Analysis of Geographical Characteristics and Risk Factors for Non-Communicable Diseases: Diabetes in Central Java Using Random Forest and SHAP

Ambar Arum Prameswari[1], Andi Prademon Yunus[1*]
[1]Telkom University, Banyumas, Indonesia
Email: ambararum@student.telkomuniversity.ac.id, andiay@telkomuniversity.ac.id
[*]Corresponding Author

Abstract

Diabetes mellitus is a growing health issue in Indonesia, particularly in Central Java Province. In 2023, it contributed significantly to public health concerns due to its increasing prevalence. This study utilizes the Random Forest algorithm to identify dominant diabetes risk factors and map their geographic distribution across the region. The analysis incorporates community lifestyle indicators such as the number of sweet drink stores, gyms, sports halls, transportation availability, karaoke venues, tourist attractions, fast food outlets, and alcohol stores in each regency/city. Results indicate that the number of tourist attractions (feature importance: 0.2817) and sweet drink stores (feature importance: 0.1502) are the primary global risk factors. Locally, tourism dominates as the key risk factor in 18 regencies/cities (51.4% of the region). The model employs optimal hyperparameter tuning to improve accuracy and utilizes SHAP techniques to evaluate the importance of local features. This study aims to enhance understanding of the key risk factors contributing to diabetes cases and their distribution in Central Java Province.

Keywords: diabetes, risk factors, Random Forest, SHAP, Central Java, non-communicable diseases

1. Introduction

Non-communicable diseases (NCDs) or chronic diseases are among the most significant global health challenges. NCDs encompass various conditions that are not transmitted between individuals but are caused by genetic, physiological, lifestyle, and environmental factors [1]. Major NCDs such as cardiovascular diseases, cancer, chronic respiratory diseases, and diabetes mellitus account for a high number of deaths globally, with an annual toll of 74% of total global deaths or approximately 41 million lives [2].

In Indonesia, NCDs were responsible for 76% of total deaths in 2019, with diabetes mellitus contributing to 7% of those fatalities [3]. Diabetes is a serious chronic condition characterized by elevated blood glucose levels due to the body's inability to produce sufficient insulin or use insulin effectively. Without proper management, diabetes can lead to severe complications such as cardiovascular diseases, kidney failure, and damage to vital organs [4].

The global prevalence of diabetes continues to rise. In 2021, the International Diabetes Federation (IDF) reported approximately 537 million people living with diabetes worldwide, a figure projected to grow to 643 million by 2030 and 783 million by 2045 [5]. Indonesia ranks among the top countries with the highest number of diabetes cases, ranking fifth globally in 2021 with 19.5 million adults affected. This number is expected to increase to 28.6 million by 2045 [5].

The prevalence of diabetes in Central Java also shows an upward trend. In 2021, there were 618,546 diabetes cases in the province, rising to 624,082 cases in 2023 [6] [7]. Although most regencies and cities in Central Java have achieved standardized healthcare services for diabetes patients, this growing trend underscores the need for improved preventive and control measures.

Diabetes mellitus is a chronic disease influenced by numerous risk factors, including unhealthy lifestyle habits such as smoking, physical inactivity, unhealthy diets, and excessive alcohol consumption [8]. These factors lead to physiological changes such as high blood pressure, obesity, elevated blood glucose levels, and high cholesterol, all of which contribute to the onset of diabetes [8].

Psychological conditions, including stress, anxiety, and depression, are also closely associated with NCDs, including diabetes [9]. Given the complexity of these risk factors, identifying dominant risk factors at the local level is essential to support more effective diabetes prevention and control efforts.

This study aims to use data-driven approaches to Identify the key risk factors contributing to diabetes prevalence in Central Java at the global (province-wide) level and analyze the dominant risk factors in each regency/city and city in Central Java.

This study employs the Random Forest algorithm as the primary method. Random Forest is a supervised learning algorithm capable of handling complex data, reducing overfitting, and efficiently determining the importance of each feature [10]. This algorithm is used to calculate feature importance globally, covering the entire Central Java region.

Additionally, this study applies SHAP (SHapley Additive exPlanations) to analyze feature importance locally for each regency/city. SHAP enables a detailed analysis of how each feature influences individual predictions, providing more specific insights[11]. The integration of SHAP with geographical mapping offers a valuable opportunity to visualize the distribution of dominant risk factors across Central Java. This approach provides critical insights into identifying the specific risk factors prevalent in each regency or city, enabling targeted prevention strategies. By focusing on the dominant risk factors within each region, this method ensures that preventive efforts are more precise, effective, and aligned with the unique needs of the local population.

2. Methodology

2.1. *Global Feature Importance*

In this study, we utilize the Random Forest algorithm to identify the dominant risk factors contributing to the prevalence of diabetes in Central Java. Random Forest, first introduced by Leo Breiman [12], is an ensemble learning method based on decision trees. Each tree in the forest relies on a subset of randomly selected variables, and the final prediction is obtained by combining the results from multiple trees [13]. This approach is known for its high accuracy, ability to handle missing data, and capacity to identify important variables [14].

Random Forest works by constructing multiple decision trees, where each tree is trained using a random subset of data and features. The decision tree algorithm makes predictions by recursively splitting the dataset based on feature values. In Random Forest, the choice of features at each node is randomized, meaning only a subset of the features is considered when determining the optimal split. This helps reduce overfitting and increases the robustness of the model [13].

One of the key advantages of Random Forest, as mentioned earlier, is its ability to estimate feature importance, which indicates how much each feature contributes to the model's prediction. In this case, we focus on identifying global feature importance, which measures the overall contribution of each feature to the prediction of diabetes prevalence across Central Java.

To build decision trees, Random Forest employs the Mean Squared Error (MSE) as a measure of impurity when splitting nodes. The goal is to find the optimal split at each node, minimizing the MSE. The formula for MSE is as follows in Eq.(1).

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \qquad (1)$$

N is the number of data points in the node, y_i is the actual value of the i-th data point, and $\hat{y}_i$ is the predicted value for the i-th data point[15].

Each time a feature is used to split the data at a node, the impurity (MSE for regression) is reduced. This reduction in impurity is accumulated for each feature across all nodes and decision trees in the Random Forest. The final feature importance is determined by averaging the accumulated impurity reduction for each feature across all trees in the forest. Features that result in the greatest reduction in impurity are considered the most important for predicting the target variable (in this case, diabetes prevalence in Central Java).

2.2. *Local Feature Importance*

In this study, we utilize SHAP (SHapley Additive exPlanations) to identify and explain the contribution of each feature towards the predictions made by the Random Forest model. SHAP is a technique used to provide a transparent explanation of machine learning model predictions by calculating the Shapley value for each feature, which represents its contribution to the model's prediction for each individual instance (data point). This local explanation provides insights into the specific impact of each feature on individual predictions, offering a deeper understanding of the model's decision-making process for each sample in the dataset [11]. The formula for the Shapley value for a feature ***j*** as follows in Eq.(2).

$$Shapley(X_j) = \sum_{S \subseteq N \setminus \{j\}} \frac{k!\,(p-k-1)!}{p!}\left(f(S \cup \{j\}) - f(s)\right) \qquad (2)$$

Where p represents the total number of features, $N\setminus\{j\}$ denotes the set of all possible feature combinations excluding feature X_j, S is any subset of features within $N\setminus\{j\}$, $f(S)$ is the model's prediction using only the features in S, and $f(S \cup \{j\})$ refers to the model's prediction when the feature X_j is added to the feature set S [16].

Once the global feature importance has been determined, the next step is to calculate the local feature importance for each regency/city in Central Java using SHAP. This process is as follows:

1. The SHAP explainer is initialized using the trained Random Forest model and the selected feature dataset.
2. The explainer computes the SHAP values, which represent the contribution of each feature to the model's prediction for every individual sample in the dataset (each regency/city).
3. For each sample (regency/city), the SHAP value is computed for every feature, and the features with the largest SHAP values are identified as the dominant risk factors for that specific prediction.
4. After determining the dominant risk factors for each regency/city, the frequency of each feature being the dominant risk factor across all regencies/cities is calculated.

2.3. *Dataset and Variables*

The dataset used in this study includes lifestyle factors of the population, divided into two types: primary data and secondary data. The primary data consists of the number of risk factors present in each regency and city in Central Java, including alcohol stores, sweet drink

outlets (such as boba and iced tea), fast food outlets, tourist attractions, fitness facilities (such as sports halls and gyms), public transportation facilities (such as terminals and bus stops), and entertainment venues (such as karaoke). This data was obtained through a scraping process on Google Maps using a tool called Instant Data Scraper Extension. The scraping process was conducted in September 2024. The secondary data used in this study includes the number of diabetes patients in Central Java as of December 2023, which was sourced from the Health Department of Central Java. By incorporating these lifestyle factors, the study offers a more localized and contextual approach to understanding the contribution of lifestyle factors to diabetes prevalence in the region.

2.4. *Model Evaluation and Geo-mapping*

To evaluate the performance of the model, two evaluation metrics were used: R-squared (R^2) and Root Mean Square Error (RMSE). RMSE measures the deviation of the model's predictions from the actual values. A lower RMSE value indicates better model performance. On the other hand, R^2 assesses how well the model explains the variance in the target data. The R^2 value ranges from 0 to 1, where a value closer to 1 indicates better explanatory power. R^2 can also be negative if the model performs poorly. The formulas used for R^2 and RMSE are provided below in Eqs. (3) and (4).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{3}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4}$$

Where n is the number of observations, y_i is the actual value for observation $\hat{y}_i$ is the predicted value for observation $\bar{y}_i$ is the mean of the actual values [17].

For geo-mapping, the goal is to provide a clearer visual representation of the contribution of each feature and its distribution across regencies and cities in Central Java. This process began by utilizing spatial data of Central Java from GeoJSON files. The data was filtered to focus only on Central Java, removing irrelevant entities such as "Waduk Kedungombo" and "WaterBody." The spatial data was then merged with SHAP data, which includes the dominant risk factors for each regency/city.

Before merging, the names of the regencies/cities in both datasets (primary and secondary dataset) were standardized to ensure correct integration. Once the data was successfully merged, a thematic map was created to display the distribution of dominant risk factors across regencies and cities. A unique color was assigned to each category of risk factors using a color map. The data was plotted on the map, with colors representing the dominant risk factor for each regency/city. The names of the regencies/cities were added as labels for easier identification. Finally, an analysis of the thematic map provided additional insights into the geographic distribution of dominant risk factors, which can inform targeted intervention strategies.

3. Results and Discussions

3.1. *Simulations settings*

This section outlines the procedure for constructing a Random Forest model to predict diabetes cases. Initially, the dataset is loaded and preprocessed, which includes handling outliers and applying logarithmic transformations to the target variable in order to mitigate the influence of extreme values. Feature selection is then performed using SelectKBest, retaining only the top k features with the most significant linear correlation to the target variable.

Subsequently, hyperparameter tuning is conducted via GridSearchCV to determine the optimal number of estimators (n_estimators). During this process, 5-fold cross-validation is employed, with R^2 as the evaluation metric. Once the best parameter combination is identified, the model is trained using the full dataset and the optimal number of decision trees.

In the Random Forest algorithm, each decision tree is constructed using bootstrap sampling, automatically handled by the algorithm. Node splitting is based on Mean Squared Error (MSE) for regression tasks, with the best split being the one that yields the highest variance reduction. Upon completion of all decision trees, their individual predictions are aggregated to generate the final model prediction. The model's performance is evaluated using RMSE and R^2, and feature importance is calculated to identify the most influential features in predicting diabetes.

The entire process was executed in Google Collaboratory, a cloud-based platform, using Python as the programming language. The system configuration included 12 GB of RAM, provided by the free-tier account, and an NVIDIA Tesla T4 GPU, which facilitated the acceleration of computations, ensuring efficient data processing and model training.

3.2. *Performance and Analysis*

In this research, we evaluated two main metrics which are R^2 and RMSE. We use K-fold cross validation with average RMSE as the main metric to measure the Random Forest model performance. This evaluation aims to give us an understanding of the hyperparameter tuning to the model performance. We tune the hyperparameter of n_estimator to get the best optimal combination on the model with GridSearchCV. Table 1 shows the evaluation result with various n_estimator with value range from 1 to 201 with the K-fold determined to 3, 5, and 7.

Table 1. Hyperparameter tuning and its impact on the performance of the Random Forest model.

n_estimators range	Best n_estimators	R^2	RMSE	CV RMSE (Avg)
CV (folds)=3				
1-51	50	0.85	0.22	0.62±0.11
51-101	85	0.86	0.21	0.57±0.19
101-151	126	0.86	0.21	0.63±0.09
151-201	156	0.86	0.21	0.63±0.10
CV (folds)=5				
1-51	50	0.85	0.22	0.58±0.09
51-101	85	0.86	0.21	0.57±0.10
101-151	**126**	**0.86**	**0.21**	**0.56±0.10**
151-201	156	0.86	0.21	0.57±0.10
CV (folds)=7				
1-51	50	0.85	0.22	0.60±0.14
51-101	85	0.86	0.21	0.60±0.14
101-151	126	0.86	0.21	0.59±0.14
151-201	156	0.86	0.21	0.59±0.14

The results in Table 1, as shown in the highlighted row, utilized an n_estimators value of 126, resulting in an RMSE of 0.21, an R^2 score of 0.86, and the lowest CV RMSE (Avg) of 0.56 with a standard deviation of 0.10, which delivered the best evaluation results.

The best-performing model was subsequently selected for use in the global feature importance analysis in Central Java. Below are the results of the global feature importance obtained using the Random Forest model.

Table 2. Global feature importance value.

Feature	Importance Value
tourism	0.2817
sweet_drinks	0.1502
gym	0.1378
fast_food	0.1368
sport_hall	0.0833
karaoke	0.0831
transportation	0.0730
alcohol_stores	0.0541

Table 2. highlights the global feature importance for diabetes incidence in Central Java. The "tourist_attraction" feature has the highest importance value (0.2817), followed by "sweet_beverages" (0.1502), and "gym" and "fast_food," which have nearly equivalent values (0.1378 and 0.1368). While the other features demonstrate lower contributions.

Following the global feature importance analysis, local feature importance was assessed using SHAP (SHapley Additive Explanations) to identify the dominant risk factors for each regency and city in Central Java. These factors were then mapped to visualize their distribution. The feature importance values were further illustrated on a map of regencies and cities in Central Java using GeoJSON. This map highlights the prevalence of diabetes diseases cases across the regions, with color intensity representing the distribution of significant risk factors.

The visualization shown in Fig.1 results reveal that the "tourism" feature emerges as the most dominant risk factor, identified in 18 regencies and cities across Central Java, making it the most frequently observed factor both locally and globally. In the global feature importance analysis using Random Forest, it holds the highest importance value of 0.2817, underscoring its significant contribution to diabetes incidence. Locally, its dominance in regions such as Batang, Blora, Boyolali, Brebes, Demak, Kendal, Klaten, Kudus, Magelang, Pati, Pekalongan, Purworejo, Rembang, Sragen, Sukoharjo, Temanggung, and Semarang city aligns with its global significance. The dual role of tourism is evident: while tourist activities can promote physical movement (e.g., walking), they also foster unhealthy dietary habits, with increased availability of sugary drinks, snacks, and fast food around tourist attractions. This highlights the intricate relationship between tourism and public health, making it a critical factor in diabetes prevention strategies.

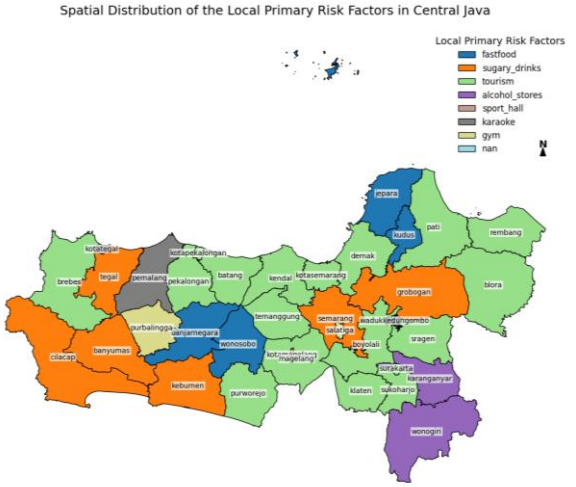

Fig.1 Spatial distribution of local primary risk factors in central java.

The "sweet_drinks" feature, although second in global importance (value: 0.1502), is dominant in only 6 regions, including Banyumas, Cilacap, Grobogan, Kebumen, Tegal, and Semarang. This suggests that while sweet drinks significantly impact diabetes incidence globally, their influence is geographically concentrated. Localized habits, such as a preference for sugary beverages in these regions, may amplify their impact. This discrepancy between global and local significance highlights the need for region-specific interventions to address varying consumption patterns and mitigate the influence of sweet drinks on diabetes.

The "gym" feature ranks third in global importance with a value of 0.1378 and is dominant in 2 regions: Purbalingga and Salatiga. While gyms symbolize access to fitness infrastructure, their limited accessibility to the broader population or unhealthy post-exercise dietary choices may mitigate their potential health benefits. Additionally, gyms tend to attract individuals who are already conscious of health, leaving larger populations unaffected. This explains why gym dominance is localized despite its relatively high global importance.

The “fast_food” feature ranks fourth in global importance with a value of 0.1368 but is dominant in just 4 regions, such as Banjarnegara, Jepara, Kudus, and Wonosobo. This uneven distribution reflects the variable prevalence of fast-food outlets, with higher concentrations in urban or semi-urban areas. Globally, fast food contributes significantly to diabetes due to its high-calorie and low-nutrient content. However, its limited local dominance suggests that other factors may play a larger role in regions with fewer fast-food establishments. This points to the importance of tailored strategies, such as promoting healthy eating habits alongside regulating fast-food expansion.

The “sports_hall” feature is identified as a dominant factor in only one region, contributing a mere 2.9% to the overall distribution. This limited dominance signifies its minimal impact on diabetes incidence at a global level, reflecting its relatively small influence on the overall variance.

In contrast, the “karaoke” feature, despite its lower global importance, is dominant in two regions—Pemalang and Tegal City. Karaoke establishments may indirectly increase diabetes risk through associated behaviors, such as stress relief activities that involve excessive alcohol consumption and smoking, particularly in venues equipped with bar facilities. These lifestyle factors may overshadow the potential stress-reducing benefits of karaoke, rendering it a locally significant yet globally minor contributor to diabetes prevalence.

Similarly, the “alcohol_store” feature holds a lower global importance value of 0.0541 but emerges as a dominant factor in two regions—Karanganyar and Wonogiri. While its contribution to diabetes is modest on a global scale, its local significance underscores the relationship between the prevalence of alcohol stores and increased alcohol consumption in these specific areas.

4. Conclusion

In this study, we employed a Random Forest (RF) model with 126 n_estimators to identify the key risk factors contributing to the prevalence of diabetes in Central Java. The feature importance analysis revealed that “tourism_spot” and “sweet_drinks” are the most significant risk factors, with “tourism_spot” having the greatest impact on diabetes incidence. The SHAP (Shapley Additive Explanations) analysis provided a deeper understanding of the local distribution of these risk factors, showing that “tourism_spot” dominates in 18 regions, while “sweet_drinks” is dominant in 6 regions. Model evaluation results indicated strong predictive accuracy, with an RMSE of 0.21, an R^2 score of 0.86, and the lowest CV RMSE (Avg) of 0.56 with a standard deviation of 0.10, delivering the best evaluation outcomes.

The visualization of dominant risk factors across regencies and cities in Central Java further highlights regional variations in the prevalence of diabetes risk factors. The map predominantly features green areas, indicating “tourism_spot” as the leading risk factor, influenced by both increased physical activity and unhealthy consumption patterns at tourist sites. Additionally, areas marked in orange, such as Banyumas, Cilacap, and Grobogan, suggest that “sweet_drinks” consumption is a dominant risk factor, underlining its significant contribution to diabetes risk. These findings emphasize the complex interplay of lifestyle and environmental factors in shaping diabetes risk across different regions.

Acknowledgments

This work was supported by Telkom University.

References

1. A. Budreviciute, S. Damiati, D. K. Sabir, K. Onder, P. Schuller-Goetzburg, G. Plakys, A. Katileviciute, S. Khoja, and R. Kodzius, "Management and Prevention Strategies for Non-communicable Diseases (NCDs) and Their Risk Factors," *Frontiers in Public Health*, vol. 8, 2020.
2. K. R. Thankappan and G. K. Mini, "Noncommunicable Diseases in the Elderly," in Handbook of Aging, Health and Public Policy: Perspectives from Asia, Singapore: Springer Nature Singapore, 2022, pp. 1-9.
3. World Health Organization, "Noncommunicable Diseases (NCD) Portal," *World Health Organization*. [Online]. Available: https://ncdportal.org/. [Accessed: May 6, 2024].
4. D. J. Magliano, E. J. Boyko, and IDF Diabetes Atlas 10th edition scientific committee, *IDF Diabetes Atlas*, 10th ed. Brussels: International Diabetes Federation, 2021, ch. 1, "What is diabetes?" [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK581938/. [Accessed: May 6, 2024].
5. International Diabetes Federation (IDF), "IDF Diabetes Atlas, 10th ed.," Brussels, Belgium, 2021. [Online]. Available: https://www.diabetesatlas.org. [Accessed: May 6, 2024].
6. Dinkes Jateng, "Profil Kesehatan Jateng 2021," *dinkesjatengprov.go.id*, 2021. [Online]. Available: https://dinkesjatengprov.go.id/v2018/dokumen/Profil_Kesehatan_2021/mobile/index.html. [Accessed: May 7, 2024].
7. Dinkes Jateng, "Profil Kesehatan Jateng 2023," *dinkesjatengprov.go.id*, 2020. [Online]. Available: https://dinkesjatengprov.go.id/v2018/dokumen/1Profil_Kesehatan_2023/mobile/index.html. [Accessed: Oct. 4, 2024].
8. World Health Organization, "Global Health Observatory (GHO) data: Noncommunicable diseases (NCD) risk factors," [Online]. Available: https://www.who.int/data/gho/data/themes/topics/topic-details/GHO/ncd-risk-factors. [Accessed: May 7, 2024].
9. X. Liu, H. Cao, H. Zhu, H. Zhang, K. Niu, N. Tang, et al., "Association of chronic diseases with depression, anxiety and stress in Chinese general population: The CHCN-BTH cohort study," *Journal of Affective Disorders*, vol. 282, pp. 1278-1287, 2021.
10. Q. Zhou, W. Lan, Y. Zhou, and G. Mo, "Effectiveness Evaluation of Anti-bird Devices based on Random Forest Algorithm," 2020 7th International Conference on Information, Cybernetics, and Computational Social

Systems (ICCSS), Guangzhou, China, 2020, pp. 743-748, doi: 10.1109/ICCSS52145.2020.9336891.
11. T. T. H. Le, H. Kim, H. Kang, and H. Kim, "Classification and explanation for intrusion detection system based on ensemble trees and SHAP method," *Sensors*, vol. 22, no. 3, p. 1154, 2022.
12. L. Breiman, "The Random Forest," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.
13. T. Zhu, "Analysis on the applicability of the random forest," in *Journal of Physics: Conference Series*, vol. 1607, no. 1, p. 012123, Aug. 2020, IOP Publishing.
14. F. R. Aszhari, Z. Rustam, F. Subroto, and A. S. Semendawai, "Classification of thalassemia data using random forest algorithm," in *Journal of Physics: Conference Series*, vol. 1490, no. 1, p. 012050, Mar. 2020, IOP Publishing.
15. M. T. Islam, M. Raihan, N. Aktar, M. S. Alam, R. R. Ema, and T. Islam, "Diabetes Mellitus Prediction using Different Ensemble Machine Learning Approaches," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, pp. 1-7, 2020.
16. Z. Li, "Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost," *Computers, Environment and Urban Systems*, vol. 96, p. 101845, 2022.
17. G. Grekousis, Z. Feng, I. Marakakis, Y. Lu, and R. Wang, "Ranking the importance of demographic, socioeconomic, and underlying health factors on US COVID-19 deaths: A geographical random forest approach," *Health & Place*, vol. 74, Art. no. 102744, 2022.

Authors Introduction

Ms. Ambar Arum Prameswari

She is currently pursuing a Bachelor of Informatics Engineering at Telkom University Purwokerto, Indonesia. Her research interests focus on machine learning and data analysis.

Dr. Andi Prademon Yunus

He is an Assistant Professor at Telkom University and he received his PhD in Engineering from Mie Unviersity, Japan. His research focuses on applied and fundamental machine learning for motion and behavior computing. He also collaborates with industry partners to develop AI-based tools for language modeling and image analytics.