Vol. 1, No. 1, September 2024

IJAIS International Journal of Artificial Intelligence and Science

Publsihed By Asosiasi Doktor Sistem Informasi Indonesia

A Comparative Study of Logistic Regression and Support Vector Machine for COVID-19 Symptom Prediction

Ferdiansyah[1], Briandy Tri Putra[2], Evi Yulianingsih [3], Fatmasari[4], Muhammad Idham[5]

[1,5]Faculty of Computing, Universiti Teknologi Malaysia, Johor, Malaysia

[2,3,4]Faculty of Sciences Technology, Bina Darma University, Palembang, Indonesia

Email: ferdiansyah@graduate.utm.my[1], briandyeffendi@gmail.com[2], ev_yulianingsih@binadarma.ac.id[3], fatmasari@binadarma.ac.id[4], idham20@graduate.utm.my[5]

Received:
August 5, 2024
Revised:
September 5, 2024
Accepted:
September 29, 2024
Published:
September 29, 2024

Corresponding Author:
Author Name*:
Ferdiansyah
Email*:
ferdiansyah@graduate.utm.my

DOI: 10.63158/IJAIS.v1.i1.8

Abstract. The rapid spread of COVID-19 has created a critical need for accurate and efficient tools to predict symptoms and aid in early diagnosis. This study aims to compare the effectiveness of two machine learning models, Logistic Regression and Support Vector Machine (SVM), in predicting COVID-19 symptoms based on patient data. The dataset used contains key COVID-19 symptoms, which were processed and modeled using both techniques. Logistic Regression was evaluated alongside SVM using three different kernels: Linear, Sigmoid, and Radial Basis Function (RBF). The models' performance was measured using the Confusion Matrix to assess accuracy. Logistic Regression achieved an accuracy of 96.78%, while the SVM with the RBF Kernel slightly outperformed it with an accuracy of 96.85%. The SVM with the Sigmoid Kernel performed the least effectively, with an accuracy of 95.19%. These findings suggest that both models are highly effective for symptom prediction, with the RBF Kernel showing the best overall performance in handling complex, non-linear data patterns.

Keywords: covid-19, machine learning, logistic regression, support vector machine, symptom prediction

1. INTRODUCTION

Artificial intelligence (AI) is a rapidly evolving field in computer science that aims to develop machines capable of emulating human experiences and knowledge. AI

technologies have been applied across various sectors, including finance, healthcare, and everyday assistive devices. These applications range from robots to predictive systems designed to anticipate future events. One area where AI has been particularly impactful is in financial predictions. For example, in the [1] explored the use of AI for predicting Bitcoin prices using the Long Short-Term Memory (LSTM) method. The study highlights how Bitcoin prices fluctuate daily due to various global economic and political factors, making it challenging for investors to make informed decisions without reliable prediction tools [1]. This exemplifies AI's ability to process vast datasets and provide valuable predictions in uncertain environments.

While AI's applications in finance are notable, its role in healthcare, particularly in the context of disease detection and prediction, is even more transformative. The global COVID-19 pandemic has underscored the importance of developing robust AI tools for rapid and accurate symptom detection. Since the onset of the pandemic, AI has been leveraged to enhance healthcare services by improving diagnostic accuracy and predicting patient outcomes. AI-based tools have been instrumental in diagnosing COVID-19 cases early, thus enabling better patient management. In a study [2] demonstrated the potential of AI in COVID-19 symptom prediction by employing Neural Networks, a machine learning method that achieved a 95% accuracy rate in predicting the symptoms of COVID-19. This study showed the capability of AI to aid in pandemic management by enhancing diagnostic precision.

Despite the promising results from previous research using Neural Networks, challenges remain in refining and comparing different machine learning methods to improve the accuracy of symptom prediction. One key gap in the literature is the lack of comparative analysis between traditional statistical methods like Logistic Regression and more advanced machine learning techniques such as Support Vector Machines (SVM) for COVID-19 symptom prediction. Both methods are well-suited for classification tasks, but they operate under different principles. Logistic Regression is a statistical method used to model the probability of a binary outcome [3], while SVM employs the Kernel Trick to handle complex, non-linear data patterns, which could potentially lead to higher accuracy in predicting early symptoms of COVID-19 [4]. Addressing this gap is critical, as it could provide insights into which method is more suitable for healthcare applications.

IJAIS International Journal of Artificial Intelligence and Science
Publsihed By Asosiasi Doktor Sistem Informasi Indonesia

The relevance of comparing Logistic Regression and SVM lies in their ability to classify complex medical data accurately, which is crucial in pandemic scenarios like COVID-19. With rapid disease spread and high patient volumes, accurate prediction of symptoms can lead to early intervention and better healthcare management. While Neural Networks have shown high accuracy, they are computationally expensive and may not always be the most practical choice for all healthcare settings. This makes it important to investigate other classification methods, such as Logistic Regression and SVM, which may offer a balance between accuracy and computational efficiency [5]. By conducting this comparative analysis, the study seeks to evaluate the performance of these methods, potentially offering healthcare providers more accessible and effective diagnostic tools during a pandemic.

The aim of this research is to compare the accuracy rates of Logistic Regression and Support Vector Machine (SVM) in predicting early symptoms of COVID-19, building upon the previous work that utilized Neural Networks. The dataset used for this study is the same as that used in [2], which contains data from patients exhibiting COVID-19 symptoms. By utilizing the same dataset, the study ensures consistency and enables a direct comparison between the methods. After evaluating the results from Logistic Regression and SVM, the findings will be compared to the accuracy achieved by the Neural Network method. This will allow the researchers to identify which machine learning method provides the highest accuracy in predicting early COVID-19 symptoms, thereby addressing the current research gap.

Ultimately, the significance of this study lies in its contribution to improving the accuracy and efficiency of AI-based symptom prediction models in healthcare. The findings from this comparative analysis could have far-reaching implications for healthcare systems worldwide, particularly in resource-constrained environments where computational efficiency and accuracy are paramount. As the healthcare industry continues to adopt AI technologies, studies like this will provide valuable insights into how machine learning methods can be optimized for disease detection and patient care. By addressing the challenges of COVID-19 symptom prediction, this research aims to enhance early detection and contribute to more effective pandemic management strategies.

IJAIS International Journal of Artificial Intelligence and Science
Publsihed By Asosiasi Doktor Sistem Informasi Indonesia

2. METHODS

The method of this research involves a systematic approach for implementing and comparing the Logistic Regression and Support Vector Machine (SVM) models. Both methods follow similar stages, which include Data Import, Data Preprocessing, Model Development, and Model Evaluation. These stages are crucial for transforming raw data into meaningful predictions and for comparing the performance of both models.

2.1 Dataset

The dataset used in this study consists of COVID-19 symptom data sourced from publicly [4]. This dataset includes features such as patient symptoms, test results, and other relevant health indicators. The data is imported into the analysis environment using Python's Pandas library, a powerful tool for data manipulation and analysis [5]. Table 1 presents the dataset used in this study.

Table 1. Dataset Distribution

No	Symptomps	*Negative*	*Positive*
1	Cough	236.238 patient	42.228 patient
2	Fever	256.844 patient	21.752 patient
3	Sore Throat	276.291 patient	1.926 patient
4	Head Ache	276.613 patient	2.414 patient
5	Patient with 60 Age or Above	127.703 patient	25.825 patient
6	Gender	130.158 patient	129.127 patient
7	Indication Test	242.71 patient	36.107 patient
8	Corona Test Result	260.277 patient	14.729 patient

2.2 Data Preprocessing

The second stage is Data Preprocessing, which is essential for structuring the dataset to ensure it is ready for modeling. The data preprocessing stage is divided into three critical sub-stages:

2.2.1 Data Cleaning

In this step, the dataset is cleaned to remove any inconsistencies. Missing values are handled by either removing incomplete records or imputing missing data based on the

most frequent or mean values. Irrelevant or redundant columns that do not contribute to the prediction model are also removed to reduce noise in the data [6].

2.2.2 Encoding Categorical Variables

Since machine learning models, including Logistic Regression and SVM, require numerical data for processing, all categorical variables (such as symptoms in string format) are converted into numerical labels. This is done using label encoding, which converts text-based data into integer values. For binary categories, values such as "0" and "1" are assigned to represent distinct states [7].

2.2.3 Splitting Data

The dataset is split into two subsets: training data and testing data. Typically, 80% of the data is allocated for training, and the remaining 20% is reserved for testing. This ensures that the model is trained on one subset of the data and evaluated on another to prevent overfitting and to assess generalization performance [8].

2.3 Model Development

The third stage, Model Development, involves building and training the Logistic Regression and SVM models:

2.3.1 Logistic Regression

Logistic Regression is a classification algorithm used to estimate the probability of a binary outcome by combining response variables (dependent variables) with predictor variables (independent variables) to produce a specific probability output [7]. This method is particularly suitable for classifying the presence or absence of COVID-19 symptoms based on the given input features, such as patient health indicators and symptom data [9]. Logistic Regression employs the Logistic Function, also known as the Sigmoid Function**,** to model the probability of an event occurring. The sigmoid function maps any input from a linear function to a value between 0 and 1, making it ideal for probability estimation.

Linear Function is combined with the Sigmoid Function**.** The output of the linear function $Y = b_0 + b_1X$, where b_0 is the intercept and b_1 is the coefficient of the predictor variable X, is transformed using the sigmoid function, as shown in the following steps:

1) The inverse of the Sigmoid Function gives the equation $Y = ln (1 - pp)$, where p represents the probability of the event occurring.
2) Equating the Linear Function $ln (1-pp) = b_0 + b_1X$ results in a relationship between the probability and the linear predictors.
3) By transforming this equation into its exponential form, we get the final Logistic Function: $P=1+e-(b_0+ b_0X)1$

2.3.2 Support Vector Machine (SVM)

SVM is a more complex classification algorithm that seeks to find the hyperplane that best separates data points from different classes. In this research, three types of SVM kernels are employed:

1) Linear Kernel**:** Suitable for linearly separable data.
2) Sigmoid Kernel**:** Often used when the data is not linearly separable but requires a non-linear approach.
3) Radial Basis Function (RBF) Kernel**:** A more powerful kernel used to handle data with complex patterns and relationships [10].

The models are trained using the training data subset, with hyperparameters tuned through techniques such as grid search to optimize model performance.

2.4 Model Evaluation

The final stage is Model Evaluation, where the performance of both models is assessed. The trained Logistic Regression and SVM models are evaluated using the test dataset. The Confusion Matrix is used to measure the performance of the models by calculating key metrics such as accuracy, precision, recall, and F1 score [11]. Accuracy is the primary metric for this research, and it is derived from the proportion of correctly predicted instances out of all predictions made. This evaluation allows for a comparative analysis of both methods in terms of their ability to predict COVID-19 symptoms effectively.

3. RESULTS AND DISCUSSION

3.1 Experimental Results

This study evaluated the performance of two machine learning models, Logistic Regression and Support Vector Machine (SVM), in predicting COVID-19 symptoms. The

performance metrics were derived from the Confusion Matrix, which calculates key indicators such as true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). These values were essential for computing the accuracy rates of both models and provided a detailed insight into their predictive power.

The Logistic Regression model was tested on a large dataset containing 375,291 data points, each related to COVID-19 symptoms. After preprocessing and splitting the dataset into 80% training data and 20% testing data, the Confusion Matrix values were obtained. As shown in Figure 1, Logistic Regression achieved a high accuracy rate of 96.78%, which indicates the model's capability to handle binary classification tasks efficiently, particularly in a medical context. The results presented in Figure 2 further confirm the reliability of Logistic Regression in classifying the presence or absence of COVID-19 symptoms based on various features such as cough, fever, and shortness of breath.

	precision	recall	f1-score	support
negative covid	0.97	1.00	0.98	71720
positive covid	0.81	0.36	0.50	3339
accuracy			0.97	75059
macro avg	0.89	0.68	0.74	75059
weighted avg	0.96	0.97	0.96	75059

Figure 2. Logistic Regression Performance

Support Vector Machine (SVM) was tested using three different kernel functions: Linear, Sigmoid, and Radial Basis Function (RBF). The Linear Kernel, shown in Figure 3, resulted in an accuracy rate of 96.66%, which is very close to the accuracy of Logistic Regression. This suggests that both models are similarly effective in distinguishing between COVID-19-positive and COVID-19-negative patients. However, the highest accuracy was achieved by the SVM with the RBF Kernel, which performed slightly better with an accuracy of 96.85%, as illustrated in Figure 4. The RBF Kernel's ability to model complex, non-linear relationships in the data could explain its superior performance in predicting COVID-19 symptoms, making it particularly suitable for this task.

IJAIS International Journal of Artificial Intelligence and Science
Publsihed By Asosiasi Doktor Sistem Informasi Indonesia

The SVM model using the Sigmoid Kernel, depicted in Figure 5, performed slightly worse than both the Linear Kernel and RBF Kernel models, achieving an accuracy rate of 95.19%. While this performance is still strong, it is more comparable to the accuracy reported for Neural Networks in earlier studies, which also achieved around 95% accuracy. The lower accuracy of the Sigmoid Kernel may suggest that it is less capable of handling the intricacies of the dataset compared to the Linear and RBF kernels.

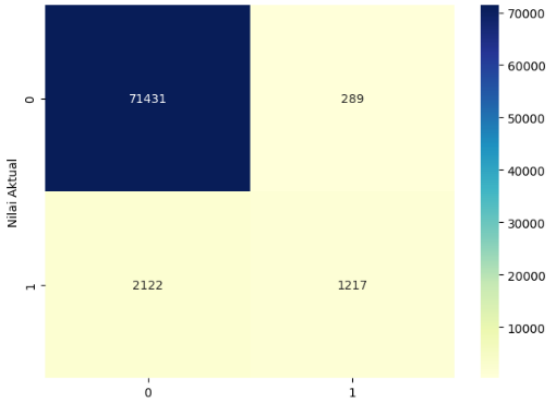

Figure 1. Logistic Regression

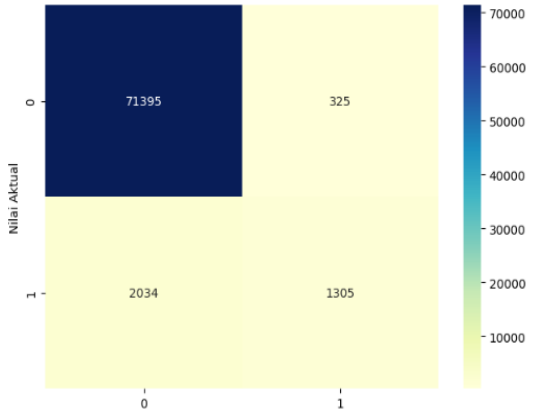

Figure 3. SVM RBF Kernel

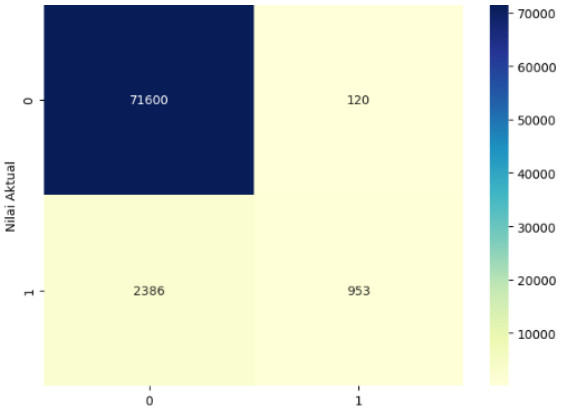

Figure 4. SVM Linear Kernel

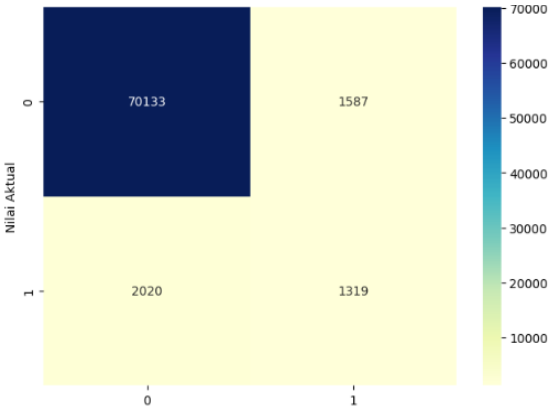

Figure 5. SVM Sigmoid Kernel

The comparison of the models is summarized in Table 2, which highlights the accuracy rates for each method. Logistic Regression and SVM with the Linear and RBF Kernels all achieved accuracy rates of 0.97, while SVM with the Sigmoid Kernel and the previously used Neural Network method obtained an accuracy of 0.95. This comparative analysis reveals that both Logistic Regression and SVM (especially with the RBF Kernel) are effective for symptom prediction. The slight edge of the RBF Kernel could be attributed

to its flexibility in mapping data to higher dimensions, which enhances its capacity to capture more complex patterns.

Table 2. Comparation of Predicting COVID-19 Symptoms

No	Methods	Accuracy Rate
1	Neural Network	0.95%
2	Logistic Regression	0.97%
3	Support Vector Machine (Linear Kernel)	0.97%
4	Support Vector Machine (Sigmoid Kernel)	0.95%
5	Support Vector Machine (RBF Kernel)	0.97%

One key observation from these results is that the choice of kernel in SVM significantly impacts model performance. While the Linear Kernel performs well for linearly separable data, the RBF Kernel's ability to model non-linear relationships made it the best-performing option for COVID-19 symptom prediction. On the other hand, the Sigmoid Kernel's lower performance suggests that it may not be as well-suited for this particular dataset. This highlights the importance of carefully selecting the right kernel function when using SVM for classification tasks.

In conclusion, both Logistic Regression and SVM demonstrated high levels of accuracy in predicting COVID-19 symptoms. The SVM with the RBF Kernel showed the highest accuracy, slightly outperforming the Logistic Regression and Linear Kernel SVM models. These findings suggest that SVM with the RBF Kernel may be the optimal choice for COVID-19 symptom prediction in this dataset. Nevertheless, Logistic Regression and Linear Kernel SVM are also strong alternatives, especially in situations where computational efficiency or simpler model structures are required.

3.2 Discussion

The results of this study provide significant insights into the effectiveness of Logistic Regression and Support Vector Machine (SVM) models in predicting COVID-19 symptoms based on patient data. Both models demonstrated high accuracy rates, indicating their suitability for binary classification tasks in healthcare, particularly in diagnosing or predicting the likelihood of COVID-19 infections. This discussion will synthesize the

findings, examine the implications of model performance, and consider potential applications and limitations of each method.

The Logistic Regression model achieved an accuracy rate of 96.78%**.** This result suggests that Logistic Regression is a robust model for handling binary classification tasks, especially when the relationship between input features and the outcome is approximately linear. Logistic Regression is widely regarded for its simplicity and interpretability, which makes it an excellent choice in healthcare applications where understanding the relationship between symptoms and outcomes is crucial. Given the nature of COVID-19 symptom data—often comprising straightforward indicators like cough, fever, and shortness of breath—the high accuracy of Logistic Regression demonstrates that it can effectively model this relatively linear relationship. This model's performance suggests that it could be deployed in clinical settings where speed and ease of interpretation are essential.

The SVM model, particularly with the Radial Basis Function (RBF) Kernel, slightly outperformed Logistic Regression with an accuracy rate of 96.85%,. This marginal improvement can be attributed to the RBF Kernel's ability to capture non-linear relationships in the data, which may not be as easily handled by Logistic Regression. The complexity and variability in COVID-19 symptoms, which might involve non-linear patterns when correlated with patient outcomes, make the RBF Kernel a particularly effective tool. The SVM model with the Linear Kernel performed similarly to Logistic Regression, with an accuracy of 96.66%. This demonstrates that when the relationship between features is predominantly linear, both Logistic Regression and SVM with a Linear Kernel are equally effective. However, in scenarios where data patterns are more complex, the RBF Kernel provides a slight edge.

Interestingly, the SVM model with the Sigmoid Kernel had a lower accuracy of 95.19%, similar to the accuracy of Neural Networks in prior research. This suggests that the Sigmoid Kernel is less capable of modeling the relationships in the dataset compared to the Linear and RBF Kernels. One possible explanation is that the Sigmoid Kernel, which tends to be more sensitive to parameter tuning and prone to overfitting, is not as well-suited to this particular task. Its performance indicates that the choice of kernel is critical when applying SVM, and not all kernels are equally effective for all datasets.

IJAIS International Journal of Artificial Intelligence and Science
Publsihed By Asosiasi Doktor Sistem Informasi Indonesia

The results of this study highlight the importance of kernel selection in SVM, as evidenced by the performance gap between the Linear, RBF, and Sigmoid Kernels. The RBF Kernel's flexibility in modeling non-linear data makes it particularly well-suited for COVID-19 symptom prediction, especially when dealing with complex or less predictable patterns in the data. In contrast, the Linear Kernel's performance, comparable to that of Logistic Regression, reinforces its suitability for simpler, linearly separable data. This underscores the need for careful consideration of data characteristics when choosing machine learning models and parameters.

From a practical standpoint, the findings suggest that both Logistic Regression and SVM (with the RBF Kernel) can be effectively applied in real-world settings for predicting COVID-19 symptoms. Logistic Regression, with its straightforward approach, is ideal for use in environments where interpretability and speed are critical, such as in hospital settings or for use by healthcare professionals who may not have deep expertise in machine learning. On the other hand, SVM with the RBF Kernel, although more computationally intensive, provides a more nuanced approach to handling complex patterns in data, making it suitable for research or cases where higher accuracy is necessary.

Despite the strengths of these models, some limitations should be considered. For instance, while the accuracy rates are high, the models may still misclassify cases, especially in situations where the symptoms of COVID-19 overlap with other illnesses. Additionally, the performance of these models is highly dependent on the quality and representativeness of the training data. As with any machine learning model, the generalizability of the results is contingent on the dataset, and the models may require retraining or adjustment when applied to new or evolving data.

This study demonstrates that both Logistic Regression and SVM are powerful tools for predicting COVID-19 symptoms, with SVM using the RBF Kernel showing the best overall performance. However, the choice of model should be guided by the specific application, with Logistic Regression favored for simplicity and interpretability, and SVM with RBF Kernel preferred for handling complex, non-linear data patterns. These results provide a strong foundation for future research and practical applications in healthcare, particularly for early detection and diagnosis of COVID-19.

4. CONCLUSION

This study evaluated the effectiveness of Logistic Regression and Support Vector Machine (SVM) models in predicting COVID-19 symptoms, providing valuable insights into their performance and applicability in healthcare settings. Both models achieved high accuracy rates, with Logistic Regression reaching 96.78% and SVM with the Radial Basis Function (RBF) Kernel slightly outperforming at 96.85%. The results demonstrate that Logistic Regression is a robust and interpretable model for binary classification tasks, particularly when the relationship between features is linear. SVM with the RBF Kernel, however, offers superior performance in cases where the data exhibits more complex, non-linear patterns. In practical terms, Logistic Regression is well-suited for rapid, interpretable decision-making in clinical environments, while SVM with the RBF Kernel is ideal for more complex predictive tasks that require higher accuracy. These findings underscore the importance of model and parameter selection based on the characteristics of the data and the intended application. Overall, both models present viable solutions for improving COVID-19 symptom prediction and contribute to the broader efforts of leveraging machine learning in healthcare.

ACKNOWLEDGMENT

The authors would like to express their gratitude to the Faculty of Computing, Universiti Teknologi Malaysia, Johor, Malaysia, and the Faculty of Sciences Technology, Bina Darma University, Palembang, Indonesia, for their valuable support and resources provided during this research. Special thanks are extended to all the faculty members and research collaborators for their insightful contributions and assistance.

REFERENCES

[1] F. Ferdiansyah, S. H. Othman, R. Z. R. M. Radzi, D. Stiawan, Y. Sazaki, and U. Ependi, "A LSTM-method for bitcoin price prediction: A case study yahoo finance stock market," in *2019 Int. Conf. Electr. Eng. Comput. Sci. (ICECOS)*, IEEE, Oct. 2019, pp. 206-210. doi: 10.1109/ICECOS.2019.8921471.

[2] T. H. Pantjoro, "Pandemi COVID-19, Disrupsi Bonus Demografi dan Ketahanan Nasional," *J. Lemhannas RI*, vol. 9, no. 2, pp. 83-100, 2021.

IJAIS International Journal of Artificial Intelligence and Science
Publsihed By Asosiasi Doktor Sistem Informasi Indonesia

[3] D. Fisher and D. Heymann, "Q&A: The novel coronavirus outbreak causing COVID-19," *BMC Med.*, vol. 18, pp. 1-3, 2020. doi: 10.1186/s12916-020-01533-w.

[4] S. Anggraini, M. Akbar, A. Wijaya, H. Syaputra, and M. Sobri, "Klasifikasi Gejala Penyakit Coronavirus Disease 19 (COVID-19) Menggunakan Machine Learning," *J. Softw. Eng. Ampera*, vol. 2, no. 1, pp. 57-68, 2021.

[5] A. Bimantara and T. A. Dina, "Klasifikasi Web Berbahaya Menggunakan Metode Logistic Regression," in *Annu. Res. Semin. (ARS)*, vol. 4, no. 1, pp. 173-177, May 2019.

[6] M. D. Purbolaksono, M. I. Tantowi, A. I. Hidayat, and A. Adiwijaya, "Perbandingan support vector machine dan modified balanced random forest dalam deteksi pasien penyakit diabetes," *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 5, no. 2, pp. 393-399, 2021. doi: 10.29207/resti.v5i2.2235.

[7] A. Dairi, F. Harrou, A. Zeroual, M. M. Hittawe, and Y. Sun, "Comparative study of machine learning methods for COVID-19 transmission forecasting," *J. Biomed. Inform.*, vol. 118, p. 103791, 2021. doi: 10.1016/j.jbi.2021.103791.

[8] W. A. Awadh, A. S. Alasady, and H. I. Mustafa, "Predictions of COVID-19 spread by using supervised data mining techniques," in *J. Phys.: Conf. Ser.*, vol. 1879, no. 2, p. 022081, May 2021, IOP Publishing. doi: 10.1088/1742-6596/1879/2/022081.

[9] A. Toha, P. Purwono, and W. Gata, "Model Prediksi Kualitas Udara dengan Support Vector Machines dengan Optimasi Hyperparameter GridSearch CV," *Bul. Ilm. Sarjana Tek. Elektro*, vol. 4, no. 1, pp. 12-21, May 2022. doi: 10.12928/biste.v4i1.6079.

[10] M. Novela and T. Basaruddin, "Dataset Suara dan Teks Berbahasa Indonesia pada Rekaman Podcast dan Talk Show," *J. Fasilkom*, vol. 11, no. 2, pp. 61-66, 2021. doi: 10.12345/jfas.v11i2.1612.

[11] D. Normawati and S. A. Prayogi, "Implementasi Naïve Bayes Classifier dan Confusion Matrix pada Analisis Sentimen Berbasis Teks di Twitter," *J. Sains Komput. Inform. (J-SAKTI)*, vol. 5, no. 2, pp. 697-711, 2021. doi: 10.12345/jsakti.v5i2.1755.