Terbit *online* pada laman web jurnal : http://innovatics.unsil.ac.id

Innovation in Research of Informatics (INNOVATICS)

| ISSN (Online) 2656-8993 |

Multi-Layer Perceptron For Diagnosing Stroke With The SMOTE Method In Overcoming Data Imbalances

M Hafidz Ariansyah[1], Sri Winarno[2], Esmi Nur Fitri[3], Helynda Mulya Arga Retha[4]*

[1,3]*Information Systems Study Program, Dian Nuswantoro University, Jl. Imam Bonjol No.207, Kota Semarang, Jawa Tengah 50131, Indonesia*
[2] *Information Technology Study Program, Dian Nuswantoro University, Jl. Imam Bonjol No.207, Kota Semarang, Jawa Tengah 50131, Indonesia*
[4] *Mathematics Study Program, IPB University, Jl. Raya Dramaga, Babakan, Kabupaten Bogor, Jawa Barat 16680, Indonesia*

[1]*112201906146@mhs.dinus.ac.id,* [2]*sri.winarno@dsn.dinus.ac.id,* [3]*112201906175@mhs.dinus.ac.id,* [4]*Helyndaretha@apps.ipb.ac.id*

ARTICLE INFORMATION

Article History:
Received: January 25, 2023
Last Revision: February 9, 2023
Published *Online*: March 5, 2023

KEY WORDS

Stroke,
Multi-Layer Perceptron,
SMOTE,
Diagnosis,
Data Imbalanced

CORRESPONDENCE

Phone: +62895634832199

E-mail: 112201906146@mhs.dinus.ac.id

ABSTRACT

Stroke is the sudden loss of brain function due to an interruption of the blood supply to the brain. Stroke is a dangerous disease that can even cause death for patients. The diagnosis of stroke must be made quickly and precisely to increase the likelihood that the patient can live a normal life again. In making a diagnosis, several factors can influence the patient to get a stroke diagnosis, including symptoms of hypertension to heart disease. From these problems, the researcher wants to classify the diagnosis of stroke so that stroke can get earlier treatment so that patients do not experience prolonged illness. The data used in this study is a stroke dataset with 4861 data labeled 0 which indicates no stroke, and 249 data labeled 1 which indicates a stroke diagnosis. This study uses the Synthetic Minority Over-sampling (SMOTE) method that will be applied to the Multi-Layer Perceptron algorithm so that researchers can get the performance of the stroke diagnosis classification model. Researchers use the SMOTE method so that the data in the classification model is balanced so that the model can make accurate predictions and avoid overfitting on the Multi-Layer Perceptron so that the accuracy in predicting stroke is better than just using an ordinary Multi-Layer Perceptron. The results of the confusion matrix analysis show that SMOTE can increase the prediction of stroke diagnosis from 12,5 % to 84,89% in optimal test.

1. INTRODUCTION

Stroke is a sudden loss of brain function due to a disruption of the blood supply to the brain [1]. A stroke is caused by an interruption of blood flow to the brain, which results in brain cell death. Impaired brain function causes symptoms including facial or limb paralysis, speech not fluent, speech not clear, possibly changes in consciousness, visual disturbances, and others [2,3]. Stroke is a disease that is the third highest cause of death in Indonesia after heart disease and cancer. Stroke attacks often come suddenly without definite signs. Stroke is a cerebrovascular (brain blood vessel) disease characterized by the death of brain tissue (cerebral infarction) that occurs due to reduced blood flow and oxygen to the brain [3]. Reduced blood flow and oxygen can be due to blockages, narrowing, or rupture of blood vessels. In Indonesia, the age characteristics of the population affected by stroke in 2018 are those aged 75 years and over occupying the first rank, followed by those aged 65-74 years, and third place at the ages of 55-64 years. The highest prevalence of stroke is in urban areas at 12.6% and 8.8% in rural areas, with the highest level of education not attending school (21.2%), and status not working (21.8%) [4]. From the results of Riskesdas (2018), it can also be seen that the prevalence of stroke in Jambi Province has increased from 3.6% in 2013 to 7% in 2018 [5].

Stroke is one of the most common diseases affecting people in Indonesia and around the world, and stroke is one of the leading causes of death [6]. In this study, researchers used machine learning (ML) to classify data to achieve the highest accuracy for building models to diagnose stroke. ML has become part of the medical field and is being used for a variety of purposes, including data analysis, diagnostic classification, and disease prediction. Although ML has brought many benefits to the healthcare sector, it also has some limitations. Because ML relies on the data available for investigation, data biases can affect the results. One example of data bias is data imbalance. Data imbalance can happen because when someone collects the data, the data obtained is not necessarily the same size. Therefore a method is needed to overcome this data bias so that the data can be modeled properly. SMOTE is the appropriate method to overcome this data imbalance [7,8] The advantages of the SMOTE method in general are that it does not cause any loss of information, avoids overfitting, builds a larger decision area, and can increase the accuracy of minority class predictions. The drawback of this method is that overgeneralization causes overlapping, it is not appropriate to use in cases that consider the importance of features [8].

One classifier that can be useful for balancing data in ML models is the Multi-Layer Perceptron. Multi-Layer Perceptron can handle data that is not linear, by using a hidden layer that can handle a non-linear representation of the input data. In addition, Multi-Layer Perceptron can also manage imbalanced classification problems by using techniques such as data resampling or using appropriate cost functions. It is what makes researchers use this classifier to classify stroke diagnoses so that the model can work well.

In this paper, researchers would model stroke data using the Multi-Layer Perceptron algorithm with the SMOTE method. The contribution of this research is applying the Multi-Layer Perceptron and the SMOTE method, the researchers hope that the model can classify well based on the features in the stroke diagnosis data so that patients get the right treatment and their life expectancy can increase.

2. RELATED RESEARCH REVIEWS

In research [9], the researchers made a comparison of the Multi-Layer Perceptron (MLP) and Support Vector Machine (SVM) methods for breast cancer classification using the Orange Data Mining application. The researcher used the Wisconsin Breast Cancer dataset. The dataset contains 569 data, which consists of 212 types of malignant cancer, 357 benign cancer, and 30 attributes that contain the characteristics of breast cancer patients. The results obtained show that in the classification of the Multi-Layer Perceptron (MLP) method with the Logistic activation function and the Adam optimization function it gives the best accuracy, precision, and recall values compared to the Support Vector Machine which is 97.7%. A study conducted by E. Chamseddine [10], this study attempts to develop an accurate model that aids clinicians in the early identification of COVID-19 using balanced data. The researchers used transfer learning (TL) to train six cutting-edge neural networks (NNs) on three separate COVID-19 datasets. The model was created to conduct a multi-classification job that distinguishes between instances of COVID-19, ordinary, and viral pneumonia. Synthetic Minority Oversampling (SMOTE) is employed on each dataset independently to overcome the imbalance. The best results are obtained by DenseNet201 and VGG-19. WCL paired with CheXNet beat the other models tested, with 98.87% accuracy, 98.21% F1 Score, 98.86% sensitivity, 99.43% specificity, 100% precision, and 99.15% AUC. In research conducted by A. F. Hardiyanti and D. Fitrianah [11], researchers used the Multilayer Perceptron and the C4.5 Algorithm to classify hospital classes in Jakarta. Based on the results, the Multilayer Perceptron produced an accuracy of 92.64%, and the C4.5 Algorithm produced an accuracy of 83.82%. Thus it can be concluded that the Multilayer Perceptron had better performance than the C4.5. Therefore, the Multilayer Perceptron algorithm can be implemented as a decision-making recommendation in assisting the Indonesian Ministry of Health in determining hospital classes.

3. METHODOLOGY

This research reveals patterns from datasets using data mining approaches, SMOTE methods, and classification. In short, this is a research model for classifying stroke diagnoses. Researchers select the Multi-Layer Perceptron algorithm from the Scikit-Learn, train the model of the algorithm with 50% - 90% training data in the data set, and analyze the model to measure the accuracy of model performance. Figure 1 shows the stages in this study which consist of five processes; data collection, dividing the dataset into training and testing, implementing the SMOTE method, implementing Multi-Layer Perceptron with default parameters, and evaluating model performance.

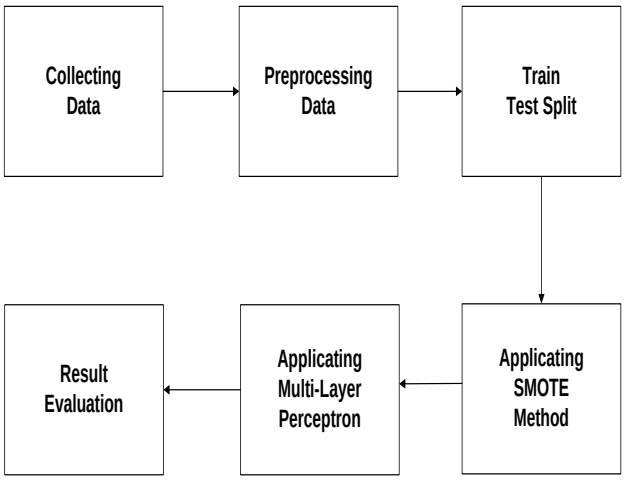

FIGURE 1. RESEARCH METHOD

3.1 Collecting Data

At this stage, the researchers determine the data be processed. Researchers use this slick Multi-Layer Perception algorithm on stroke datasets from the Kaggle machine-learning repository to observe the best available models [12]. The dataset in this study consists of 5110 data

consisting of 10 features and a label. Table 1 shows the research dataset.

TABLE 1. RESEARCH DATASET

Features	Value	Value	…	Value
Sex	Male	Female	…	Female
Age	67	61	…	49
Hypertension	No	No	…	No
Heart Disease	Yes	No	…	No
Ever Married	Yes	Yes	…	Yes
Work	Private	Self-Employed		Private
Residence	Urban	Rural	…	Urban
Avg Glucose	228.69	202.21		186.21
BMI	36.6	32.5	…	34.4
Smoking	Formerly	Never		Smokes
Label	Yes	Yes	….	Yes

*Source = https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

3.2 Pre-Processing Data

The next step is data cleansing. Data cleaning is the activity of cleaning up irrelevant or noisy data. These data may be missing, invalid, or in the form of typos. Cleanup is done by deleting data as previously described [13]. Data cleansing impacts data processing by reducing data volume and complexity [14]. No data is lost in this process. It is a sign that the dataset is valid. Data that has gone through the cleaning process enters the data transformation phase. Data is simplified to fit the data mining process. Table 2 shows the data ready to be classified.

TABLE 2. CLEAN DATASET

Features	Value	Value	…	Value
Sex	1	0	…	0
Age	67	61	…	49
Hypertension	0	0	…	0
Heart Disease	1	0	…	0
Ever Married	1	1	…	1
Work	3	4	…	3
Residence	1	0	…	1
Avg Glucose	228.69	202.21		186.21
BMI	36.6	32.5	…	34.4
Smoking	0	1		2
Label	1	1	….	1

3.3 SMOTE

The SMOTE approach employs the oversampling concept, which entails adding data from the minor class so that the sum is balanced with data from the majority class [15,16]. The SMOTE approach is used to deal with class imbalances [17,18]. The Synthetic Minority Over-Sampling Approach (SMOTE) offers good results and helps deal with unbalanced classes that encounter over-fitting in the minority class over-sampling technique. SMOTE generates a synthetic minority class instance that acts in the feature space rather than the data space. Smote produces new synthetic instances by extending the current minority samples with random samples derived from the k values of nearest neighbors by replicating the minority class examples. With synthetic results on more examples of minority groups, thus can expand their decision area for minorities. Figure 2 shows SMOTE visualization.

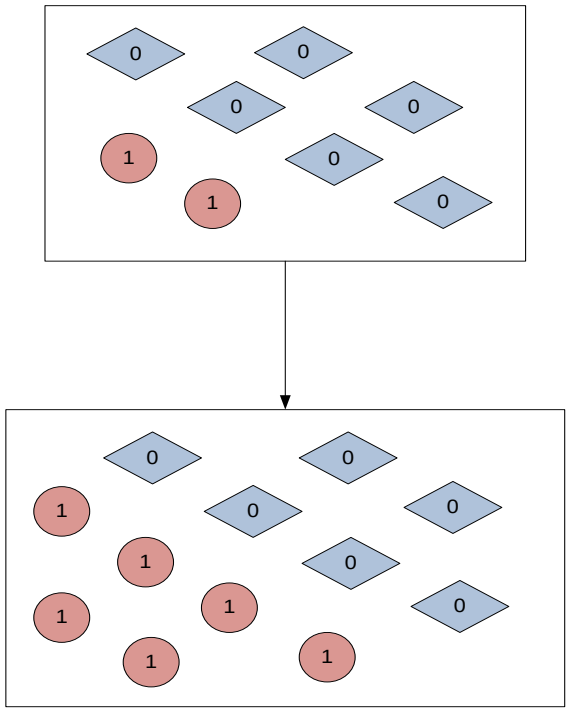

FIGURE 2. SMOTE METHOD VISUALIZATION

3.4 Train-Test Split

This method of separating training and testing provides more accurate results on new or untrained data [19]. Test data is not used to train the model, so the model does not know the results of the data [20]. It is called an out-of-sample test. This data separation is carried out after the application of the SMOTE method to the data so that the data becomes balanced. In this section, the researchers split the dataset into five scenarios (50:50, 60:40, 70:30, 80:20, 90:10) comparing training and testing.

3.5 Multi-Layer Perceptron

Multi-Layer Perceptron (MLP) is a feed-forward artificial neural network with one or more hidden layers [21]. MLP consists of an input layer that is a collection of neurons for data input, at least one hidden layer as computational neurons, and one output layer as storage neurons for computational results [22]. In MLP, there are two important parameters, namely the activation function and the optimization function. The activation function determines the output at the node of the input element. The optimization function is used to determine the most suitable weight based on input and output. The performance of the MLP network classification will depend on the network structure and training algorithm. Figure 3 shows MLP's architecture.

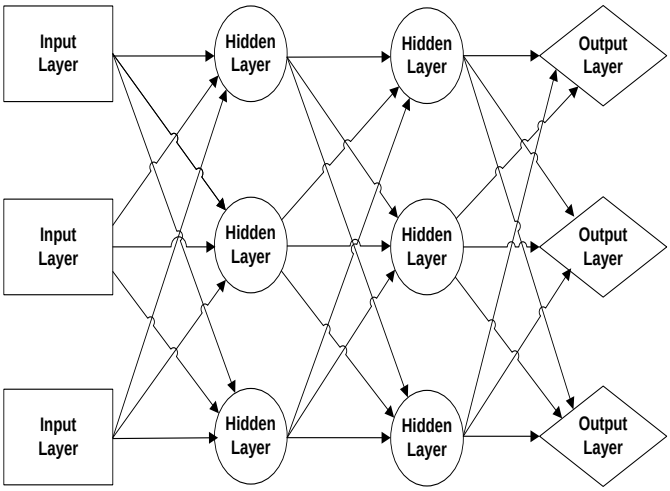

FIGURE 3. MLP ARCHITECTURE

3.6 Confusion Matrix

The confusion matrix is a table that states the classification of the correct number of test data and the wrong number of test data. An example of a confusion matrix for binary classification is shown in Figure 4 [23,24].

	Positive	Negative	Predicted Values
Positive	TP	FP	
Negative	FN	TN	
Actual Values			

FIGURE 4. CONFUSION MATRIX

Explanation :
TP (True Positive) = the number of class 1 that is correctly classified as class 1
TN (True Negative) = the number of class 0 which is correctly classified as class 0
FP (False Positive) = number of class 0 incorrectly classified as class 1
FN (False Negative) = number of class 1 which is incorrectly classified as class 0

4. RESULT AND DISCUSSION

4.1 Balancing Data

The researchers balanced the data using the SMOTE method. At this stage, the dataset will be checked first for the distribution of the labels. The result is for label 0 (Not a stroke) it has 4861 data out of 5110 data, while for label 1 (Stroke) it has 249 data. Figure 5 shows the label distribution of the dataset.

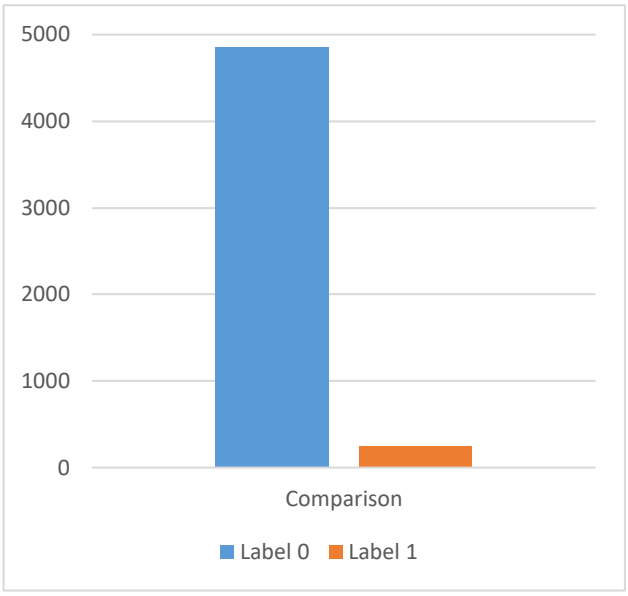

FIGURE 5. ORIGINAL DATA LABEL DISTRIBUTION

Next, the researchers applied the SMOTE method to the dataset so that the distribution of the data was balanced between label 0 and label 1. The researchers got a 1:1 ratio for labels 0 and 1 with a total of 9722 data. Figure 6 shows the results of this process.

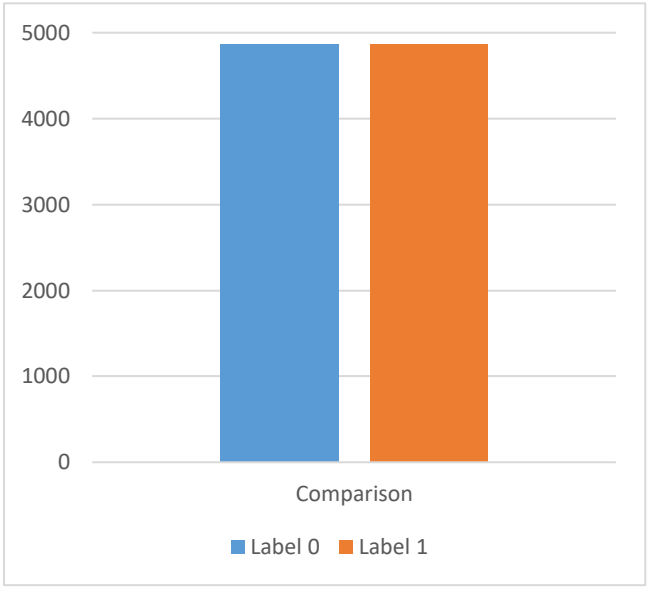

FIGURE 6. SMOTE DATA LABEL DISTRIBUTION

4.2 Set Algorithm Parameters

After obtaining a dataset with balanced data, researchers apply the Multi-Layer Perceptron algorithm for model classification. Researchers applicated the default parameters because the default parameters are parameters that have been set according to the standards of the algorithm. Table 3 shows the parameters of the Multi-Layer Perceptron used in the modeling process.

TABLE 3. MLP DEFAULT PARAMETER

Parameters	Default
Hidden Layer Size	100
Activation	relu
Solver	adam
Alpha	0.0001
Batch Size	auto
Learning Rate	constant
Learning Rate Init	0.001
Power T	0.5
Max Iter	200
Shuffle	True
Random State	None
Tol	1e-4
Verbose	False
Warm Start	False
Momentum	0.9
Nesterovs moment	True
Early Stopping	False
Validation Fract	0.1
Beta 1	0.9
Beta 2	0.999
Epsilon	1e-8
N Iter No Change	10
Max Fun	15000

4.3 Evaluation

4.3.1. 50:50 Scenario

In this scenario, researchers used 50% of the dataset to train the model. Figure 7 is the result of applying the algorithm to the original data while Figure 8 is the result of applying the algorithm to the data using the SMOTE method.

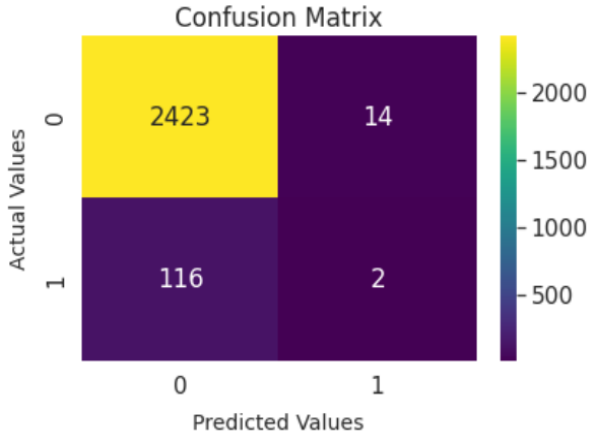

FIGURE 7. ORIGINAL DATA CONFUSION MATRIX (50:50)

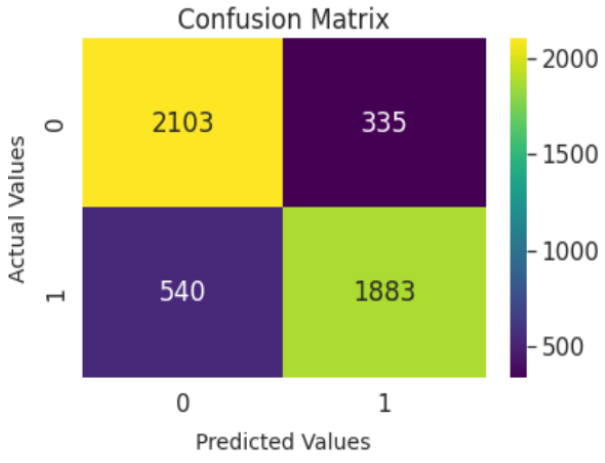

FIGURE 8. SMOTE METHOD CONFUSION MATRIX (50:50)

Figures 7 and 8 show a big difference in the accuracy of the predicted data on label 1 (STROKE). In Figure 7, it can be seen that label 1 is only 2 correct data out of the 16 predicted data, so the accuracy value of the prediction results is only 12.5 %. Whereas in Figure 8, the results of the correct predictions for label 1 amount to 1883 out of 2218 data, so the accuracy value of the prediction results for label 1 is 84.89 %.

4.3.2. 60:40 Scenario

In this scenario, researchers used 60% of the dataset to train the model. Figure 9 is the result of applying the algorithm to the original data while Figure 10 is the result of applying the algorithm to the data using the SMOTE method.

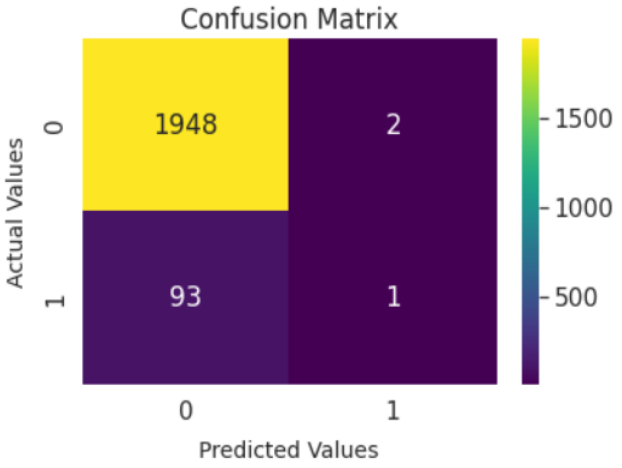

FIGURE 9. ORIGINAL DATA CONFUSION MATRIX (60:40)

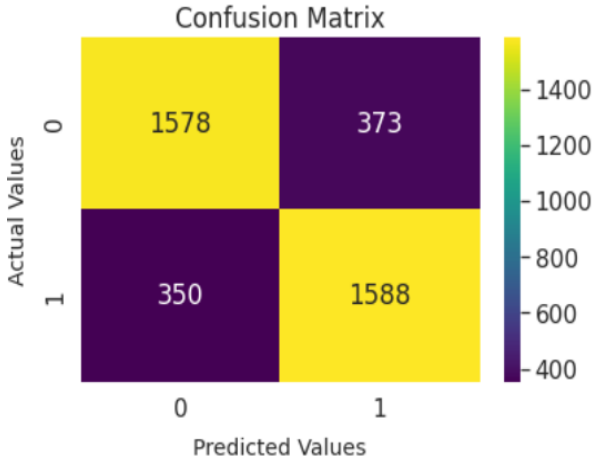

FIGURE 10. SMOTE METHOD COMNFUSION MATRIX (60:40)

Figures 9 and 10 show that there is a large difference in the accuracy of predicted data for label 1. In Figure 9, label 1 contains correct data for only one of the three predicted data. Therefore, the prediction result has an accuracy value of only 33,33%. In Figure 10, the correct prediction result for label 1 is 1588 out of 1961, and the prediction result for label 1 has an accuracy value of 80.97%.

4.3.3. 70:30 Scenario

In this scenario, researchers used 70% of the dataset to train the model. Figure 11 is the result of applying the algorithm to the original data while Figure 12 is the result

of applying the algorithm to the data using the SMOTE method.

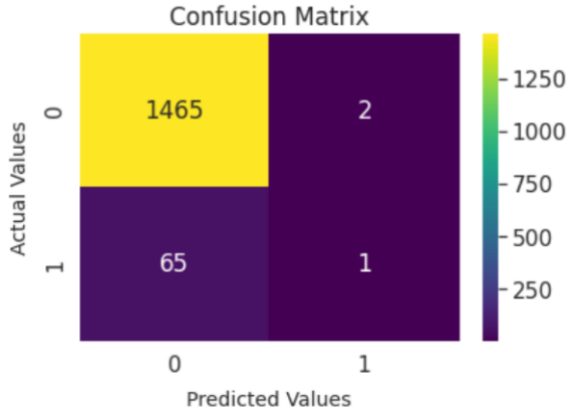

FIGURE 11. ORIGINAL DATA CONFUSION MATRIX (70:30)

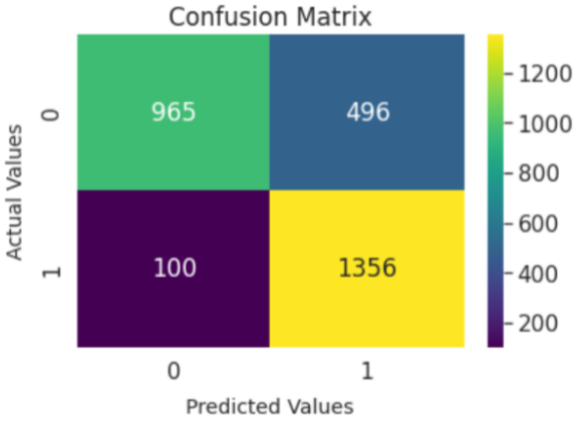

FIGURE 12. SMOTE METHOD CONFUSION MATRIX (70:30)

Figures 11 and 12 show a little difference in the accuracy of the predicted data on label 1. In Figure 11, it can be seen that label 1 is only one correct data out of the three predicted data, so the accuracy value of the prediction results is only 33,33 %. In Figure 12, the results of the correct predictions for label 1 amount to 1356 out of 1852 data, so the accuracy value of the prediction results for label 1 is 73.22%.

4.3.4. 80:20 Scenario

In this scenario, the researchers trained the model using 80% of the dataset. Figure 13 is the result of applying the algorithm to the original data, and Figure 14 is the result of applying the algorithm to the data using the SMOTE method.

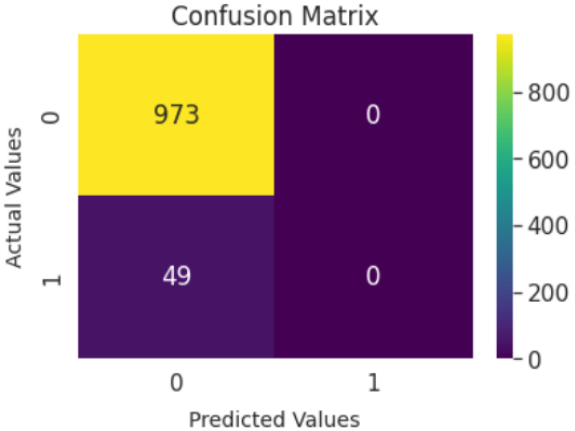

FIGURE 13. ORIGINAL DATA CONFUSION MATRIX (80:20)

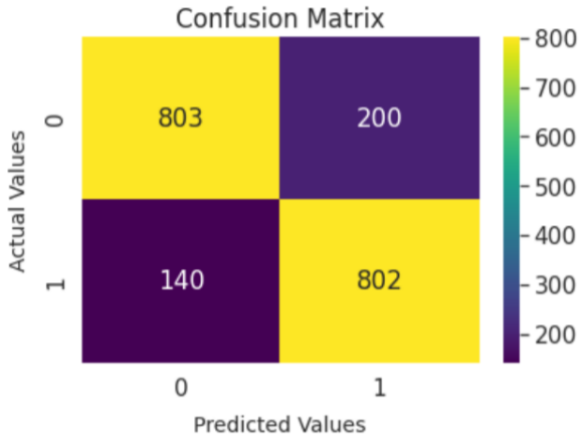

FIGURE 14. SMOTE METHOD CONFUSION MATRIX (80:20)

Figures 13 and 14 show a big difference in the accuracy of the predicted data on label 1. In Figure 13, it can be seen that label 1 does not have the correct data, so the accuracy value of the prediction results is 0%. In Figure 14, the results of the correct predictions for label 1 amount to 802 out of 1002 data, so the accuracy value of the prediction results for label 1 is 80.04%.

4.3.5. 90:10 Scenario

In this scenario, the researchers trained the model using 90% of the dataset. Figure 15 is the result of applying the algorithm to the original data, and Figure 16 is the result of applying the algorithm to the data using the SMOTE method.

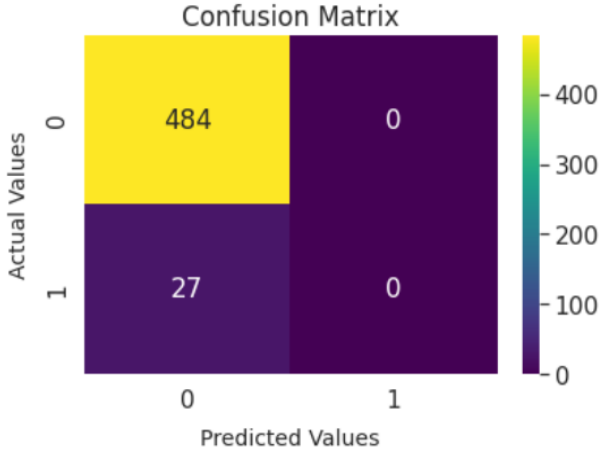

FIGURE 15. ORIGINAL DATA CONFUSION MATRIX (90:10)

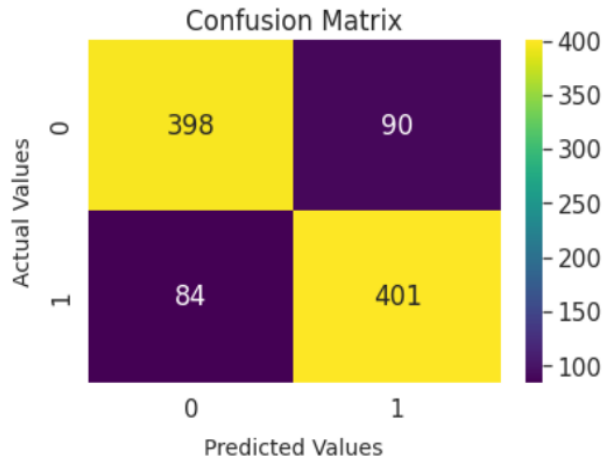

FIGURE 16. SMOTE METHOD CONFUSION MATRIX (90:10)

Figures 15 and 16 show a big difference in the accuracy of the predicted data on label 1. In Figure 15, it can be seen that label 1 does not have the correct data, so the accuracy value of the prediction results is 0%. In Figure 1, the results of the correct predictions for label 1 amount to 401 out of 491 data, so the accuracy value of the prediction results for label 1 is 81.67%.

4.4 Overall Evaluation

After passing various tests, the researchers collated the test results for accuracy comparison. This comprehensive review aims to find the best model to predict stroke diagnosis. The model with original data is compared with the model with the SMOTE method. In Figure 17, researchers often show pictures of actual models, and in Figure 18, researchers present an accuracy model for predicting stroke diagnosis.

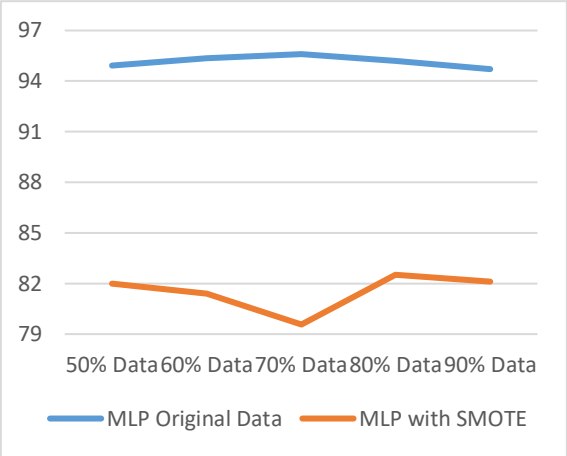

FIGURE 17. COMPARISON ACCURACY MODEL (OVERALL)

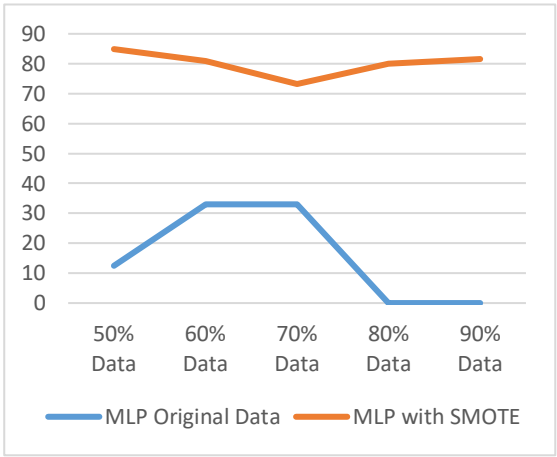

FIGURE 18. COMPARISON STROKE DIAGNOSTIC ACCURACY

From the comparison results in Figures 17 and 18, MLP with original data is superior overall, but it is inadequate at diagnosing stroke. It happens because there is an imbalance in the data between labels 1 and 0. Then MLP with SMOTE produces lower accuracy than MLP with original data, but MLP with SMOTE can diagnose stroke very well. The proof is the accuracy value in Figure 18. MLP with SMOTE can do modeling well because the data is balanced. The best model for diagnosing stroke is found in the MLP model with SMOTE with 50% data training, and this model achieves an accuracy of up to 84.89% in diagnosing stroke.

5. CONCLUSION

The conclusion is that the SMOTE method is very helpful in dealing with unbalanced data and Multi-Layer Perceptron is a very good algorithm because this algorithm can balance the accuracy of training and testing. The results for the accuracy value of the Multi-Layer Perceptron (MLP) algorithm with original data have a higher accuracy value overall. The optimal value for Multi-Layer Perceptron (MLP) with original data is a 70:30 data split with an accuracy value of 95.6%. However, in predicting stroke diagnosis, the model is unable to predict accurately because it has an accuracy of 33.33% stroke diagnosis. As for the Multi-Layer Perceptron (MLP) with the SMOTE method, the researchers found the optimal value on the 50:50 data split. Although the Multi-Layer Perceptron (MLP) with the SMOTE method has a lower overall accuracy value than the original data, which is 82%, this algorithm can achieve a stroke diagnosis accuracy of 84.89%. It happens because the data used in training and validation are balanced data. There is no inequality between labels 1 and 0, so the model can run very well. For further research, researchers hope that the accuracy in predicting stroke diagnoses can increase. Future researchers can use feature selection so that the data input process does not take a long time and use another algorithm to find the better accuracy in diagnosing stroke.

REFERENCES

[1] R. Perna, L. Harik, "The role of rehabilitation psychology in stroke care described through case examples", *NeuroRehab*, vol.46, no.2, pp. 195-204, 2020, 10.3233/NRE-192970.

[2] T. G. Rahayu, "Hubungan Pengetahuan dan Sikap Keluarga Dengan Risiko Kejadian Stroke Berulang", *JIKP*, vol.9, no.2, pp. 140–146, 2020.

[3] E. Ernawati, S. Sovia, and D. Nomiko, "Family Coaching terhadap Pelaksanaan Tugas Kesehatan Keluarga pada Klien Stroke", *JKS*, vol.6, no.1, pp. 109-116, 2022, https://doi.org/10.31539/jks.v6i1.3847.

[4] Kementrian Kesehatan Republik Indonesia, "Hasil utama Riskesdas 2018", In Kementerian Kesehatan Badan Penelitian dan Pengembangan Kesehatan, 2018, https://kesmas.kemkes.go.id/assets/upload/dir_519d41d8cd98f00/files/Hasil- riskesdas-2018_1274.pdf, Accessed in 02 Jan 2023.

[5] Dinas Kesehatan Provinsi Jambi, "Profile Health Department of Health Jambi Province", 2020, http://dinkes.jambiprov.go.id/file/informasi_publik/MTYxNTE2NDQyOA_Wkt1615164428_XtLnBkZg.pdf, Accessed in 02 Jan 2023.

[6] R. E. Pambudi, S. Sriyanto, and F. Firmansyah, "Klasifikasi Penyakit Stroke Menggunakan Algoritma Decision Tree C. 45", *TEKNIKA*, vol.16,

no.2, pp. 221-226, 2022, https://doi.org/10.5281/zenodo.7535865.

[7] D. Elreedy, and A. F Atiya, "A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance", *Information Sciences*, vol. 505, pp. 32-64, 2019, https://doi.org/10.1016/j.ins.2019.07.070.

[8] X. W. Liang, A. P. Jiang, T. Li, Y. Y. Xue, and G. T. Wang, "LR-SMOTE—An improved unbalanced data set oversampling based on K-means and SVM", *KBS*, vol. 196, pp. 105845, 2020, https://doi.org/10.1016/j.knosys.2020.105845.

[9] J. Kusuma, B. H. Hayadi, W. Wanayumini, And R. Rosnelly, "Komparasi Metode Multi Layer Perceptron (MLP) dan Support Vector Machine (SVM) untuk Klasifikasi Kanker Payudara", *MIND*, vol. 7, no.1, pp. 51-60, 2022, https://doi.org/10.26760/mindjournal.v7i1.51-60.

[10] E. Chamseddine, N. Mansouri, M. Soui, and M. Abed, "Handling class imbalance in COVID-19 chest X-ray images classification: Using SMOTE and weighted loss", *Applied Soft Computing*, vol. 129, p. 109588, 2022, https://doi.org/10.1016/j.asoc.2022.109588.

[11] A. F. Hardiyanti, and D. Fitrianah, "Perbandingan Algoritma C4. 5 dan Multilayer Perceptron untuk Klasifikasi Kelas Rumah Sakit di DKI Jakarta", *InComTech*, vol.11, no.3, 198-209, 2021, 10.22441/incomtech.v11i3.10632.

[12] F. Fedesoriano, "Stroke Prediction Dataset", Kaggle, 2021, https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset, Accessed in 20 Dec 2023.

[13] Q. A'yuniyah, E. Tasia, N. Nazira, P. F. Pratama, M. R. Anugrah, J. Adhiva, and M. Mustakim, "Implementasi Algoritma Naïve Bayes Classifier (NBC) untuk Klasifikasi Penyakit Ginjal Kronik", *JSON*, vol. 4, no. 1, 72-76, 2022, http://dx.doi.org/10.30865/json.v4i1.4781.

[14] Y. Zhang, M. Safdar, J. Xie, J. Li, M. Sage, and Y. F. Zhao, "A systematic review on data of additive manufacturing for machine learning applications: the data quality, type, preprocessing, and management", *JIM*, pp. 1-36, 2022, https://doi.org/10.1007/s10845-022-02017-9.

[15] H. Hairani, K. E. Saputro, and S. Fadli, "K-means-SMOTE untuk menangani ketidakseimbangan kelas dalam klasifikasi penyakit diabetes dengan C4. 5, SVM, dan Naive Bayes", *JTSK*, vol. 8, no.2, pp. 89-93, 2020, https://doi.org/10.14710/jtsiskom.8.2.2020.89-93.

[16] F. D. Astuti, and F. N. Lenti, "Implementasi SMOTE untuk mengatasi Imbalance Class pada Klasifikasi Car Evolution menggunakan K-NN", *JUPITER*, vol. 13, no. 1, pp. 89-98, 2021.

[17] X. W. Liang, A. P. Jiang, T. Li, Y. Y. Xue, and G. T. Wang, "LR-SMOTE—An improved unbalanced data set oversampling based on K-means and SVM", *KBS*, vol. 196, pp. 105845, 2020, https://doi.org/10.1016/j.knosys.2020.105845.

[18] J. Sun, H. Li, H. Fujita, B. Fu, and W. Ai, "Class-imbalanced dynamic financial distress prediction based on Adaboost-SVM ensemble combined with SMOTE and time weighting", *Info Fusion*, vol. 54, pp. 128-144, 2020, https://doi.org/10.1016/j.inffus.2019.07.006.

[19] T. Pan, J. Zhao, W. Wu, and J. Yang, "Learning imbalanced datasets based on SMOTE and Gaussian distribution", *Info Sci*, vol. 512, pp. 1214-1233, 2020, https://doi.org/10.1016/j.ins.2019.10.048.

[20] R. Y. Choi, A. S. Coyner, J. Kalpathy-Cramer, M. F. Chiang, and J. P. Campbell, "Introduction to machine learning, neural networks, and deep learning", *TVST*, vol. 9, no.2, pp. 14-14, 2020, https://doi.org/10.1167/tvst.9.2.14.

[21] M. Desai, and M. Shah, "An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and Convolutional neural network (CNN)", *Clinical eHealth*, vol. 4, pp. 1-11, 2021, https://doi.org/10.1016/j.ceh.2020.11.002.

[22] S. Nosratabadi, S. Ardabili, Z. Lakner, C. Mako, and A. Mosavi, "Prediction of food production using machine learning algorithms of multilayer perceptron and ANFIS", *Agriculture*, vol. 11, no. 5, pp. 408, 2021, https://doi.org/10.3390/agriculture11050408.

[23] M. H. Ariansyah, S. Winarno, and A. Salam, "STB Sentiment Analysis Classification Multiclass Modeling Using Calibrated Classifier With SGDC Tuning As Basis and Sigmoid Method", *International Journal of Computer and Information System (IJCIS)*, vol. 4, no. 1, pp. 1-7, 2023, https://doi.org/10.29040/ijcis.v4i1.107.

[24] J. Xu, Y. Zhang, and D. Miao, "Three-way confusion matrix for classification: A measure driven view", Info sci, vol. 507, pp. 772-794, 2020, https://doi.org/10.1016/j.ins.2019.06.064.

AUTHOR BIODATA

M. Hafidz Ariansyah
Student at Dian Nuswantoro University
Information System, Data Analyst

Sri Winarno, Ph.D
Head of Information Technology Department at Dian Nuswantoro University
Field : Education and Data Mining

Esmi Nur Fitri
Student at Dian Nuswantoro University
Information System, Data Analyst

Helynda Mulya Arga Retha
Student at IPB University
Mathematics, Data Analyst