Jurnal CoreIT, Vol.11, No.1, June 2025
ISSN 2460-738X (Print)
ISSN 2599-3321 (Online)

Optimizing Student Depression Prediction Using Particle Swarm Optimization and Random Forest

Mukhammad Khoirul Effendi[1], Sriyanto[2*], Suhendro Yusuf Irianto, Chairani Fauzi, Yelfi Vitriani
[1234]Magister Teknik Informatika, IBI Darmajaya, Indonesia
[5]Teknik Informatika, Universitas Islam Negeri Sultan Syarif Kasim Riau, Indonesia
khoiruleffendi24@gmail.com, sriyanto@darmajaya.ac.id, suhendro@darmajaya.ac.id, chairani@darmajaya.ac.id, yelfi.vitriani@uin-suska.ac.id

Abstract. Student mental health is a growing concern due to increasing academic pressure, social demands, and economic factors affecting their well-being. Depression, a common issue among students, significantly impacts academic performance and overall quality of life. Therefore, early detection and accurate prediction of student mental health conditions are essential to provide timely interventions. This study aims to improve the accuracy of depression prediction among university students by integrating Particle Swarm Optimization (PSO) for feature selection with Random Forest (RF) as the classification model. The dataset used is the Student Depression Dataset from Kaggle, consisting of 27,900 respondents with 18 features related to demographic, academic, and psychological factors. Data preprocessing includes handling missing values, normalization, categorical encoding, and feature selection using PSO. The model is trained and evaluated using 10-Fold Cross-Validation. Experimental results show that PSO-optimized Random Forest outperforms the standard Random Forest model. The optimized model achieves an accuracy of 84.08%, precision of 82.79%, recall of 77.79%, and an AUC-ROC score of 0.912, improving classification performance. These findings demonstrate that PSO effectively enhances feature selection, leading to better classification accuracy. This study contributes to the development of a more accurate and efficient machine learning model for detecting student depression. By optimizing feature selection, this approach reduces computational complexity while maintaining high predictive performance. Future research can explore hybrid optimization techniques such as Genetic Algorithm (GA) or Differential Evolution (DE) to further enhance model generalization across different datasets.

Keywords: Machine Learning, Mental Health, Particle Swarm Optimization, Random Forest, Student Depression

Received March 2025 / **Revised** April 2025 / **Accepted** May 2025

INTRODUCTION

Student mental health is an issue that is getting more and more attention because students' mental health is an aspect that can affect their academic achievement and psychological well-being. The transition from high school to college is a challenging period that often leads to stress and depression, especially for new students who face increased academic and social pressure[1]. Academic factors such as final projects can also be the main trigger for mental health problems, especially for final year students who have to complete their thesis independently, and early detection of mental health disorders can be an effective preventive step in reducing negative psychological impacts on students.[2].

Several previous studies have proposed various machine learning-based methods to classify and predict depression rates among college students. Research by Sawangarreerak and Thanathamathee [3] developed a model of predicting student depression using the Random Forest algorithm with sampling techniques to handle data imbalances. The results show that a combination of sampling techniques and ensemble learning can improve accuracy in predicting students' mental health. However, the study still faces challenges in selecting optimal features to improve model efficiency.

Another study by Abrori and Fatah [4] used Decision Tree in depression classification with RapidMiner software. This model shows high accuracy, reaching 97.50%, but is prone to overfitting complex datasets. Meanwhile, Budiman et al. [5] examined a Naïve Bayes-based approach to detecting indications of depression based on social media text analysis. The results of their study showed that the combination of TF-IDF and Complement Naïve Bayes (CNB) provided the best performance with an F-score of 91.98%. This approach is effective in text analysis, but it is less than optimal in handling data with heterogeneous features.

DOI: 10.24014/coreit.v11i1.35954

In addition, research by Aziz et al. [6] applied the text mining-based Support Vector Machine (SVM) method for depression classification, with high accuracy results of up to 100%. However, this model requires complex and less flexible parameter tuning to datasets with diverse numerical and categorical features. Rahayu et al. [7] compared several machine learning algorithms, including Decision Tree, Random Forest, Naïve Bayes, and K-Nearest Neighbor (KNN) in the classification of texts related to depression and anxiety. The results of their research show that Random Forest has the best performance with an accuracy rate of 96%, making it one of the most reliable models for mental health predictions.

Based on previous studies, it can be seen that the machine learning method has provided a promising approach in detecting students' mental health conditions. However, most studies still have limitations in terms of optimal feature selection and handling of datasets with complex and unbalanced features. In addition, some models have limitations in computational efficiency and interpretability. Therefore, this study proposes a combined approach of Particle Swarm Optimization (PSO) for feature selection and Random Forest as a classification model to improve the accuracy of student mental health predictions. With this approach, it is hoped that a more efficient, accurate, and well-interpretable model can be obtained to support early intervention in handling student mental health.

METHODS

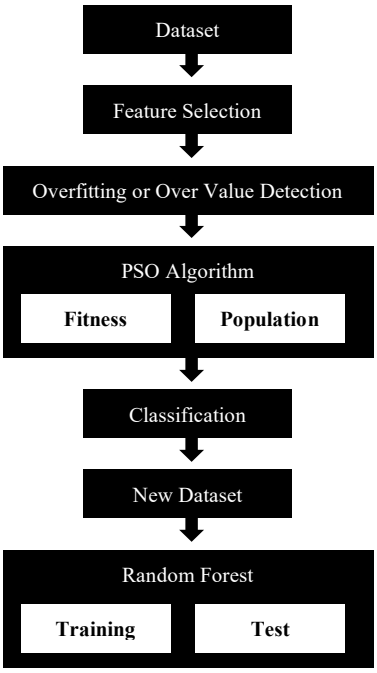


Figure 1. Step Method

Data Collection

This study used the Student Depression Dataset from Kaggle, which consisted of 27,900 respondents with 18 attributes that included demographic, academic, and psychological factors of students. The main attributes used included age, gender, academic stress, learning satisfaction, social media addiction, and family history of mental illness. This data is processed through several stages.

Table 1. Student Depression Dataset Feature Table

Feature	Data Type	Description
Id	String	A student's unique identification number
Gender	Categorical	Gender (male/female)

Age	Numerical	The Age of the Students in the Year
City	Categorical	City of residence
Profession	Categorical	Profession
Academic Pressure	Numerical (1-10)	Academic stress levels
Work Pressure	Numerical (1-10)	Work stress level
CGPA	Numerical	Cumulative average score
Study Satisfaction	Numerical (1-10)	Learning satisfaction level
Job Satisfaction	Numerical (1-10)	Job satisfaction level
Sleep Duration	Categorical	Average sleep duration
Dietary Habits	Categorical	Eating habits
Degree	Categorical	College Majors
Have you ever had suicidal thoughts?	Categorical	Suicidal Thoughts (Yes/No)
Work/Study Hours	Numerical (0-12)	Average Learning Duration
Financial Stress	Numerical (1-10)	Financial Stress Levels
Family History of Mental Illness	Categorical	Family history of mental illness (Yes/No)
Depression	Categorical	Depressed status (Yes/No) - Target Variable

Data Preprocessing

In this stage, the data is analyzed The presence of missing values in the dataset can reduce the amount of information learned by the machine learning model during the training stage, which ultimately negatively affects the accuracy of the classification[8]. Data cleansing is an important process in handling lost data to ensure the integrity and quality of data analysis[9]. Data cleansing is not just about filling in the gaps in the dataset, but also about understanding and selecting the most appropriate techniques to ensure that the data used in clinical analysis or decision-making is accurate and reliable[10]. In this study, the researcher overcame missing values with the mean imputation method.

Data normalization is an important step in the data processing process that aims to improve the accuracy of the prediction model[11] Numerical features are normalized using Min-Max Scaling to range from [0,1] Min-max model normalization is done by changing each value in a feature to improve model performance [12].

Transforming category data into numerical data is an important step in many machine learning applications[13]. Encoding Categorical Data Variables such as gender, relationship status, and eating habits are converted into numerical forms with One-Hot Encoding. One-Hot Encoding consistently achieves the highest accuracy across the various machine learning algorithms evaluated, although it requires longer processing times, especially when feature cardinality is high[13].

The selection of the right features can make a significant contribution to improving the efficiency and effectiveness of big data analytics in the performance of machine learning models[14]. Use Pearson Correlation Coefficient to eliminate irrelevant or redundant features in further analysis with high linear correlation[15].

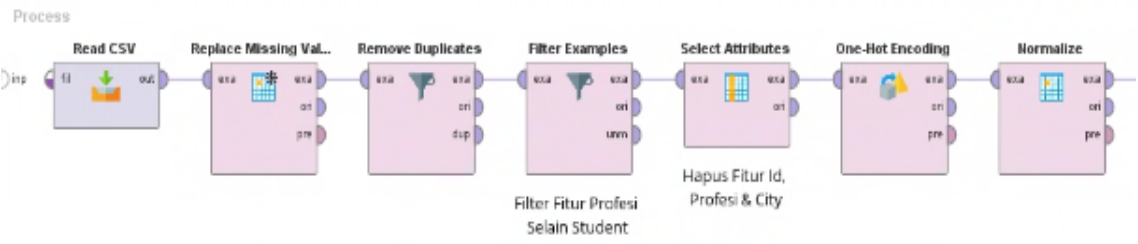


Figure 2. Data Preprocessing

Feature Selection with Particle Swarm Optimization (PSO)
Particle Swarm Optimization (PSO) is a metaheuristic optimization algorithm inspired by the social behavior of flocks, such as birds or fish, in search of optimal solutions in the search room[16]. In PSO, each particle represents a set of hyperparameter values that are evaluated based on specific criteria, such as accuracy or prediction error values. With the mechanisms of exploration and exploitation, these particles move towards optimal solutions based on individual and collective experiences in the population[17]. Feature selection was conducted using Particle Swarm Optimization (PSO) to select the best subset of features that improve the accuracy of students' mental health predictions. The PSO algorithm works with:

Initialize a random particle population in the feature search space[18].

Determine fitness function based on the accuracy of the Random Forest model using a selected feature subset[19].

Update particle position and velocity based on individual best values (p_{best}) dan global (g_{best}) Using the equation

$$v_i^{t+1} = w\ v_i^t + c_1\ r_1\ (pbest_i - x_i^t) + c_2\ r_2\ (gbest\ - x_i^t)$$

$$x_i^{t+1} = x_i^t + v_i^{t+1}$$

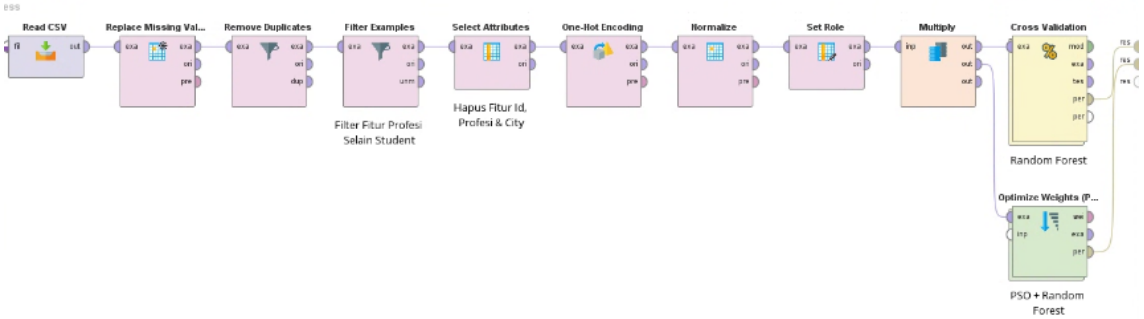
With w as inertia weight, c_1, c_2 as an acceleration coefficient, dan r_1, r_2 as a random number between 0 and 1[20]. Iterations are carried out to convergence, i.e. when the maximum number of iterations is reached or there is no significant change in the value fitness[19].

Classification with Random Forest
After obtaining a subset of the best features from the PSO, the Random Forest (RF) model is used for classification with the following steps:

Constructing RF Classifications: The model is formed with n_estimators selected through hyperparameter tuning[20].

The 10-Fold Cross-Validation technique was used to divide the data into 10 subgroups, where one subgroup was used as test data and the rest as training data[21]. By utilizing 10-Fold Cross-Validation, machine learning models show increased accuracy and reliability[22]

Proper Prediction and Validation The proper evaluation of Machine Learning models is essential to ensure security, fairness, and reliability in their use[23]. The model is tested with test data to obtain evaluation metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. All of these metrics are used in the study to evaluate the performance of machine learning models, with the aim of providing a more comprehensive picture of the model's reliability and usefulness in real-world contexts[23].



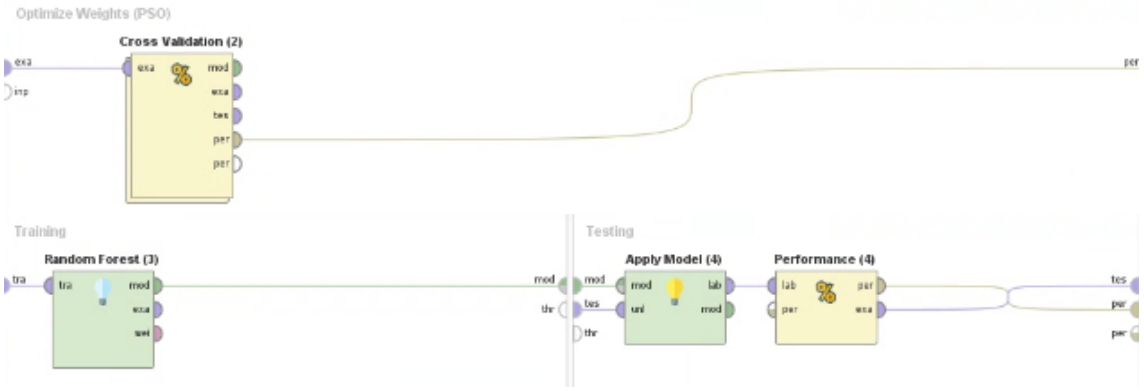


Figure 3. Testing the Random Forest Algorithm Model and
Random Forest algorithm optimized with PSO

Model Evaluation

Model evaluation was carried out by comparing Random Forest with and without PSO. Some of the key metrics used are:

Accuracy : Measures how often the model makes correct predictions.
Precision : The correct accuracy of positive predictions of all positive predictions made.
Recall : The model's ability to detect all actual positive examples.
AUC-ROC : Measures model performance at various thresholds, with areas below the ROC curve higher indicating better model performance in distinguishing between positive and negative classes[24].

RESULT AND DISCUSSION

This study aims to improve the accuracy of students' mental health predictions by optimizing feature selection using Particle Swarm Optimization (PSO) on the Random Forest (RF) model. The results of the experiment showed that the integration of PSO with RF provided improved model performance compared to RF without feature optimization.

Table 2. Model Performance Comparison

	Akurasi (%)	**Precision (%)**	**Recall (%)**	**AUC-ROC**
Random Forest without PSO	83.66 ± 05	82.35 ± 1.24	77.14 ± 1.00	0.909 ± 0.006
Random Forest with PSO	84.08 ± 0.57	82.79 ± 1.01	77.79 ± 1.25	0.912 ± 0.00404

The increase in accuracy by 0.42%, precision by 0.44%, and recall by 0.65% indicates that PSO is able to optimize the selection of features that are more relevant to the prediction target. Higher recalls showed that models with PSO were better at detecting cases of student depression compared to models without feature selection. The increased precision suggests that models with PSO are able to reduce errors in classifying individuals who are not depressed as depressed.

Compared to previous studies using Decision Tree and Naïve Bayes methods, the methods proposed in this study provide better results in terms of model accuracy and stability. In addition, this approach is also superior to the Random Forest method without feature selection, as it can improve modeling efficiency without significantly adding complexity.

However, there are some limitations in this study. The relatively small increase in accuracy suggests that PSO can still be explored further with more optimal parameters or combined with other methods such as Genetic Algorithm (GA). In addition, the increased computing time due to PSO iterations is one of the challenges that need to be considered. Therefore, future research can examine more efficient optimization techniques or test models on datasets with more diverse characteristics to improve model generalization.

CONCLUSION

This study has shown that the use of Particle Swarm Optimization (PSO) as a feature selection method in Random Forest can improve the accuracy in predicting students' mental health. The results of the experiment proved that the model with PSO optimization provided higher accuracy, precision, and recall than Random Forest without feature selection, although the improvement was relatively small. Nonetheless, this approach offers efficiency in data processing by eliminating less relevant features, allowing the model to work more optimally. The impact of this research is to contribute to the development of a machine learning-based student depression early detection system, which can help educational institutions and professionals in identifying students who are at risk of mental disorders. However, challenges in PSO parameter selection and increased computational time indicate the need for further exploration of other optimization techniques. Therefore, further research is recommended to combine PSO with other methods, such as Genetic Algorithm (GA) or Differential Evolution (DE), and test it on more diverse datasets to ensure broader model generalization.

REFERENCES

[1] C. V. LOTULUNG and I. G. PURNAWINADI, "Deteksi Dini Depresi Mahasiswa Baru Jurusan Keperawatan," *PAEDAGOGY J. Ilmu Pendidik. dan Psikol.*, vol. 4, no. 2, pp. 179–185, 2024, doi: 10.51878/paedagogy.v4i2.3042.

[2] M. K. Sari and E. A. Susmiatin, "Deteksi Dini Kesehatan Mental Emosional pada Mahasiswa," *J. Ilm. STIKES Yars. Mataram*, vol. 13, no. 1, pp. 10–17, 2023, doi: 10.57267/jisym.v13i1.226.

[3] S. Sawangarreerak and P. Thanathamathee, "Random forest with sampling techniques for handling imbalanced prediction of university student depression," *Inf.*, vol. 11, no. 11, pp. 1–13, 2020, doi: 10.3390/info11110519.

[4] S. Abrori and Z. Fatah, "Implementasi Metode Decission Tree Dalam Mengklasifikasi Depresi Menggunakan Rapidminer," vol. 5, no. 2, pp. 123–132, 2025.

[5] A. Budiman, J. C. Young, and A. Suryadibrata, "Implementasi Algoritma Naïve Bayes untuk Klasifikasi Konten Twitter dengan Indikasi Depresi," *J. Inform. J. Pengemb. IT*, vol. 6, no. 2, pp. 133–138, 2021, doi: 10.30591/jpit.v6i2.2419.

[6] F. Aziz, P. Ishak, and S. Abasa, "Klasifikasi Depresi Menggunakan Support Vector Machine: Pendekatan Berbasis Data Text Mining," *J. Pharm. Appl. Comput. Sci.*, vol. 2, no. 2, pp. 33–38, 2024, doi: 10.59823/jopacs.v2i2.53.

[7] K. Rahayu, V. Fitria, D. Septhya, R. Rahmaddeni, and L. Efrizoni, "Klasifikasi Teks untuk Mendeteksi Depresi dan Kecemasan pada Pengguna Twitter Berbasis Machine Learning," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 3, no. 2, pp. 108–114, 2023, doi: 10.57152/malcom.v3i2.780.

[8] A. Palanivinayagam and R. Damaševičius, "Effective Handling of Missing Values in Datasets for Classification Using Machine Learning Methods," *Inf.*, vol. 14, no. 2, pp. 1–15, 2023, doi: 10.3390/info14020092.

[9] P. A. Popoola, J. R. Tapamo, and A. G. H. Assounga, "Effective and Efficient Handling of Missing Data in Supervised Machine Learning," *Data Sci. Manag.*, 2024, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666764924000663

[10] H. Wang, J. Tang, M. Wu, X. Wang, and T. Zhang, "Application of machine learning missing data imputation techniques in clinical decision making: taking the discharge assessment of patients with spontaneous supratentorial intracerebral hemorrhage as an example," *BMC Med. Inform. Decis. Mak.*, vol. 22, no. 1, pp. 1–14, 2022, doi: 10.1186/s12911-022-01752-6.

[11] M. Sholeh, D. Andayati, and R. Y. Rachmawati, "Data Mining Model Klasifikasi Menggunakan Algoritma K-Nearest Neighbor Dengan Normalisasi Untuk Prediksi Penyakit Diabetes," *TeIKa*, vol. 12, no. 02, pp. 77–87, 2022, doi: 10.36342/teika.v12i02.2911.

[12] Wenny, "Normalisasi Data Kependudukan Dengan Model Min Max Dan Algoritma K-Means Untuk Pengelompokkan Tingkat Ekonomi Masyarakat," *Bull. Inf. Syst. Res.*, vol. 2, no. 2, pp. 63–73, 2024, [Online]. Available: https://journal.grahamitra.id/index.php/bios

[13] A.-I. Udilaˇ, A. Ionescu, and A. Katsifodimos, "Encoding Methods for Categorical Data: A Comparative Analysis for Linear Models, Decision Trees, and Support Vector Machines," 2023, [Online]. Available: http://repository.tudelft.nl/.

[14] W. Albattah, R. U. Khan, M. F. Alsharekh, and S. F. Khasawneh, "Feature Selection Techniques for Big Data Analytics," *Electron.*, vol. 11, no. 19, 2022, doi: 10.3390/electronics11193177.

[15] K. Mei, M. Tan, Z. Yang, and S. Shi, "Modeling of Feature Selection Based on Random Forest Algorithm and Pearson Correlation Coefficient," *J. Phys. Conf. Ser.*, vol. 2219, no. 1, 2022, doi:

10.1088/1742-6596/2219/1/012046.
[16] R. M. Ubaidilah, “Performance Comparasion of Adaboost and PSO Algorithms for Cervical Cancer Classification Using KNN Algorithm,” vol. 3321, no. X, pp. 65–74, doi: 10.24014/coreit.v10i2.31711.
[17] A. C. S. Alexita, P. Kusumaningtyas, and ..., “Optimasi Algoritma Random Forest Menggunakan Pso Untuk Klasifikasi Kanker Payudara Dengan Citra Mammograms,” *Tek. STTKD J. ...*, 2025, [Online]. Available: https://jurnal.sttkd.ac.id/index.php/ts/article/view/1346
[18] Y. Zhang and Z. Tang, “PSO-weighted random forest for attractive tourism spots recommendation,” *Futur. Gener. Comput. Syst.*, vol. 127, pp. 421–425, 2022, doi: https://doi.org/10.1016/j.future.2021.09.029.
[19] M. Ajdani and H. Ghaffary, “ Introduced a new method for enhancement of intrusion detection with random forest and PSO algorithm ,” *Secur. Priv.*, vol. 4, no. 2, pp. 1–10, 2021, doi: 10.1002/spy2.147.
[20] Kurniabudi *et al.*, “Improvement of attack detection performance on the internet of things with PSO-search and random forest,” *J. Comput. Sci.*, vol. 64, no. April, p. 101833, 2022, doi: 10.1016/j.jocs.2022.101833.
[21] S. M. Malakouti, M. B. Menhaj, and A. A. Suratgar, “The usage of 10-fold cross-validation and grid search to enhance ML methods performance in solar farm power generation prediction,” *Clean. Eng. Technol.*, vol. 15, no. February, p. 100664, 2023, doi: 10.1016/j.clet.2023.100664.
[22] S. M. Malakouti, “Babysitting hyperparameter optimization and 10-fold-cross-validation to enhance the performance of ML methods in predicting wind speed and energy generation,” *Intell. Syst. with Appl.*, vol. 19, no. March, p. 200248, 2023, doi: 10.1016/j.iswa.2023.200248.
[23] B. Hutchinson, N. Rostamzadeh, C. Greer, K. Heller, and V. Prabhakaran, “Evaluation Gaps in Machine Learning Practice,” *ACM Int. Conf. Proceeding Ser.*, pp. 1859–1876, 2022, doi: 10.1145/3531146.3533233.
[24] D. Rajput, W. J. Wang, and C. C. Chen, “Evaluation of a decided sample size in machine learning applications,” *BMC Bioinformatics*, vol. 24, no. 1, pp. 1–17, 2023, doi: 10.1186/s12859-023-05156-9.