




Feature Selection and Hybrid Sampling with Machine Learning Methods for Health Data Classification



Hairani Hairani^{1,2}, Triyanna Widiyaningtyas^{1*}, Didik Dwi Prasetya¹

¹ Department of Electrical Engineering and Informatics, Universitas Negeri Malang, Malang 65145, Indonesia

² Department of Computer Science, Universitas Bumigora, Mataram 83127, Indonesia

Corresponding Author Email: triyannaw.ft@um.ac.id

Copyright: ©2024 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/ria.380419>

ABSTRACT

Received: 26 February 2024

Revised: 28 April 2024

Accepted: 26 June 2024

Available online: 23 August 2024

Keywords:

hybrid sampling, stroke classification, feature selection correlation, health data imbalance

This study aims to improve the performance of classification algorithms in dealing with unbalanced and high-dimensional health in stroke prediction by integrating correlation feature selection and hybrid sampling techniques. Several previous studies that used machine learning methods to predict stroke still had less than optimal accuracy. This is because stroke data has several problems, including missing values, many attributes, and data imbalance can cause a decrease in the performance of the classification method. Therefore, this research uses an integrated approach to feature selection and hybrid sampling. The objective of the feature selection technique is to identify important attributes within stroke data. After that, the SMOTE-Enn hybrid sampling approach is utilized to address data imbalance. The research findings indicate that employing correlation-based feature selection along with SMOTE-Enn and the Random Forest algorithm leads to improved performance compared to no sampling with the SVM and XGBoost methods, with an increase in accuracy of 3%, recall of 91.3%, and AUC of 45.2%. Thus, the proposed method performed better than recent stroke classification studies.

1. INTRODUCTION

The increasing availability of health data related to disease in electronic medical records presents an interesting opportunity for pattern analysis using machine learning [1]. However, the existing collection of disease data in electronic medical records is often underutilized, resulting in a problem of being rich in data but poor in knowledge [2, 3]. Nevertheless, this dataset can be leveraged to extract valuable knowledge patterns, such as early detection of diseases based on symptoms [4-6] like stroke [7]. Stroke is one of the leading causes of disability and death globally [1].

With the increasing availability of comprehensive health datasets, there is a huge opportunity to leverage machine learning [8] to predict stroke, potentially saving lives through early detection. However, utilizing health data to accurately predict stroke has many challenges. In this study, machine learning is applied to predict stroke disease early by seeking methods that provide a very high level of accuracy to minimize errors in the prediction. By utilizing machine learning models, it is possible to predict early on whether a patient has the potential to suffer from stroke or not based on the inputted stroke symptoms [9, 10]. The problems with stroke data are that it has missing values, there are many of features, and the data is unbalanced. A large number of features can have a negative impact on the performance of classification methods, when the number of relevant features is less than irrelevant features [11]. Besides that, many features have an impact on long computing times [12]. Selection of influential features

can have a positive influence on the performance of the classification method and the computing time can be fast [13]. Apart from that, the problem with stroke data is that the data is unbalanced. In the stroke data, the number of stroke classes is 209 instances (minority class) less than the non-stroke class of 4699 instances (majority class). This can cause the results of the classification method to be biased and make it difficult to classify stroke class compared to non-stroke class [14]. In other words, the stroke class can be classified as a non-stroke class, because this class is less represented in the stroke dataset. Therefore, this research needs to solve the problem of many features and unbalanced data simultaneously, in order to improve the performance of the classification method more optimally. Previous researchers have already conducted stroke disease prediction using machine-learning approaches. Ahammad [15] employed various machine learning methods with feature selection and achieved an accuracy rate of 97%. Ray et al. [16] used the chi-square feature selection method to determine relevant attributes in stroke disease data. It was found that using chi-square feature selection significantly improved accuracy and reduced computation time in the model used. Gupta and Raheja [17] utilized the Random Forest method and adaptive synthetic sampling (adasyn) with an accuracy of 97.6%. Yin et al. [18] predicted stroke using multiple machine learning methods and data sampling, achieving an accuracy of 71.2% and a recall of 80.3%. Ashrafuzzaman et al. [19] employed the deep learning Convolutional Neural Network (CNN) method for stroke disease prediction with an accuracy of 95.5%. Research [20]

compared the Random Forest method with KNN, where KNN obtained a higher accuracy rate of 95.7% compared to Random Forest.

Wang et al. [21] solves the problem of imbalanced data in stroke data using SMOTE and then performs classification with Random Forest. The outcome shows that the SMOTE method combined with Random Forest attains an accuracy of 70.29% and a precision of 70.05%. Islam et al. [22] used the Random Forest algorithm and various sampling techniques to predict stroke, achieving precision, recall and F1 score of 96% respectively. Abd Mizwar et al. [23] predicts stroke using the Extreme Gradient Boosting method with an accuracy of 96%, specificity of 87.7%, and recall of 19.7%. Various machine learning techniques were evaluated for predicting stroke disease [24]. The results showed that Neural Networks outperformed Random Forest, KNN, Naive Bayes, and SVM, achieving an accuracy of 95% and a recall of 100%. Bandi et al. [25] compares several machine learning methods for stroke disease prediction. According to the findings of the study, the Random Forest technique outperforms other machine learning methods. Dev et al. [26] uses machine learning and PCA for stroke disease prediction, where the Neural Networks method with PCA performs better than Random Forest and Decision Tree with an accuracy of 75%, recall of 68%, and f1-measure of 73%. Hairani and Priyanto [27], Hairani et al. [28] used a combined sampling strategy utilizing Random Forest to address imbalanced data issues in diabetes disease. The research uses the hybrid sampling methods of SMOTE-Enn and SMOTE-Tomek Link. Research findings show that the use of the SMOTE-Enn technique combined with Random Forest provides better results compared to the use of the SMOTE-Tomek Link method.

Several previous studies, such as studies Gupta and Raheja [17], Yin et al. [18], Ashrafuzzaman et al. [19], Wang et al. [21], have only focused on solving the problem of imbalanced data using the oversampling approach, which has the drawback of generating a lot of noise data in the artificially created minority data. On the other hand, studies of Ahammad [15], Ray et al. [16], Dev et al. [26] have focused more on the use of feature selection to improve the performance of the classification methods used without addressing the problem of imbalanced data. Thus, several gaps or challenges can be addressed by considering the unresolved issues in previous research, namely: (1) its suboptimal performance that can be enhanced, (2) the unresolved issue of numerous attributes and imbalanced stroke data simultaneously, which can affect the performance of the methods. To improve the weaknesses of previous research, this research proposes two approaches simultaneously, namely feature selection and hybrid sampling in solving the problem of stroke data which has many attributes and imbalance data. Where, the proposed method is a novelty that has not been used by previous research referred to in predicting stroke. Hence, this research aims to enhance the performance of classification methods such as Random

Forest, SVM, and XGBoost in stroke disease classification by integrating correlation feature selection techniques and hybrid sampling. By identifying the most relevant features and balancing the dataset through hybrid sampling, the research goal to improve how classification algorithms deal with imbalanced and high-dimensional health data, with the potential to create more precise and effective models for healthcare applications like diagnosing stroke disease.

2. MATERIALS AND METHODS

Figure 1 demonstrates the research stages, starting with the acquisition of a stroke disease dataset from the UCI Machine Learning Repository. This dataset was selected due to its frequent use in prior studies as experimental material [15-18]. This dataset consists of 5110 data points with 10 input and 1 output feature, as shown in Figure 1. The features of the stroke dataset are Gender, Heart Disease, Age, Hypertension, Work Type, Ever Married, Residence Type, Smoking Status, BMI, Avg Glucose Level (mg/dL), and Stroke. Data samples can be seen in Table 1.

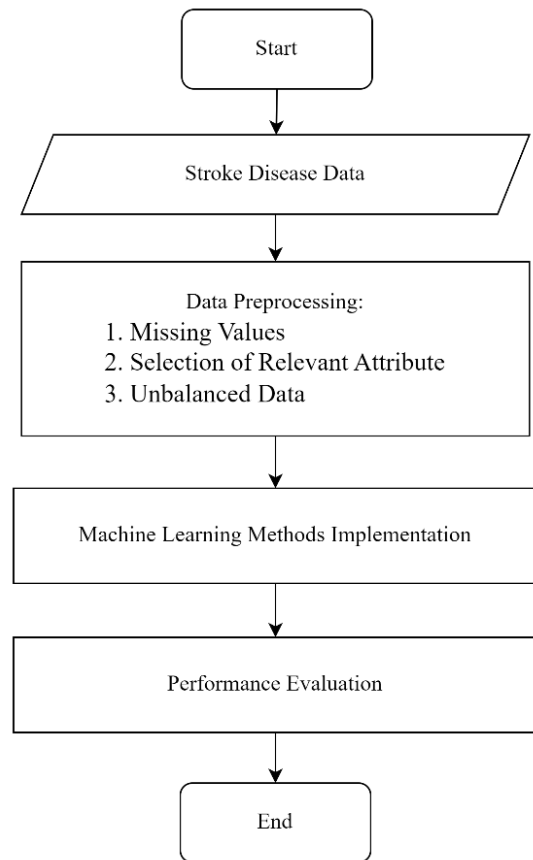


Figure 1. Research flow

Table 1. Example of stroke disease dataset

No	Age	Gender	Heart Disease	Residence Type	BMI	Hypertension	Ever Married	Work Type	Avg Glucose Level	Smoking	Stroke
1	67	Male	1	Urban	36	0	Yes	Private	228.36	Formerly Smoked	1
...
5110	44	Female	0	Urban	26.2	0	Yes	Govt_job	85.28	Unknown	0

Preprocessing data is a critical step in machine learning to achieve optimal results in classification methods by improving data quality. In this study, the processed data is stroke disease data with several issues, such as missing values, many attributes, and imbalanced data. Therefore, the data preprocessing stage focuses on handling missing values, many attributes, and imbalanced data. Data with missing values will be removed as the percentage of missing values is less than 5%, which does not significantly impact the method's performance [29]. This study uses a correlation-based feature selection method to identify features that impact stroke prediction. This method evaluates the relationship between independent features and the class within a dataset to determine the most relevant features. Eq. (1) is used as the formula to calculate the correlation.

$$r = \frac{n(\sum XY) - (\sum X)(\sum Y)}{\sqrt{(n\sum X^2 - (\sum X)^2)(n\sum Y^2 - (\sum Y)^2)}} \quad (1)$$

The correlation coefficient is represented by r , and n denotes the number of data points. Y stands for the dependent variable, whereas X represents the independent variable.

After performing feature selection using correlation, the next step is to address the imbalanced data in stroke disease data. The approach used to address this issue is hybrid sampling using SMOTE-Enn. The SMOTE-Enn method combines the SMOTE method with edited nearest neighbors (Enn) to reduce noise by randomly selecting instances and removing majority class instances close to the minority class. The SMOTE-Enn method works by balancing the data using the SMOTE method, which generates noise. Then, the noise data from SMOTE is removed using the Enn method. The Enn method removes noise instances, which are majority-class instances close to the minority class, to minimize classification errors and reduce the occurrence of noise. The working concept of SMOTE-Enn is shown in Figure 2.

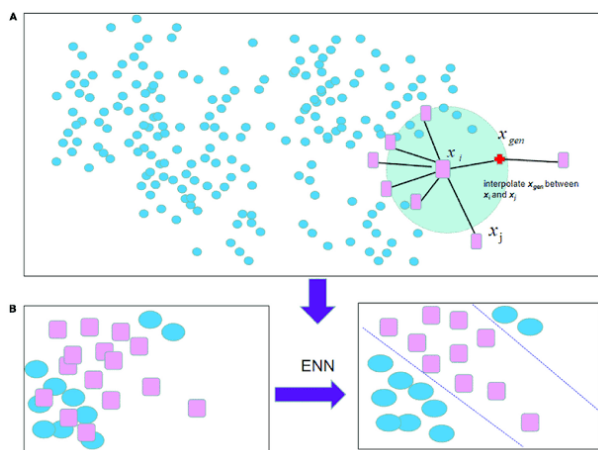


Figure 2. SMOTE-Enn working concept [30]

In stroke disease data, several attributes with nominal types such as gender, residence type, work type, category married convs, and smoking status were found, requiring transformation. Transformation is used to convert nominal attribute types into numerical representations. This is necessary to perform feature selection with correlation and address the issue of imbalanced data using hybrid sampling. The transformation results of the *gender*, *residence type*, *work type*, *category married convs*, and *smoking status* attributes are shown in Tables 2 to 6.

Table 2. Gender attribute transformation result

No	Gender Category	Conversion
1	Male	1
2	Female	2

Table 3. Transformation result of residence type attribute

No	Residence Type Category	Conversion
1	Urban	1
2	Rural	2

Table 4. Transformation result of work type attribute

No	Work_Type Category	Conversion
1	Never_worked	5
2	Private	4
3	Self-employed	3
4	Govt_job	2
5	Children	1

Table 5. Transformation result of category married convs attribute

No	Married_Convs Category	Conversion
1	Yes	1
2	No	2

Table 6. Transformation result of smoking status attribute

No	Smoking_Status Category	Conversion
1	Unknown	4
2	Formerly Smoked	3
3	Never Smoked	2
4	Smokes	1

This study employs several classification methods for predicting stroke disease: Random Forest, SVM, and XGBoost. Random Forest is an ensemble learning based on decision trees [29], which creates a group of decision trees from randomly selected subsets. Each decision tree provides a prediction, and then voting is performed for these predictions. The best or most accurate prediction is selected based on the majority voting and considered the final prediction.

SVM is a classification technique designed to identify the optimal hyperplane by maximizing the separation margin between classes [31] and works on linear and nonlinear data. SVM uses soft margins and feature space to transform nonlinear data into linear data.

One of the well-known variants of the gradient boosting algorithm is XGBoost, which emphasizes speed and accuracy [32]. XGBoost is an ensemble learning method that utilizes gradient boosting with decision trees to achieve maximum scalability [33]. In terms of base classifiers, XGBoost focuses solely on decision trees. Various error functions can be used to control the complexity level of these trees [34].

In the performance evaluation stage, machine learning methods such as Random Forest, SVM, and XGBoost are tested to assess their accuracy in predicting stroke after the data is divided into training and testing sets using the 10-fold cross-validation technique. The performance of these methods is evaluated using three metrics: accuracy, recall, and AUC, which are obtained from the confusion matrix table. Accuracy [35], recall, and AUC are calculated using Eqs. (2) to (4) [36-38].

$$accuracy = \frac{TP+TN}{TP+FN+TN+FP} \quad (2)$$

$$recall = \frac{TP}{TP+FN} \quad (3)$$

$$AUC = \frac{(recall * specificity)}{2} \quad (4)$$

3. RESULTS AND DISCUSSION

This section explains the research results achieved based on the research flow shown in Figure 1. This study used data on stroke disease obtained from the UCI Machine Learning Repository with a total of 5110 instances. However, there are several issues with the stroke data, such as missing values or incomplete data, numerous attributes, and imbalanced data. 201 missing values, or approximately 3.9% of the total data, can be removed. After that, there is one instance in the gender attribute that has a different value from the others, namely "other," so that instance is deleted. In removing missing values and unfamiliar data, the number of instances of stroke disease becomes 4908. There are 4699 instances from the non-stroke class (the majority class) and 209 instances from the stroke class (the minority class).

The following task is to identify essential characteristics in stroke disease classification to enhance the effectiveness of the employed classification technique. The approach for selecting features in stroke disease data relies on evaluating the correlation between input attributes and the class. Figure 3 shows that the most dominant features in stroke disease data are age, hypertension, heart disease, and average glucose level, with correlation values greater than 0.1 ($r >= 0.1$). Four attributes will be processed in the next stage.

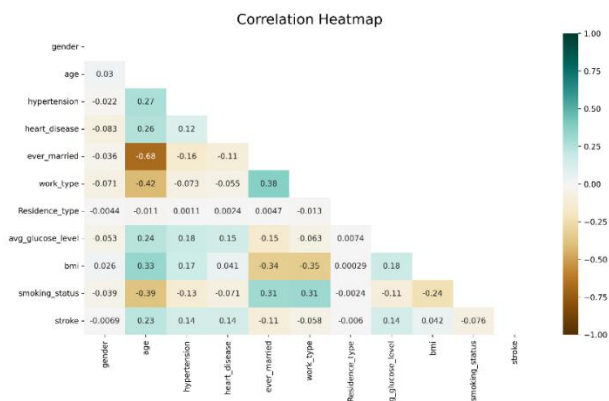


Figure 3. Correlation of input attributes with class

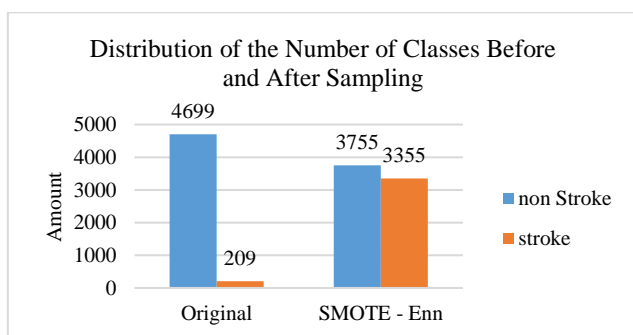


Figure 4. Distributions of number of stroke disease classes before and after sampling

The next step is to balance data using the hybrid sampling method, namely SMOTE-Enn. Data balancing using hybrid sampling involves adding minority data using SMOTE and then removing majority data that is considered noise and overlapping until the data is balanced using SMOTE-Enn. In Figure 4, the results of data balancing using the Hybrid sampling method SMOTE-Enn can be seen. Figure 3 shows that the hybrid sampling method SMOTE-Enn results in class balance, with 3355 instances of the minority class (stroke) and 3755 instances of the majority class (non-stroke).

The next step involves implementing several classification methods, namely Random Forest, SVM, and XGBoost, to classify stroke diseases based on the correlation feature selection and balanced data. In applying classification methods, stroke disease data is divided into testing and training data using the 10-fold cross-validation technique, which is repeatedly used as the testing and training data for each fold, for the classification results on the testing data, the Random Forest, SVM, and XGBoost methods are used in each fold. Then, the average of these results is calculated based on the obtained confusion matrix table.

In Table 7, the XGBoost method successfully predicted 4683 instances of the non-stroke class out of 4699 instances. Meanwhile, only 2 out of 209 were accurately classified as stroke class. The XGBoost method with SMOTE-Enn predicted 3607 instances out of 3755 as the non-stroke class and 3026 instances out of 3355 as the stroke class.

In Table 7, the SVM method without data sampling successfully predicted 4699 instances of the non-stroke class out of 4699 instances. None of the stroke class instances were successfully classified out of 209. When using the SVM method with SMOTE-Enn, only 3427 instances out of 3755 were predicted as a non-stroke class, and 2665 instances out of 3355 were classified as stroke class.

In Table 7, the Random Forest method without data sampling successfully predicted 4645 instances out of 4699 instances classified as non-stroke. However, out of the total of 209 stroke instances, only 16 instances were successfully classified. Using Random Forest with SMOTE-Enn method predicted 3714 instances of the non-stroke class out of 3755 instances and successfully classified 3285 instances of the stroke class out of a total of 3355 instances.

Table 8 illustrates that employing feature selection classification methods without sampling leads to high accuracy rates but lower recall and AUC values. This occurs because the classification method prioritizes the majority class (non-stroke) over the minority class (stroke). However, by balancing the data using hybrid sampling techniques like SMOTE-Enn, the classification method's performance improves notably in terms of recall and AUC values. In summary, utilizing the SMOTE-Enn method with Random Forest yields superior performance compared to no sampling. The study's findings show that feature selection via correlation and SMOTE-Enn with Random Forest achieves 98.4% accuracy, 98.9% recall, and a 98.4% AUC, surpassing XGBoosting and SVM methods, which is consistent with prior research showing that SMOTE-Enn can improve the performance of classification methods compared to without sampling [39-41]. The proposed method needs to be compared with the latest research to demonstrate the level of performance improvement compared to existing methods (see Table 9).

Table 9 illustrates the superior performance (accuracy and AUC) of the proposed method than several previous studies.

In general, the performance of classification methods can be improved through feature selection and data sampling [42-44]. Our study found that the use of feature selection and hybrid sampling of SMOTE-Enn provides increased performance, especially recall and AUC for all classification methods used in stroke prediction. SMOTE-ENN (Synthetic Minority Over-sampling Technique-Edited Nearest Neighbors) is a hybrid method that combines SMOTE to generate synthetic samples and ENN to clean the data set by removing noise. This technique aims to improve model performance by addressing class imbalance and improving data quality. In general, removing noise from majority classes adjacent to minorities can improve data quality, this can be seen from the performance results obtained with classification methods such as increasing recall and AUC. While feature selection methods

based on correlation and SMOTE-Enn with Random Forest exhibit superior performance in classifying stroke disease, they still have limitations. Additional research is required to address the shortcomings of the Enn method, particularly in removing samples from the majority class that are near the minority class using three nearest neighbors. The removal of three examples from the majority class close to one minority class can result in the loss of numerous majority class samples, potentially leading to valuable information loss [45]. Another weakness is the suboptimal noise removal process in the SMOTE-Enn method [40]. It is advisable to enhance this method by removing the minority class identified as the closest noise to the majority class, adjusting the number of nearest neighbors to avoid excessive data loss [46].

Table 7. Confusion matrix results of classification method on stroke disease data

Feature Selection	Data Sampling	Classification Method	Actual	Prediction		
				Non-stroke	Stroke	
Correlation	Original	XGBoost	non-stroke	4683	16	
			Stroke	207	2	
	SMOTE- Enn		non-stroke	3607	148	
			Stroke	329	3026	
	Original		SVM	non-stroke	4699	0
				Stroke	209	0
	SMOTE- Enn	non-stroke		3427	328	
		Stroke		690	2665	
	Original	Random Forest		non-stroke	4645	54
				Stroke	193	16
	SMOTE- Enn		non-stroke	3714	41	
			Stroke	70	3285	

Table 8. Performance of stroke classification methods with hybrid sampling and feature selection

Feature Selection	Data Sampling	Methods	Accuracy	Recall	AUC
Correlation	Original	XGBoost	95.4%	1.0%	50.3%
		SVM	95.7%	0.0%	50.0%
	SMOTE - Enn	Random Forest	94.9%	7.6%	53.2%
		XGBoost	93.3%	96.1%	93.1%
		SVM	85.7%	91.3%	85.4%
		Random Forest	98.4%	98.9%	98.4%

Table 9. Comparison of the proposed method with recent previous research

No	Researchers	Methods	Dataset	Accuracy	Recall	AUC
1	Ahammad [15]	Feature selection with XGBoost	stroke disease	97%	-	-
2	Ray et al. [16]	Chi-square with Decision Tree	stroke disease	96.8%	-	-
3	Gupta and Raheja [17]	Adasyn with Random Forest	stroke disease	97.67%	-	-
4	Yin et al. [18]	SMOTE-Enn with Logistic Regression	stroke disease	71.2%	80.3%	-
5	Ashrafuzzaman et al. [19]	Feature selection with CNN	stroke disease	95.5	100%	-
6	Islam et al. [22]	SMOTE with Random Forest	stroke disease	-	96%	-
7	Dev et al. [26]	PCA with Neural Networks	stroke disease	75%	68%	-
8	Proposed method	Correlation and SMOTE-Enn with Random Forest	stroke disease	98.4%	98.9%	98.4%

4. CONCLUSIONS

This study uses a combined approach of feature selection and hybrid sampling to predict stroke. The feature selection method is used to identify relevant attributes in the stroke data, followed by the application of the SMOTE-Enn hybrid sampling method to balance the data. Utilizing hybrid feature selection and sampling methods will improve the performance of classification techniques, increasing metrics such as accuracy, recall, and AUC. Compared with previous research,

the proposed method obtains the best performance in predicting stroke, this is due to the relevant attributes resulting from attribute selection, then balanced by hybrid sampling SMOTE-Enn. On average, the feature selection method and SMOTE-Enn with Random Forest outperform the absence of sampling with SVM and XGBoost methods, resulting in a 3% increase in accuracy, 91.3% increase in recall, and 45.2% increase in AUC. For future research, it is suggested to eliminate minority data considered as noise to minimize the amount of deleted data. Additionally, the potential noise

caused by overlapping can be addressed using a clustering approach.

REFERENCES

- [1] Feigin, V.L., Norrving, B., Mensah, G.A. (2017). Global burden of stroke. *Circulation Research*, 120(3): 439-448. <https://doi.org/10.1016/B978-0-323-69424-7.00014-4>
- [2] Eichler, H.G., Bloechl-Daum, B., Broich, K., Kyrle, P. A., Oderkirk, J., Rasi, G., Ivo, R.S., Schuurman, A., Senderovitz, T., Slawomirski, L., Wenzl, M., Paris, V. (2019). Data rich, information poor: Can we use electronic health records to create a learning healthcare system for pharmaceuticals?. *Clinical Pharmacology & Therapeutics*, 105(4): 912-922. <https://doi.org/10.1002/cpt.1226>
- [3] Dash, S., Shakyawar, S.K., Sharma, M., Kaushik, S. (2019). Big data in healthcare: Management, analysis and future prospects. *Journal of Big Data*, 6(1): 1-25. <https://doi.org/10.1186/s40537-019-0217-0>
- [4] Widiyaningtyas, T., Zaeni, I.A.E., Jamilah, N. (2020). Diagnosis of fever symptoms using naive bayes algorithm. In *Proceedings of the 5th International Conference on Sustainable Information Engineering and Technology*, pp. 23-28. <https://doi.org/10.1145/3427423.3427426>
- [5] Elmunsyah, H., Mu'awanah, R., Widiyaningtyas, T., Zaeni, I.A., Dwiyanto, F.A. (2019). Classification of employee mental health disorder treatment with k-nearest neighbor algorithm. In *2019 international Conference on Electrical, Electronics and Information Engineering (ICEEIE)*, pp. 211-215. <https://doi.org/10.1109/ICEEIE47180.2019.8981418>
- [6] Wirawan, I.M., Widiyaningtyas, T., Siti, N.B. (2019). Nutritional status of infants classification by calculating anthropometry through C4. 5 algorithm. In *2019 International Conference on Electrical, Electronics and Information Engineering (ICEEIE)*, pp. 216-219. <https://doi.org/10.1109/ICEEIE47180.2019.8981427>
- [7] Sarwar, T., Seifollahi, S., Chan, J., Zhang, X., Aksakalli, V., Hudson, I., Verspoor, K., Cavedon, L. (2022). The secondary use of electronic health records for data mining: Data characteristics and challenges. *ACM Computing Surveys (CSUR)*, 55(2): 1-40. <https://doi.org/10.1145/3490234>
- [8] Nia, N.G., Kaplanoglu, E., Nasab, A. (2023). Evaluation of artificial intelligence techniques in disease diagnosis and prediction. *Discover Artificial Intelligence*, 3(1): 5. <https://doi.org/10.1007/s44163-023-00049-5>
- [9] Alanazi, E.M., Abdou, A., Luo, J. (2021). Predicting risk of stroke from lab tests using machine learning algorithms: Development and evaluation of prediction models. *JMIR Formative Research*, 5(12): e23440. <https://doi.org/10.2196/23440>
- [10] Mridha, K., Ghimire, S., Shin, J., Aran, A., Uddin, M.M., Mridha, M.F. (2023). Automated stroke prediction using machine learning: An explainable and exploratory study with a web application for early intervention. *IEEE Access*, 11: 52288-52308. <https://doi.org/10.1109/ACCESS.2023.3278273>
- [11] Mamdouh Farghaly, H., Abd El-Hafeez, T. (2023). A high-quality feature selection method based on frequent and correlated items for text classification. *Soft Computing*, 27(16): 11259-11274. <https://doi.org/10.1007/s00500-023-08587-x>
- [12] Pudjihartono, N., Fadason, T., Kempa-Liehr, A.W., O'Sullivan, J.M. (2022). A review of feature selection methods for machine learning-based disease risk prediction. *Frontiers in Bioinformatics*, 2: 927312. <https://doi.org/10.3389/fbinf.2022.927312>
- [13] Chen, R.C., Dewi, C., Huang, S.W., Caraka, R.E. (2020). Selecting critical features for data classification based on machine learning methods. *Journal of Big Data*, 7(1): 52. <https://doi.org/10.1186/s40537-020-00327-4>
- [14] Utimura, L., Costa, K., Scherer, R. (2022). Real-time application of OPF-based classifier in Snort IDS. In *Optimum-Path Forest*, pp. 55-93. <https://doi.org/10.1016/B978-0-12-822688-9.00011-6>
- [15] Ahammad, T. (2022). Risk factor identification for stroke prognosis using machine-learning algorithms. *Jordanian Journal of Computers and Information Technology*, 8(3): 282-296.
- [16] Ray, S., Alshouli, K., Roy, A., AlGhamdi, A., Agrawal, D.P. (2020). Chi-squared based feature selection for stroke prediction using AzureML. In *2020 International Engineering, Technology and Computing (IETC)*, pp. 1-6. <https://doi.org/10.1109/IETC47856.2020.9249117>
- [17] Gupta, S., Raheja, S. (2022). Stroke prediction using machine learning methods. In *2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pp. 553-558. <https://doi.org/10.1109/Confluence52989.2022.9734197>
- [18] Yin, Q., Ye, X., Huang, B., Qin, L., Ye, X., Wang, J. (2023). Stroke risk prediction: Comparing different sampling algorithms. *International Journal of Advanced Computer Science and Applications*, 14(6): 1074-1081. <https://doi.org/10.14569/IJACSA.2023.01406115>
- [19] Ashrafuzzaman, M., Saha, S., Nur, K. (2022). Prediction of stroke disease using deep CNN based approach. *Journal of Advances in Information Technology*, 13(6): 604-613. <https://doi.org/10.12720/jait.13.6.604-613>
- [20] Kamal, N.P., Saraswathi, S. (2023). Comparing the random forest algorithm with k-nearest neighbor algorithm for a novel stroke prediction. *Journal of Survey in Fisheries Science*, 10: 2296-2303.
- [21] Wang, M., Yao, X., Chen, Y. (2021). An imbalanced-data processing algorithm for the prediction of heart attack in stroke patients. *IEEE Access*, 9: 25394-25404. <https://doi.org/10.1109/ACCESS.2021.3057693>
- [22] Islam, M.M., Akter, S., Rokunojjaman, M., Rony, J.H., Amin, A., Kar, S. (2021). Stroke prediction analysis using machine learning classifiers and feature technique. *International Journal of Electronics and Communications Systems*, 1(2): 17-22. <https://doi.org/10.24042/ijecs.v1i2.10393>
- [23] Abd Mizwar, A.R., Sunyoto, A., Arief, M.R. (2022). Stroke prediction using machine learning method with extreme gradient boosting algorithm. *MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, 21(3): 595-606. <https://doi.org/10.30812/matrik.v21i3.1666>
- [24] Pamungkas, Y., Wibawa, A.D., Cahya, M.D. (2022). Electronic medical record data analysis and prediction of stroke disease using Explainable Artificial Intelligence (XAI). *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, 7(4). <https://doi.org/10.22219/kinetik.v7i4.1535>

- [25] Bandi, V., Bhattacharyya, D., Midhunchakkravarthy, D. (2020). Prediction of brain stroke severity using machine learning. *Revue d'Intelligence Artificielle*, 34(6): 753-761. <https://doi.org/10.18280/RIA.340609>
- [26] Dev, S., Wang, H., Nwosu, C.S., Jain, N., Veeravalli, B., John, D. (2022). A predictive analytics approach for stroke prediction using machine learning and neural networks. *Healthcare Analytics*, 2: 100032. <https://doi.org/10.1016/j.health.2022.100032>
- [27] Hairani, H., Priyanto, D. (2023). A new approach of hybrid sampling SMOTE and ENN to the accuracy of machine learning methods on unbalanced diabetes disease data. *International Journal of Advanced Computer Science and Application*, 14(8): 585-890. <https://doi.org/10.14569/ijacsa.2023.0140864>
- [28] Hairani, H., Anggrawan, A., Priyanto, D. (2023). Improvement performance of the random forest method on unbalanced diabetes data classification using Smote-Tomek Link. *JOIV: International Journal on Informatics Visualization*, 7(1): 258-264. <https://doi.org/10.30630/joiv.7.1.1069>
- [29] Madley-Dowd, P., Hughes, R., Tilling, K., Heron, J. (2019). The proportion of missing data should not be used to guide decisions on multiple imputation. *Journal of Clinical Epidemiology*, 110: 63-73. <https://doi.org/10.1016/j.jclinepi.2019.02.016>
- [30] Gao, Q., Jin, X., Xia, E., Wu, X., Gu, L., Yan, H., Xia, Y., Li, S. (2020). Identification of orphan genes in unbalanced datasets based on ensemble learning. *Frontiers in Genetics*, 11: 820. <https://doi.org/10.3389/fgene.2020.00820>
- [31] Bhati, B.S., Rai, C.S. (2020). Analysis of support vector machine-based intrusion detection techniques. *Arabian Journal for Science and Engineering*, 45(4): 2371-2383. <https://doi.org/10.1007/s13369-019-03970-z>
- [32] Korstanje, J. (2021). Gradient boosting with XGBoost and LightGBM. In *Advanced Forecasting with Python*. Berkeley, CA: Apress, pp. 193-205. https://doi.org/10.1007/978-1-4842-7150-6_15
- [33] Chen, T., Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, pp. 785-794. <https://doi.org/10.1145/2939672.2939785>
- [34] Bentéjac, C., Csörgő, A., Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, 54: 1937-1967. <https://doi.org/10.1007/s10462-020-09896-5>
- [35] Hairani, H., Widiyaningtyas, T. (2024). Augmented rice plant disease detection with convolutional neural networks. *INTENSIF: Jurnal Ilmiah Penelitian dan Penerapan Teknologi Sistem Informasi*, 8(1): 27-39. <https://doi.org/10.29407/intensif.v8i1.21168>
- [36] Farou, Z., Aharrat, M., Horváth, T. (2023). A comparative study of assessment metrics for imbalanced learning. In *European Conference on Advances in Databases and Information Systems*, pp. 119-129. https://doi.org/10.1007/978-3-031-42941-5_11
- [37] Rezvani, S., Wang, X. (2023). A broad review on class imbalance learning techniques. *Applied Soft Computing*, 143: 110415. <https://doi.org/10.1016/j.asoc.2023.110415>
- [38] Saifudin, I., Widiyaningtyas, T. (2024). Systematic literature review on recommender system: Approach, problem, evaluation techniques, datasets. *IEEE Access*, 12: 19827-19847. <https://doi.org/10.1109/ACCESS.2024.3359274>
- [39] Khushi, M., Shaukat, K., Alam, T.M., Hameed, I.A., Uddin, S., Luo, S., Yang, X., Reyes, M.C. (2021). A comparative performance analysis of data resampling methods on imbalance medical data. *IEEE Access*, 9: 109960-109975. <https://doi.org/10.1109/ACCESS.2021.3102399>
- [40] Ependi, U., Rochim, A.F., Wibowo, A. (2023). A hybrid sampling approach for improving the classification of imbalanced data using ROS and NCL methods. *International Journal of Intelligent Engineering and Systems*, 16(3): 345-361. [10.22266/ijies2023.0630.28](https://doi.org/10.22266/ijies2023.0630.28)
- [41] Sasada, T., Liu, Z., Baba, T., Hatano, K., Kimura, Y. (2020). A resampling method for imbalanced datasets considering noise and overlap. *Procedia Computer Science*, 176: 420-429. <https://doi.org/10.1016/j.procs.2020.08.043>
- [42] Putra, L.G.R., Marzuki, K., Hairani, H. (2023). Correlation-based feature selection and Smote-Tomek Link to improve the performance of machine learning methods on cancer disease prediction. *Engineering & Applied Science Research*, 50(6): 577-583. <https://doi.org/10.14456/easr.2023.59>
- [43] Ramos-Pérez, I., Arnaiz-González, Á., Rodríguez, J.J., García-Osorio, C. (2022). When is resampling beneficial for feature selection with imbalanced wide data? *Expert Systems with Applications*, 188: 116015. <https://doi.org/10.1016/j.eswa.2021.116015>
- [44] Sun, Y., Que, H., Cai, Q., Zhao, J., Li, J., Kong, Z., Wang, S. (2022). Borderline smote algorithm and feature selection-based network anomalies detection strategy. *Energies*, 15(13): 4751. <https://doi.org/10.3390/en15134751>
- [45] Wang, K., Tian, J., Zheng, C., Yang, H., Ren, J., Li, C., Han, Q., Zhang, Y. (2021). Improving risk identification of adverse outcomes in chronic heart failure using SMOTE+ ENN and machine learning. *Risk Management and Healthcare Policy*, 2453-2463. <https://doi.org/10.2147/RMHP.S310295>
- [46] Kim, K. (2021). Noise avoidance SMOTE in ensemble learning for imbalanced data. *IEEE Access*, 9: 143250-143265. <https://doi.org/10.1109/ACCESS.2021.3120738>