Journal of Information System Exploration and Research Vol. 3, No. 1, January 2025, pp. 1-10

Journal of Information System Exploration and Research

https://shmpublisher.com/index.php/joiser

p-ISSN 2964-1160 | e-ISSN 2963-6361

The Asthma Classification Using an Adaptive Boosting Model with SVM-SMOTE Sampling

Ahmad Ubai Dullah[1*], Aditya Yoga Darmawan[2], Putri Utami[3], Jumanto[4]

[1,2,3,4]*Department of Computer Science, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Indonesia*

DOI: https://doi.org/10.52465/joiser.v3i1.486

Received 14 October 2024; Accepted 28 January 2025; Available online 28 January 2025

Article Info

Keywords:
AdaBoost;
SVM-SMOTE;
Asthma classification;
Symptoms

Abstract

Asthma is a disease that affects the human respiratory tract, characterized by inflammation and narrowing of the respiratory tract such as wheezing, coughing, and shortness of breath. The causes of asthma can come from genetics, lifestyle, and a bad environment. Diagnosis made to asthma patients is very influential on the severity and treatment carried out. However, the diagnosis process may not be able to precisely determine asthma patients because the diagnosis is influenced by the classification of asthma based on the symptoms that appear. Therefore, this study proposes an asthma disease classification model that is optimized using a sampling method to balance the data. The proposed classification model uses the Adaptive Boosting algorithm with a sampling technique using SVM-SMOTE to help balance the data. The results obtained from the experiment achieved an accuracy of 98.60%. This result shows that the proposed model is more accurate and optimal in performing classification when compared to previous research.

1. Introduction

Asthma is a chronic respiratory disease characterized by airway inflammation and narrowing such as wheezing, coughing, and shortness of breath. The disease exhibits heterogeneity, as different phenotypes and endotypes impact individual severity and treatment response [1]. With the increasing incidence of asthma globally, it appears that a combination of lifestyle, environmental, and genetic variables contribute significantly to the development of this disease [2], [3]. More sophisticated methods are needed for more accurate diagnosis and proper management of asthma although current methods such as spirometry and clinical assessment, have been widely used. However, these approaches often cannot optimally address the complexity of asthma clinical variations [4].

Shortness of breath, coughing, and wheezing are some of the symptoms of asthma that often interfere with patients' daily activities. In addition, some things such as allergies, air pollution, and intense physical exercise, can cause asthma episodes [5]. People who have asthma should be aware of

[*] ***Corresponding Author:***

Ahmad Ubai Dullah,
Computer Science Department, Faculty of Mathematics and Natural Sciences,
Universitas Negeri Semarang,
Semarang, Indonesia.
Email : Ubaid@students.unnes.ac.id

these triggers and take the necessary precautions, such as using an inhaler or staying away from dangerous situations [6]. People with asthma can live more active and productive lives if their symptoms are well managed, but they must always be vigilant when their symptoms flare up [7]. Successful treatment strategies can be developed with the help of early diagnosis so that patients can experience fewer attacks and a higher standard of living. Support from friends and family is also very important for people with asthma to maintain motivation and lead a healthy life [8].

Previous research has been conducted to assist in the classification of asthma diseases to improve the diagnosis made. The application of machine learning in this study showed promising results, with algorithms capable of analyzing patient data and identifying relevant patterns. Like the previous research, this research model uses a multiclass ensemble model with a Support Vector Machine (SVM), Tunnel Boring Machine (TBM), K-Nearest Neighbor (KNN), Gaussian Naive Bayes (GNB), eXtreme Gradient Boosting (XGB), Random Forest (RF) [9]. The results of this study are recall 87%, specificity 94%, precision 89%, accuracy 92%. The results obtained show that this approach is effective in improving the accuracy of asthma diagnosis, as well as providing insight into the factors that affect the patient's condition. However, the results obtained can still be improved so that the diagnosis made can improve and be more accurate in determining the patient's treatment process.

Therefore, this study was conducted to find a more accurate and effective research model for asthma classification. The more accurate the classification, the better the diagnosis results. This research uses a machine learning model for the learning model. The model used is AdaBoost, which has the advantage of being able to train weak classification models by modifying weights based on classification errors, resulting in a strong model [10], [11]. As a result, it is expected that this model will perform better in determining asthma severity and assist medical professionals in helping patients get the right treatment.

2. Literature Review

This research was conducted to find an accurate classification model for classifying asthma disease. The topic in this research has been researched before, but using different classification methods so that it can bring new and useful innovations and knowledge. Previous research was conducted by Zne Jung Lee et al. [8] in 2024, using the extreme gradient boosting classification model. Before being classified, the data is processed first with feature selection to find important features and augmentation data to improve the dataset. Augmentation data is used to create sample data and to increase the size of the training data so that the data becomes balanced. The boost classification model that has been enhanced with GAN (Generative Adversarial Networks) has good classification results in its performance. The accuracy obtained is 94.03% and the AUC is 0.929. Based on these results, there is still a possibility to develop a model that can improve the diagnosis results to be even higher. This study will improve accuracy with a different approach.

There is research conducted by Piyush Bhardwaj et al. [12] in 2023, using several machine learning models for asthma classification. The machine learning models used include random forest, support vector machine, extreme gradient boosting, extra tree classifier, logistic regression, and adaptive boosting. The model is a good algorithm for classification. There are nine features in the data used as parameters. The best results obtained in the study include 77.8% (SVM) and 76.2% (LR). Although SVM and LR show good performance, there is still a possibility to optimize the performance of the model by using a larger dataset and better data sampling techniques.

The research conducted by Syed Zohaib Hassan Naqvi [13] in 2020, the research was conducted using a dataset containing 300 data from the Pakistan Institute of Medical Sciences. The data is processed with several pre-processing methods such as feature extraction, segmentation, and normalization. Then trained with a research model that uses the Support vector machine model and several other models as a comparison. The comparison models include LD, KNB, and GNB. The results obtained reached 96.70% accuracy with the SVM Linear model.

This research was conducted with the consideration of developing the best machine-learning model for asthma classification from existing research. The methods and data used are quite different in this study. The dataset used is patient examination data diagnosed with asthma or not, with several features used as parameters in classification. The research model used is an adaptive boosting [14] model with data sampling using SVM-SMOTE [15] to help balance the data. The Adaboost model has been used in previous studies but the difference is the use of SVM-SMOTE as a method for sampling.

3. Method

This research uses a machine learning model, AdaBoost. The ability of this model is that it can be used to train weak classification models by modifying the weights based on classification errors, resulting in a strong model. As a result, it is expected that this model has better performance in determining the severity of asthma and assisting medical personnel in helping patients get the right treatment.

The method used in this research is the implementation of data processing and model building for the evaluation of the research model. The flow in the research can be seen in the flowchart Figure 1.

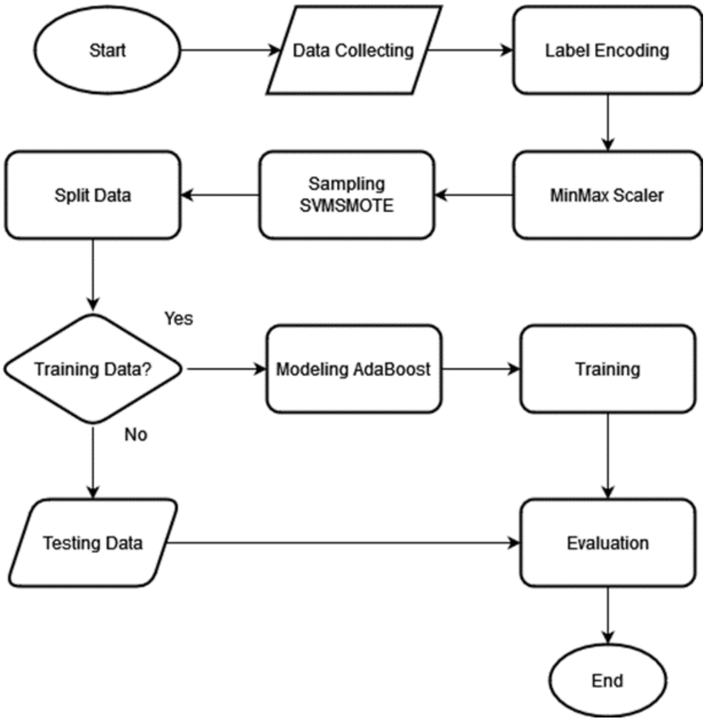

Figure 1. Flowchart Proposed Method

3. 1. Data Collecting

The dataset used is based on the Kaggle public dataset website. The dataset used consists of 29 columns and 2,392 data. The data contains patient information data, lifestyle factors, environmental and allergy factors, medical history, symptoms, and diagnosis.

3. 2. Label Encoding

Label encoding is the process of converting unique values in a column into a numeric data type. Because some machine learning models can only process numeric type data. In this study, it is used to convert the responsible doctor column into a numeric data type so that the research model can process the data more optimally and the results obtained are also better.

3. 3. Min-Max Scaler

Minmax scaler is a data normalization method used to scale features in the range of 0 and 1. In this study minmax scaler is used in column X data. Because the data in column X has different range values. The calculation in the min-max scaler can be seen in the formula (1):

$$X_{new} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

Explanation:
X = The original value
X_{new} = The new Normalized value
X_{min} = The Minimum value of the feature
X_{max} = The Maximum value of the feature

3. 4. Sampling Using SVM-SMOTE

SMOTE (Synthetic Minority Over-sampling Technique) [16]–[19] is a method to increase the amount of data by making samples from the class that has the lowest data with synthetic data. The number of data samples will be increased according to the class that has the most data so that the data becomes balanced. SVM-SMOTE [20] is a variant of SMOTE that adopts the principle of Support Vector Machine (SVM) to overcome class imbalance in the data. The advantage of SVM-SMOTE is that it is smarter in determining the lowest sample that needs to be sampled. This reduces overfitting because it only adds data to minority classes that are difficult to separate.

In this study, SVM-SMOTE [21] is used to balance the Diagnosis column data (class label) because the distribution of the data is very unbalanced. There are 2268 data with data label 0 and 124 data with data label 1. So, it is necessary to do oversampling by using SVM-SMOTE to balance the data. The distribution of data in the diagnosis column data is seen in Figure 2.

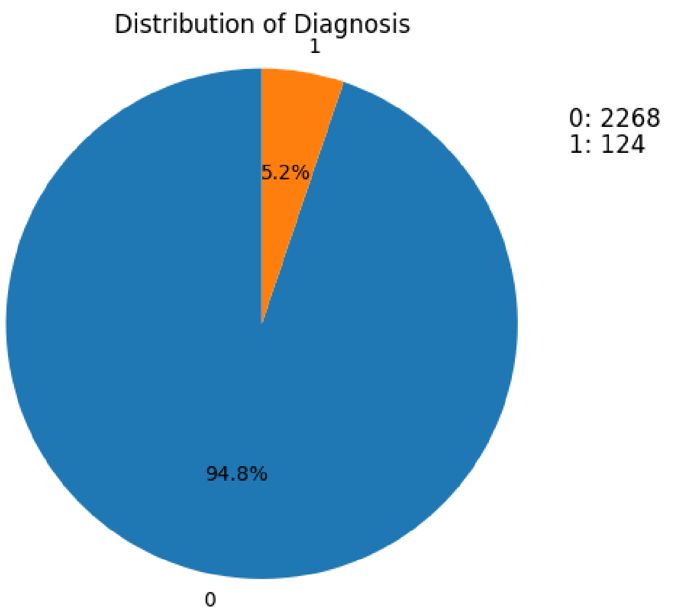

Figure 2. Distribution of Diagnosis Column

3. 5. Modeling

This study uses the Adaptive Boosting (AdaBoost) machine learning classification model. One of the machine learning algorithms, called Adaboost [22], [23], comes from the ensemble library. It is used to increase accuracy by combining a few weak models into a strong model. A weak model Conversely, a strong model has better performance. An Adaboost focuses on data points that are closely predicted by the model estimator (weak model). The advantage of this Adaboost model is that it is very easy to build, effective in increasing accuracy, and does not overfit [24]. However, it has limitations due to its sensitivity to outliers and its slow computing speed for large data.

The weak model used as an estimator in this study is the random forest algorithm [25]. The random forest algorithm is one of the machine learning algorithms that can be used for classification and regression. The way random forest works is to train several decision trees independently to make predictions. There are several parameters used in the Adaboost classifier [26] such as n estimators to determine the number of weak models to be combined. Then there is a learning rate to regulate the effect of weak models on the final result. In this study using n estimators as many as 100 and a learning rate of 0.1. More details can be seen in Figure 3 which is an overview of the Adaboost model.

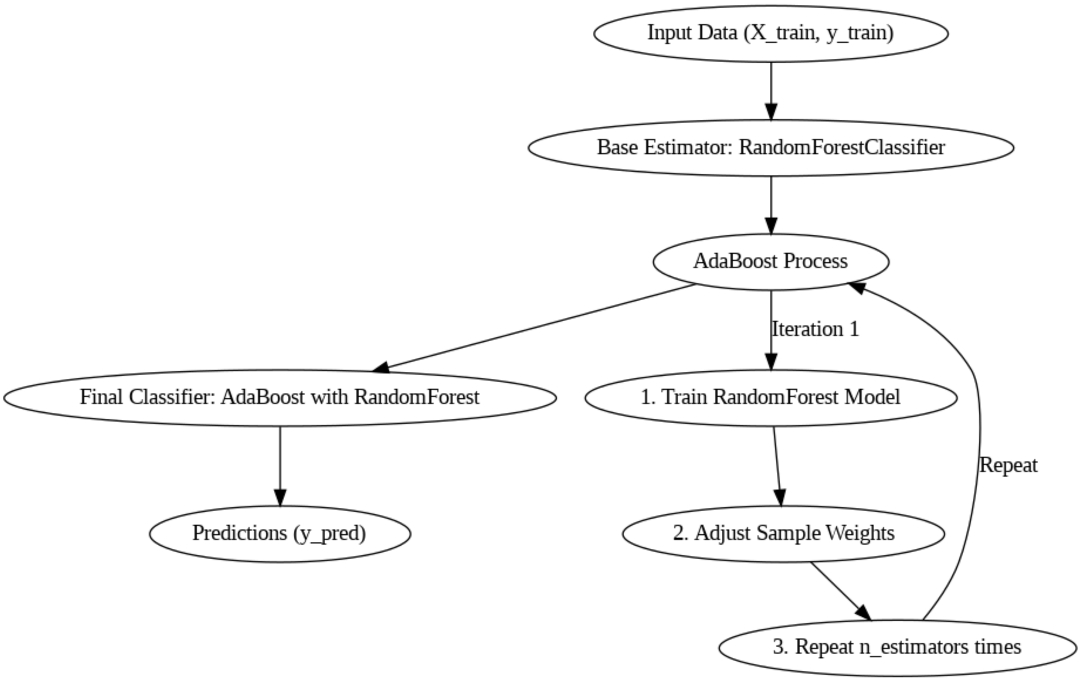

Figure 3. Model Adaptive Boosting

3. 6. Evaluation

In this study, the model evaluation that is used consists of two primary methods, which are the classification report and the confusion matrix [27], to determine the model's overall performance. The confusion matrix provides information in detail on the number of accurate and reasonable predictions for each class, while the classification report presents important metrics such as accuracy, F1 score, recall, and precision. Accuracy tends to decrease the true prediction value from all predictions, whereas precision indicates a few positive predictions that are quite relevant. Recall [28] measures how well the model can interpret all available positive data, and the F1-score [29] is the average of accuracy and recall, providing a mutually reinforcing effect. Combining these metrics yields a comprehensive evaluation of the model's classification performance. In addition, the AUC [30] (Area Under the Curve) of the ROC (Receiver Operating Characteristic) is also used to assess the model's ability to distinguish between positive and negative data points; a higher threshold indicates a more favorable discriminant capacity. Combining these metrics yields a comprehensive evaluation of the model's classification performance.

In analyzing the results using the confusion matrix, there is a mathematical formula that can be seen in formulas (2), (3), and (4):

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \times 100 \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F1\ Score = 2 \times \frac{precision \ \times recall}{precision + recall} \tag{5}$$

Description:

TN (True Negative): True data that is categorized as negative
TP (True Positive): True data that is categorized as positive.
FP (False Positive): False data that is categorized as positive
FN (False Negative): False data that is categorized as negative

4. Results and Discussion

4.1. Results

The results of this study display the results of data processing as pre-processing data before being trained by the research model. Pre-processing includes label encoding, min-max scaler, and sampling with SVM-SMOTE. Then the results of training and testing the model in the form of a classification report and confusion matrix.

a. Result of Label Encoding

In this study, label encoding is used to change the value in a column that has an object data type, namely the Doctor in Charge column. This column contains data about doctor payments. The results of label encoding can be seen in Table 1.

Table 1. Result label encoding

Before Label Encoding	After Label Encoding
dr_Confid	0
dr_Confid	0
dr_Confid	0
dr_Confid	0

The data in the Doctor in Charge column containing the dr_config value is changed to a value of 0. All the data has been converted to numeric form.

b. Result Min-Max Scaler

Data normalization in this study using a min-max scaler. Data processed with a min-max scaler is data that includes X data (independent variables). X data is normalized to facilitate the model training process. Examples of data normalization results can be seen in Table 2.

Table 2. Result normalization data using Min-Max scaler

Before normalization data		After normalization data	
Age	BMI	Age	BMI
63	15.848744	0.783784	0.032738
26	22.757042	0. 283784	0.309582
57	18.395396	0. 702703	0.134793
40	38.515278	0.472973	0.941078

c. Result Sampling using SVM-SMOTE

Sampling using SVM-SMOTE aims to balance the data because based on the label column the distribution of the data is less balanced so it needs to be sampled with SVM-SMOTE. The results of sampling with SVM-SMOTE can be seen in Figure 4.

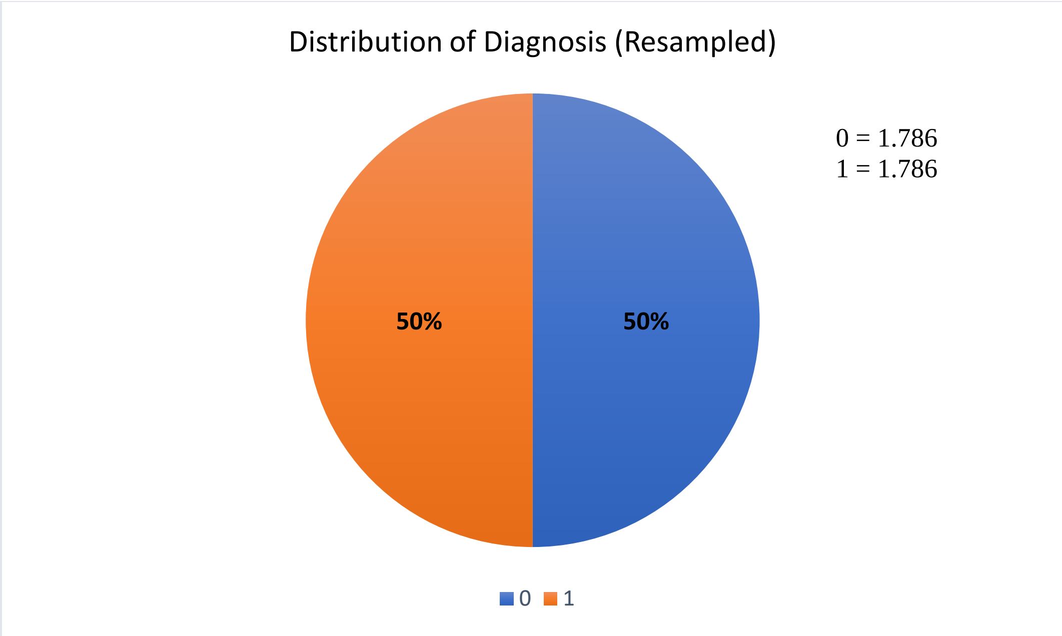

Figure 4. Result Sampling Using SVM-SMOTE

After the process of balancing the sample data, with the comparison of data distribution as shown in Figure 4, the number of samples becomes 3,572 with a class percentage of 50:50. Thus, the performance of the model will be optimal for determining data classification.

d. Result model

The results obtained from training and testing the adaptive boosting model will be in the form of classification reports and confusion matrix. This research uses a ratio of 80:20 for training and testing data. The confusion matrix results can be seen in Figure 5.

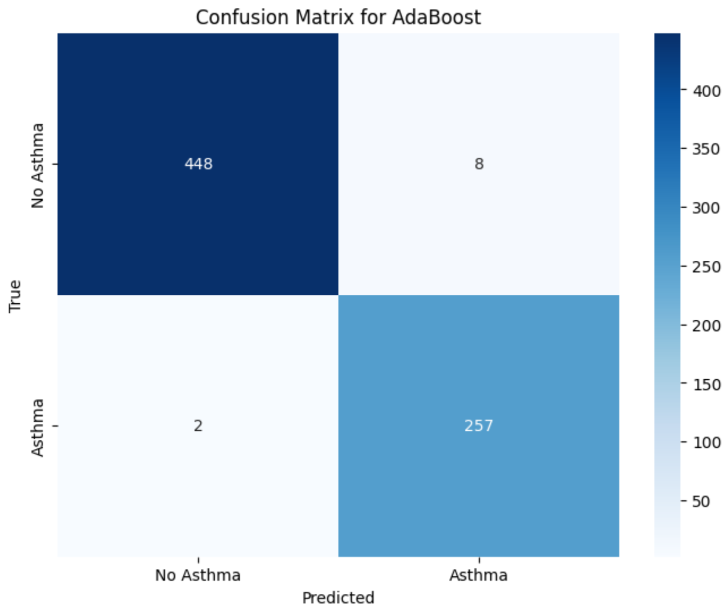

Figure 5. Confusion Matrix Model AdaBoost

Based on the results of the confusion matrix above, the adaptive boosting model can classify asthma well. Asthma prediction results that match the label are 257 data with an error of 2 data. The classification of non-asthma labels according to the label is 448 data with a prediction error of 8 data. This shows the performance of the model is very good. The evaluation results can be seen in Table 3 about the classification report.

Table 3. Classification report

	Precision	Recall	F1-score	Support
No Asthma	1.00	0.98	0.99	456
Asthma	0.97	0.99	0.98	295
Accuracy			0.99	715
Macro avg	0.98	0.98	0.98	715
Weighted avg	0.99	0.99	0.99	715

The Classification Report results show that the classification model performed very well in detecting both Asthma and No Asthma classes, with precision, recall, and F1-score above 0.97 for both classes, respectively. For No Asthma, the recall of 0.98 indicated that 98% of the “Asthma” cases were correctly detected. For Asthma, the precision and recall are also very good, 0.97 and 0.99 respectively. The weighted average precision, recall, and F1-score are all at 0.99, with an overall accuracy of 0.99 or 0.9860 (98.60%), indicating the model is very reliable in predicting Asthma and No Asthma conditions.

4.2. Discussion

In this study, the results obtained by the Adaptive Boosting research model with data balancing pre-processing using SVM-SMOTE will be compared with previously existing research models. The results of this comparison will show the performance of the proposed model when compared to existing models. Model comparison can be seen in Table 4.

Table 4. The Comparison with previous model

Author	Year	Research Model	Result
Roghaye Khasha [9]	2021	Ensemble Learning, Support Vector Machine, Tree-Based Model, Multinominal Logistic Regression, K-Near Neighbor, Gaussian Naïve Bayes, Extreme Gradient Boosting, Random Forest.	Accuracy: 92 %
Zne Jung Lee [8]	2024	Extreme Gradient Boosting	Accuracy: 92.03%
Piyush Bhardwaj[12]	2023	Random Forest, Support Vector Machine, Extreme Gradient Boosting, Extra Tree Classifier, Logistic Regression, And Adaptive Boosting	Accuracy LR: 76.2% Accuracy SVM: 77.8%
Proposed Method	**2024**	**Adaptive Boosting with SVM-SMOTE**	**Accuracy: 98,60%**

Based on the comparison table above, the proposed model shows better and optimal results when compared to previous research models. The proposed model has also been used in previous research. Namely in research by Piyush Bhardwaj in 2023, the highest accuracy results were achieved by the support vector machine with an accuracy of 77.8%. Therefore, the proposed model uses SVM-SMOTE sampling in pre-processing to balance the data so that the results achieved are more accurate and optimal reaching an accuracy of 98.60%. For this reason, the proposed research model is better and more accurate in asthma classification.

5. Conclusion

Asthma is a disease that affects the human airways and is characterized by inflammation and narrowing of the airways, such as wheezing, coughing, and shortness of breath. The causes of asthma can be genetic, lifestyle, or environmental. The diagnosis made to asthma patients is very influential on the severity and treatment carried out. However, the diagnosis process may not be able to accurately determine asthma patients because the diagnosis is influenced by the classification of asthma based on the symptoms that appear. Therefore, a classification model is proposed to classify asthma based on its symptoms. The effectiveness of the AdaBoost model reinforced with the SVM-SMOTE technique for asthma classification was successfully demonstrated in this study with high levels of accuracy, precision, recall, and F1-score. The sampling method used in this study can handle data imbalance so that the classification results become more accurate. Compared to models in previous studies, the use of AdaBoost and SVM-SMOTE is superior with an overall accuracy of 98.60%. For future research, it is better to try using other models or combining models using a larger and more diverse dataset so that it can have a more significant and more accurate impact in the medical world, especially in asthma classification.

References

[1] H. C. Kuo, B. S. Lin, Y. Di Wang, dan B. S. Lin, "Development of Automatic Wheeze Detection Algorithm for Children with Asthma," *IEEE Access*, vol. 9, hal. 126882–126890, 2021, doi: 10.1109/ACCESS.2021.3111507.

[2] Y. Wu, J. Zhang, Y. Chen, J. Wang, W. Shi, dan Q. Zhang, "Ubi-Asthma: Toward Ubiquitous Asthma Detection Using the Smartwatch," *IEEE Internet Things J.*, vol. 10, no. 13, hal. 11576–11587, 2023, doi: 10.1109/JIOT.2023.3243188.

[3] D. Bagnasco *et al.*, "Severe asthma: One disease and multiple definitions," *World Allergy Organ. J.*, vol. 14, no. 11, 2021, doi: 10.1016/j.waojou.2021.100606.

[4] M. A. Awal *et al.*, "An Early Detection of Asthma Using BOMLA Detector," *IEEE Access*, vol. 9, hal. 58403–58420, 2021, doi: 10.1109/ACCESS.2021.3073086.

[5] R. Haba, G. Singer, S. Naftali, M. R. Kramer, dan A. Ratnovsky, "A remote and personalised novel approach for monitoring asthma severity levels from EEG signals utilizing classification algorithms," *Expert Syst. Appl.*, vol. 223, no. February, hal. 119799, 2023, doi: 10.1016/j.eswa.2023.119799.

[6] J. L. M. Amaral, A. G. Sancho, A. C. D. Faria, A. J. Lopes, dan P. L. Melo, "Differential diagnosis of asthma and restrictive respiratory diseases by combining forced oscillation measurements, machine learning and neuro-fuzzy classifiers," *Med. Biol. Eng. Comput.*, vol. 58, no. 10, hal. 2455–2473, 2020, doi: 10.1007/s11517-020-02240-7.

[7] D. Aulia, R. Sarno, S. C. Hidayati, dan M. Rivai, "Optimization of the Electronic Nose Sensor Array for Asthma Detection Based on Genetic Algorithm," *IEEE Access*, vol. 11, no. June, hal. 74924–74935, 2023, doi: 10.1109/ACCESS.2023.3291451.

[8] Z. J. Lee, M. R. Yang, dan B. J. Hwang, "A Sustainable Approach to Asthma Diagnosis: Classification with Data Augmentation, Feature Selection, and Boosting Algorithm," *Diagnostics*, vol. 14, no. 7, 2024, doi: 10.3390/diagnostics14070723.

[9] R. Khasha, M. M. Sepehri, dan N. Taherkhani, "Detecting asthma control level using feature-based time series classification," *Appl. Soft Comput.*, vol. 111, hal. 107694, 2021, doi: 10.1016/j.asoc.2021.107694.

[10] S. Pang, Y. Zhang, M. Ding, X. Wang, dan X. Xie, "A Deep Model for Lung Cancer Type Identification by Densely Connected Convolutional Networks and Adaptive Boosting," *IEEE Access*, vol. 8, hal. 4799–4805, 2020, doi: 10.1109/ACCESS.2019.2962862.

[11] E. Rizkyani, N. Aliffiyanti Iskandar, dan N. Chamidah, "Klasifikasi dalam Mendeteksi Penyakit Kanker Payudara dengan Menggunakan Metode Random Forest dan Adaboost," *Semin. Nas. Mhs. Ilmu Komput. dan Apl. Jakarta-Indonesia*, no. September, hal. 335–343, 2021, [Daring]. Tersedia pada: www.kaggle.com

[12] P. Bhardwaj, A. Tyagi, S. Tyagi, J. Antão, dan Q. Deng, "Machine learning model for classification of predominantly allergic and non-allergic asthma among preschool children with asthma hospitalization," *J. Asthma*, vol. 60, no. 3, hal. 487–495, 2023, doi: 10.1080/02770903.2022.2059763.

[13] S. Z. H. Naqvi, M. Arooj, S. Aziz, M. U. Khan, M. A. Choudhary, dan M. N. Ul Hassan, "Spectral Analysis of Lungs sounds for Classification of Asthma and Pneumonia Wheezing," *2nd Int. Conf. Electr. Commun. Comput. Eng. ICECCE 2020*, no. June, 2020, doi: 10.1109/ICECCE49384.2020.9179417.

[14] S. V. Razavi-termeh, A. Sadeghi-niaraki, dan S. M. Choi, “Spatial modeling of asthma-prone areas using remote sensing and ensemble machine learning algorithms,” *Remote Sens.*, vol. 13, no. 16, hal. 1–22, 2021, doi: 10.3390/rs13163222.

[15] S. S. Bagui, D. Mink, S. C. Bagui, dan S. Subramaniam, “Determining Resampling Ratios Using BSMOTE and SVM-SMOTE for Identifying Rare Attacks in Imbalanced Cybersecurity Data,” *Computers*, vol. 12, no. 10, 2023, doi: 10.3390/computers12100204.

[16] T. H. S. Li, H. J. Chiu, dan P. H. Kuo, “Hepatitis C Virus Detection Model by Using Random Forest, Logistic-Regression and ABC Algorithm,” *IEEE Access*, vol. 10, no. August, hal. 91045–91058, 2022, doi: 10.1109/ACCESS.2022.3202295.

[17] H. F. El-Sofany, “Predicting Heart Diseases Using Machine Learning and Different Data Classification Techniques,” *IEEE Access*, vol. 12, no. August, hal. 106146–106160, 2024, doi: 10.1109/ACCESS.2024.3437181.

[18] R. Rofik, F. Artiyani, dan D. A. A. Pertiwi, “Breast Cancer Diagnosis Utilizing Artificial Neural Network (ANN) Algorithm for Integrating Multi-Omics Data and Clinical Features,” *J. Inf. Syst. Explor. Res.*, vol. 2, no. 2, hal. 61–68, 2024, doi: 10.52465/joiser.v2i2.249.

[19] A. R. Safitri dan M. A. Muslim, “Improved Accuracy of Naive Bayes Classifier for Determination of Customer Churn Uses SMOTE and Genetic Algorithms,” *J. Soft Comput. Explor.*, vol. 1, no. 1, hal. 70–75, 2020, doi: 10.52465/joscex.v1i1.5.

[20] R. Nithya, T. Kokilavani, dan T. L. A. Beena, “Balancing cerebrovascular disease data with integrated ensemble learning and SVM-SMOTE,” *Netw. Model. Anal. Heal. Informatics Bioinforma.*, vol. 13, no. 1, 2024, doi: 10.1007/s13721-024-00447-4.

[21] A. S. Almajid, “Multilayer Perceptron Optimization on Imbalanced Data Using SVM-SMOTE and One-Hot Encoding for Credit Card Default Prediction,” *J. Adv. Inf. Syst. Technol.*, vol. 3, no. 2, hal. 67–74, 2022, doi: 10.15294/jaist.v3i2.57061.

[22] M. Mamun, A. Farjana, M. Al Mamun, dan M. S. Ahammed, “Lung cancer prediction model using ensemble learning techniques and a systematic review analysis,” *2022 IEEE World AI IoT Congr. AIIoT 2022*, hal. 187–193, 2022, doi: 10.1109/AIIoT54504.2022.9817326.

[23] A. Lestari, “Increasing Accuracy of C4.5 Algorithm Using Information Gain Ratio and Adaboost for Classification of Chronic Kidney Disease,” *J. Soft Comput. Explor.*, vol. 1, no. 1, hal. 32–38, 2020, doi: 10.52465/joscex.v1i1.6.

[24] E. Listiana, R. Muzayanah, M. A. Muslim, dan E. Sugiharti, “Optimization of support vector machine using information gain and adaboost to improve accuracy of chronic kidney disease diagnosis,” *J. Soft Comput. Explor.*, vol. 4, no. 3, hal. 152–158, 2023.

[25] E. Bolat, H. Yildirim, S. Altin, dan E. Yurtseven, “a Comprehensive Comparison of Machine Learning Algorithms on Diagnosing Asthma Disease and Copd,” *PONTE Int. Sci. Res. J.*, vol. 76, no. 3, 2020, doi: 10.21506/j.ponte.2020.3.17.

[26] W. Xu *et al.*, “Automatic pediatric congenital heart disease classification based on heart sound signal,” *Artif. Intell. Med.*, vol. 126, no. December 2021, hal. 102257, 2022, doi: 10.1016/j.artmed.2022.102257.

[27] M. Lopez-Perez, P. Morales-Alvarez, L. A. D. Cooper, R. Molina, dan A. K. Katsaggelos, “Deep Gaussian Processes for Classification With Multiple Noisy Annotators. Application to Breast Cancer Tissue Classification,” *IEEE Access*, vol. 11, no. January, hal. 6922–6934, 2023, doi: 10.1109/ACCESS.2023.3237990.

[28] M. Qorib, T. Oladunni, M. Denis, E. Ososanya, dan P. Cotae, “Covid-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset,” *Expert Syst. Appl.*, vol. 212, hal. 118715, 2023, doi: https://doi.org/10.1016/j.eswa.2022.118715.

[29] F. Jáñez-Martino, R. Alaiz-Rodríguez, V. González-Castro, E. Fidalgo, dan E. Alegre, “Classifying spam emails using agglomerative hierarchical clustering and a topic-based approach,” *Appl. Soft Comput.*, vol. 139, hal. 110226, 2023, doi: 10.1016/j.asoc.2023.110226.

[30] M. A. Heuvelmans *et al.*, “Lung cancer prediction by Deep Learning to identify benign lung nodules,” *Lung Cancer*, vol. 154, no. January, hal. 1–4, 2021, doi: 10.1016/j.lungcan.2021.01.027.