International Journal of Management and Data Analytics (IJMADA)

Int. J. Management and Data Analytics, Vol. 3 (1), 1-6
ISSN: 2816-9395
Journal Homepage: http://ijmada.com

Hypertension Detection Using Passive-Aggressive Algorithm with the PA-I And PA-II Methods

M. Hafidz Ariansyah[1], Sri Winarno[2]

[1] *Department of Information System, Dian Nuswantoro University, Indonesia*
[2] *Department of Information Technology, Dian Nuswantoro University, Indonesia*

Received: Jan 16, 2023. Revised: March 13, 2023. Published: March 31, 2023

ARTICLE INFO
Keywords:
Diagnostic; Classification; Hypertension; Model; Passive-Aggressive

ABSTRACT
Hypertension is a primary factor in diseases such as stroke, heart failure, myocardial infarction, atrial fibrillation, peripheral arterial disease, and aortic dissection. Early detection of hypertension from medical history is very urgent for the first treatment of patients so that the patient's life expectancy increases, increases the effectiveness of treatment, reduces treatment costs, and reduces the severity of hypertension. Researchers get detection results using a branch of AI technology, namely machine learning to find new knowledge from data and find patterns to make diagnoses. Researchers use machine learning that can explore large amounts of data sets to produce knowledge that is beneficial to science. In this paper, the researchers used the Passive-Aggressive algorithm and the PA-I and PA-II methods to make a model for the diagnosis of hypertension. This algorithm can work well for learning by transforming data and dealing with unbalanced classification problems. PA-I shows stable accuracy of test data with a value of 80.3 - 84.15%, and PA-II shows accuracy instability with a value of 71.41 - 82.41%. From these results, PA-I shows that the model is good in diagnosing hypertension patients because its accuracy is stable and high enough. The results also show that the model is not overfitting, and the new data can be predicted well in line with the training data because, on the results of training accuracy, PA-I shows an accuracy of 81.6 - 84.56% while PA-II shows an accuracy of 71.6 - 82.71%.

1. INTRODUCTION

Blood pressure is the force that blood uses as it circulates through the arterial walls of the human body [1]. There are two types of blood pressure in humans: systolic and diastolic [2,3]. The pressure in blood vessels when the heart is beating is called systolic pressure, and diastolic pressure occurs between heartbeats. Hypertension is diagnosed with systolic blood pressure above 140 mmHg [4,5] and diastolic blood pressure above 90 mmHg [6,7]. Hypertension is one of the main factors leading to several diseases like stroke, heart failure, myocardial infarction, atrial fibrillation, and peripheral arterial disease up to aortic dissection.

The incidence of hypertension in the world reaches 1.13 billion people, of which 31% are at risk in adults, and continues to increase by 5.1% compared to the global prevalence in 2000-2010 [8]. In Indonesia, hypertension occurs in the age group of 31-44 years with a percentage of 31.6%, 45-54 of 45.3%, and 55-64 of 55.2% [9]. The high rate of hypertension makes the medical world have to take action as early as possible so that this disease does not get worse. The importance of handling during the golden hour will also have an impact on the patient's recovery quickly so that the patient will be able to live an ordinary life as before.

The demand for machine learning techniques that can solve immediate construction problems in real time has increased significantly. Online learning is a set of machine learning algorithms that can consistently train models from input data. In addition, online learning techniques have been used as an alternative learning method to reduce computational costs for large-scale problems. Among the popular online learning algorithms, Passive-Aggressive (PA) [10] offers a suitable solution for this type of task. This technique is a simple formulation of a classification framework that aims to find new linear models of the data in the form of weight vectors that are close to the current one but provide the correct classification of the existing data. It is a compromise between a passive state where the algorithm updates the model as little as possible and an aggressive state where it tries to fit the current instance correctly.

In previous research studies, Passive-Aggressive is usually used in text classification because this algorithm is proven to be one of the best in text classification besides SGDC. Research [11] shows the work of four machine learning algorithms namely Logistic Regression, Decision Tree, Random Forest, and Passive-Aggressive Classifier

for the Kaggle dataset to predict fake news on social media. The prediction accuracy using the Logistic Regression algorithm is 98.81%, the Random Forest algorithm can predict with an accuracy of 99.01%, the Decision Tree can predict with an accuracy of 99.69%, and the Passive-aggressive Classifier provides an accuracy of 92.50%. Research [12] shows challenge of SQL injection detection is solved using machine learning algorithms. The researchers used classification techniques to classify incoming communications as a SQL injection or plain text. The researchers used the Naive Bayes classifier, the Passive-Aggressive classifier, SVM, CNN, and Logistic Regression. The Nave Bayes classifier machine learning model has 95% accuracy, the Passive-Aggressive classifier has 79% accuracy, SVM has 79% accuracy, and the Logistic Regression has 92% accuracy.

Based on previous research, Passive-aggressive is more often used in research with text classification, but in this paper, the researchers use this algorithm for classification with tabular data. The performance measurement model in this paper uses accuracy, precision, recall, and specificity in predicting hypertension diagnoses. This model can be used as a prediction system when there is new data to be tested for hypertension diagnosis.

2. METHODOLOGY

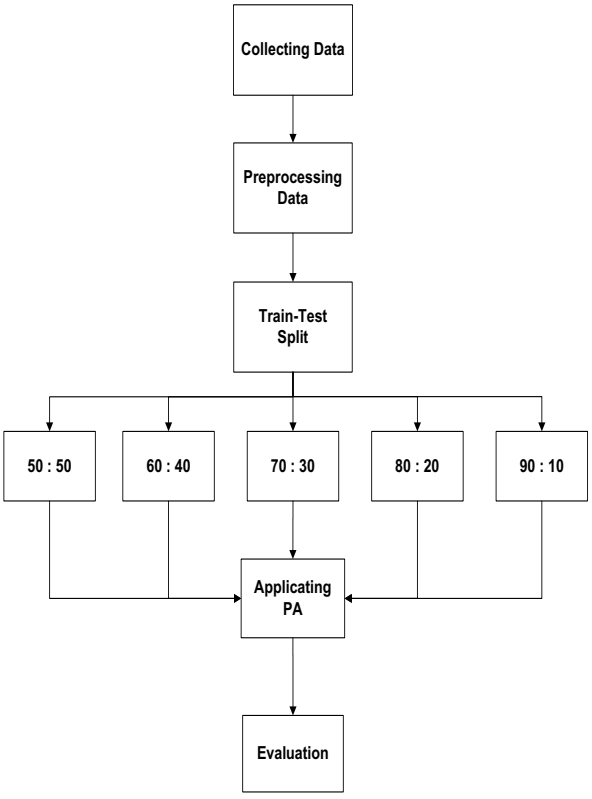

Figure 1. Research Method

From Fig. 1, researchers carried out six stages in classification. These stages are data collection, pre-processing data, splitting data into training and testing data, applying the Passive-Aggressive algorithm with the PA-I and PA-II methods, and evaluating the model to find the best model for diagnosing hypertension.

A. Collecting Data

This research utilizes datasets from Kaggle that can be used to determine the performance of the hypertension diagnosis model [13]. This dataset has two classes, namely hypertension and not. This data has a total of 26083 data, 14274 data labeled as hypertension (1), and 11809 data labeled as not (2). Table 1 shows the dataset of this study, and Fig. 2 shows the distribution of the labels.

TABLE I. DATASET

Features	Value	…	Value
Age	57	…	64
Sex	Man	…	Woman
Chest pain	Non-Anginal	…	Atypical-Angina
Trestbp	145	…	130
Chol	233	…	250
Fbs	> 120mg/dl	…	<120mg/dl
Restecg	Normal	…	S-T abnormal
Thalach	150	…	187
Exang	Yes	…	Yes
Oldpeak	2.3	…	3.5
Slope	Upsloping	…	Upsloping
Ca	0	…	0
Thal	1	…	2
Label	Hypertension	…	Hypertension

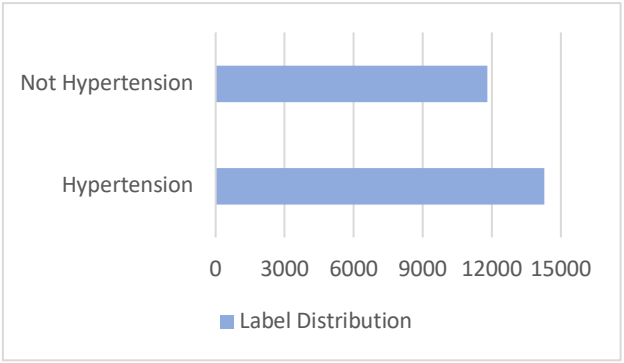

Figure 2. Distribution of Labels in Dataset

B. Pre-processing Data

The next step is data pre-processing so that the data is ready for modeling. At this stage, there are different ways to process the data as cleaning out irrelevant and noisy data (missing records or typos) [14]. As mentioned earlier, pre-processing is done by cleaning data. Data cleaning affects data processing by reducing data volume and complexity

[15]. In this stroke dataset, 0.0958% of the sex data is missing. The researcher then fills in the sex data so that this data can be modeled well. Through this process, researchers can gain additional benefits in taxonomic modeling.

C. Train-Test Split

Train-test split is a simple scenario for testing a classification model by dividing the portion of the dataset into two parts, namely the training data and the test data [16,17]. The goal is to test the classification model under different dataset circumstances [18]. In this study, researchers ran five scenarios, namely: 50% training data and 50% test data, 60% training data and 40% test data, 70% training data and 30% test data, 80% training data and 20% test data, 90% training data and 10% test data. Fig. 3 shows the distribution of the data used in the training model, while Fig. 4 shows the distribution of the data used in the test model.

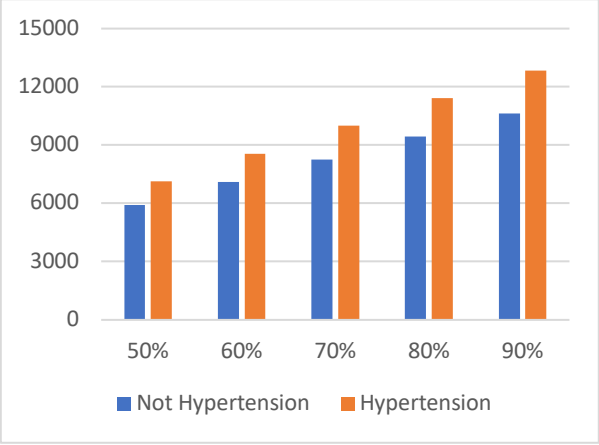

Figure 3. Distribution of Training Data

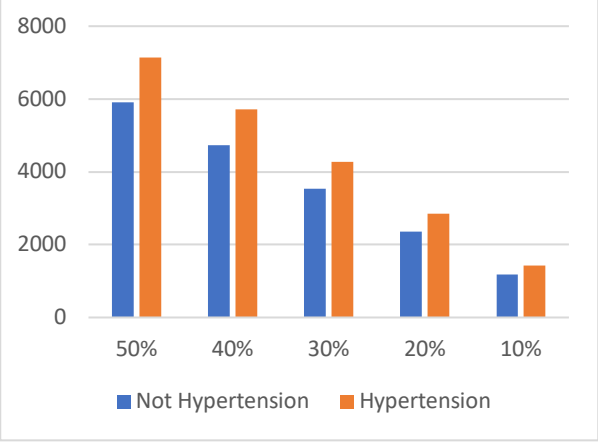

Figure 4. Distribution of Testing Data

D. Classification

Classification is one of the primary tasks in data mining [19]. Classification is a technique in data mining to group data based on the data attached to the sample data. In the classification, there are two stages: the training stage and the testing stage [20]. The training stage is a process when the algorithm builds a classification model from training data whose class label is already known. The testing phase is a step to apply the model to the test data so that the class can be known [21]. The benefit of data classification is to organize data so that data becomes easier to process into information because it is well organized according to each category.

E. Passive-Aggressive

Passive-Aggressive algorithms are a group of algorithms for large-scale learning [22]. It is conceptually similar in that it does not require a learning rate. But unlike Perceptron, it contains C regularization parameters. Passive-Aggressive Classifier can be used with loss=hinge (PA-I) or loss=hinge_squared (PA-II) for classification [10]. Passive-Aggressive Regressor can be used for regression with loss = epsilon_insensitive (PA-I) or loss = squared_epsilon_insensitive (PA-II) [10]. In this paper, researchers used PA-I and PA-II in the classification method.

F. Evaluation

Researchers use accuracy, precision, recall and specificity to evaluate models from machine learning. Accuracy, precision, recall and specificity represent predictions and the actual conditions of the data generated by the algorithm [23-25]. Accuracy is the comparison between correct predictions and overall predictions. Equation 1 shows the accuracy's formula.

$$\text{Accuracy} = \frac{\text{The number of predictions are correct}}{\text{Number of testing/training data}} \quad (1)$$

Precision is the ratio of correct positive predictions to all positives. Equation 2 shows the precision's formula.

$$\text{Precision} = \frac{\text{Number of correctly predicted hypertension data}}{\text{Overall predictive outcome of hypertension}} \quad (2)$$

Recall is the ratio of correct positive predictions compared to all positive correct data. Equation 3 shows the recall's formula.

$$\text{Recall} = \frac{\text{Number of correctly predicted hypertension data}}{\text{Overall correct data on hypertension}} \quad (3)$$

Specificity is the correctness of predicting a negative compared to the overall negative data. That is, the number of correctly predicted non-hypertension data

divided by data of non-hypertension. Equation 4 shows the specificity's formula.

$$\text{Specificity} = \frac{\text{Accurate prediction of non-hypertension data}}{\text{The number of testing data is not hypertension}} \quad (4)$$

3. Analysis and Discussion

A. Application of Passive-Aggressive

In applying the algorithm, the researcher uses the PA-I and PA-II methods which are set using the default parameters of Passive-Aggressive. There are 14 parameters used for this research. Table 2 shows the parameters used in this study.

TABLE II. PASSIVE-AGGRESSIVE PARAMETERS

Parameters	Value
C	1.0
Fit_intercept	True
Max_iter	1000
Tol	1e-3
Early_stopping	False
Validation_fraction	0.1
n_iter_no_change	5
Shuffle	True
Verbose	0
N-jobs	None
Random_state	None
Warm_start	False
Class_weight	None
Average	False

B. Accuracy Performance Evaluation

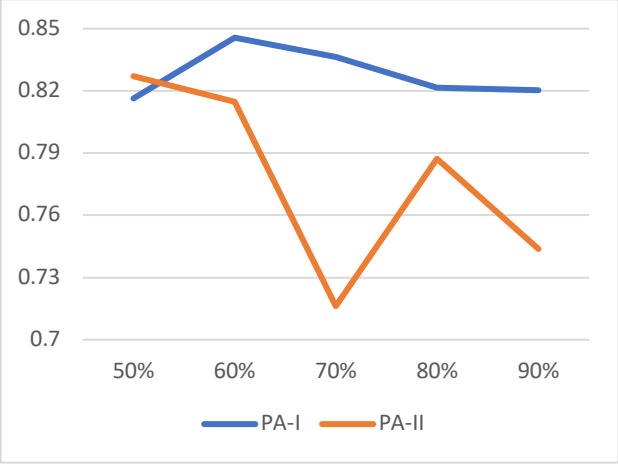

Figure 5. Training Accuracy

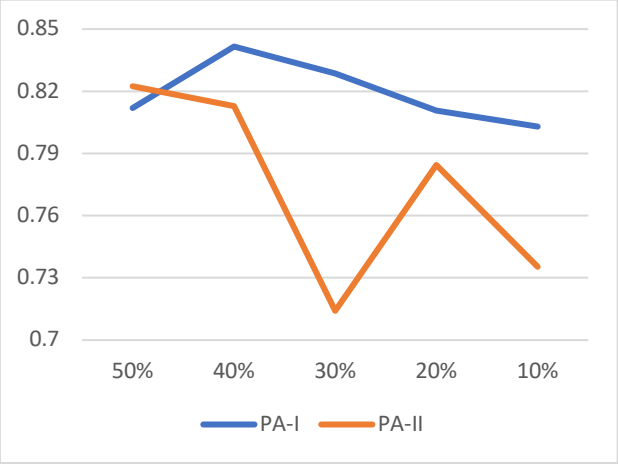

Figure 6. Testing Accuracy

Fig. 5 and 6 above show the accuracy of the training and testing processes on the dataset. The PA-I method shows stable training and testing accuracy in Fig. 5 and 6. The PA-I also does not experience that the model is overfitting. This method does not experience a situation where the data used for training is the best. So during the testing process using different data, the accuracy does not differ much between the training and testing processes. The PA-II method also does not experience overfitting. However, in the PA-II, the model accuracy shows instability because the training and testing accuracy fluctuates significantly. In terms of accuracy, the PA-I model is the best model for diagnosing hypertension.

C. Precision Performance Evaluation

Figure 7. Precision Accuracy

Fig. 7 above proves that in the precision side, PAI-I is still superior to PA-II. Even though the precision values do not show stability, the precision values of the PA-II method can still catch up with the precision values of the PA-I. It

also proves that the probability of a correct diagnosis is still greater in PA-I than in PA-II.

D. Recall Performance Evaluation

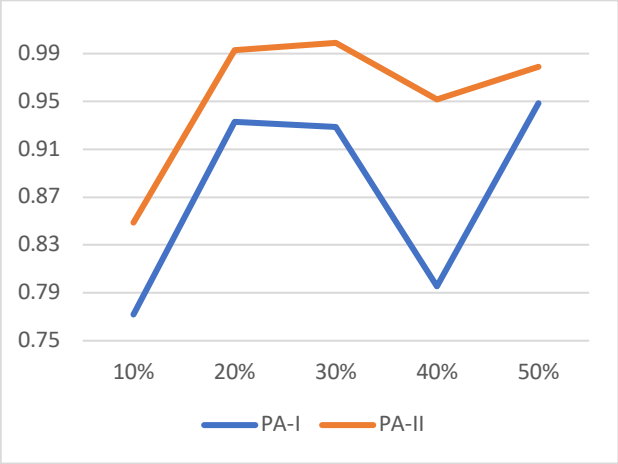

Figure 8. Recall Accuracy

Fig. 8 above shows in the recall side, PAI-II is superior to PA-I. Even though the performance value is not very stable, PA-II shows an almost perfect value when testing data of 30% of the dataset. The highest PA-II value is 0.999, which means that this model has a good prediction ratio for correct hypertension diagnoses.

E. Specificity Performance Evaluation

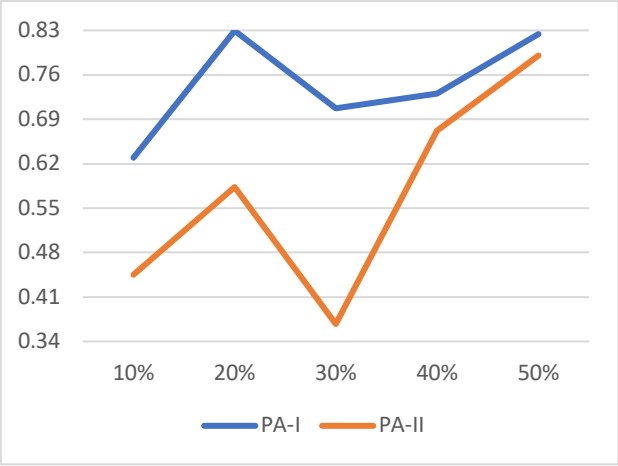

Figure 9. Specificity Accuracy

Fig. 9 above shows that in terms of specificity, PAI-I is again superior to PA-II. In the performance evaluation with specificity, PA-II shows extraordinary instability. It is evidenced by the decline in the specificity value to 0.36748. In this performance analysis, PA-I can predict patients who do not have hypertension, while PA-II is considered poor in predicting patients who do not have hypertension.

4. CONCLUSION

This study concludes that in general, the Passive-Aggressive Algorithm with the PA-I method works well in the hypertension diagnosis dataset, which is indicated by better model performance than the PA-II method. The highest accuracy value of Passive-Aggressive reached 84.15% with a comparison of training and test data of 60:40, while the highest precision and specificity values were 84.86 and 82.9% with a comparison of 80:20. But in terms of recall performance, the PA-I method is not better than PA-II. PA-II achieves an optimal value of 99% at a comparison of training and test data of 70:30.

This study shows that this algorithm works well because the algorithm does not experience overfitting, which means that the accuracy of the training and testing models is balanced and only has slight differences.

In future research, we suggest trying the Passive-Aggressive algorithm on this dataset model by adding feature selection or dimension reduction so that the classification process can run faster and better to increase the efficiency of the algorithm model.

REFERENCES

[1] M. Nour, and K. Polat, (2020). Automatic classification of hypertension types based on personal features by machine learning algorithms. Math. Probs. in Eng, pp 1-13.

[2] E. Susanti, and D. Anggara, (2021). Pengaruh Slow Deep Breathing Terhadap Tekanan Darah Sistolik Dan Diastolik Pada Penderita Hipertensi: Literature Review. JKM, vol. 1, no. 1, pp. 8-16.

[3] P. Dewi, P. Purwono, and S. D. Kurniawan, (2022). Pemanfaatan Teknologi Machine Learning pada Klasifikasi Jenis Hipertensi Berdasarkan Fitur Pribadi. Smart Comp, vol.11, no.3, pp. 377-387.

[4] A. C. Telaumbanua and Y. Rahayu, (2017). Penyuluhan Dan Edukasi Tentang Penyakit Hipertensi. J. Abdimas Saintika, vol. 3, no. 1, p. 119.

[5] S. Umemura, H. Arima, S. Arima, K. Asayama, Y. Dohi, Y. Hirooka, and N. Hirawa, (2019). The Japanese Society of Hypertension guidelines for the management of hypertension (JSH 2019). Hypertension Research, vol. 42, no.9, pp. 1235-1481.

[6] Y. Nursakinah and A. Handayani, (2021). Faktor-Faktor Risiko Hipertensi Diastolik Pada Usia Dewasa Muda. J. Pandu Husada, vol. 2, no. 1, p. 21.

[7] G. Lippi, J. Wong, and B. M. Henry, (2020). Hypertension and its severity or mortality in Coronavirus Disease 2019 (COVID-19): a pooled analysis. Pol Arch Intern Med, vol. 130, no.4, pp. 304-309.

[8] Y. T. G. Arum, (2019). Hipertensi pada Penduduk Usia Produktif (15-64 Tahun). Hig J. Pub. Heal. Res. Dev., vol. 3, no. 3, pp. 84–94.

[9] A. Syntya, (2021). Hypertension and heart disease: literature review. J. Ilm. Permas J. Ilm. STIKES Kendal, vol. 11, no. 4, pp. 541–550.

[10] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer, (2006). Online passive aggressive algorithms. JMLR, vol. 7, pp. 551 - 585.

[11] M. Chugh, D. Arora, M. Singh, M. Shobhit, and M. Ronak, (2021). Media Manipulation Detection System Using Passive Aggressive. IJIRCST, pp. 2347-5552.

[12] S. A. Krishnan, A. N. Sabu, P. P. Sajan, and A. L. Sreedeep, (2021). SQL Injection Detection Using Machine Learning. Rev Geint-Gest. Inov E Tecno, vol. 11, no.3, pp. 300-310.

[13] CDC. (2015). Diabetes, Hypertension and Stroke Prediction. Kaggle, https://www.kaggle.com/datasets/prosperchuks/health-dataset?select=hypertension_data.csv [Accessed in 24 Dec 2022].

[14] Q. A'yuniyah, E. Tasia, N. Nazira, P. F. Pratama, M. R. Anugrah, J. Adhiva, and M. Mustakim, (2022). Implementasi Algoritma Naïve Bayes Classifier (NBC) untuk Klasifikasi Penyakit Ginjal Kronik. JSON, vol. 4, no. 1, pp. 72-76.

[15] Y. Zhang, M. Safdar, J. Xie, J. Li, M. Sage, and Y. F. Zhao, (2022). A systematic review on data of additive manufacturing for machine learning applications: the data quality, type, preprocessing, and management. J of Intel Manu, pp. 1-36.

[16] R. Y. Choi, A. S. Coyner, J. Kalpathy-Cramer, M. F. Chiang, and J. P. Campbell, (2020). Introduction to machine learning, neural networks, and deep learning. Trans Vis Sci & Tech, vol. 9, no. 2, p. 14.

[17] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, (2019). Machine learning algorithm validation with a limited sample size. PloS one, vol. 14, no. 11, p. e0224365.

[18] M. Aamir, and S. M. A. Zaidi, (2021). Clustering based semi-supervised machine learning for DDoS attack classification. JKSUCIS, vol. 33, no. 4, pp. 436-446.

[19] H. Fernando, and J. Marshall, (2020). What lies beneath: Material classification for autonomous excavators using proprioceptive force sensing and machine learning. Auto in Const, vol. 119, p. 103374.

[20] C. A. Ramezan, T. A. Warner, and A. E. Maxwell, (2019). Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification. Rem Sens, vol. 11, no. 2, p. 185.

[21] M. A. Moreno-Ibarra, Y. Villuendas-Rey, M. D. Lytras, C. Yáñez-Márquez, and J. C. Salgado-Ramírez, (2021). Classification of diseases using machine learning algorithms: A comparative study. Mathematics, vol. 9, no. 15, p. 1817.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, and E. Duchesnay, (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, vol. 12, pp. 2825-2830.

[23] M. H. Ariansyah, S. Winarno, and A. Salam, (2023). STB Sentiment Analysis Classification Multiclass Modeling Using Calibrated Classifier With SGDC Tuning As Basis and Sigmoid Method. IJCIS, vol. 4, no. 1, pp. 1-7.

[24] J. Xu, Y. Zhang, and D. Miao, (2020). Three-way confusion matrix for classification: A measure driven view. Information sciences, vol. 507, pp. 772-794.

[25] C. Fatichah, and D. Purwitasari, (2017). Deteksi Gempa Berdasarkan Data Twitter Menggunakan Decision Tree, Random Forest, dan SVM. JTITS, vol. 6, no. 1, pp. A159-A162.