Improving Stroke Detection with Hybrid Sampling and Cascade Generalization

Widya Putri Nurmawati[1*], Indahwati[2], Farit Mochamad Afendi[3]

[1,2,3]*Department of Statistics, Faculty of Mathematics and Natural Science, IPB University, Indonesia*

*corr_author: widyaputrinurmawati@apps.ipb.ac.id

Abstract - The prevalence of stroke in Indonesia has increased. One survey in Indonesia that contains information about the health conditions of the Indonesian people is the Indonesian Family Life Survey (IFLS). The proportion of respondents who had a stroke and non-stroke in IFLS5 showed an imbalance with an extreme level of imbalance; hence, this research aims to overcome this problem with SMOTE, SMOTE-Tomek Link, and SMOTE-ENN; then, the balanced dataset is classified using the ensemble and cascade approaches to improve the detection of stroke risk and to identify the important variables. However, the stroke respondents were still challenging to classify after imbalance class handling, presumably because of the large amount of data before and after balancing. The solution is to balance the training data with various percentages. The results showed the best percentage is applied to 5% of the training data, balanced by the SMOTE-ENN, and the ensemble method with the cascade approach increases the sensitivity and balanced accuracy values. Random forest and logistic regression combine models that produce the best performance, with a classification tree as the final model. The important variables obtained from this combination are the addition of probability from random forest, logistic regression, history of hypertension, age, and physical activity.

Keywords: Ensemble, IFLS, imbalanced, cascade, stroke.

I. INTRODUCTION

The second leading cause of death worldwide is stroke [1]. Based on the Basic Health Research results, the prevalence of stroke in Indonesia increased by 7% to 10.9% in 2018 [2]. A person who has a stroke is not only caused by one risk factor but also several factors. The risk factors for stroke are heart disease, hypertension, and diabetes mellitus [3]. Several studies have shown that sociodemographic factors, such as gender, body mass index, education, and marital status, have a relationship with stroke [4]–[7]. Moreover, lifestyle factors such as physical activity and smoking habits are also significant risk factors in detecting someone at risk of stroke [8]-[9]. One of the surveys in Indonesia that contains information about the health conditions and sociodemographic of the Indonesian people is the Indonesian Family Life Survey (IFLS). Research related to Indonesian Family Life Survey wave 5 (IFLS5) stoke data has been carried out by [10], who analysed the relationship between age, gender, blood pressure classification, diabetes, heart disease, and the incidence of stroke in patients with hypertension using the chi-square test.

The proportion of respondents who had a stroke and non-stroke in IFLS5 showed an imbalance with an extreme level of imbalance, where the percentage of imbalance ratio is 99:1. One method of dealing with the imbalance problem through a data-level approach is recommended by [11]. The training dataset uses two main types of basic data sampling: under-sampling and oversampling. One of the oversampling techniques widely used in research is the synthetic minority oversampling technique (SMOTE). This method generates synthetic data with the concept of k-nearest minority class neighbors [12]. This method's disadvantage is blindly generalizing the region of a minority class without considering a majority class. Hence, this research proposes to use a hybrid sampling technique that combines the SMOTE oversampling technique with the Tomek Link and ENN under-sampling techniques. Research using the SMOTE-ENN and SMOTE-Tomek Link methods was carried out by several researchers [13]-[14], which showed that the combination of oversampling and under-sampling methods is significantly better than using just one method.

The final prediction from ensemble techniques is voting, averaging, and stacking; another method is cascading, which was introduced by [15]. This method works by adding new attributes in the form of probability obtained from predictions from response variables. Research related to the cascade method has been conducted by [16] to apply the cascade method to improve the ability of breast cancer detection with mammography that implements the Bayesian Network algorithm. Aziz et al. used the cascade method with a combination of basic classification methods, including logistic regression, neural networks, support vector

machines, bagging, boosting, and random forest [17]. The results showed that the cascade method combined with the proposed basic classification methods increased the accuracy of breast cancer and adult data. Previous research about the cascade method only focused on improving the accuracy value, which is a high accuracy that does not necessarily indicate that the model can predict the minority class if the data held is imbalanced. So, the research focuses on increasing not only accuracy but also sensitivity and balanced accuracy. The high sensitivity means machine learning can predict minority data. Minority data in this study is important to predict because the number of strokes is less than non-stroke. This research aims to compare each combination of models in the ensemble method with the cascade method on data with imbalanced classes and identify important variables for detecting stroke risk.

II. METHOD

A. Data

The data for analysis were taken from the Indonesian Family Life Survey (IFLS) wave 5 in 2014-2015 conducted by RAND Labor and Population. IFLS was conducted in 13 provinces in Indonesia, including DKI Jakarta, West Java, East Java, South Kalimantan, South Sulawesi, South Sumatra, West Nusa Tenggara, Central Java, D.I Yogyakarta, Bali, North Sumatra, West Sumatra and Lampung [18]. The sampling method in IFLS applies stratified random sampling with the province as the stratum, which is taken randomly from each province. The variables used in the research are age, gender, marital status, education, residence, behavioural risk factors such as physical activity, smoking habits, body mass index and a history of diseases such as hypertension, heart disease, and diabetes mellitus.

B. Research Procedure

The procedure of analysis of the research is shown in Fig. **1**. Firstly, preprocessing. Secondly, modelling: Handling imbalances in training data using SMOTE with several scenarios for the percentage; then, undersampling was carried out using the Tomek Links and ENN methods, and each data set was repeated ten times. Next, cascade modeling with the final model of the classification tree and two basic models; the basic model used is bagging, boosting, random forest, and logistic regression; then, optimization hyperparameter with 5-fold cross-validation. Thirdly, evaluating the model by comparing the value of accuracy, sensitivity, and balanced accuracy of each combination algorithm.

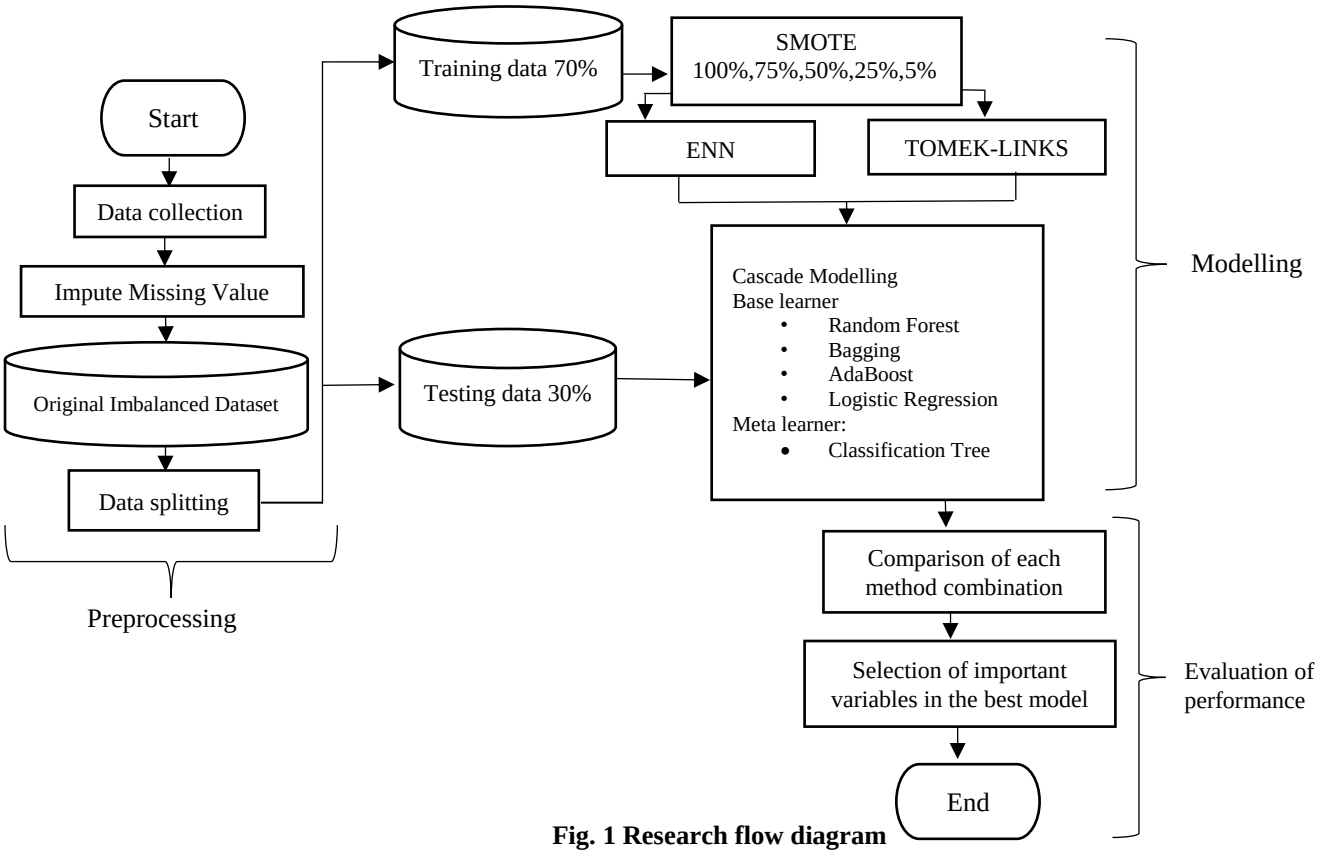


Fig. 1 Research flow diagram

C. Resampling Technique

The research uses SMOTE, SMOTE Tomek-Link, and SMOTE-ENN. SMOTE was proposed by [12] to deal with data imbalance problems by generating synthetic data. SMOTE works using the k-nearest neighbor algorithm to generate synthetic data. The steps of SMOTE algorithm are: (1) Choose a case of the minority class randomly; (2) Decide the number of nearest numbers (k), and identify k-nearest neighbors; (3) Randomly select one of the neighbors and create a synthetic sample randomly selected between the two data. The following (1) expresses the synthesized sample x_{new}.

$$x_{new} = x + (\hat{x} - x) * rand[0,1] \quad (1)$$

x represents a minority class sample, $\hat{x}$ is one of k nearest neighbors, and $rand[0,1]$ represents a random number between 0 and 1. The weakness of this method is the process of generating synthetic data without regard to the area around the minority class, causing synthetic data to be in any area, including the majority class. To solve this problem, the research proposes a hybrid sampling algorithm combining SMOTE as oversampling and Tomek Link and ENN are used to under-sampling.

SMOTE Tomek Link was developed by Batista *et al*. [19]. Tomek Links is the distance between majority and minority classes, E_i and E_j, and $d(E_i, E_j)$ is the distance between E_i dan E_j. A pair (E_i, E_j) is called Tomek-Link if there is no sample at E_1, such that (2):

$$d(E_i, E_1) < d(E_i, E_j) \text{ or } d(E_j, E_1) < d(E_i, E_j) \quad (2)$$

The SMOTE-Tomek Link algorithm's first step is over-sampling the original data set with SMOTE. Then Tomek link is identified and removed, producing a balanced data set with well-defined class clusters. SMOTE-ENN is another hybrid resampling method which combines SMOTE and ENN [20]. The idea of this method is like SMOTE-Tomek Link. So, the step of this method, first SMOTE, is applied to create synthetic data of minority class samples, then using ENN to remove any samples whose class differs from that of at least two of its three nearest neighbours.

D. Cascade Generalization

Cascade generalization is a method to combine predictions from several classification models. The cascade method works to add a new attribute in the form of a prediction probability from the response variable. The data that has been added is then remodelled using a different method; then, predictions are made with the final model with the data that has been added. The classification tree is used to construct meta-level classifier, and $L_1$and L_2 are used as the basic-level classifier. For example, data D has j predictor variables, and Y is response variable with binary classes, namely 1 and 0. The modelling steps are as follows [17]:

1. Build a basic-level classifier with $L_1$and L_2 algorithms using D.
2. Create two new predictor variables W_i where W_i is $\hat{P}(Y = 1)$ derived from the model using the L_i algorithm; $i = 1,2$.
3. Modelling with classification tree with Y as the response variable and predictor variable $\{X_1, \dots, X_j\} \cup \{W_1, W_2\}$.

Ref. [15] suggest combining classifiers with different behaviours. An algorithm with low bias is used for the meta-level classifier, and an algorithm with low variance is used for the basic level model, so in the research, the basic model used are bagging, boosting, random forest, and logistic regression. Two algorithms were used for each combination so that there are 12 combinations.

E. Performance Evaluation

A comparison of the performance evaluation of the classification method was carried out based on accuracy, sensitivity, and balanced accuracy. Those three measurements can be obtained through the confusion matrix, as in (3-5) with TP: True Positive; TN: True Negative; FP: False Positive; FN: False Negative.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (3)$$

$$Sensitivity = \frac{TP}{TP+FN} \times 100\% \quad (4)$$

$$Balanced\ Accuracy = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \times 100\% \quad (5)$$

Sensitivity measures the percentage of observations of respondents who have a stroke classified correctly from all observations from respondents who have had a stroke. Balanced accuracy is a better metric to use with imbalanced data. It calculates for positive and negative classes and does not mislead with imbalanced data.

III. RESULT AND DISCUSSION

A. Data Exploration

The data exploration shows that three predictor variables indicate a missing value. The problem of missing value in this survey was caused by several things, such as the question did not apply to the respondent, the respondent refusing to answer the

question, or the respondent did not know the answer and the interviewer's error. This study handled missing data using Multivariate Imputation by Chained Equations (MICE). The variables of weight and height are continuous distributed, so these variables are imputed using predictive mean matching. Meanwhile, the education variable has a discrete distribution with an ordinal scale, so it is imputed using the proportional odds model.

The number of respondents in the IFLS5 data is 31,173 individuals. These respondents filled out complete information about a stroke diagnosis by a doctor/paramedic. The results of data exploration showed that 185 respondents had a stroke (1%), while the respondent non-stroke were 30,988 people (99%). This proportion shows an imbalance in the data, so the problem is difficult to identify the minority class because the number of observations is too small. An exploration of explanatory variables is shown in Table I.

The research showed that the percentage of respondents who had a stroke was more men (0.62%) compared to women (0.57%). In terms of marital status, the percentage of married respondents is more affected by stroke (0.62%) compared to respondents who are not married (0.53%). Most respondents who had a stroke lived in urban, as many as 122 individuals (65.95%), while in rural, 63 individuals (34.05%). It can happen because of lifestyle factors and consumption of people in urban areas tend to be more at risk than those in rural areas. Respondents who smoked and had a stroke were 70 individuals (0.63%), while those who smoked and did not have a stroke were 11,077 individuals. Smokers have an increased risk of stroke of 1.5 to 2 times overall compared to nonsmokers [8].

The respondent's medical history used in this research was hypertension, heart disease and diabetes mellitus. The number of respondents who had a stroke in this study was 185, with 127 individuals (68.65%) having a history of hypertension. So, most respondents with a stroke had a history of high blood pressure or hypertension. A systolic blood pressure of more than 140 mmHg and a diastolic blood pressure of more than 90 mmHg are indicators of hypertension. [21]. Fig. 2 shows the characteristics of respondents with numerical explanatory variables.

TABLE I
CHARACTERISTICS OF RESPONDENTS

Characteristics	Stroke n (%)	No Stroke n (%)	Total n (%)
Sociodemographic	185 (1)	30.988 (99)	
Gender			
Male	90 (0.62)	14.458 (99.38)	14.548(100)
Female	95 (0.57)	16.530 (99.43)	16.625(100)
Marital status			
Married	140 (0.62)	22.474 (99.38)	22.614(100)
Not Married	45 (0.53)	8.514 (99.47)	8.559(100)
Education			
Elementary School/Equivalent	84 (0.87)	9.598 (99.13)	9682(100)
Junior High School/Equivalent	40 (0.64)	6.193 (99.36)	6.233(100)
Senior High School/Equivalent	33 (0.31)	10.607 (99.69)	10.640(100)
Vocational	6 (0.52)	1.138 (99.48)	1144(100)
Bachelor	17 (0.53)	3.162 (99.47)	3.179(100)
Postgraduate	3 (1.49)	198 (98.51)	201(100)
Other	2 (2.13)	92 (97.87)	94(100)
Residence			
Rural	63 (0.49)	12.739 (99.51)	12.802(100)
Urban	122 (0.66)	18.249 (99.34)	18.371(100)
Smoking Habit			
Smoked	70 (0.63)	11.077 (99.37)	11.147(100)
Not Smoked	115 (0.57)	19.911 (99.43)	20.026(100)

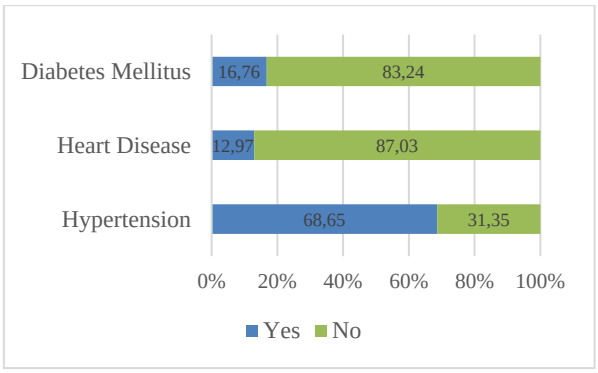


Fig. 2 Stacked bar respondent's disease history based on stroke

In the research, the average age of respondents who had a stroke was around 56 years, which is in line with research conducted by [22], which showed the highest prevalence of stroke was at the age of 50 to 59 years. Meanwhile, respondents with non-stroke had an average age of about 37 years. The physical activity score shows that the average respondent with a stroke is a respondent who tends to do a mild activity. In contrast, the respondent non-stroke does more physical activity with an average physical activity score of 3, or physical activity tends to be moderate. Increasing physical exercise can aid in lowering the risk of stroke and stroke-related death. The average BMI of respondents with stroke was 24.36 kg/m2, while respondent non-stroke was 23.46 kg/m2, so it can be concluded that the body mass index of respondents with a stroke was higher than that of respondents non-stroke. According to the BMI threshold of the Ministry of Health of the Republic of Indonesia, a value of less than 25 kg/m2 indicates that BMI is still in the normal category.

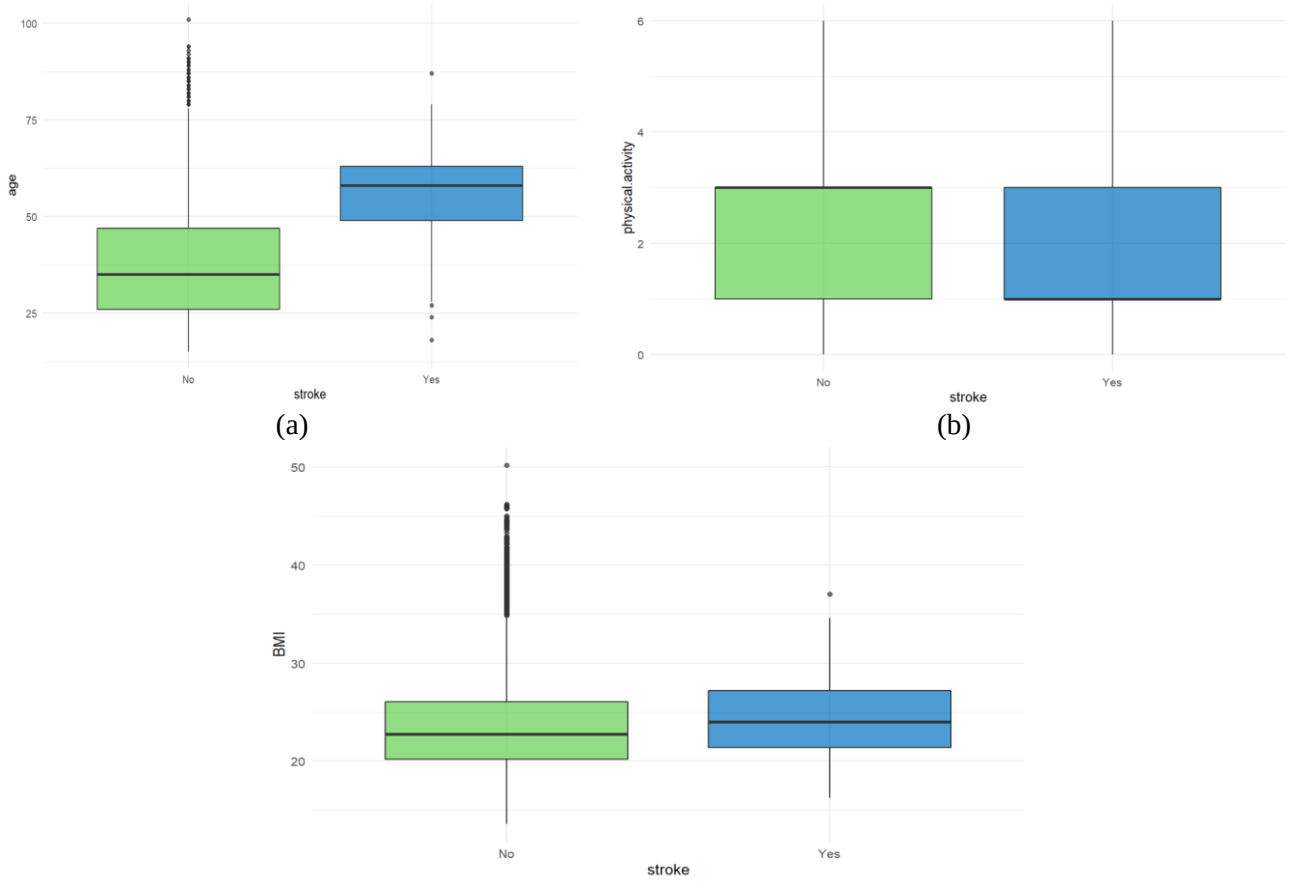


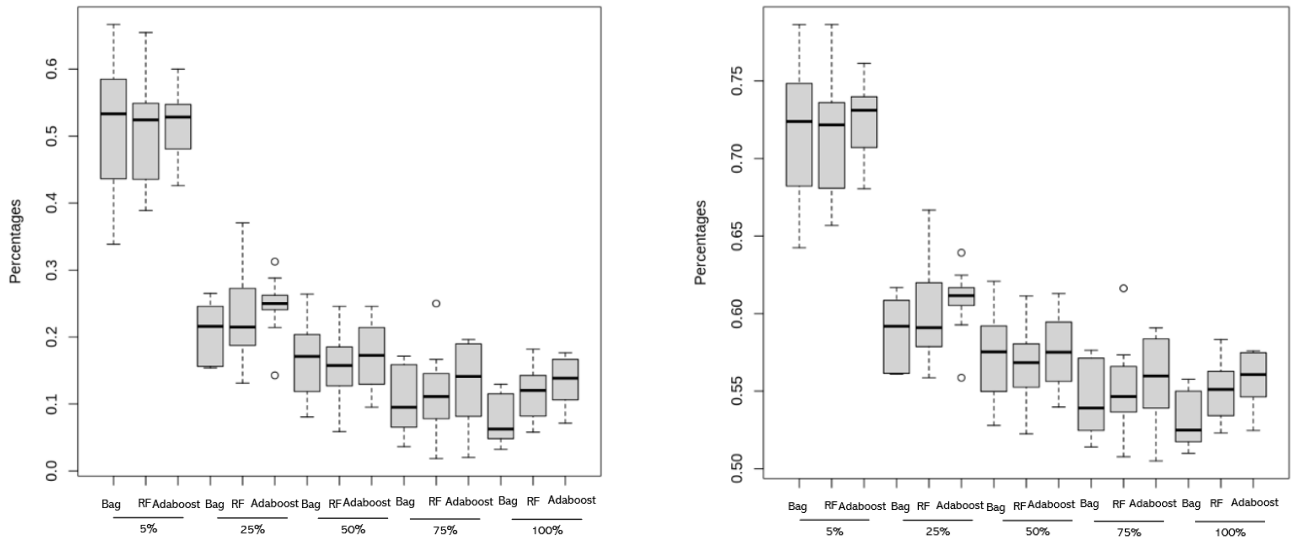
Fig. 3 *Boxplot* (a) age, (b) phisical activity score, (c) BMI based on stroke

B. Handling Dataset Imbalance

In this research, the original dataset is highly imbalanced with an imbalance ratio of 99:1. This level of imbalance is considered extreme, so the problem is that it is difficult to identify the minority class. The performance results using all training data show a relatively low sensitivity, under 20%, which means only a few respondents can classify strokes when all training data is balanced with the SMOTE. When all training data is balanced, the plots of the two classes overlap each other, making it difficult to distinguish between the minor and the major class. This also causes the sensitivity of various methods to be relatively low. The solution is to balance the training data using SMOTE with various percentages ranging from 100% to 5% in 25% intervals, resulting in 5 datasets for each data set repeated ten times.

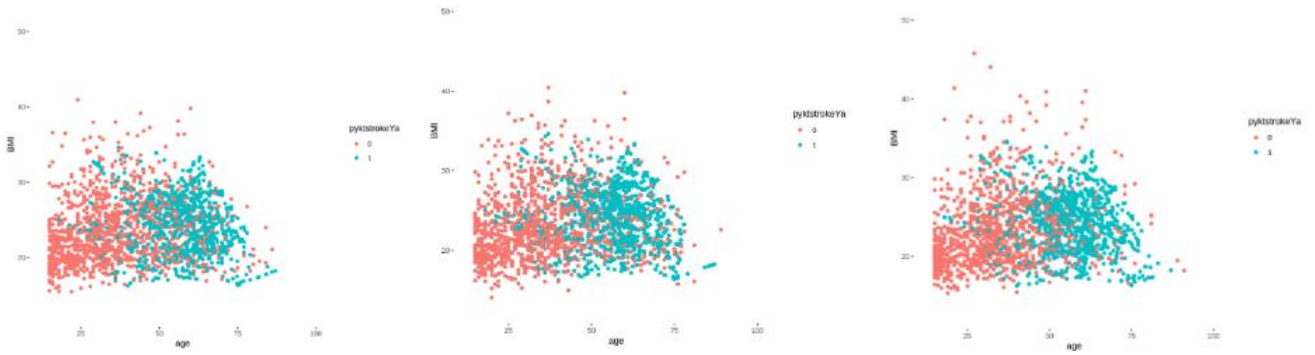
The highest sensitivity and balanced accuracy values were obtained based on various percentages when the ensemble method was applied to 5% of the training data. This percentage shows that the major and minor classes can be easily separated, so it is easy to classify the two classes compared to the other percentages. Although not perfectly separated, only a few major and minor data accumulate on top of each other. Based on the results of the highest sensitivity and balanced accuracy performance, the classification model is built based on 5% of the training data to classify respondents in the testing data. Fig. 5(a) shows that 5% of the training data generated using SMOTE resulted in 2,116 respondents. The application of SMOTE increases the number of respondents who had a stroke. The resulting dataset contains 1,115 respondent non-stroke (negative cases) and 1001 respondents who had a stroke (positive cases).

SMOTE generates random synthetic data with the concept of k-nearest neighbors included in the major class area. Fig. 5(a) shows that some minor classes overlap in the major class, and to overcome this problem, under-sampling was carried out with Tomek Link and ENN. Fig. 5(b) is the dataset resampled by SMOTE Tomek-Links consisting of 1,040 negative and 1,001 positive cases. The proportion of positive cases is 49%; this proportion was not perfectly rectified but was improved compared to the original imbalanced dataset. The dataset resampled by SMOTE-ENN and total observation is smaller than the dataset rebalanced by SMOTE because SMOTE-ENN generates synthetic positive case and remove some negative cases. Fig. 5(c) shows that the data generated with SMOTE-ENN looks cleaner, and the major and minor data that overlap each other are reduced.



(a) (b)

Fig. 4 Comparison of (a) sensitivity and (b) balanced accuracy at various training dataset percentages



(a) (b) (c)

Fig. 5 Balancing dataset using (a) SMOTE, (b) SMOTE-Tomek Link, (c) SMOTE-ENN

C. Comparison of Ensemble Method with Cascade Generalization

The basic model used in this research are bagging, adaboost, random forest, and logistic regression; then, k-fold cross validation was used for hyperparameter tuning. The hyperparameters used in the research for random forests are *mtry* (the number of predictor variables when creating a node), which is 9, and *ntree* (the number of trees formed) is 500. While the hyperparameter for bagging is the number of bootstrap repetitions, which is 25, and adaboost uses *mfina*l (the number of iterations in the boosting process) is 150, and the *maxdepth* (the limit to stop further node separation when the specified tree depth has been reached) is 3.

Fig. 6 presents the original and balanced data analyzed using the ensemble and cascade methods. The results show that the accuracy of all models on the original data is high when compared to the data that has been balanced. The resulting accuracy value is relatively high, above 80%. However, high accuracy does not necessarily mean the model can predict the minority class if the data is imbalanced. The sensitivity value in this research needs to be considered because respondents who have had a stroke are selected to be a positive class, so it is essential to predict because the number of strokes is less than non-stroke.

The sensitivity in Table II shows that all ensemble methods have a value of 0, which means that none of the minority classes have been classified correctly. None of the respondents who had a stroke were classified into the stroke class. In the research, respondents with stroke are important minority data to be identified, so the imbalance problem needs to be overcome to improve the resulting classification performance. After the balancing process, the sensitivity value increased in all balancing methods. The highest average sensitivity is shown in the data balanced with SMOTE-ENN. This method shows a cleaner data plot compared to SMOTE, making distinguishing major and minor classes easier. In addition to looking at the accuracy in classifying stroke respondents, this research also pays attention to the balanced accuracy value, which shows the accuracy in predicting stroke and non-stroke respondents (Table III).

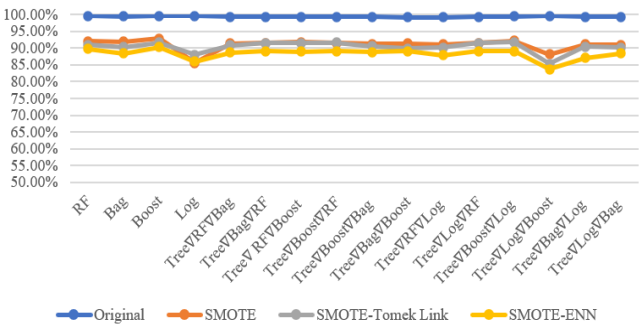


Fig. 6 Comparison of performance evaluation of accuracy

RF=Random Forest; Bag=Bagging; Boost=Boosting; Log= Logistics Regression; Tree=Classification Tree; TL= Tomek Link; ENN=Edited Nearest Neighbor.

TABLE II
COMPARISON OF PERFORMANCE EVALUATION OF SENSITIVITY

Model	Original (%)	SMOTE		SMOTE-TOMEK Link		SMOTE-ENN	
		Average (%)	STDV	Average (%)	STDV	Average (%)	STDV
RF	0	51.84	0.0847	57.47	0.0864	61.94	0.0660
Bag	0	51.37	0.0976	55.20	0.0591	63.28	0.0532
Boost	0	51.87	0.0489	52.99	0.0494	57.66	0.0684
Log	0	64.08	0.0491	62.24	0.1208	68.56	0.0857
Cascade Model							
Tree∇RF∇Bag	2.38	53.23	0.0475	53.07	0.1069	63.64	0.0782
Tree∇Bag∇RF	7.14	52.07	0.0556	55.26	0.1555	66.15	0.1094
Tree∇ RF∇Boost	7.14	53.61	0.0578	56.95	0.1356	66.71	0.0975
Tree∇Boost∇RF	0.00	52.07	0.0556	55.63	0.1492	66.15	0.1094
Tree∇Boost∇Bag	2.38	55.39	0.0559	54.86	0.1096	64.04	0.0754
Tree∇Bag∇Boost	4.76	54.68	0.0525	60.11	0.1016	64.05	0.0762
Tree∇RF∇Log	2.38	53.05	0.0669	60.79	0.1685	68.65	0.0968
Tree∇Log∇RF	0.00	52.07	0.0556	54.90	0.1619	66.15	0.1094
Tree∇Boost∇Log	0.00	47.94	0.0824	53.82	0.1606	64.02	0.0812
Tree∇Log∇Boost	0.00	60.52	0.0732	63.41	0.1330	68.49	0.0608
Tree∇Bag∇Log	2.38	53.40	0.0534	53.90	0.1123	67.84	0.0660
Tree∇Log∇Bag	2.38	49.74	0.0803	48.25	0.1055	60.74	0.1038

RF=Random Forest; Bag=Bagging; Boost=Boosting; Log= Logistics Regression; Tree=Classification Tree; TL= Tomek Link; ENN=Edited Nearest Neighbor.

TABLE III
COMPARISON OF PERFORMANCE EVALUATION OF BALANCED ACCURACY

Model	Original (%)	SMOTE		SMOTE-TOMEK Link		SMOTE-ENN	
		Average (%)	STDV	Average (%)	STDV	Average (%)	STDV
RF	51.15	72.08	0.0418	74.29	0.0428	75.92	0.0325
Bag	50.00	71.78	0.0452	72.82	0.0277	75.93	0.0235
Boost	50.00	72.48	0.0241	72.41	0.0223	74.06	0.0330
Log	50.00	77.84	0.0226	75.21	0.0512	77.38	0.0303
Cascade Model							
Tree∇RF∇Bag	51.03	72.39	0.0205	72.07	0.0485	76.24	0.0367
Tree∇Bag∇RF	53.42	71.92	0.0269	73.47	0.0749	77.69	0.0508
Tree∇ RF∇Boost	53.44	72.79	0.0294	74.34	0.0638	77.90	0.0442
Tree∇Boost∇RF	49.88	71.92	0.0269	73.74	0.0704	77.69	0.0508
Tree∇Boost∇Bag	51.03	73.43	0.0244	72.79	0.0485	76.48	0.0354
Tree∇Bag∇Boost	52.21	73.17	0.0257	75.08	0.0484	76.65	0.0355
Tree∇RF∇Log	51.02	72.22	0.0315	75.56	0.0775	78.30	0.0430
Tree∇Log∇RF	49.88	71.92	0.0269	73.32	0.0774	77.69	0.0508
Tree∇Boost∇Log	49.95	70.23	0.0412	72.91	0.0708	76.96	0.0332
Tree∇Log∇Boost	49.98	74.43	0.0292	75.48	0.0516	77.29	0.0409
Tree∇Bag∇Log	51.10	72.38	0.0218	72.23	0.0537	77.54	0.0341
Tree∇Log∇Bag	51.03	70.45	0.0376	69.35	0.0451	74.65	0.0502

RF=Random Forest; Bag=Bagging; Boost=Boosting; Log= Logistics Regression; Tree=Classification Tree; TL= Tomek Link; ENN=Edited Nearest Neighbor.

The highest average balanced accuracy is shown in the data balanced with SMOTE-ENN, and some cascade methods show more balanced accuracy values than the base method. Random forest and logistic regression with a classification tree as the final model are the best performance. The sensitivity value means the percentage of individuals predicted to have a stroke compared to all respondents who have a stroke. The increase in the

sensitivity value in the cascade method means that the cascade model effectively increases the predictions of respondents who have a stroke. Based on the results of research by [15], a method with low variance is recommended to be used as a base classifier and low bias as a meta level classifier. The success and failure of the cascade method in this research can occur because of the bias-variance characteristics of the method.

Important variables influence determining whether a person is at risk of a stroke. The best method obtained is a combination of random forest and logistic regression. The important variables obtained using this method are the probability of the logistic regression, random forest, and the variables of history of hypertension, physical activities, and age. The addition of a new variable from the probability of response variable, which is added to the cascade method process, becomes an important variable in predicting someone who has a stroke, meaning that by providing additional information in the form of probability calculated by the previous ensemble method then re-modeled by the cascade method helps in improving the performance of the model.

Hypertension is a history of disease that puts a person at risk for stroke. Hypertension or high blood pressure is a condition in repeated measurements where systolic blood pressure is more than 140 mmHg, and diastolic blood pressure is more than 90 mmHg. If hypertension is uncontrolled and handled correctly, it can cause various complications, including stroke [23]. In general, there are many mechanisms by which hypertension affects the occurrence of stroke, one of which is described in the study of [24] that hypertension causes hypoperfusion, which is a condition that causes nutrient intake to the organs to experience nutritional deficiencies if this condition continues it will cause thickening and remodeling of blood vessels which can lead to an increased risk of stroke.

In addition to hypertension, age is a consistent variable that is an important variable in all methods. In this study, the average age of respondents who had a stroke was around 56 years; this is parallel with research conducted by [22], which showed the highest prevalence of stroke was at the age of 50 to 59 years. Increasing age is a process that is followed by a reduce in the physiological function of every organ in the body; this decrease in function makes everyone more susceptible to infection or what we call inflammation. Inflammation is the body's response to an injury that aims to repair the damaged tissue. However, in old age, this process has been disrupted, which can result in a pile of remnants of the process that can clog the blood vessels, which can eventually lead to stroke [25].

IV. CONCLUSION

The results showed that the problem of data imbalance in the research was a problem that needed to be solved first because none of the respondents who had a stroke was classified into the stroke class. After implementing the imbalance handling methods, namely SMOTE, SMOTE-Tomek Link, and SMOTE-ENN, the stroke respondents were still difficult to classify, presumably because of the large amount of data before and after balancing. So, the solution is to balance the training data with various percentages, and the best sampling percentage is when the ensemble method is applied to 5% of the training data, balanced by the SMOTE-ENN. The results showed that the ensemble method with the cascade approach increases the sensitivity and balanced accuracy values. It means that the cascade model effectively classifies respondents who have had a stroke. Random forest and logistic regression with a classification tree as the final model are the combination of models that produce the best performance. The important variables obtained from this combination are the addition of probability from random forest, logistic regression, history of hypertension, age, and physical activity.

REFERENCES

[1] D. Kuriakose and Z. Xiao, “Pathophysiology and treatment of stroke: Present status and future perspectives,” *Int. J. Mol. Sci.*, vol. 21, no. 20, pp. 1–24, 2020, doi: 10.3390/ijms21207609.

[2] Kemenkes RI, “Hasil Riset Kesehatan Dasar Tahun 2018,” *Kementrian Kesehat. RI*, vol. 53, no. 9, pp. 1689–1699, 2018.

[3] Kemenkes RI, “Stroke Dont Be The One.” p. 10, 2018.

[4] Y. Wu and Y. Fang, “Stroke prediction with machine learning methods among older chinese,” *Int. J. Environ. Res. Public Health*, vol. 17, no. 6, pp. 1–11, 2020, doi: 10.3390/ijerph17061828.

[5] M. Shiozawa *et al.*, “Association of body mass index with ischemic and hemorrhagic stroke,” *Nutrients*, vol. 13, no. 7, pp. 1–13, 2021, doi: 10.3390/nu13072343.

[6] C. A. Jackson, C. L. M. Sudlow, and G. D. Mishra, “Education, sex and risk of stroke: a prospective cohort study in New South Wales, Australia,” *BMJ Open*, vol. 8, no. 9, p. e024070, Sep. 2018, doi: 10.1136/bmjopen-2018-024070.

[7] Q. Liu *et al.*, “Association between marriage and outcomes in patients with acute ischemic stroke,” *J. Neurol.*, vol. 265, no. 4, pp. 942–948, 2018, doi: 10.1007/s00415-018-8793-z.

[8] B. Pan, X. Jin, L. Jun, S. Qiu, Q. Zheng, and M. Pan,

"The relationship between smoking and stroke A meta-analysis," *Med. (United States)*, vol. 98, no. 12, pp. 1–8, 2019, doi: 10.1097/MD.0000000000014872.

[9] S. Ghozy *et al.*, "Physical activity level and stroke risk in US population: A matched case–control study of 102,578 individuals," *Ann. Clin. Transl. Neurol.*, vol. 9, no. 3, pp. 264–275, 2022, doi: 10.1002/acn3.51511.

[10] A. Hidayati, S. Martini, and L. Y. Hendrati, "Determinan Kejadian Stroke pada Pasien Hipertensi (Analisis Data Sekunder IFLS 5)," *J. Kesehat. Glob.*, vol. 4, no. 2, pp. 54–65, 2021, doi: 10.33085/jkg.v4i2.4794.

[11] B. W. Yap, K. A. Rani, H. A. Abd Rahman, S. Fong, Z. Khairudin, and N. N. Abdullah, "An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets," in *Lecture Notes in Electrical Engineering*, 2014, vol. 285, pp. 13–22, doi: 10.1007/978-981-4585-18-7_2.

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[13] F. Yang, K. Wang, L. Sun, M. Zhai, J. Song, and H. Wang, "A hybrid sampling algorithm combining synthetic minority over-sampling technique and edited nearest neighbor for missed abortion diagnosis," *BMC Med. Inform. Decis. Mak.*, vol. 22, no. 1, pp. 1–14, 2022, doi: 10.1186/s12911-022-02075-2.

[14] T. Kimura, "Customer Churn Prediction With Hybrid Resampling and Ensemble Learning," *J. Manag. Inf. Decis. Sci.*, vol. 25, no. 1, pp. 1–23, 2022.

[15] J. Gama and P. Brazdil, "Cascade Generalization," *Mach. Learn.*, vol. 41, no. 3, pp. 315–343, 2000, doi: 10.1023/A:1007652114878.

[16] K. A. Nugroho, N. A. Setiawan, and T. B. Adji, "Cascade generalization for breast cancer detection," in *Proceedings - 2013 International Conference on Information Technology and Electrical Engineering: "Intelligent and Green Technologies for Sustainable Development", ICITEE 2013*, 2013, pp. 57–61, doi: 10.1109/ICITEED.2013.6676211.

[17] A. A. Aziz, Indahwati, and B. Sartono, "Improving prediction accuracy of classification model using cascading ensemble classifiers," in *IOP Conference Series: Earth and Environmental Science*, Jul. 2019, vol. 299, no. 1, p. 012025, doi: 10.1088/1755-1315/299/1/012025.

[18] J. Strauss, F. Witoelar, and B. Sikoki, "User's Guide for the Indonesia Family Life Survey, Wave 5: Volume 2," RAND Corporation, 2016. doi: 10.7249/WR1143.2.

[19] G. E. Batista, A. L. Bazzan, and M. C. Monard, "Balancing Training Data for Automated Annotation of Keywords: a Case Study," *WOB*, vol. 3, pp. 10–18, 2003.

[20] G. E. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 20–29, Jun. 2004, doi: 10.1145/1007730.1007735.

[21] S. Singh, R. Shankar, and G. P. Singh, "Prevalence and Associated Risk Factors of Hypertension: A Cross-Sectional Study in Urban Varanasi," *Int. J. Hypertens.*, vol. 2017, pp. 1–10, 2017, doi: 10.1155/2017/5491838.

[22] I. Setyopranoto *et al.*, "Prevalence of Stroke and Associated Risk Factors in Sleman District of Yogyakarta Special Region, Indonesia," *Stroke Res. Treat.*, vol. 2019, pp. 1–8, May 2019, doi: 10.1155/2019/2642458.

[23] A. Yonata and A. S. P. Pratama, "Hipertensi sebagai Faktor Pencetus Terjadinya Stroke," *J. Major.*, vol. 5, no. 3, pp. 17–21, 2016.

[24] M. J. Cipolla, D. S. Liebeskind, and S. L. Chan, "The importance of comorbidities in ischemic stroke: Impact of hypertension on the cerebral circulation," *J. Cereb. Blood Flow Metab.*, vol. 38, no. 12, pp. 2129–2149, 2018, doi: 10.1177/0271678X18800589.

[25] T. W. Buford, "Hypertension and aging," *Ageing Res. Rev.*, vol. 26, no. 10, pp. 96–111, Mar. 2016, doi: 10.1016/j.arr.2016.01.007.