Journal of Electronics, Electromedical Engineering, and Medical Informatics
Multidisciplinary: Rapid Review: Open Access Journal **Vol. 5, No. 4, August 2023, pp: 261-270; eISSN: 2656-8632**

RESEARCH ARTICLE **OPEN ACCESS**

Manuscript received July 27, 2023; revised August 20, 2023; accepted September 21, 2023; date of publication October 30, 2023
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeemi.v5i4.328

How to cite: Muhamad Fawwaz Akbar, Muhammad Itqan Mazdadi, Muliadi, Triando Hamonangan Saragih, and Friska Abadi, "Implementation of Information Gain Ratio and Particle Swarm Optimization in the Sentiment Analysis Classification of Covid-19 Vaccine Using Support Vector Machine, vol. 5, no. 4, pp. 261-270, October 2023.

Implementation of Information Gain Ratio and Particle Swarm Optimization in the Sentiment Analysis Classification of Covid-19 Vaccine Using Support Vector Machine

Muhamad Fawwaz Akbar, Muhammad Itqan Mazdadi, Muliadi, Triando Hamonangan Saragih, and Friska Abadi

Computer Science Department, Lambung Mangkurat University, Banjarbaru, South Kalimantan, Indonesia

Corresponding author: Muhammad Itqan Mazdadi (e-mail: mazdadi@ulm.ac.id).

This work was supported by Lambung Mangkurat University for providing resources and support.

ABSTRACT In the current digital era, sentiment analysis has become an effective method for identifying and interpreting public opinions on various topics, including public health issues such as COVID-19 vaccination. Vaccination is a crucial measure in tackling this pandemic, but there are still a number of people who are skeptical and reluctant to receive the COVID-19 vaccine. This public perception is largely influenced by, including information received from social media and online platforms. Therefore, sentiment analysis of the COVID-19 vaccine is one way to understand the public's perception of the COVID-19 vaccine. This research has the purpose to enhance the classification performance in sentiment analysis of COVID-19 vaccines by implementing Information Gain Ratio (IGR) and Particle Swarm Optimization (PSO) on the Support Vector Machine (SVM). With a dataset of 2000 entries consisting of 1000 positive labels and 1000 negative labels, validation was performed through a combination of data splitting with an 80:20 ratio and stratified 10-Fold cross-validation. Applying the basic SVM, an accuracy of 0.794 and an AUC value of 0.890 were obtained. Integration with Information Gain Ratio (IGR) feature selection improved the accuracy to 0.814 and an AUC of 0.907. Furthermore, through the combination of SVM based on PSO and IGR, the accuracy significantly improved to 0.837 with an AUC of 0.913. These results demonstrate that the combination of feature selection techniques and parameter optimization can enhance the performance of sentiment classification towards COVID-19 vaccines. The conclusions drawn from this research indicate that the integration of IGR and PSO positively contributes to the effectiveness and predictive capability of the SVM model in sentiment classification tasks.

INDEX TERMS Covid-19 vaccine, Information Gain Ratio, Particle Swarm Optimization, SVM

I. INTRODUCTION

WHO declared Coronavirus Disease 2019 (COVID-19) a global pandemic on March 11, 2020. COVID-19 is brought on by the novel coronavirus Sars-CoV-2, which is considered highly dangerous due to its rapid and easy transmission. Various steps were taken to overcome this pandemic. One of them is the manufacture of vaccines [1]. Vaccination is a crucial measure in tackling this pandemic, but there are still a number of people who are skeptical and reluctant to receive the COVID-19 vaccine. Public perception of the COVID-19 vaccine can be influenced by various factors, including information received from social media and online platforms.

Therefore, sentiment analysis of the COVID-19 vaccine is one way to understand the public's perception of the COVID-

19 vaccine. Sentiment analysis is a strategy used to determine the presence of tweets on social media platform twitter that are positive and negative in nature. Through sentiment analysis, we can observe what users on Twitter frequently tweet about [2]. The aim is to comprehend and depict individuals' feelings or opinions towards a topic, product, or service. The results of sentiment analysis on the COVID-19 vaccine can provide valuable insights for organizations or governments in formulating policies related to the COVID-19 vaccine. SVM is a machine learning method that may be utilized for the analysis of sentiment in data.

SVM is a model for supervised learning that calls for sequential SVM training and subsequent testing. SVM classification attempts to separate the data space using either nonlinear or linear classification between different classes. The concept of SVM classification involves a hyperplane that acts as a separator between two classes of data (positive and negative aspects) [3],[4]. However, the use of SVM in sentiment analysis can give rise to several issues such as overfitting and low accuracy. To address these weaknesses of SVM, parameter optimization and feature selection can be employed.

Feature selection is one of the methods commonly used to narrow down or reduce dimensions among attributes, resulting in a few attributes considered relevant for the classification process [5]. Information gain is one of the calculations for choosing the best highlights for a feature set [6]. The Gain Ratio is a modified and improved approach to feature selection based on Information Gain, aimed at reducing its inherent bias by accounting for the number of possible attribute outcomes. In the feature selection process of gain ratio, improvement is achieved by considering intrinsic information of attributes. [5],[7]. In a research conducted by [8] examining the utilization of SVM, IG and IGR, an accuracy of 0.662 and AUC of 0.533 was achieved with IG, an accuracy of 0.691 and AUC of 0.584 with IGR. Feature selection plays a crucial role in improving machine learning model performance by reducing the dimensionality of the dataset and selecting the most informative attributes.

The particle swarm optimization algorithm stands out as a prominent optimization tool that can be employed for decision-making processes, facilitating the search for the optimal solution during data mining and classification endeavors. PSO is also a swarm intelligence algorithm, a field of computational systems inspired by collective intelligence [9]. In the research conducted by [10] the NB algorithm bringing about an accuracy of 0.738 and AUC of 0.712. Additionally, an accuracy of 0.808 and AUC of 0.739 was achieved with the addition of the algorithm for PSO. The similar aftereffects of opinion examination arrangement methods in this context demonstrate that the optimized NB using PSO outperforms the regular NB.

In this study, Information Gain Ratio (IGR) was chosen for feature selection to address bias issues and assess attribute relevance effectively. Particle Swarm Optimization (PSO) was employed for optimization due to its efficient and global search capabilities. These selections were made to enhance the accuracy of sentiment analysis regarding COVID-19 vaccines, aligning with the study's goals.

This research makes several contributions to the field of sentiment analysis. Firstly, it IGR as a feature selection technique, refining the feature set and potentially enhancing sentiment analysis classification accuracy. Secondly, it optimizes SVM model parameters using PSO, effectively improving the model's performance. However, the primary and most crucial contribution of this research lies in its unwavering focus on improving sentiment analysis accuracy, specifically in the context of COVID-19 vaccine sentiment data, by synergistically combining IGR and PSO.

II. METHOD

The research procedure used in this study is as follows. FIGURE 1 illustrates the research workflow.

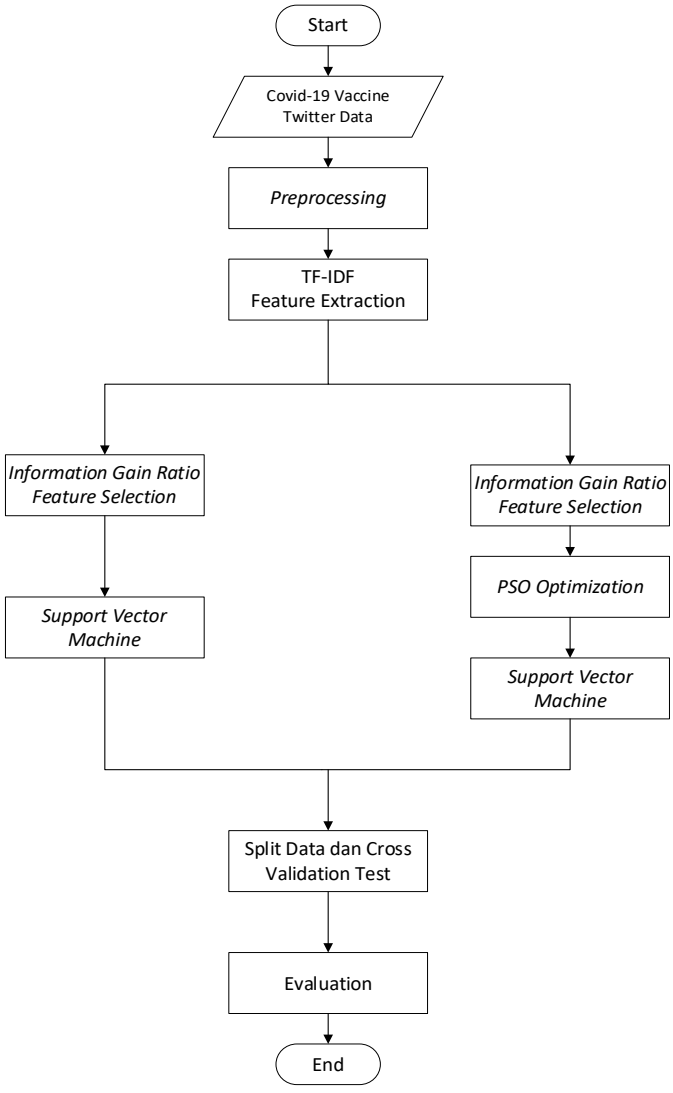

FIGURE 1. Research Flow

The progression of this exploration begins with collecting the "*vaksin covid*" dataset from Twitter within the time range of 1 January 2022 to 1 January 2023. Following this, labeling is performed using Vader. Subsequently, feature extraction is conducted using TF-IDF, and after feature extraction, feature selection is carried out using Information Gain Ratio. Following this, the research moves into the SVM-PSO optimization phase with testing validation that involves combined Split Data with an 80:20 ratio and Cross Validation Stratified with k-fold = 10. For performance measurement, the Confusion Matrix and AUC methods are employed.

A. DATA COLLECTION

The data for this research was collected using data scraping with the snscrape library, using the search term "*vaksin covid*". This dataset can be downloaded from this link https://github.com/FwzAr/DataVaccinesOnTwitter. The data used for the classification process consisted of 2000 tweets within the time range of 1 January 2022 to 1 January 2023, with 1000 positive and 1000 negative distributed equally. The data was then assigned two types of labels: Positive and Negative, using the Vader library. An example of the data used can be found in TABLE 1.

TABLE 1
Detail Dataset

No	Tweet	Label
1	*@Ndoro68295960 @StopPlandemit @bernardwee13 @ftrrii1 @DaraSagita99 @babydoge62 @BuKasunNdeso @MprAldo Harusnya nakes &amp lembaga terkait berikan 2 solusi sebagai alternative cara mencegah &amp mengobati covid. Bukan cuma terus menyuarakan 1 solusi doang AYO VAKSIN!""*	Postive
2	*@RazibSyah undi BN covid terhapus kerana vaksin sampai tahap dos 4.5.6*	Postive
3	*@vierda @mascarponecizz Pemikiran yg aneh. Vaksin COVID lebih berbahaya dari penyakit COVID. Kok orang seneng tebak-tebakan ngarang gitu ya hahaa..*	Postive
4	*@dr_koko28 Vaksin copit mmg mungkin tdk menyebabkan copit, tapi vaksin yg bahkan sudah booster-pun tdk bisa memcegah org terkena covid, BETUL???*	Negative
5	*@StopPlandemit yg masih percaya covid itu otaknya udah rusak,, udah sakau,, kebanyakan vaksin kek lagi ngedrugs.*	Negative

B. PREPROCESSING

Data preprocessing is a foundational and multifaceted process in Natural Language Processing (NLP), serving as the critical bridge between raw text data and meaningful insights. It encompasses several essential stages, each contributing to the quality and effectiveness of NLP applications [11]. The process begins with labeling, where data is categorized into distinct labels, typically 'positive' and 'negative,' crucial for tasks like sentiment analysis. Following labeling, cleaning is undertaken to meticulously remove extraneous elements such as URLs, hashtags, usernames, and special characters, reducing noise and enhancing data quality. Case folding standardizes letter casing, ensuring uniformity and simplifying subsequent text processing. Word normalization rectifies non-standard word variants, promoting consistency in the text data. Tokenization breaks down text into constituent words or tokens, facilitating structured analysis. Stopword removal eliminates non-informative words, enhancing the efficiency of NLP algorithms. Stemming further simplifies text by equating various word forms to their base form. Collectively, these preprocessing steps empower NLP practitioners to extract valuable insights and knowledge from text data with precision and efficiency, underlining the pivotal role of data preprocessing in NLP's continued evolution [12].

C. EXTRACTION FEATURE

TF-IDF

In text mining, feature extraction is a crucial step Since it furnishes details regarding the texts, such as the highest and lowest term recurrence for each record. Choosing connected elements and deciding scope effect for machine learning. Additionally, a positive aspect of the training model was its ability to extract informationl [13]. TF-IDF is an ordinarily referred to and it is utilized as weighting method its presentation actually even equivalent with newer techniques. The term weighting considers documents as the factors. The feature selection process is the key preprocessing step necessary for indexing the documents [14].The formula for TF-IDF is defined as equation (1):

$$\mathrm{TF-IDF} = \mathrm{TF} \times \mathrm{IDF} = \mathrm{TF}_{t,d}\ \mathrm{X}\ \frac{|\mathrm{D}|}{\mathrm{DF}_t} \quad (1)$$

$\mathrm{TF}_{t,d}$ denoting the term frequency, which signifies how often a specific word "t" appears within a particular document "d". DF_t corresponds to the document frequency, representing the total count of documents within a corpus that contain the word "t". D refers to the overall count of accessible documents in the corpus [11].

D. SELECTION FEATURE

INFORMATION GAIN RATIO

The term used to describe the shift in class entropy from the previous state to the state where the value of the attribute is identified is referred to as information gain (IG). IG is used to indicate feature relevance in this setting [15]. The formula for Information Gain (IG) is defined as equation (2) [16]:

$$IG(X;Y) = H(X) - H(X \mid Y) \quad (2)$$

Information Gain tends to favor attributes with more branches, which can lead to overfitting. To address the limitations of IG, Gain Ratio is employed to assess separation attributes [6]. The result obtained by dividing information gain by intrinsic

information is known as the gain ratio. Furthermore, when dealing with a large number of branching features, the gain ratio is a change in IG designed to reduce its inherent bias. When choosing a feature, the gain ratio takes into account the size as well as the number of branches. Equation describes the Gain Ratio (GR) formula (3):

$$GR = \frac{IG}{H(X)} \quad (3)$$

The value of the Gain Ratio always falls between 0 and 1. If the value of GR is 1, it means that all knowledge about X leads to Y, and if it is 0, it means that there is no connection between X and Y [17].

E. SUPPORT VECTOR MACHINE (SVM)

SVM, are one of the most effective approaches to text classification [17]. The SVM algorithm is commonly employed in classification and regression assignments [18]. A SVM display it is a depiction of instances as points in space, strategically arranged to ensure that the models of distinct categories are segregated by a meaningful gap that is maximally wide. Within a high-dimensional or potentially infinite-dimensional space, a support vector machine constructs a hyperplane or a collection of hyperplanes that serve the purpose of classification, relapse, or other tasks [19].

The mathematical foundation of SVMs is rooted in the formulation of a hyperplane defined by the equation (4) where ω weight vector orthogonal to the hyperplane, and b signifies the bias term, which controls the hyperplane's offset from the origin. SVMs aim to maximize the margin, which is the distance between the hyperplane and the nearest data points of each class, termed support vectors. Mathematically, SVM solve an optimization problem by minimizing ½ |(| ω |)|^2 subject to constraints equation (5) for all data points, with y_i representing the class labels. This formulation ensures not only accurate classification but also a maximally wide margin, making SVMs a robust choice for various machine learning tasks.

$$\omega \cdot x + b = 0 \quad (4)$$

$$y_i(x_i + b) \geq 1 \quad (5)$$

[20],[21].

The goal of SVM is to separate two classes in a dataset with a maximum margin. SVM can generate an effective and accurate classification model, even for complex datasets or those with high dimensions. In SVM, a hyperplane (separating plane) is formed to separate classes within the dataset, and the calculation of the maximum margin aims to amplify the division distance between classes data points and the hyperplane. SVM can also employ kernels to project information into a higher-layered include space, thus enabling the separation of non-linearly separable classes within the dataset. Fundamentally, non-linear SVM is a solution to the linear SVM problem by applying a kernel function in a high-dimensional feature space. In other words, SVM enables the maximization of a model's generalization capability [22]. The definitions of linear and non-linear SVM equations can be observed in TABLE 2.

TABLE 2
SVM Linear and Non-Linear

SVM	Kernel Type	Formula
SVM Linear	Linier	$K(x,y) = x.y$
SVM Non-Linear	Polynomial	$K(x,y) = (x.y + 1)^p$
	Gaussian RBF	$K(x,y) = e^{-\|\|X-Y\|\|^2/2\sigma^2}$

F. PARTICLE SWARM OPTIMIZATION (PSO)

PSO draws inspiration from the cooperative and social behaviors exhibited by multiple species as they seek to fulfill their needs within the search space. In order to determine the particles' subsequent positions Within the search space, this algorithm is directed by personal experience P_{best}, overall experience G_{best}, and and the current motion of particles [23]. Simply put, PSO is a method of evolutionary computation akin to the genetic algorithm (GA) in which The commencement of the process by populating the search space with random solutions. initiates the search for solutions [24]. The following are the PSO steps:

These steps begin with the meticulous setup of PSO parameters, followed by the utilization of a fitness function to evaluate each particle's associated cost. Concurrently, close attention is given to the values of P_{best} dan G_{best}. Subsequently, particle velocities are updated using equation (6), which considers correction factors c_1 and c_2 along with random variables r_1 and r_2. Alongside this velocity update, the position of each particle is modified according to equation (7), ensuring a dynamic exploration of the solution space. These integral steps collectively define the systematic operation of the PSO algorithm in its quest for optimized solutions [25].

$$v_i(t) = wv_i(t-1) + c_1r_1 + c_2r_2(x_{gbest}(t) - x_t(t)) \quad (6)$$

$v_i(t)$ represents a particle's velocity at time t as it moves within the solution space. $wv_i(t-1)$ is the previous velocity, capturing prior momentum. c_1 influences a particle's behavior based on its personal best-known position. r_1, a random variable between 0 and 1, adds exploration randomness. c_2 influences behavior based on the global best-known position $x_{gbest}(t)$. Similarly, r_2, another random variable, introduces stochasticity. Lastly, $(x_{gbest}(t) - x_t(t))$ shows the difference between the global best-known position and the particle's current position at time t [25].

$$x_i(t+1) = x_i(t) + v_i(t) \quad (7)$$

The variable $x_i(t+1)$ represents the updated position of the $i-th$ particle at the next time step, while $x_i(t)$ indicates its current position at time t. This equation essentially describes how the particle's position is adjusted by adding its current velocity $v_i(t)$ [25].

G. CONFUSION MATRIX

Classification algorithms, which encounter numerous challenges in model construction, require data pre-processing stages, especially for high-dimensional data [26]. The Confusion Matrix's performance evaluation metrics and accuracy are both important considerations when choosing the best classification algorithm [27]. Machine learning classification algorithms use a statistical measurement known as the Confusion Matrix to determine the model's accuracy [28]. A classification model's performance is depicted by evaluation metrics. The evaluation metrics that are used to comprehend the algorithm's performance and efficiency are the most important aspect of classification [29]. The Confusion Matrix is a table that can be Produced for a classifier when dealing with binary datasets and can serve to characterize the classifier's effectiveness. The Matrix table can be observed in TABLE 3 [30].

TABLE 3
Confusion Matrix

Classification	Predicted Class	
	Positive	Negative
Actual : Positive	TP(True Postive)	FN (False Negative)
Actual : Negative	FP (False Positive)	TN (True Negative)

This matrix is derived from the terms: True Positives (TP): Both predicted and actual outcomes are positive. True Negatives (TN): The prediction is negative, while the actual is positive. False Positive (FP): The prediction is positive and the actual is negative. False Negative (FN): Both predicted and actual outcomes are negative. Accuracy is the proportion of true positive and true negative classifications of all documents defined as (8) [29]:

$$accuracy = \frac{TN + TP}{TN + FP + FN + TP} \quad (8)$$

H. AREA UNDER THE CURVE (AUC)-RECEIVER OPERATING CHARACTERISTIC (ROC)

A visual instrument for the comparison of classification models is the ROC. In the 1970s, its application grew to include the interpretation of medical test results in the biomedical field. In the past few years, the domains of machine learning and data have witnessed significant developments mining research have made extensive use of its analysis methods [31],[32]. The discrete classifier solely provides predictions about because the ROC curve is a two-dimensional graph, the category the tested object belongs to. There are four potential outcomes: False positives, false negatives, and true positives [33]. Here is a guide to classify testing accuracy using the AUC value [34]. The categories for AUC values can be seen in TABLE 4.

TABLE 4
Accuracy of Classification Results Based on AUC Value

AUC Value	Category
0,90 - 1,00	Excellent classification
0,80 - 0,90	Good classification
0,70 - 0,80	Fair classification
0,60 - 0,70	Poor classification
0,50 - 0,60	Failure classification

III. RESULT

In this section, the evaluation results of three model variations: SVM, SVM+IGR, and SVM-PSO+IGR are presented. As an initial step, a series of data pre-processing stages are conducted, encompassing data cleaning, word normalization, tokenization, elimination of common words (stopwords), as well as stemming processes. The details of these pre-processing outcomes can be observed in TABLE 5.

TABLE 5
Dataset Result

No	Tweet	Label
1	*nakes lembaga kait solusi alternative cegah obat covid suara solusi doang ayo vaksin*	Postive
2	*undi covid hapus ranah vaksin tahap dos*	Postive
3	*pikir aneh vaksin covid bahaya sakit covid orang senang tebak tebak arang*	Postive
4	*vaksin covid covid vaksin booster cegah orang kena covid*	Negative
5	*percaya covid otak rusak sakau vaksin obat*	Negative

Based on the pre-processing stages carried out, the data is subsequently further processed through the feature extraction phase by adopting the TF-IDF method. This approach is applied to characterize text based on the significance of specific words within the context of the entire document collection. Through this feature extraction process, a total of 4,313 unique features representing the dataset are successfully obtained. The detailed feature representation based on TF-IDF can be seen in TABLE 6.

TABLE 6
Extraction Feature Result

aman	*covid*	….	*manfaat*	*vaksin*	Label
0	0.172	….	0.970	0.172	Positive
0	0.055	….	0	0.055	Positive
….	….	….	….	….	….
0	0.049	….	0	0.049	Negative

Following the feature extraction that resulted in a total of 4,313 features from the dataset of COVID-19 vaccine-related tweets, the subsequent step taken is the implementation of feature selection techniques using the Information Gain Ratio (IGR) approach. Through this approach, with a percentile parameter of 80% as the selection criterion, the IGR method successfully filters and determines features that possess high relevance and informativeness for analysis. As a result of this selection process, the feature dimension undergoes significant reduction, with only 3450 features remaining and poised for use in further analysis. The outcomes of the feature selection can be observed in TABLE 7.

TABLE 7
Feature Selection Result

Attribute	Weight
covid	0.096
vaksin	0.096
perintah	0.090
booster	0,088
vaksinasi	0.084

After the feature selection process yields 3450 essential features, the next step involves optimizing SVM model by integrating the PSO technique and utilizing a rbf kernel. In this optimization phase, default parameter values are used, specifically $c1$=0,5, $c2$=0,5, and w=0,9. To ensure optimal convergence in exploring the solution space, the PSO configuration is set with a population of 10 particles and a total of 30 iterations.

This research proceeds with the stages of model performance evaluation. To ensure an objective and comprehensive evaluation, a combined validation approach is employed, wherein To perform both training and testing, the dataset is split in an 80:20 ratio, and a stratified 10-fold cross-validation is applied approach is used. This approach ensures that the label distribution in each fold is representative of the label distribution in The complete dataset is utilized for the purpose of determining how well the model performance and dependability, the primary evaluation metrics employed are the confusion matrix and the AUC. By utilizing these metrics, this research seeks to obtain a comprehensive understanding of the model's capability to classify COVID-19 vaccine-related tweet data.

1. RESEARCH OUTCOMES: SUPPORT VECTOR MACHINE

In this research, the experimental results are obtained through the application of the SVM method. As part of the evaluation process, the data is divided using an 80:20 splitting scheme, and an additional stratified 10-Fold cross-validation is implemented to ensure the model's integrity. After the validation stage, the model's performance evaluation is carried out utilizing the confusion matrix and AUC-ROC. The evaluation results are presented in TABLE 8:

TABLE 8
Accuracy And AUC SVM

Model	Accuracy	AUC
SVM	0.794	0,890

The results from the SVM model, with an 80:20 data split ratio and stratified 10-fold cross-validation, yield an accuracy of 0.794 and an AUC of 0.890.

TABLE 9
Confusion Matrix SVM

Classification	Predicted Class	
	Positive	Negative
Actual : Postive	518	47
Actual : Negative	282	753

The accuracy can be calculated using the results of the confusion matrix in TABLE 9:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Accuracy = 0.794$$

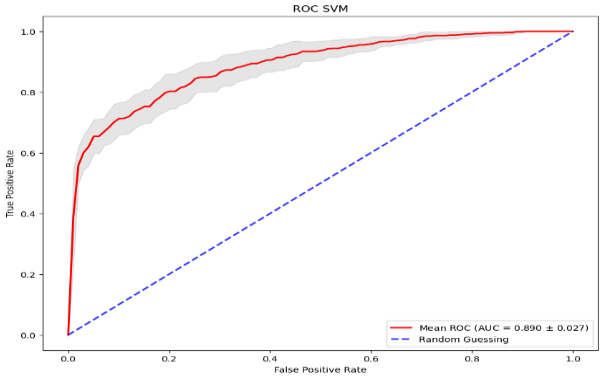

FIGURE 2. ROC Curve of SVM

During the testing of the SVM method, a The ROC curve, which can be seen in FIGURE 2, also yielded an AUC of 0.890. The AUC value falls within the "Good Classification" range due to its positioning within the value range of 0.80 to 0.90.

2. RESEARCH OUTCOMES: SUPPORT VECTOR MACHINE + IGR

In this research, the experimental outcomes are obtained through the implementation of the Support Vector Machine (SVM) method combined with IGR. From this process, a total of 3450 features are selected through the feature selection process. To ensure the integrity and reliability of the model, evaluation is carried out using two approaches: data splitting with an 80:20 ratio and stratified 10-Fold cross-validation. Following these stages, The model's effectiveness evaluation

is conducted by employing the confusion matrix and AUC-ROC to acquire an in-depth understanding of the model's classification capabilities. The evaluation results are presented in TABLE 10:

TABLE 10
Accuracy and AUC SVM+IGR

Model	Accuracy	AUC
SVM + IGR	0.814	0.907

The results from the SVM+IGR model, with an 80:20 data split ratio and stratified 10-fold cross-validation, yield an accuracy of 0.814 and AUC of 0.907.

TABLE 11
Confusion Matrix SVM+IGR

Classification	Predicted Class	
	Positive	Negative
Actual : Postive	566	63
Actual : Negative	234	737

The accuracy can be determined as follows from the confusion matrix's results in TABLE 11:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Accuracy = 0.814$$

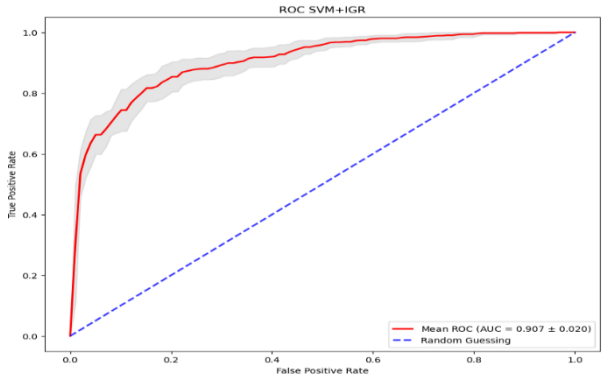

FIGURE 3. ROC Curve of SVM+IGR

During the testing of the SVM+IGR method, Furthermore, a ROC curve was produced, as depicted in FIGURE 3, which yielded an AUC of 0.907. This AUC score falls within the "Excellent Classification" range due to its positioning within the value range of 0.90 to 1.00.

3. RESEARCH OUTCOMES: SVM-PSO + IGR

In this research, the experimental outcomes are obtained through the implementation of the Support Vector Machine technique that includes SVM-PSO combined with IGR. This method is executed with default parameters: coefficient $c1$=0,5, $c2$=0,5, and inertia weight w=0,9. During the optimization process, 10 particles are employed, and iterations are performed 30 times. For evaluation purposes, the data is partitioned using an 80:20 splitting scheme, and a stratified 10-Fold cross-validation is applied to obtain a more comprehensive overview of the model's performance across various data subsets. Based on this procedure, the model's effectiveness evaluation is measured utilizing the confusion matrix and AUC-ROC, providing an in-depth depiction of the model's capacity to classify data. Evaluation results are presented in TABLE 12.

TABLE 12
Accuracy and AUC SVM-PSO+IGR

Model	Accuracy	AUC
SVM-PSO+IGR	0.837	0.913

The results from the SVM-PSO+IGR model, with an 80:20 data split ratio and stratified 10-fold cross-validation, yield an accuracy of 0.837 and AUC of 0.913.

TABLE 13
Confusion Matrix SVM-PSO+IGR

Classification	Predicted Class	
	Positive	Negative
Actual : Postive	627	88
Actual : Negative	173	712

The accuracy can be determined as follows from the confusion matrix's results in TABLE 13:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Accuracy = 0.837$$

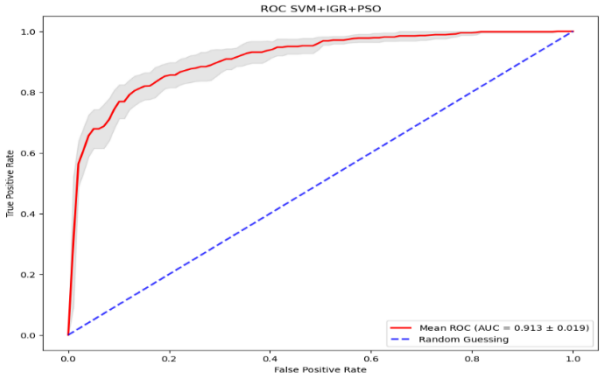

FIGURE 4. ROC Curve of SVM-PSO+IGR

During the testing of the SVM-PSO+IGR method, a Additionally, an ROC curve was constructed and can be observed in FIGURE 4, resulting in an AUC score of 0.913. This AUC measurement falls within the "Excellent Classification" range due to its positioning within the value range of 0.90 to 1.00.

IV. DISCUSSION

This research utilizes a dataset comprising 2000 tweets related to the COVID-19 vaccine, with a balanced label distribution of 1000 positive and 1000 negative samples as the classification foundation. The validation process is conducted through a combination of data splitting based on an 80:20 ratio and stratified 10-Fold cross-validation to ensure a fair representation of both classes. The classification method adopted in this research is Support Vector Machine (SVM). The Information Gain Ratio technique is applied to identify and eliminate noise in the dataset. As an optimization step, Particle Swarm Optimization (PSO) with default parameters *c*1=0,5, *c*2=0,5, inertia weight *w*=0,9, *n*_particle=10, and 30 iteration, is implemented to perform tuning and discover the optimal parameters for the RBF kernel in SVM.

In this analysis, employing the basic SVM model, the evaluation results obtained from the confusion matrix and AUC for the classification of COVID-19 vaccine-related tweet data reveal an accuracy of 0.794 and an AUC of 0.890. According to the reference AUC values, these results can be categorized as a good classification. This signifies that the SVM model possesses a significant capability to distinguish between positive and negative sentiments within the dataset of COVID-19 vaccine-related tweets.

The implementation of SVM and IGR on the selected 3450 features reveals an improved performance in classifying COVID-19 vaccine-related tweet data, yielding an accuracy of 0.814 and an AUC value of 0.907, as shown in the evaluation results. Referring to the AUC value reference, these outcomes can be categorized as excellent classification. Consequently, the implication is that the integration of SVM with IGR significantly contributes to enhancing the model's capability to classify sentiment within the dataset of COVID-19 vaccine-related tweets.

By implementing the combination of SVM-PSO and IGR on the selected 3450 features and conducting tuning using Particle Swarm Optimization (PSO) with default parameters *c*1=0,5, *c*2=0,5, inertia weight *w*=0,9, *n_particle*=10, and 30 iterations, the evaluation results exhibit a significant enhancement in the classification performance of COVID-19 vaccine-related tweet data, with an achieved accuracy of 0.837 and an AUC value of 0.913. Based on the reference AUC scale, the performance of this model can be categorized as excellent classification. This signifies that the integration of SVM, PSO, and IGR effectively enhances the precision in sentiment identification within the dataset of COVID-19 vaccine-related tweets. The evaluation results are presented in TABLE 14 and FIGURE 5.

TABLE 14
AUC Classification of Covid Vaccine-Related Tweet Data

Model	Accuracy	AUC
SVM	0.794	0.890
SVM+IGR	0.814	0.907
SVM-PSO+IGR	0.837	0.913

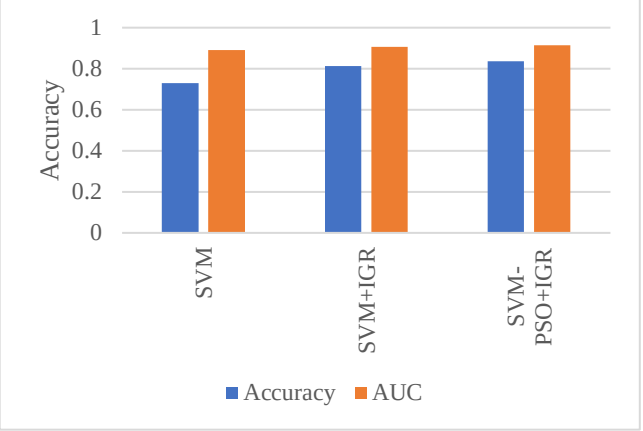

FIGURE 5. Comparision Accuracy and AUC

The evaluation results of the SVM-PSO+IGR, when compared with other studies conducted by [8] and [10]. In study conducted by [8] utilizing SVM-IGR, reported an accuracy of 0.691 and an AUC of 0.584. [10] on the other hand, using the NB-PSO approach, achieved a slightly improved accuracy of 0.808 and an AUC of 0.739. However, the present study, employing the combined SVM-PSO+IGR methodology, has yielded outstanding results, with an accuracy of 0.837 and an AUC of 0.913. These findings highlight the substantial performance gains the methodology has brought to the field, underscoring the significance of the research in the context of other studies. The comparision results are presented in TABLE 15 and FIGURE 6.

TABLE 15
Comparision With Other Studies

Model	Accuracy	AUC
NB-PSO	0.808	0.739
SVM+IGR	0.691	0.584
SVM-PSO+IGR	0.837	0.913

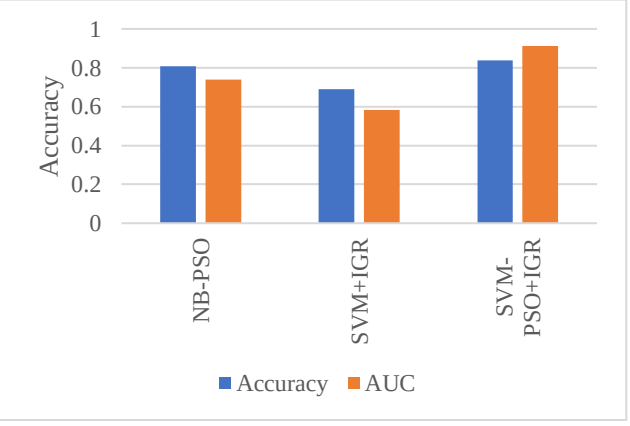

FIGURE 6. Comparision With Other Studioes

In this research, an improvement is observed in the SVM classification outcomes following the implementation of

feature selection using IGR and parameter optimization through PSO. Based on the evaluation, the SVM model that has been optimized with PSO and enhanced with Information Gain Ratio (IGR) feature selection demonstrates superior performance in terms of both accuracy and AUC in comparison to other analyzed models. This indicates the effectiveness of the integration between IGR feature selection and PSO optimization in enhancing the efficiency and precision of the SVM model in data classification.

Limitation of the conducted research lies in the use of default parameters during the optimization process for SVM-PSO+IGR. The decision to rely on default parameters, without comprehensive parameter tuning may have influenced the achieved results. While these default settings produced an accuracy of 0.837 and an AUC of 0.913, it's important to acknowledge that optimization algorithms, including particle swarm optimization, often require fine-tuning of parameters to maximize their effectiveness. Consequently, the study may not have fully explored the potential of the SVM-PSO+IGR approach with different parameter configurations. Future research could benefit from a more extensive investigation into parameter tuning to assess its impact on model performance and further refine the methodology.

The research outcomes achieved through the SVM-PSO+IGR approach, while utilizing default parameters, hold noteworthy implications for the field of optimization and machine learning. These results suggest that this methodology has the potential to provide valuable contributions to model performance. It emphasizes the importance of further research endeavors focusing on comprehensive parameter tuning to optimize the approach fully. Researchers and practitioners should consider the benefits of parameter tuning when implementing this methodology in various applications to maximize its effectiveness and adapt it to specific problem domains. This study underscores the significance of fine-tuning parameters in machine learning applications and encourages further investigations into parameter optimization techniques to enhance model performance and broaden the applicability of this approach.

V. CONCLUSION

Based on the research results conducted on the dataset of COVID-19 vaccine-related tweets, it was discovered that the basic SVM classification model yields an accuracy of 0.794 and an AUC value of 0.890, which falls under the category of "Good Classification". Meanwhile, through the integration of feature selection using IGR on 3450 selected features, the model's performance improves, achieving an accuracy of 0.814 and an AUC of 0.907, thereby entering the "Excellent Classification" category. Furthermore, the optimization of SVM based on PSO along with Improved IGR yields an accuracy of 0.837 and an AUC of 0.913, both falling within the "Excellent Classification" category. The enhancement achieved through IGR amounts to 0.020 for accuracy and 0.017 for AUC. Meanwhile, the combined approach of SVM-PSO+IGR provides an accuracy improvement of 0.043 and an AUC improvement of 0.023. From this analysis, it can be inferred that the integration of IGR and PSO significantly contributes to enhancing the effectiveness of classifying tweets related to the COVID-19 vaccine.

REFERENCES

[1] M. Rahardi, A. Aminuddin, F. F. Abdulloh, and R. A. Nugroho, "Sentiment Analysis of Covid-19 Vaccination using Support Vector Machine in Indonesia," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 6, pp. 534–539, 2022, doi: 10.14569/IJACSA.2022.0130665.

[2] P. H. Prastyo, A. S. Sumi, A. W. Dian, and A. E. Permanasari, "Tweets Responding to the Indonesian Government's Handling of COVID-19: Sentiment Analysis Using SVM with Normalized Poly Kernel," *J. Inf. Syst. Eng. Bus. Intell.*, vol. 6, no. 2, p. 112, 2020, doi: 10.20473/jisebi.6.2.112-122.

[3] K. R. Kavitha, A. Gopinath, and M. Gopi, "Applying improved SVM classifier for leukemia cancer classification using FCBF," *2017 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2017*, vol. 2017-Janua, pp. 61–66, 2017, doi: 10.1109/ICACCI.2017.8125817.

[4] S. Zahoor and R. Rohilla, "Twitter Sentiment Analysis Using Machine Learning Algorithms: A Case Study," *Proc. - 2020 Int. Conf. Adv. Comput. Commun. Mater. ICACCM 2020*, pp. 194–199, 2020, doi: 10.1109/ICACCM50413.2020.9213011.

[5] P. P. R., V. M.L., and S. S., "Gain Ratio Based Feature Selection Method for Privacy Preservation," *ICTACT J. Soft Comput.*, vol. 01, no. 04, pp. 201–205, 2011, doi: 10.21917/ijsc.2011.0031.

[6] R.-H. Dong, H.-H. Yan, and Q.-Y. Zhang, "An Intrusion Detection Model for Wireless Sensor Network Based on Information Gain Ratio and Bagging Algorithm," *Int. J. Netw. Secur.*, vol. 22, no. 2, pp. 218–230, 2020, doi: 10.6633/IJNS.202003.

[7] B. Prasetiyo, Alamsyah, M. A. Muslim, and N. Baroroh, "Evaluation of feature selection using information gain and gain ratio on bank marketing classification using naïve bayes," *J. Phys. Conf. Ser.*, vol. 1918, no. 4, 2021, doi: 10.1088/1742-6596/1918/4/042153.

[8] T. A. H. Tengku Mazlin, R. Sallehuddin, and M. Y. Zuriahati, "Utilization of Filter Feature Selection with Support Vector Machine for Tumours Classification," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 551, no. 1, 2019, doi: 10.1088/1757-899X/551/1/012062.

[9] D. D. R. Ochoa, L. A. P. Domínguez, E. A. M. Gómez, and D. L. Cruz, "PSO, a Swarm Intelligence-Based Evolutionary Algorithm as a Decision-Making Strategy: A Review," 2022.

[10] S. Panggabean, W. Gata, and T. A. Setiawan, "Analysis of Twitter Sentiment Towards Madrasahs Using Classification Methods," *J. Appl. Eng. Technol. Sci.*, vol. 4, no. 1, pp. 375–389, 2022, doi: 10.37385/jaets.v4i1.1088.

[11] S. Vijayarani, M. J. Ilamathi, M. Nithya, A. Professor, and M. P. Research Scholar, "Preprocessing Techniques for Text Mining - An Overview," vol. 5, no. 1, pp. 7–16.

[12] Y. Kang, Z. Cai, C. W. Tan, Q. Huang, and H. Liu, "Natural language processing (NLP) in management research: A literature review," *J. Manag. Anal.*, vol. 7, no. 2, pp. 139–172, 2020, doi: 10.1080/23270012.2020.1756939.

[13] T. Wen and Z. Zhang, "Effective and extensible feature extraction method using genetic algorithm-based frequency-domain feature search for epileptic EEG multiclassification," *Med. (United States)*, vol. 96, no. 19, pp. 1–11, 2017, doi: 10.1097/MD.0000000000006879.

[14] M. Ramya and J. A. Pinakas, "Different Type of Feature Selection for Text Classification," *Int. J. Comput. Trends Technol.*, vol. 10, pp. 102–107, Apr. 2014, doi: 10.14445/22312803/IJCTT-V10P118.

[15] E. Odhiambo Omuya, G. Onyango Okeyo, and M. Waema Kimwele, "Feature Selection for Classification using Principal Component Analysis and Information Gain," *Expert Syst. Appl.*,

vol. 174, no. January, p. 114765, 2021, doi: 10.1016/j.eswa.2021.114765.
[16] M. Shakil Pervez and D. Md. Farid, "Literature Review of Feature Selection for Mining Tasks," *Int. J. Comput. Appl.*, vol. 116, no. 21, pp. 30–33, 2015, doi: 10.5120/20462-2829.
[17] A. H. Mohammad, "Comparing two feature selections methods (Information gain and gain ratio) on three different classification algorithms using arabic dataset.," *J. Theor. Appl. Inf. Technol.*, vol. 96, no. 6, pp. 1561–1569, 2018.
[18] L. Gunawan, M. S. Anggreainy, L. Wihan, Santy, G. Y. Lesmana, and S. Yusuf, "Support vector machine based emotional analysis of restaurant reviews," *Procedia Comput. Sci.*, vol. 216, no. 2022, pp. 479–484, 2022, doi: 10.1016/j.procs.2022.12.160.
[19] A. P. Gopi, R. N. S. Jyothi, V. L. Narayana, and K. S. Sandeep, "Classification of tweets data based on polarity using improved RBF kernel of SVM," *Int. J. Inf. Technol.*, vol. 15, no. 2, pp. 965–980, 2023, doi: 10.1007/s41870-019-00409-4.
[20] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer New York, NY, 2006. doi: 10.1007/978-3-030-57077-4_11.
[21] M. A. Chandra and S. S. Bedi, "Survey on SVM and their application in image classification," *Int. J. Inf. Technol.*, vol. 13, no. 5, 2021, doi: 10.1007/s41870-017-0080-1.
[22] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," *Neurocomputing*, vol. 408, no. xxxx, pp. 189–215, 2020, doi: 10.1016/j.neucom.2019.10.118.
[23] M. Y. Cho and T. T. Hoang, "Feature Selection and Parameters Optimization of SVM Using Particle Swarm Optimization for Fault Classification in Power Distribution Systems," *Comput. Intell. Neurosci.*, vol. 2017, 2017, doi: 10.1155/2017/4135465.
[24] D. J. Kalita and S. Singh, "SVM Hyper-parameters optimization using quantized multi-PSO in dynamic environment," *Soft Comput.*, vol. 24, no. 2, pp. 1225–1241, 2020, doi: 10.1007/s00500-019-03957-w.
[25] R. Indraswari and A. Z. Arifin, "RBF KERNEL OPTIMIZATION METHOD WITH PARTICLE SWARM OPTIMIZATION ON SVM USING THE ANALYSIS OF INPUT DATA'S MOVEMENT," *J. Ilmu Komput. dan Inf.*, vol. 10, no. 1, p. 36, 2017, doi: 10.21609/jiki.v10i1.410.
[26] M. R. A.-G. Ahmed and A. M. Abdalla, "Enhancing Hybrid Intrusion Detection and Prevention System for Flooding Attacks Using Decision Tree," in *International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE)*, 2019, no. September, pp. 1–4.
[27] U. Ahmad, H. Asim, M. T. Hassan, and S. Naseer, "Analysis of Classification Techniques for Intrusion Detection," *3rd Int. Conf. Innov. Comput. ICIC 2019*, no. Icic, 2019, doi: 10.1109/ICIC48496.2019.8966675.
[28] A. A. Salih and A. M. Abdulazeez, "Evaluation of Classification Algorithms for Intrusion Detection System: A Review," *J. Soft Comput. Data Min.*, vol. 02, no. 01, pp. 31–40, 2021, doi: 10.30880/jscdm.2021.02.01.004.
[29] T. Saranya, S. Sridevi, C. Deisy, T. D. Chung, and M. K. A. A. Khan, "Performance Analysis of Machine Learning Algorithms in Intrusion Detection System: A Review," *Procedia Comput. Sci.*, vol. 171, no. 2019, pp. 1251–1260, 2020, doi: 10.1016/j.procs.2020.04.133.
[30] M. Navim and R. Pankaja, "Performance Analysis of Text Classification Algorithms using Confusion Matrix," *Int. J. Eng. Tech. Res. IJETR*, vol. 6, no. 4, pp. 75–78, 2016.
[31] J. Hernández-Orallo, "ROC curves for regression," *Pattern Recognit.*, vol. 46, no. 12, pp. 3395–3411, 2013, doi: 10.1016/j.patcog.2013.06.014.
[32] D. Lin, L. Sun, K. A. Toh, J. B. Zhang, and Z. Lin, "Twin SVM with a reject option through ROC curve," *J. Franklin Inst.*, vol. 355, no. 4, pp. 1710–1732, 2018, doi: 10.1016/j.jfranklin.2017.05.003.
[33] C. Y. Lee and W. C. Lin, "Induction Motor Fault Classification Based on ROC Curve and t-SNE," *IEEE Access*, vol. 9, pp. 56330–56343, 2021, doi: 10.1109/ACCESS.2021.3072646.
[34] F. Gorunescu, *Data Mining : Concepts, Models and Techniques*. Berlin: Germany: Springer-Verlag Berlin Heidelberg, 2011.

Authors Biography

Muhamad Fawwaz Akbar was born in Sampit, Central Kalimantan. Since 2018, he has pursued academic as a student of Computer Science Department at Lambung Mangkurat University. His current area of research lies within field of data science. Additionally, his final project entailed conducting research that centered around the classification of text within messages sourced from social media. This research contributes to the better understanding of public sentiment and opinions surrounding the Covid-19 vaccine.

Muhammad Itqan Mazdadi finished his bachelor's degree in Computer Science from Universitas Lambung Mangakurat in 2013. Subsequently, in 2017, he obtained his master's degree from the Department of Informatics at Islamic University of Indonesia, Yogyakarta. Following that, he joined Lambung Mangkurat University as a lecturer in Computer Science. Currently, he also serves as the secretary of the Computer Science Department at his workplace as a lecturer. His current area of research revolves around Data Science, IoT, Network, and Digital forensic.

Muliadi is a lecturer in the Department of Computer Science, Lambung Mangkurat University. His research interest is focused on Data Science. Before becoming a lecturer, he completed his bachelor's degree at STMIK Akakom Yogyakarta in 2004. After that, he pursued a master's degree in Computer Science at Gajah Mada University in 2009. Currently, he also serves as a manager of a scientific journal at his workplace as a lecturer. The research field he is involved in is Data Science.

Triando Hamonangan Saragih is a lecturer in Department of Computer Science, Lambung Mangkurat University. His research interest is focused on Data Science. He completed his bachelor's degree in Informatics at Brawijaya University, Malang in 2016. After that, he pursued a master's degree in Computer Science Brawijaya University, Malang in 2018. The research field he is involved in is Data Science.

Friska Abadi finished his bachelor's degree in Computer Science from Lambung Mangkurat University in 2011. Subsequently, in 2016, he obtained his master's degree from the Department of Informatics at STIMIK Amikom, Yogyakarta. Following that, he joined Lambung Mangkurat University as a lecturer in Computer Science. Currently, he holds the position of head of the software engineering laboratory. His current area of research revolves around software engineering.