SMARTICS Journal, Vol.12 No.01 2026. p36-43
ISSN online: 2476-9754, ISSN print: 2623-0429
DOI : https://doi.org/10.21067/smartics.v12i1.13919

Prediction Of Theft Security Level Using Naive Bayes In West Merapi Police Sector Area

Sutria Rahmi[a], Indah Permatasari[b], Evi Purnamasari[c*]
[a,b,c] Universitas Indo Global Mandiri
[*]correspondence : evi.ps@uigm.ac.id

Abstrak*—Theft remains one of the most prevalent forms of criminal activity in the jurisdiction of the West Merapi Police Sector, significantly impacting public safety and community security. Historically, the handling and securing of this region has been reactive in nature, lacking a predictive system capable of estimating theft risks preventively. This study aims to develop a predictive model for regional security levels related to theft cases using a machine learning approach. The data utilized in this research comprises secondary data obtained from 459 theft case reports documented by the West Merapi Police Sector from 2021 to 2024. Ten relevant variables were selected as features, while three security level categories (Low, Medium, and High) served as target classes. Data preprocessing included data cleaning, variable transformation, and label encoding. The Naive Bayes algorithm was employed with a 70% training data and 30% testing data split. The results demonstrated that the Naive Bayes method achieved an accuracy of 76.09% in predicting regional security levels. The model exhibited optimal performance for the High security level class, while the Low class showed lower performance due to imbalanced data distribution. This research demonstrates that police case report data can be effectively utilized to support data-driven risk analysis and has the potential to serve as a decision-making tool for preventive measures by law enforcement agencies.

***Index Terms*—Naive Bayes; prediction; security level; theft; machine learning.**

Abstrak*-Pencurian merupakan salah satu bentuk tindak kriminal yang masih sering terjadi di wilayah hukum Polsek Merapi Barat dan berpengaruh terhadap tingkat keamanan masyarakat. Selama ini, proses penanganan dan pengamanan wilayah bersifat reaktif dan belum didukung oleh sistem prediksi yang mampu memperkirakan tingkat risiko pencurian secara preventif. Penelitian ini bertujuan untuk membangun model prediksi tingkat keamanan wilayah terhadap kasus pencurian menggunakan pendekatan machine learning. Data yang digunakan merupakan data sekunder dari 459 laporan kasus pencurian Polsek Merapi Barat periode 2021-2024. Sepuluh variabel relevan dipilih sebagai fitur dan tiga kategori tingkat keamanan (Rendah, Sedang, Tinggi) digunakan sebagai kelas target. Algoritma Naive Bayes digunakan dengan pembagian data latih 70% dan data uji 30%. Hasil pengujian menunjukkan akurasi sebesar 76,09%. Model menunjukkan performa terbaik pada kelas tingkat keamanan Tinggi, sedangkan kelas Rendah memiliki performa lebih rendah akibat distribusi data tidak seimbang.

***Kata Kunci*—Naive Bayes; prediksi; tingkat keamanan; pencurian; machine learning.**

I. INTRODUCTION

Theft crime constitutes one of the most common forms of criminal offense used to assess the security level of a region. The fluctuation in theft cases can reflect the effectiveness of security systems implemented and the extent to which regional security is maintained[1]. Well-maintained security is essential for creating a safe, comfortable, and prosperous environment for the community. Conversely, disturbances to security stability, such as increasing theft cases, can affect various aspects of community life, from the comfort of daily activities to the overall decline in quality of life[2].

The number of theft cases occurring in a region becomes an important indicator in assessing its security level. Theft is often used as a reference to determine whether an area is safe or vulnerable to crime. When theft rates are high, communities tend to feel threatened and anxious, which ultimately can decrease their quality of life[3]. The discomfort caused by this insecurity affects many aspects of life, from disrupted social interactions to declining economic development in the area[4]. Communities living in insecurity tend to have difficulty conducting their activities normally, which can hinder the creation of an environment that supports progress[5].

Several factors influence the occurrence of theft in a region, including the economic condition of the local community. When unemployment rates are high and employment opportunities are limited, individuals often feel compelled to commit crimes such as theft to meet their living needs. Additionally, low education levels also contribute to increasing crime rates, as less educated individuals have more limited alternatives for better livelihoods. Other factors contributing to theft include weak law enforcement, social inequality, and inadequate security facilities in an area[6].

The West Merapi Police Sector jurisdiction is no exception to theft cases, with many reports received regarding such incidents. To optimize existing resources for maintaining order and protecting the community from criminal acts, a system capable of processing and analyzing data efficiently is needed. One solution that can be used is data analysis technology capable of predicting theft levels. However, until now, the West Merapi Police Sector in Lahat Regency, South Sumatra, still faces limitations in data analysis technology for predicting theft levels in the region. The unavailability of this technology causes difficulties for local police in identifying patterns and trends of crime, which impacts their difficulty in anticipating potential security threats[7].

The main problem underlying this research is the inability of existing systems to identify patterns and trends in theft cases effectively. Without appropriate data analysis technology, police face difficulties in predicting and anticipating potential crimes that may occur, hindering more optimal prevention and countermeasures. Therefore, this research aims to develop a prediction system using the Naive Bayes method that can analyze historical data of theft cases. With this system, it is expected that police can map areas and times with high theft risk and take more targeted and effective preventive measures[8].

Previous research by Kurniawan et al.[9] showed that this method can detect increases in criminality with fairly high accuracy, especially in cases of aggravated theft. However, this method also has weaknesses, being less sensitive in detecting decreases in crime rates, with lower recall values. Other research by Marbun and Prianggono [10] showed that the K-Nearest Neighbors (KNN) algorithm provides superior results compared to Naive Bayes in terms of accuracy, precision, and recall, and is more effective in recognizing theft patterns based on location, time, and type of incident. However, although KNN shows better performance, this research still chose the Naive Bayes method because of its simplicity and efficiency in handling more limited data, especially at the local level which has limited historical data.

II. RESEARCH METHOD

This research uses quantitative methods with an experimental approach through the application of Machine Learning algorithms to classify regional security levels based on historical data of theft case reports. This approach was chosen because the research aims to produce a predictive model that can map village security levels into three categories: Low, Medium, and High based on previous data patterns. The research stages were conducted sequentially and systematically from data collection to model evaluation as shown in Fig. 1.

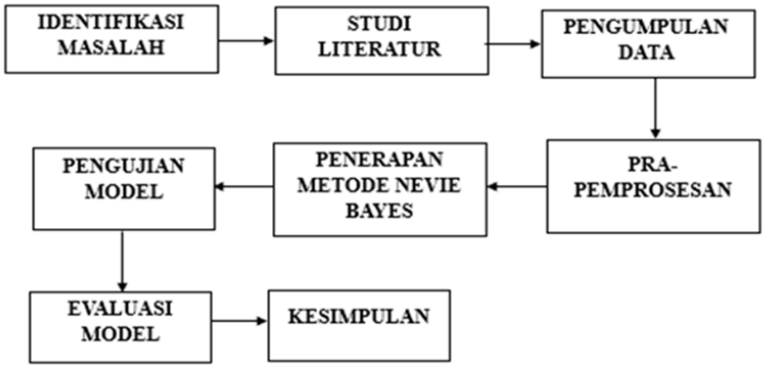

Fig. 1. Research stages diagram

A. Data Collection

The type of data used in this research is secondary data obtained through documentation studies of theft case report archives in the jurisdiction of West Merapi Police Sector, Lahat Regency, during the period 2021 to 2024. The data collected includes various important information such as report dates, incident times, locations, types of theft, identities of victims or reporters, and incident status. The collected data reflects factors that can affect security levels and will be analyzed to provide more accurate predictions about potential crimes in the region. The total data used in this research amounts to 459 cases after going through the data preprocessing process with 10 variables used.

B. Research Stages

The research stages in Fig. 1 are described as follows:

1. Data Collection: This initial stage involves gathering all necessary data as well as preparing hardware and software to be used for analysis. The data is secondary data obtained from several theft case report files in Excel format from the West Merapi Police Sector.
2. Data Integration: The datasets from multiple years are combined into a single dataset so that the data used is more complete and integrated. This merging process is performed before further data processing stages.
3. Data Preprocessing: This stage is conducted to ensure data is in clean and ready-to-use condition. The process includes checking for empty values, filling in missing data, and transforming categorical data into numerical form. Additionally, separation between feature data and target data is performed so that the modeling process can be carried out properly.
4. Data Splitting: The preprocessed data is then divided into training data and testing data. Training data is used to build a predictive model, while testing data is used to measure the model's ability to make predictions on data it has never seen before. The division uses a 70:30 ratio with 321 training data and 138 testing data.
5. Model Training: At this stage, the model training process is conducted using the Naive Bayes method. The model is trained using training data to learn the relationship patterns between data features and target classes based on probability calculations.
6. Prediction and Evaluation: The trained model is then used to make predictions on testing data. The prediction results are then evaluated to determine the model's performance in predicting security levels. Evaluation is conducted using metrics such as accuracy and classification report.
7. Results Analysis: The results analysis stage is conducted by interpreting the evaluation values obtained. This analysis aims to determine the success level of the method used and discuss the prediction results produced by the model.

C. Research Variables

The research variables are divided into two categories: input variables (X) and output variables (Y). The input variables consist of eight attributes that influence security levels, including Report Date, Incident Date, Incident Time, Village Name, Type of Theft Method, Stolen Items, Estimated Loss, and Case Status. The output variable is the Security Level with three categories: Low, Medium, and High. Table I presents the distribution of target variable classes.

Table I. Distribution Of Target Variable Classes

Class	Number of Samples	Percentage
High (3)	235	51.2%
Medium (2)	152	33.1%
Low (1)	72	15.7%
Total	**459**	**100%**

D. Naive Bayes Method

Naive Bayes is a classification method based on Bayes' Theorem developed by British scientist Thomas Bayes. This method is used to predict the probability of a class based on available features with the assumption that each feature is independent of one another[11]. This method is well-known in text classification and other applications due to its simplicity, speed, and ability to provide good results even with limited data [12]. The Bayes' Theorem formula is as follows:

$$P(C|X) = P(X|C) \times P(C) / P(X)$$

Where X is the feature or variable from data with unknown cases, C is the hypothesis about case data X to be predicted, P(C) is the probability of hypothesis C (prior probability), P(X) is the initial probability for each case data, P(C|X) is the probability of hypothesis C based on condition X (posterior probability), and P(X|C) is the probability of X based on condition in hypothesis C. The main advantage of Naive Bayes is its ability to handle large datasets quickly. The model training process involves calculating the frequency of feature occurrences in each class and calculating their probabilities[13].

E. Data Preprocessing

Data preprocessing is a crucial stage in this research to ensure that the data used is in good, clean, and consistent condition before being used in the modeling process. This stage is conducted to improve data quality so that the prediction results obtained can be more accurate. The data preprocessing process includes data cleaning, handling missing values, duplicate data, and invalid data. Additionally, categorical variables such as report time, village name, theft method, and stolen items need to be converted into numerical form so they can be processed further. This conversion process is done using categorical data transformation techniques such as label encoding, so that each category is represented in numerical form[14].

Furthermore, the loss amount variable which was initially in text form is normalized into numerical form so it can be used as one of the attributes in the prediction process. Meanwhile, date and time report information is transformed into certain categories to facilitate the model learning process in recognizing incident patterns based on time. After all data preprocessing stages are completed, the clean and structured data is ready to be used in the predictive model building process using the Naive Bayes method to produce more optimal security level predictions[15].

III. RESULTS AND DISCUSSION

The Naive Bayes method was applied to the preprocessed dataset consisting of 459 data with 10 attributes. The data was divided into 321 training data and 138 testing data. The training data was used to build a classification model with security level as the target variable, while other attributes served as predictor variables. The model training process was successfully executed and produced a well-formed Naive Bayes model without obstacles at the execution stage. The success of this process indicates that the data used has met the requirements of the applied method.

After the Naive Bayes model was built using training data, the next stage was testing the model using testing data. This stage aims to determine the model's ability in predicting regional security levels based on data never seen before. Testing was conducted by providing testing data and observing the correspondence between the predicted labels produced and the actual labels in the data. The model testing results are presented in Table II.

Table II. Model Testing Results

Class	Precision (%)	Recall (%)	F1-Score (%)	Support (%)
Low	88%	52%	65%	29%
Medium	75%	61%	68%	44%
High	74%	97%	84%	65%
Accuracy	**76.09%**	-	-	-

Based on the evaluation results in Table II, the Naive Bayes method achieved an accuracy of 76.09% in predicting security levels. This accuracy value indicates that the majority of testing data can be correctly classified by the model. For the Low security level class, precision of 88%, recall of 52%, and F1-score of 65% were obtained. The relatively high precision value indicates that the model's predictions for the Low class are quite accurate. However, the still low recall value indicates that not all data actually belonging to the Low class were successfully identified by the model.

For the Medium security level class, the model produced precision of 75%, recall of 61%, and F1-score of 68%. These results indicate that the model's performance in the Medium class is at a fairly good level and more balanced compared to the Low class. Meanwhile, for the High security level class, the model showed the most optimal performance with precision of 74%, recall of 97%, and F1-score of 84%. The very high recall value indicates that almost all data in the High security level class can be recognized well by the model.

A. Confusion Matrix Analysis

Confusion matrix is used to describe the prediction results of the Naive Bayes model in more detail in grouping the security levels of theft cases. Through this confusion matrix, the amount of data predicted correctly and incorrectly classified in each security level category can be seen. This information helps in evaluating model performance for each class as shown in Fig. 2.

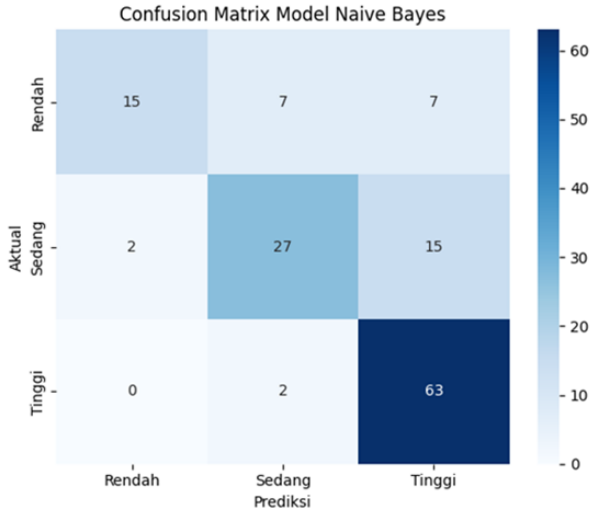

Fig. 2. Confusion matrix of Naive Bayes model prediction results

In Fig. 2, rows represent the actual classes, while columns show the prediction results produced by the Naive Bayes model. Values located on the diagonal indicate the amount of data successfully classified correctly, while values outside the diagonal indicate errors in the classification process. For the Low security level category, the model was able to correctly classify 15 data. However, there were 7 data incorrectly predicted as Medium class and 7 other data as High class. This condition indicates that the model still experiences difficulty in distinguishing data in the Low class, especially due to pattern similarities with other security level classes.

For the Medium security level category, the number of data successfully predicted correctly was 27 data. Additionally, there were 2 data incorrectly predicted as Low class and 15 data wrongly classified as High class. The relatively large prediction errors in this class indicate that data characteristics in the Medium security level are relatively overlapping, making it difficult for the model to recognize patterns consistently. Meanwhile, the best performance was shown in the High security level category, where 63 data were successfully predicted correctly. The errors occurring in this class were relatively small, with only 2 data predicted as Medium class and no data predicted as Low class. This indicates that data patterns in the High security level class are clearer and more stable, making them easier to learn and recognize by the Naive Bayes model.

B. Model Performance Distribution

The distribution of model performance across different security level classes can be visualized to better understand how the Naive Bayes model performs for each category. Fig. 3 shows the performance metrics comparison between precision, recall, and F1-score for each class.

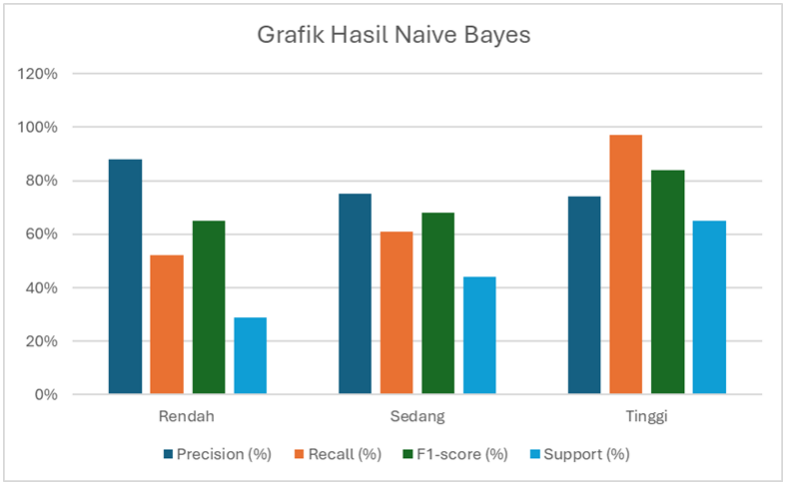

Fig. 3. Model performance distribution across security level classes

Fig. 3 shows that model performance varies for each class. In the High class, the recall value appears highest compared to other classes, reaching 97%, which means the model can recognize almost all actual data in that class correctly. The precision and F1-score values in this class also show stable performance, each at around 74% and 84%. This indicates that the model is more optimal in detecting the High class, caused by larger support data compared to other classes, so the model obtains more representative information during the training process.

Conversely, model performance in the Low class shows lower recall values at around 52%, although the precision value in that class is quite high at 88%. This condition indicates that although most of the model's predictions for the Low class are correct, the model did not successfully detect all actual data belonging to that class. In the Medium class, all three metrics are at medium levels, namely precision 75%, recall 61%, and F1-score 68%, reflecting the characteristics of the transition class that has feature similarities with both other classes. Overall, this graph shows that the model works better on classes with larger amounts of data, and is more challenging on classes with smaller amounts of data.

The class imbalance in the dataset significantly affects the model's performance. The High class, being the majority class with 51.2% of the data, achieves the best performance because the model has more representative information during the training process. Conversely, the Low class, being the minority class with only 15.7% of the data, shows lower recall because the model has difficulty learning patterns from limited examples. This phenomenon is common in multi-class classification with imbalanced data distribution and can be addressed through techniques such as oversampling, undersampling, or SMOTE in future research.

The findings indicate that Naive Bayes is suitable for use as an initial prediction method for security levels because it can capture probabilistic patterns in security report data. However, this research also indicates opportunities for performance improvement through several approaches such as class balancing (oversampling, undersampling, or SMOTE), additional feature extraction, or selection of methods with class weighting mechanisms. Thus, this research contributes to providing evidence that security level prediction can be done automatically with support from historical criminal incident data and machine learning methods.

IV. CONCLUSION

Based on the research objectives and testing results conducted, several conclusions can be drawn as follows: First, the Naive Bayes model built was able to produce predictions of regional security levels

with fairly good performance. The test results showed that the model achieved an accuracy of 76.09% with the best performance in the High security level class. This finding indicates that the model is able to identify regions with high risk levels based on historical data patterns.

Second, theft case report data can be used as an information basis for building a regional security level prediction model. The secondary data obtained from the West Merapi Police Sector covers 459 report entries in the 2021-2024 period and contains relevant variables in describing the characteristics of theft incidents. This data has been proven to be computationally processed after going through the preprocessing stage.

Third, the data preprocessing stage plays an important role in preparing the dataset for modeling needs. The cleaning process, format alignment, variable transformation, and security level label determination produced a dataset with 10 main variables that can be used as input for the classification model. Fourth, the uneven data distribution between classes affects the classification results. The Low class has the least amount of data so it produces lower prediction performance compared to the Medium and High classes. This condition indicates that dataset characteristics affect model stability.

For future research, it is recommended to compare with other classification methods to determine the most suitable method for predicting security levels based on theft case data. Additionally, the results of this research can be further developed into a decision support system that can help police monitor and predict regional security levels more effectively.

V. ACKNOWLEDGMENT

The authors would like to express sincere gratitude to the West Merapi Police Sector, Lahat Regency, South Sumatra, for providing the theft case report data that made this research possible. We also extend our appreciation to the Study Program of Informatics Engineering, Faculty of Computer Science, Indo Global Mandiri University, for the academic support and facilities provided throughout this research.

VI. REFERENCES

[1] M. S. Alfandi and Z. Fatah, “Penerapan Data Mining Menggunakan Metode K-Means Clustering Untuk Analisa Penjualan Toko Umama Hijab Kaliwates Jember,” *J. JISSI (Jurnal Ris. Sist. Informasi)*, vol. 1, no. 4, pp. 94–102, 2024, [Online]. Available: https://journal.smartpublisher.id/index.php/jissi

[2] Drs. Imron Rosyadi, SH.,MH et al., *Victim Precipitation dalam Tindak Pidana Pencurian (Sebuah Pendekatan Viktimologi)*. 2020.

[3] indah permata Sari, “Cultural Aspects Influencing the Application of E- Learning : A Literature Review Cultural Aspects Influencing the Application of E-Learning : A Literature Review,” 2019, doi: 10.1088/1742-6596/1235/1/012026.

[4] D. K. Kadali, R. N. V. J. Mohan, M. C. Naik, and Y. Bokka, “Crime Data Analysis Using Naive Bayes Classification and Least Square Estimation with MapReduce,” *Int. J. Comput. Methods Exp. Meas.*, vol. 12, no. 3, pp. 289–295, 2024, doi: 10.18280/ijcmem.120309.

[5] Z. Erico and M. Yasin, “Pemberdayaan Keamanan Dan Kesejahteraan Melalui Penggunaan Cctv, Digital Marketing, Website Desa Yang Interaktif,” *Pros. Patriot Mengabdi*, vol. 2, pp. 36–42, 2023, [Online]. Available: https://conference.untag-sby.ac.id/index.php/spm/article/view/2874%0Ahttps://conference.untag-sby.ac.id/index.php/spm/article/download/2874/1600

[6] M. N. Fadillah and R. N. Sukmana, “Penerapan Algoritma Naïve Bayes Classifier Untuk Analisis Sentimen Kelangkaan Minyak Goreng Pada Media Sosial Twitter,” *Infotronik J. Teknol. Inf. dan Elektron.*, vol. 7, no. 2, p. 82, 2022, doi: 10.32897/infotronik.2022.7.2.1716.

[7] U. Ghani, P. Toth, and D. Fekete, “Incorporating Survey Perceptions of Public Safety and Security Variables in Crime Rate Analyses for the Visegrád Group (V4) Countries of Central Europe,” *Societies*, vol. 12, no. 6, 2022, doi: 10.3390/soc12060156.

[8] H. Hamdiyah, “Analisis Unsur-Unsur Tindak Pidana Pencurian: Tinjauan Hukum,” *J. Tahqiqa J. Ilm. Pemikir. Huk. Islam*, vol. 18, no. 1, pp. 98–108, 2024, doi: 10.61393/tahqiqa.v18i1.216.

[9] B. Dwi Kurniawan, R. Heriansyah, and Z. Romegar Mair, “Analisis Prediksi Terhadap Peningkatan Tindak Pidana Dengan Metode Naive Bayes Berdasarkan Laporan Kriminalitas,” *JATI (Jurnal Mhs. Tek. Inform.*, vol. 9, no. 3, pp. 4234–4241, 2025, doi:

10.36040/jati.v9i3.13620.
[10] Nandito Marbun and Jarot Prianggono, “Analisis Klasifikasi Tindak Kejahatan Pencurian dengan Algoritma K-Nearest Neighbor dan Naive Bayes di Polres Buol,” *J. Ekon. Manaj. Sist. Inf.*, vol. 6, no. 5, pp. 3350–3365, 2025, doi: 10.38035/jemsi.v6i5.5175.
[11] E. Purnamasari, “Prediksi Tingkat Kepuasan Dalam Pembelajaran Daring Menggunakan Algoritma Naive Bayes,” vol. 5, no. 4, pp. 153–159, 2023, doi: 10.60083/jidt.v5i4.431.
[12] O. Peretz, M. Koren, and O. Koren, “Naive Bayes classifier – An ensemble procedure for recall and precision enrichment,” *Eng. Appl. Artif. Intell.*, vol. 136, no. PB, p. 108972, 2024, doi: 10.1016/j.engappai.2024.108972.
[13] H. Wahyudi and A. Abdirrohman, “Pengaruh Faktor Ekonomi, dan Penyelesaian Tindak Pidana Terhadap Tingkat Kejahatan Pencurian di Pulau Sumatera,” *J. Stud. Ilmu Sos. dan Polit.*, vol. 1, no. 2, pp. 129–142, 2022, doi: 10.35912/jasispol.v1i2.1407.
[14] A. Informatics and A. Info, “Pendekatan Data - Driven untuk Pengembangan Model Prediksi Tingkat Kemiskinan di Provinsi Indonesia,” vol. 8, no. 1, pp. 84–92, 2025.
[15] R. Heriansyah, S. Puspasari, M. H. Irfani, I. Permatasari, and E. Purnamasari, “Improving Concrete Mix Type Recognition Accuracy Using ANN and GLCM Features,” pp. 96–110, 2025.