresh synthesized instance could be estimated by applying the function (7), allowing for the creation of synthetic data based on combining these two minority examples.

$$S_i = x_i + (XZ_i - x_i)\lambda \tag{7}$$

In the equation, λ is a unique raw numeric range of values between 0 and 1. Si is the newly created synthetic example, while x_i and XZ_i represent two minority examples originating from the same environment. The process of creating this synthetic example is conducted through a linear combination of x_i and XZ_i, where λ serves as the weight in that combination, facilitating variation in generating synthetic data.

2.7. RF Algorithm

The RF is a prediction algorithm within the ensemble learning category [41], based on the principle of combining several predictions to form a stronger unity. The RF model, an improvement over the DT, incorporates various DT in its construction. The prediction process with the RF model begins with a random bootstrap from the training data, followed by the construction of several decision trees. As a final step, the RF model applies a voting mechanism to the predictions of different determination trees to determine the final prediction output. In the Diabetes Mellitus (DM) context of disease prediction research, the RF model has proven superior performance compared to seven other machine learning models [22]. Furthermore, RF has also been applied in various other studies, including disease prediction [29] and consumer behavior [42].

3. Results and Discussion

This research evaluated the performance of several machine learning models by comparing prediction results after addressing data imbalance. The models' performance was measured using metrics from the confusion matrix [43].

3.1. Feature Correlation Analysis

Understanding the heatmap correlation feature process is crucial before making predictions because it provides information about the relationships among features within a dataset. This identification of relationships among features aids in analyzing the correlations between them, enabling an understanding of whether there are linear relationships among individual features in the dataset. The heatmap correlation assists in several ways, such as understanding the patterns of relationships which offer insights into how features interact and influence each other, and it can be used to select the most relevant features to be incorporated in the predictive model, potentially enhancing the model's performance. The feature correlation is visualized using a heatmap, as shown in Fig. 2.

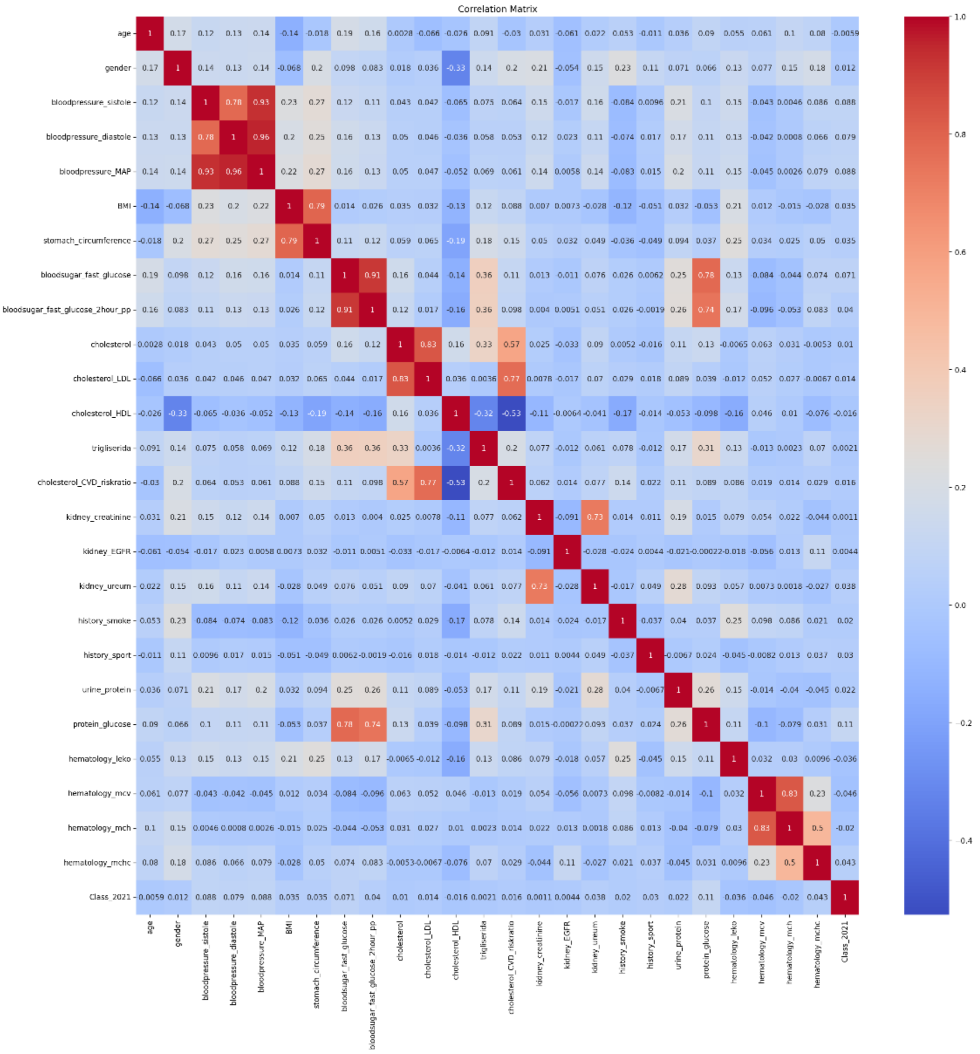

Fig. 2. Heatmap Correlation

This heatmap highlights the five feature pairs with the strongest correlations: (1) "MAP blood pressure" and "diastolic blood pressure," showing the score for 0.96; (2) "fasting blood sugar 2-hour pp" and "fasting blood sugar," exhibiting the score for 0.91; (3) "LDL cholesterol" and "cholesterol," along with "hematology mch" and "hematology mcv," both demonstrating the score for 0.83; (4) "fasting blood sugar" and "urine protein," with a score measured at 0.78; and (5) "CVD risk ratio cholesterol" and "cholesterol," reflecting a correlation of 0.77.

Additionally, the outcomes of the entropy-based feature importance analysis are presented in Fig. 3. Prior to making predictions, evaluating feature importance through entropy is essential, as it offers critical insights into the relevance of each feature in the model. This analytical step helps in understanding how any features contribute to the predictivity of the model, enabling a more targeted focus on the most influential features. It also aids in potentially refining the model by selecting key features to improve performance or simplify its structure.

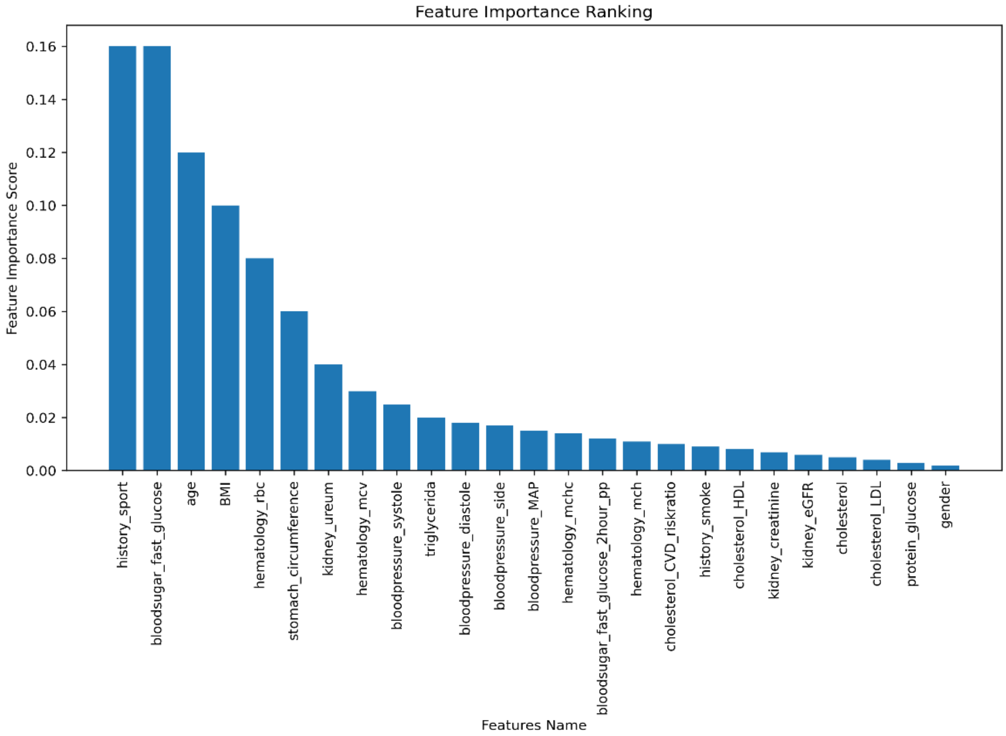

Fig. 3. Feature Rank Entropy

Features that significantly influence predictions will exhibit high entropy before data splitting and low entropy afterward. Therefore, features that can significantly reduce uncertainty become critically important to identify. Features that significantly impact predictions affect the model's precision, stability, and generalization. Thus, understanding the importance of features provides valuable insights for optimizing models used for prediction tasks.

Based on Fig. 3, the top five features that significantly influence the stroke class are (1) sports history, (2) fasting blood sugar, (3) age, (4) Body Mass Index (BMI), and (5) hematology leko. On the other hand, the five features with the lowest ranking are (1) urine protein, (2) gender, (3) protein glucose, (4) LDL cholesterol, and (5) cholesterol. Using the entropy feature importance method along with RF resulted in an accuracy of 98.7%, but precision was only 49.3%, recall 50%, and f1-score 49.6%. The RF method produced the most superior metrics values compared to seven other machine learning models when using feature importance. This occurred due to the ability of the RF method to identify and utilize the most relevant features in making predictions. Here are some additional explanations based on the feature ranking in Fig. 3.

- Features with the Greatest Influence. These features have a significant impact on stroke prediction:

  - Exercise history (history_sport): The most influential feature indicates that exercise habits are strongly related to stroke risk. Regular physical activity can reduce the risk of stroke by improving cardiovascular health.
  - Fasting blood sugar (bloodsugar_fast_glucose): High fasting blood sugar levels indicate a risk of diabetes, which is a major factor in cardiovascular disease and stroke.
  - Age (age): Stroke risk increases with age due to reduced elasticity of blood vessels and a higher likelihood of conditions like hypertension and diabetes.
  - Body Mass Index (BMI): A high BMI is linked to obesity, hypertension, and diabetes— which are both main potential risk factors for stroke.
  - Hematology leko (hematology_leko): Possibly related to white blood cell count, which can be an indicator of chronic inflammation or other health conditions associated with stroke.
- Features with Moderate Influence
  - Waist circumference (stomach_circumference): An indicator of central obesity, which is linked to a higher risk of metabolic syndrome and stroke.
  - Kidney urea (kidney_ureum): Reflects kidney function, as kidney problems are often associated with hypertension and cardiovascular disease.
  - Hematology MCV (hematology_mcv): The average size of red blood cells, which may be linked to anemia or other blood disorders affecting oxygen supply to the brain.
  - Diastolic & systolic blood pressure (bloodpressure_diastole, bloodpressure_sistole): Hypertension is one of the main risk factors for stroke.
  - Triglycerides (trigliserida): High triglyceride levels are associated with atherosclerosis, which can lead to ischemic stroke.
- Features with Low Influence. These features have a smaller impact on stroke prediction, possibly due to variability in the dataset or weaker correlations.
  - Total cholesterol, LDL, and HDL levels: Although high cholesterol can be a stroke risk factor, it may not show a strong correlation in this dataset.
  - Kidney creatinine (kidney_creatinine) & EGFR: Indicators of kidney function, but they may not directly impact stroke risk in this model.
  - Gender (gender): While men and women may have different stroke risks, its impact is smaller compared to other medical factors.
  - Urine protein (urine_protein): This could indicate kidney problems but is not a key factor in this dataset.

3.2. ADASYN Result Analysis

In this section, an analysis of the application of ADASYN is carried out, as well as an analysis of the use of a combination of important features with ADASYN, as shown in Fig. 4.

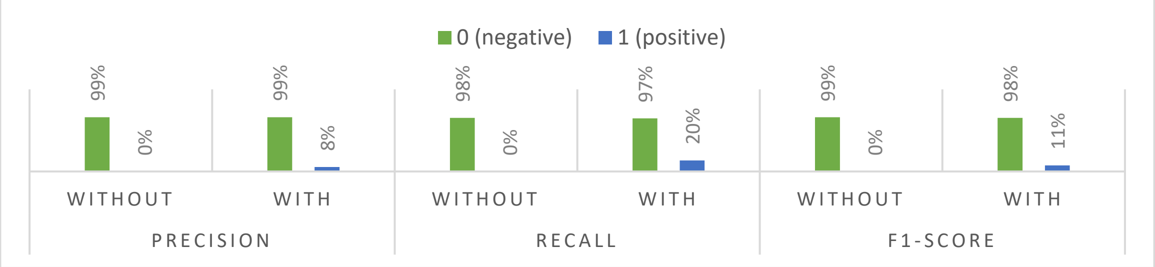

Fig. 4. Performance Prediction Comparison of Each Dataset Class With and Without ADASYN Implementation

Based on Fig. 4, the label "with" in the image above represents the ADASYN and RF (our proposed model) the precision metric in the positive category improved from 0% to 8%. This result indicates that the utilization of ADASYN has enhanced the model's quality in accurately recognizing the positive category. Concurrently, the recall metric associated with the positive category also showed an escalation from 0% to 20%. These findings suggest that the performance of ADASYN has supported the model in detecting more instances that fall into the absolute positive category. Additionally, the F1-score related to the positive category demonstrated an increase from 0% to 11%. This outcome indicates that the adoption of ADASYN has improved the balance between precision and recall for the positive category. Overall, these results demonstrate that the use of ADASYN has enhanced the model's performance in identifying the minority class (positive) through improvements in all confusion matrix for the positive class. This means that ADASYN is effective in overcoming the issue of class imbalance and enhancing the models' capacity to classify the minority class, by generating synthetic samples of the minority class. Thus, ADASYN facilitates learning patterns from classes that are underrepresented by the model, consequently refining the model's capability to detect elements from the minority class.

A significant contribution of this study is the integration of ADASYN with feature importance analysis using entropy values within the RF model. Feature importance analysis is critical in understanding which variables contribute most to stroke prediction. Our study identifies sports history, fasting blood sugar, age, and BMI as the most influential features, with entropy scores of 0.16, 0.16, 0.15, and 0.11, respectively.

The combination of ADASYN with feature importance offers two advantages:

- Enhanced Model Sensitivity: ADASYN increases the representation of stroke cases, enabling the model to learn patterns from minority class samples better. As a result, the importance of certain features (e.g., sports history and fasting blood sugar) becomes more prominent after balancing the dataset.
- Improved Feature Discrimination: The entropy-based ranking demonstrates that features identified as important in the imbalanced dataset remain relevant even after ADASYN is applied. This suggests that the synthetic data generated by ADASYN effectively preserves the underlying relationships in the dataset, allowing the model to make more informed predictions.

Feature selection based on entropy values enhances model performance by identifying the most relevant predictors and eliminating less informative features that may introduce noise. This study found that selecting the top-ranked features based on entropy improved precision and recall. For example, when using all features without selection, the RF achieved an F1 score of 49.6%. However, after selecting only the top-ranked features (sports history, fasting blood sugar, age, and BMI), the F1 score increased to 63.9%, representing a significant performance boost. This improvement demonstrates that entropy-based feature selection helps the model focus on the most impactful variables, reducing computational complexity and improving predictive accuracy.

The effectiveness of ADASYN in this study stems from the fact that the dataset contains only eight positive cases out of 1333 total entries, leading to an extreme imbalance (Imbalance Ratio = 0.00603). ADASYN assigns more synthetic samples to areas of the feature space where the minority class is underrepresented. Given the small number of stroke cases, ADASYN helps by placing new synthetic samples closer to actual positive cases, making the model more likely to learn the distinguishing characteristics of stroke patients. The generated synthetic samples aligned with features such as sports history, blood sugar levels, age, and BMI, which were ranked as the most significant factors in stroke prediction using entropy analysis. By expanding the decision space for these key features, ADASYN helps balance the learning process.

Although ADASYN improves minority class prediction, it is important to recognize its limitations, especially on more complex data sets: An Overfitting Risk: If the synthetic samples are too similar to existing examples of minority classes, the model might memorize them instead of generalizing to unseen data. This risk increases when the dataset size is small, and the synthetic points are heavily concentrated in a limited feature space [38]. Distortion of Data Distribution: Unlike techniques such as SMOTE, which creates synthetic samples along feature-space lines between minority class instances, ADASYN prioritizes regions with high-class imbalance. This may lead to an unevenly distributed decision boundary, potentially distorting the natural distribution of stroke occurrences [35]. Feature Importance Misalignment: ADASYN assumes that minority samples should be generated where a class imbalance is highest. However, if this assumption does not align with true feature importance, it may introduce synthetic noise rather than informative samples. For example, if synthetic instances are generated in a region of feature space where stroke occurrences are naturally unlikely, it could degrade model performance [34]. Scalability Challenges: ADASYN's effectiveness may decrease due to difficulties in defining appropriate feature neighborhoods in highly imbalanced datasets with a complex feature space. For instance, if some features have highly nonlinear relationships with stroke risk, ADASYN might struggle to generate synthetic samples that correctly capture these interactions [39].

Alternative or complementary approaches can be considered to mitigate these limitations: Hybrid Techniques: Combining ADASYN with techniques such as Tomek Links or ENN could help filter out synthetic noise while retaining meaningful minority class samples. Feature Engineering for Synthetic Data Validation: Before training, an additional validation step can be introduced to compare synthetic instances with real minority cases to assess the consistency of the generated samples. Alternative Imbalance Handling Methods: While ADASYN benefited this dataset, future work could explore generative models (e.g., Variational Autoencoders or GANs) that create more robust synthetic samples by preserving the dataset's natural feature distributions.

3.3. Comparison Machine Learning Models

In this section, a comparison is made between the proposed RF model and several other machine-learning models. Results of the Prediction are shown in Table 2.

Table 2. Result of the Prediction

Model	Acc (%)	Prec (%)	Rec (%)	F1-Score (%)
DT	96	53.3	58.4	54.5
RF (our proposed model)	98.7	74.4	63.9	57.7
AdaBoost	97.5	59.8	63.9	57.7
XGBoost	97.5	66.5	59.2	57.7
LR	88.2	64.4	51.7	57.7
NB	78	51.7	50.7	51.3
SVM	90.2	50.5	55.5	48.5
MLP	84.7	62.6	49.8	48.9

The experiments in Table 2 include applying strategies to cope with unbalanced data, including filling in null values, assessing feature importance ratings, and implementing various machine learning models to compare their performance against the proposed RF. These results show that the RF surpasses other models in all aspects of performance measurement.

Specifically, the DT model, as shown in Fig. 4, after implementing strategies to handle imbalanced data, demonstrated an increase in precision by 3.9%, recall by 9.2%, and f1-score by 5.2%. However, there was a decrease in accuracy by 1.5%, indicating that the application of ADASYN allowed the RF model to make more accurate predictions by reducing bias in the data. Although the DT model recorded higher accuracy compared to ensemble models such as AdaBoost and XGBoost, the lower f1 score for DT indicates the shortcomings of this model in achieving a balanced precision and recall compared to AdaBoost and XGBoost.

Using feature importance in RF enables the model to identify and leverage the most informative features in separating target classes, resulting in more accurate predictions. This is evidenced by a relatively high precision rate (74.4%), indicating that most of the correct expectation created by the modeling is relevant, and a fairly good recall rate (59.8%) in identifying most of the positive cases in the dataset. These results demonstrate that the importance of feature application in RF has positively impacted the quality of predictions. Meanwhile, other models exhibit varied performance. AdaBoost and XGBoost, although high in accuracy, display slightly lower precision, recall, and F1-Score values than RF. This indicates that feature importance in these models also has a positive impact, but not to the extent achieved by RF. On the other hand, LR, NB, SVM, and MLP perform less well compared to RF and other ensemble models. This is due to these models' lesser capability to handle complex or imbalanced features in the dataset. Meanwhile, RF and other ensemble models using feature importance have effectively addressed this issue.

The Logistic Regression (LR) and Naïve Bayes (NB) models recorded high recall but low precision, indicating a tendency for these models to mistakenly classify negative cases as positive. On the other hand, the Multi-Layer Perceptron (MLP) model showed unsatisfactory performance, with an f1-score that was one of the lowest among the tested models. This indicates that MLP is less effective in handling imbalanced data despite the application of data imbalance handling techniques using ADASYN. Table 3 show Tukey HSD (Honestly Significant Difference) testing.

Table 3. Tukey HSD (Honestly Significant Difference) Testing

Group1	Group2	Mean Difference	p-value	H0
AdaBoost	DT	-2.2	0.999999708	Rejected
AdaBoost	LR	-4	0.999981975	Rejected
AdaBoost	MLP	-5.9	0.999753214	Rejected
AdaBoost	NB	-10.075	0.992670987	Rejected
AdaBoost	RF	6.45	0.99955621	Rejected
AdaBoost	SVM	-3.75	0.999988403	Rejected
AdaBoost	XGBoost	0	1	Rejected
DT	LR	-1.8	0.999999928	Rejected
DT	MLP	-3.7	0.999989421	Rejected
DT	NB	-7.875	0.998392289	Rejected
DT	RF	8.65	0.997102627	Rejected
DT	SVM	-1.55	0.999999974	Rejected
DT	XGBoost	2.2	0.999999708	Rejected
LR	MLP	-1.9	0.999999895	Rejected
LR	NB	-6.075	0.999700633	Rejected
LR	RF	10.45	0.990902379	Rejected
LR	SVM	0.25	1	Rejected
LR	XGBoost	4	0.999981975	Rejected
MLP	NB	-4.175	0.999975869	Rejected
MLP	RF	12.35	0.976453942	Rejected
MLP	SVM	2.15	0.999999751	Rejected
MLP	XGBoost	5.9	0.999753214	Rejected
NB	RF	16.525	0.897664163	Rejected
NB	SVM	6.325	0.999609646	Rejected
NB	XGBoost	10.075	0.992670987	Rejected
RF	SVM	-10.2	0.992114121	Rejected
RF	XGBoost	-6.45	0.99955621	Rejected
SVM	XGBoost	3.75	0.999988403	Rejected

Based on Table 3, the Tukey HSD test helps identify which pairs of groups have statistically significant differences. In the context of disease prediction, the Tukey HSD test can compare the performance of various predictive models or diagnostic methods to determine the most effective [44]. According to the Tukey HSD test source, the differences between the tested models (e.g., AdaBoost, DT, RF, XGBoost, etc.) are not highly significant. However, RF often achieves high accuracy and more excellent stability compared to DT, AdaBoost, or SVM, especially when working with datasets that contain complex features. Additionally, RF is more effective than SVM or LR in handling large datasets with numerous features. Moreover, RF is less likely to overfit than DT, as it utilizes the bagging mechanism [45].

4. Conclusion

This study aimed to forecast the risk of stroke in individuals within the next year by utilizing the General Checkup (GCU) dataset. The findings reveal that the proposal model improved prediction accuracy by 98.7% and f1-score by 63.9%. Additionally, the research successfully identified key features in the dataset with strong correlations, such as MAP and diastolic blood pressure, while establishing the order of feature importance. Notably, sports history was found to influence the stroke variable significantly. Addressing dataset imbalance was critical in enhancing prediction performance, as an f1 score of any offered RF model improved by 14.2% after implementing ADASYN. The study further examines stroke risk factors, highlighting sports history, blood sugar (fast glucose), age, and BMI as the most influential. Sports history and blood sugar (fast glucose) recorded the highest entropy importance scores at 0.16, followed by age at 0.15 and BMI at 0.11. These insights can aid in formulating effective strategies for stroke prevention and treatment. Future Research, While this study successfully employed a modified RF model to predict stroke risk using clinical checkup data, several limitations and opportunities for future research can be identified: 1) Data Quality and Missing Values. Although this study addressed missing values using mean imputation, the approach may not be the most effective in all cases. Future research may also explore a more sophisticated method of imputation, such as multi-imputation or imputation based on machine learning, to boost the accuracy of addressing the missing data; 2.) Feature Importance and Data Imbalance. The study used ADASYN to handle imbalanced data and entropy to select features, which significantly improved model performance. However, other techniques for addressing data imbalance exist, such as SMOTE or SMOTE-Tomek. Exploring these methods' effectiveness, individually or in combination with ADASYN, could yield further improvements in stroke prediction accuracy; 3) Model Generalization. The dataset used in this study was limited to GCU clinical data from 2019-2021. Expanding the dataset to include data from diverse populations or other geographical locations could improve the model's generalizability. Additionally, incorporating data from other clinical risk factors or advanced diagnostic methods (e.g., genetic or imaging data) may enhance prediction; 4) Advanced Machine Learning Techniques. Although RF outperformed other models in this study, further exploration of deep learning techniques or hybrid models that combine the strengths of different algorithms (e.g., a combination of RF with neural networks or boosting methods) could be beneficial in improving predictive accuracy, especially for complex and non-linear relationships in the data; 5) Clinical Relevance of Predictive Features. Further research could delve deeper into the clinical relevance of the features identified in this study, such as sports history, blood sugar levels, and BMI. Understanding how these features influence stroke risk in a clinical setting could help refine stroke prevention strategies; 6) Evaluation Metrics and Decision Support. While confusion matrix were used in this study, additional evaluation metrics, such as AUC-ROC or precision-recall curves, could be considered to comprehensively evaluate model performance. Developing decision support systems based on these models could facilitate real-time clinical decision-making. By addressing these areas, future research could significantly contribute to the development of more accurate, robust, and clinically applicable stroke prediction

Declarations

Author contribution. Adiwijaya: conceptualization, formal analysis, methodology, writing – original draft, validation, and funding acquisition. Nur Ghaniaviyanto Ramadhan: data curation, investigation, programming, visualization, writing, editing, project administration

Funding statement. This research was supported by Telkom University, grant number: 030/LIT07/PPM-LIT/2024.

Conflict of interest. The authors declare no conflict of interest.

Additional information. No additional information is available for this paper.

References

[1] M. of Health, "Basic Health Research," May 2013. [Online]. Available at: https://repository.badankebijakan.kemkes.go.id/id/eprint/4467/1/Laporan_riskesdas_2013_final.pdf.

[2] Balitbangkes, "National Riskesdas Report 2018," *Lembaga Penerbit Balitbangkes*. p. hal 156, 2018, [Online]. Available at: https://repository.badankebijakan.kemkes.go.id/id/eprint/3514/1/Laporan Riskesdas 2018 Nasional.pdf.

[3] M. U. Emon, M. S. Keya, T. I. Meghla, M. M. Rahman, M. S. Al Mamun, and M. S. Kaiser, "Performance Analysis of Machine Learning Approaches in Stroke Prediction," in *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Nov. 2020, pp. 1464–1469, doi: 10.1109/ICECA49313.2020.9297525.

[4] G. Sailasya and G. L. A. Kumari, "Analyzing the Performance of Stroke Prediction using ML Classification Algorithms," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 6, pp. 539–545, 2021, doi: 10.14569/IJACSA.2021.0120662.

[5] M. Kaur, S. R. Sakhare, K. Wanjale, and F. Akter, "Early Stroke Prediction Methods for Prevention of Strokes," *Behav. Neurol.*, vol. 2022, no. 1, pp. 1–9, Apr. 2022, doi: 10.1155/2022/7725597.

[6] A. Alshammari, N. Atiyah, H. Alaboodi, and R. Alshammari, "Identification of stroke using deepnet machine learning algorithm," *Int. J. Med. Eng. Inform.*, vol. 15, no. 5, pp. 416–429, 2023, doi: 10.1504/IJMEI.2023.133083.

[7] E. Dritsas and M. Trigka, "Stroke Risk Prediction with Machine Learning Techniques," *Sensors*, vol. 22, no. 13, p. 4670, Jun. 2022, doi: 10.3390/s22134670.

[8] C. Kokkotis *et al.*, "An Explainable Machine Learning Pipeline for Stroke Prediction on Imbalanced Data," *Diagnostics*, vol. 12, no. 10, p. 2392, Oct. 2022, doi: 10.3390/diagnostics12102392.

[9] L. I. Santos *et al.*, "Decision tree and artificial immune systems for stroke prediction in imbalanced data," *Expert Syst. Appl.*, vol. 191, p. 116221, Apr. 2022, doi: 10.1016/j.eswa.2021.116221.

[10] M. R. Thanka, K. S. Ram, S. P. Gandu, E. B. Edwin, V. Ebenezer, and P. Joy, "Comparing Resampling Techniques in Stroke Prediction with Machine and Deep Learning," in *2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS)*, Jun. 2023, pp. 1415–1420, doi: 10.1109/ICSCSS57650.2023.10169237.

[11] M. Dahiya, N. Mishra, S. Agarwal, and Z. Parveen, "Predicting the occurrence of Ischemic stroke by Gradient Boost Approaches," in *2023 4th International Conference on Intelligent Engineering and Management (ICIEM)*, May 2023, pp. 1–4, doi: 10.1109/ICIEM59379.2023.10166287.

[12] S. D. Abdullahi and S. A. Muhammad, "Early Prediction of Cerebrovascular Disease using Boosting Machine Learning Algorithms to Assist Clinicians," *J. Appl. Sci. Environ. Manag.*, vol. 26, no. 6, pp. 1031–1037, Jun. 2022, doi: 10.4314/jasem.v26i6.6.

[13] Y. Wu and Y. Fang, "Stroke Prediction with Machine Learning Methods among Older Chinese," *Int. J. Environ. Res. Public Health*, vol. 17, no. 6, p. 1828, Mar. 2020, doi: 10.3390/ijerph17061828.

[14] S. Rahman, M. Hasan, and A. K. Sarkar, "Prediction of Brain Stroke using Machine Learning Algorithms and Deep Neural Network Techniques," *Eur. J. Electr. Eng. Comput. Sci.*, vol. 7, no. 1, pp. 23–30, Jan. 2023, doi: 10.24018/ejece.2023.7.1.483.

[15] X. Li, D. Bian, J. Yu, M. Li, and D. Zhao, "Using machine learning models to improve stroke risk level classification methods of China national stroke screening," *BMC Med. Inform. Decis. Mak.*, vol. 19, no. 1, p. 261, Dec. 2019, doi: 10.1186/s12911-019-0998-2.

[16] T. Liu, W. Fan, and C. Wu, "A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset," *Artif. Intell. Med.*, vol. 101, p. 101723, Nov. 2019, doi: 10.1016/j.artmed.2019.101723.

[17] P. Govindarajan, R. K. Soundarapandian, A. H. Gandomi, R. Patan, P. Jayaraman, and R. Manikandan, "Classification of stroke disease using machine learning algorithms," *Neural Comput. Appl.*, vol. 32, no. 3, pp. 817–828, Feb. 2020, doi: 10.1007/s00521-019-04041-y.

[18] G. Fang, W. Liu, and L. Wang, “A machine learning approach to select features important to stroke prognosis,” *Comput. Biol. Chem.*, vol. 88, p. 107316, Oct. 2020, doi: 10.1016/j.compbiolchem.2020.107316.

[19] T. Tazin, M. N. Alam, N. N. Dola, M. S. Bari, S. Bourouis, and M. Monirujjaman Khan, “Stroke Disease Detection and Prediction Using Robust Learning Approaches,” *J. Healthc. Eng.*, vol. 2021, no. 1, pp. 1–12, Nov. 2021, doi: 10.1155/2021/7633381.

[20] C. Fernandez-Lozano *et al.*, “Random forest-based prediction of stroke outcome,” *Sci. Rep.*, vol. 11, no. 1, p. 10071, May 2021, doi: 10.1038/s41598-021-89434-7.

[21] A. A. Gozali, “Hypertension Multi-Year Prediction and Risk Factors Analysis Using Decision Tree,” in *2023 10th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, Aug. 2023, pp. 76–82, doi: 10.1109/ICITACEE58587.2023.10277644.

[22] N. G. Ramadhan, Adiwijaya, W. Maharani, and A. A. Gozali, “Prediction of Diabetes Mellitus in the Upcoming Year using SMOTE and Random Forest,” in *2023 International Conference on Data Science and Its Applications (ICoDSA)*, Aug. 2023, pp. 316–321, doi: 10.1109/ICoDSA58501.2023.10277223.

[23] A. Uddin, X. Tao, C.-C. Chou, and D. Yu, “Are missing values important for earnings forecasts? A machine learning perspective,” *Quant. Financ.*, vol. 22, no. 6, pp. 1113–1132, Jun. 2022, doi: 10.1080/14697688.2021.1963825.

[24] A. Purwar and S. K. Singh, “Hybrid prediction model with missing value imputation for medical data,” *Expert Syst. Appl.*, vol. 42, no. 13, pp. 5621–5631, Aug. 2015, doi: 10.1016/j.eswa.2015.02.050.

[25] S. Dev, H. Wang, C. S. Nwosu, N. Jain, B. Veeravalli, and D. John, “A predictive analytics approach for stroke prediction using machine learning and neural networks,” *Healthc. Anal.*, vol. 2, p. 100032, Nov. 2022, doi: 10.1016/j.health.2022.100032.

[26] K. Patidar, R. K. Gour, A. Dixit, M. Verma, and A. K. Pal, “An Improved Method for the Data Cluster Based Feature Selection and Classification,” in *2023 International Conference for Advancement in Technology (ICONAT)*, Jan. 2023, pp. 1–6, doi: 10.1109/ICONAT57137.2023.10080669.

[27] S. Buyrukoğlu and A. AKBAŞ, “Machine Learning based Early Prediction of Type 2 Diabetes: A New Hybrid Feature Selection Approach using Correlation Matrix with Heatmap and SFS,” *Balk. J. Electr. Comput. Eng.*, vol. 10, no. 2, pp. 110–117, Apr. 2022, doi: 10.17694/bajece.973129.

[28] F. Viton, M. Elbattah, J.-L. Guerin, and G. Dequen, “Heatmaps for Visual Explainability of CNN-Based Predictions for Multivariate Time Series with Application to Healthcare,” in *2020 IEEE International Conference on Healthcare Informatics (ICHI)*, Nov. 2020, pp. 1–8, doi: 10.1109/ICHI48887.2020.9374393.

[29] N. G. Ramadhan, A. -, and A. Romadhony, “Preprocessing Handling to Enhance Detection of Type 2 Diabetes Mellitus based on Random Forest,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 7, pp. 223–228, Sep. 2021, doi: 10.14569/IJACSA.2021.0120726.

[30] M. Zakariah, S. A. AlQahtani, and M. S. Al-Rakhami, “Machine Learning-Based Adaptive Synthetic Sampling Technique for Intrusion Detection,” *Appl. Sci.*, vol. 13, no. 11, p. 6504, May 2023, doi: 10.3390/app13116504.

[31] M. Mujahid *et al.*, “Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering,” *J. Big Data*, vol. 11, no. 1, p. 87, Jun. 2024, doi: 10.1186/s40537-024-00943-4.

[32] S. A. Alex, J. Jesu Vedha Nayahi, and S. Kaddoura, “Deep convolutional neural networks with genetic algorithm-based synthetic minority over-sampling technique for improved imbalanced data classification,” *Appl. Soft Comput.*, vol. 156, p. 111491, May 2024, doi: 10.1016/j.asoc.2024.111491.

[33] R. M. Munshi, “Novel ensemble learning approach with SVM-imputed ADASYN features for enhanced cervical cancer prediction,” *PLoS One*, vol. 19, no. 1, p. e0296107, Jan. 2024, doi: 10.1371/journal.pone.0296107.

[34] P. Gnip, L. Vokorokos, and P. Drotár, “Selective oversampling approach for strongly imbalanced data,” *PeerJ Comput. Sci.*, vol. 7, p. e604, Jun. 2021, doi: 10.7717/peerj-cs.604.

[35] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, "Introduction to KDD and Data Science," in *Learning from Imbalanced Data Sets*, Cham: Springer International Publishing, 2018, pp. 1–17, doi: 10.1007/978-3-319-98074-4_1.

[36] S. Rana, R. Kanji, and S. Jain, "Comprehensive Analysis of Oversampling Techniques for Addressing Class Imbalance Employing Machine Learning Models," *Recent Adv. Comput. Sci. Commun.*, vol. 18, p. 95 , Dec. 2024, doi: 10.2174/0126662558347788241127051934.

[37] A. Balaram and S. Vasundra, "Prediction of software fault-prone classes using ensemble random forest with adaptive synthetic sampling algorithm," *Autom. Softw. Eng.*, vol. 29, no. 1, p. 6, May 2022, doi: 10.1007/s10515-021-00311-z.

[38] Haibo He, Yang Bai, E. A. Garcia, and Shutao Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, Jun. 2008, pp. 1322–1328, doi: 10.1109/IJCNN.2008.4633969.

[39] J.-B. Wang, C.-A. Zou, and G.-H. Fu, "AWSMOTE: An SVM-Based Adaptive Weighted SMOTE for Class-Imbalance Learning," *Sci. Program.*, vol. 2021, no. 1, pp. 1–18, May 2021, doi: 10.1155/2021/9947621.

[40] Q. Wang, W. Cao, J. Guo, J. Ren, Y. Cheng, and D. N. Davis, "DMP_MI: An Effective Diabetes Mellitus Classification Algorithm on Imbalanced Data With Missing Values," *IEEE Access*, vol. 7, pp. 102232–102238, 2019, doi: 10.1109/ACCESS.2019.2929866.

[41] L. Breiman, "Random Forests," *Mach. Learn. 2001 451*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[42] G. Alfian *et al.*, "Customer Shopping Behavior Analysis Using RFID and Machine Learning Models," *Information*, vol. 14, no. 10, p. 551, Oct. 2023, doi: 10.3390/info14100551.

[43] A. A. Gozali, "Multi-Years Diabetes Prediction Using Machine Learning and General Check-Up Dataset," in *2023 11th International Conference on Information and Communication Technology (ICoICT)*, Aug. 2023, vol. 2023-Augus, pp. 98–103, doi: 10.1109/ICoICT58202.2023.10262699.

[44] S.-C. Chang *et al.*, "The Comparison and Interpretation of Machine-Learning Models in Post-Stroke Functional Outcome Prediction," *Diagnostics*, vol. 11, no. 10, p. 1784, Sep. 2021, doi: 10.3390/diagnostics11101784.

[45] N. Komal Kumar, D. Vigneswari, M. Vamsi Krishna, and G. V. Phanindra Reddy, "An Optimized Random Forest Classifier for Diabetes Mellitus," in *Advances in Intelligent Systems and Computing*, vol. 813, Springer, Singapore, 2019, pp. 765–773, doi: 10.1007/978-981-13-1498-8_67.