Indonesian Journal of Data and Science

Volume 6 Issue 1
ISSN 2715-9936
https://doi.org/10.56705/ijodas.v6i1.214

Research Article

Implementation of Support Vector Machine Algorithm for Classification of Study Period and Graduation Predicate of Students

Sumiyatun [1,*], Yagus Cahyadi [2], Edi Faizal [3]

[1] Universitas Teknologi Digital Indonesia, Daerah Istimewa Yogyakarta 55198, Indonesia, sumiyatun@utdi.ac.id
[2] Universitas Teknologi Digital Indonesia, Daerah Istimewa Yogyakarta 55198, Indonesia, yagus.cahyadi@utdi.ac.id
[3] Universitas Teknologi Digital Indonesia, Daerah Istimewa Yogyakarta 55198, Indonesia, edifaizal@utdi.ac.id

Correspondence should be addressed to Sumiyatun; sumiyatun@utdi.ac.id

Received 15 December 2024; Accepted 20 March 2025; Published 31 March 2025

© Authors 2025. CC BY-NC 4.0 (non-commercial use with attribution, indicate changes).
License: https://creativecommons.org/licenses/by-nc/4.0/ — Published by Indonesian Journal of Data and Science.

Abstract:

Introduction: Accurately predicting the duration of study and graduation predicates in higher education is essential for improving academic outcomes and decision-making. This study aims to classify students' study period and graduation predicates in the Information Systems program at UTDI using the Support Vector Machine (SVM) algorithm. **Methods**: A dataset of 500 student records containing academic and demographic variables—including GPA, age, semesters, and graduation predicates—was processed through data cleaning, normalization, and feature selection. Study duration was categorized into three classes: short (≤4 years), medium (4–6 years), and long (>6 years). An SVM with a linear kernel was applied, and the model was evaluated using accuracy, precision, recall, and F1-score. **Results**: The SVM model achieved perfect classification for study duration, with 100% accuracy, precision, recall, and F1-score across all categories. For graduation predicate classification, the model attained 95.18% accuracy. While it performed well overall, it faced some difficulty distinguishing between "Cum Laude" and "Very Satisfactory" due to overlapping GPA ranges. The analysis identified GPA as the most influential feature in both classification tasks, while age and the number of semesters played supporting roles. **Conclusions**: The SVM model demonstrates strong capability in classifying study duration and graduation predicates, offering valuable insights for academic management. Although performance was high, especially for study period prediction, further refinement is suggested to enhance classification in overlapping categories. Future work may benefit from larger, more balanced datasets and exploration of advanced models to increase prediction reliability.

Keywords: Educational Data, Graduation Predicates Classification, Machine Learning, Study Duration, SVM Algorithm.

1. Introduction

Higher education is one of the most important components in the development of a nation. Undergraduate students represent a key group in higher education, playing a vital role in human resource development. To achieve a high level of education, students must complete a series of courses and meet specific academic requirements in accordance with the applicable curriculum.

One important parameter in evaluating student success is the duration of study required to complete an undergraduate program. During this period, students must achieve adequate academic performance to earn specific graduation predicates. However, not all students can complete their undergraduate program on time or with the desired predicate. Managing and monitoring the study duration and graduation predicates of students is a complex task for higher education institutions. In this context, the use of data analysis methods and statistical modeling can be invaluable tools.

Universitas Teknologi Digital Indonesia (UTDI), a private university in Yogyakarta, evolved from STMIK Akakom. Since its establishment, UTDI has produced numerous graduates who have contributed to various fields,

particularly the information technology industry. Currently, UTDI offers nine academic programs. Over time, as a university operates longer, the number of graduates (alumni) increases. Data on alumni and active students, including both continuing and new students, are stored on the university's servers. This data grows over time, both in quantity and variety. Such data is a valuable asset that can be utilized for various purposes, including predicting graduation levels. A common approach for this is data mining techniques.

Data mining is a technique for extracting valuable and hidden information from large collections of data (databases), revealing previously unknown but interesting patterns. It is a process of collecting and processing data to extract important insights. Data mining has three primary objectives: explanatory, confirmatory, and exploratory. It also encompasses several methods such as Association, Classification, Regression, and Clustering. Another definition describes data mining as the search and analysis of large datasets to identify meaningful patterns and rules.

Data processing is a crucial aspect of information technology. Effectively and efficiently processed data can produce accurate information. Various data processing methods have been developed, one of which is classification. Classification is a method of grouping data based on their features or characteristics.

Classification is one of the essential tasks in data mining [1]. A classification model is built from a training dataset with predefined classes. Classification consists of two phases: the learning phase and the classification phase. The learning phase involves building the classification model, while the classification phase applies the model to predict the class label of new data [2], [3]. Classification functions to group objects into several classes and maintain classification rules to predict unknown class labels [4]–[6].

One algorithm that can be used is Support Vector Machine (SVM). This method has been widely applied in various fields, including machine learning and data analysis [7]. The basic principle of SVM is linear classification, which has been further developed to address non-linear problems by integrating the kernel trick into high-dimensional workspaces. SVM can classify both linear and non-linear data [8]–[10]. In the context of this research, SVM is used to classify students into groups based on their study duration and predict the graduation predicate they are likely to achieve.

The main background of this study is to assist higher education institutions in identifying issues related to study duration management and student graduation prediction. With a better understanding of the factors influencing study duration and graduation predicates, higher education institutions can take more effective measures to improve educational quality and minimize dropout rates.

The application of the SVM algorithm in classifying study duration and graduation predicates for undergraduate students is a relevant and valuable step towards improving the efficiency and effectiveness of higher education.

2. Method:

The research methodology is designed to analyse student data using a machine learning-based approach, particularly for classifying study duration based on academic and demographic attributes. This approach aims to gain deeper insights into the factors influencing students' study duration in higher education. The research process is divided into several stages, including data collection, data pre-processing, feature selection, model selection, model training, and model evaluation [11], [12], as illustrated in **Figure 1**.

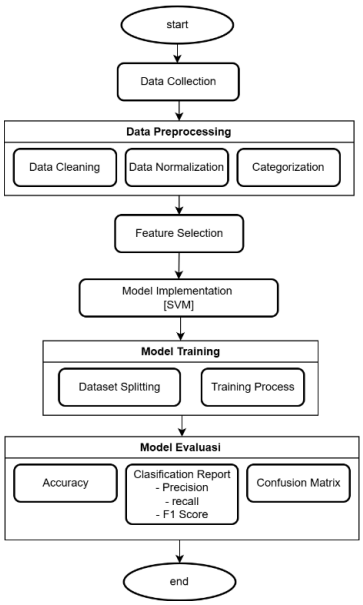

Figure 1. General Research Design Stages

Data Collection

The dataset used in this study consists of 500 student entries, encompassing several attributes relevant to the analysis, including:

a. Student Identification Number: A unique identifier for each student, although it is not utilized as a feature in the model for this analysis.
b. Age: The age of the students at the time of data collection, which can provide insights into the demographic influence on study duration.
c. GPA (Grade Point Average): An indicator of students' academic performance, reflecting their achievements throughout their studies.
d. Semester: The number of semesters completed by the students at the time of data collection, which is also related to the duration of their studies.
e. Study period: The length of time students has spent completing their studies, categorized into several groups based on predetermined time criteria.
f. Graduation Predicate: The graduation category assigned to students based on their GPA (e.g., very satisfactory, satisfactory, cum laude).

This dataset provides an overview of the relationship between students' academic and demographic characteristics and their study period categories. Each attribute is expected to contribute to building a model capable of predicting study period categories based on the available data.

Data Pre-processing

Data pre-processing is a crucial initial step to ensure that the data used in the model training process is in the correct format and ready for use [13], [14]. The pre-processing steps carried out are as follows:

a. Data Cleaning

At this stage, several actions are taken to ensure the data used does not contain elements that could compromise the analysis results, such as missing values or duplicates. Specific steps include:

- Removing Missing Values: Data entries with missing or null values in key attributes, such as GPA, age, and semesters, are either removed or imputed.

- Removing the Student Identification Number Column: The Student Identification Number column, being purely an identifier, does not contribute to the analysis or the model and is thus removed from the dataset.

b. Data Normalization

Numerical attributes, such as GPA, Age, and Semester, are expected to have a uniform scale to ensure that no attribute disproportionately influences the model due to large differences in value ranges. Therefore, normalization is applied to these attributes, scaling them into the range [0,1] [15]–[17]. The normalization is performed using Equation (1).

$$\boldsymbol{X}_{norm} = \frac{\boldsymbol{X} - \boldsymbol{X}_{min}}{\boldsymbol{X}_{max} - \boldsymbol{X}_{min}} \quad (1)$$

Where $\boldsymbol{X}$ represents the original value of the attribute, while $\boldsymbol{X_{min}}$ and $\boldsymbol{X_{max}}$ are the minimum and maximum values of the attribute, respectively. Thus, each numerical attribute used in the model will fall within a uniform range, enabling the algorithm to perform learning more effectively.

c. Study Period Categorization

The study period of students, which is the target variable in this research, will be categorized into three classes as follows:

- Short: Study period ≤ 4 years
- Medium: Study period between 4 and 6 years
- Long: Study period > 6 years

The categorization of study period is performed using a mathematical equation, which can be explained by Equation (2):

$$\textbf{Study Period Category} = \begin{cases} \textbf{Short}, & \textbf{if Study Period} \leq \textbf{4} \\ \textbf{Medium}, & \textbf{if Study Period 4} < \textbf{Study Period} \leq \textbf{6} \\ \textbf{Long}, & \textbf{if Study Period} > \textbf{6} \end{cases} \quad (2)$$

This categorization aims to group students based on their study period, making it easier to analyze patterns of different study period.

Feature Extraction

The features selected to build the study period classification model are attributes considered relevant in influencing the duration of students' studies. The chosen features include:

- GPA: As an indicator of academic performance, GPA is a critical factor in predicting students' study period.
- Age: A student's age can affect their study period. Older students may tend to have longer study period.
- Semester: The number of semesters completed reflects the student's progress in their studies.
- Graduation Predicate: The graduation category based on GPA also provides information on the relationship between academic performance and study period.

Model Implementation

The model selected for classifying study period is Support Vector Machine (SVM) with a linear kernel. SVM is an effective algorithm for classifying data with high dimensionality and can optimally separate data by constructing a hyperplane [18], [19]. The decision function of the SVM with a linear kernel can be expressed by Equation (3).

$$\boldsymbol{f(x) = w^T x + b = 0} \quad (3)$$

Where w is the weight vector, x is the feature input vector, and b is the bias. The main goal of SVM is to find the hyperplane that maximizes the margin between two different classes, with the expectation that the model can make more accurate predictions [14].

Model Training

The SVM model training process consists of two main stages: dataset splitting and the training process itself.

a. Dataset Splitting
The dataset is divided into two parts:
- Training Data: 80% of the dataset is used to train the model.
- Testing Data: 20% of the dataset is used to test the model's performance.

This division aims to ensure that the model learns from representative data and is tested on unseen data to ensure good model generalization.

b. Training Process
The SVM model is trained using the training data with selected features such as GPA, Age, Semester, and Graduation Predicate. This process involves finding the optimal values for the weight vector w and bias b to separate the data into the appropriate classes.

Model Evaluation

Model evaluation is carried out to measure how well the model can classify students' study period. The evaluation metrics used include:

a. Accuracy
Accuracy measures how many predictions are correct compared to the total number of predictions made [20], [21]. The accuracy function is expressed as in Equation (4).

$$\boldsymbol{Accuracy} = \frac{Number\ of\ Correct\ Predictions}{\boldsymbol{Total\ Predictions}} \tag{4}$$

b. Classification Report
In the classification report, several other metrics are calculated, such as:
- Precision: Measures the model's accuracy in predicting each class [22]. The precision function is expressed as in Equation (5).

$$\boldsymbol{Precision} = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{5}$$

- Recall: Measures the model's ability to detect each class [22]. The recall function is expressed as in Equation (6).

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{6}$$

- F1-Score: The harmonic mean of precision and recall, providing a balanced view between the two [23]–[25]. The F1-Score function is expressed as in Equation (7).

$$F1\text{-}Score = \mathbf{2} \times \frac{Precision \times Recall}{Precision + Recall} \tag{7}$$

c. Confusion Matrix
The confusion matrix is used to visualize the model's performance by showing the distribution of predictions against actual data. This matrix helps identify classification errors between the different classes. The confusion matrix function is expressed as in Equation (8).

$$Confusion\ Matrix = \begin{pmatrix} True\ Positives & False\ Positives \\ False\ Negatives & True\ Negatives \end{pmatrix} \tag{8}$$

The easiest way to follow the paper formatting rules is to use the format provided in this document. Save the file with a different name, then type your paper's content into it.

3. Results and Discussion

Results

The evaluation results for the study period classification model are presented in **Table 1**, with an accuracy of 100%.

Table 1. Evaluation of Study Period Classification Model

Study Period Category	Precision	Recall	F1-Score	Support
Long	1.00	1.00	1.00	14
Medium	1.00	1.00	1.00	36
Short	1.00	1.00	1.00	33
Accuracy	1.00			83
Macro Average	1.00	1.00	1.00	83
Weighted Average	1.00	1.00	1.00	83

The study period classification model demonstrates excellent performance with an accuracy of 100%. This indicates that the model can perfectly predict the study period categories. An analysis of the classification report shows that the precision, recall, and F1-score values are all at the maximum level (1.00) for all study period categories: Long, Medium, and Short.

Table 2. Evaluation of Graduation Predicate Classification Model

Graduation Predicate	Precision	Recall	F1-Score	Support
Cum Laude	1.00	0.73	0.84	11
Satisfactory	0.98	1.00	0.99	42
Very Satisfactory	0.91	0.97	0.94	30
Accuracy	0.95			83
Macro Average	0.96	0.90	0.92	83
Weighted Average	0.95	0.95	0.95	83

The graduation predicate classification model achieved a high accuracy of 95.18%, as shown in Table 2, indicating that the model is very effective in classifying students' graduation predicates. However, there are variations in performance across the different graduation predicate categories. For the Cum Laude category, the precision is 1.00, indicating that all predictions of Cum Laude are correct. However, the recall is 0.73, meaning only 73% of the actual Cum Laude data is correctly predicted, which suggests challenges in detecting students with the Cum Laude predicate. The F1-score for this category is 0.84, representing a compromise between precision and recall. This lower F1-score compared to other categories highlights difficulties in consistently classifying students with the Cum Laude predicate. In contrast, the Satisfactory category demonstrates excellent performance, with a precision of 0.98, indicating that most predictions for this category are correct, and a recall of 1.00, showing that all actual data in the Satisfactory category is correctly predicted. This results in an F1-score of 0.99, reflecting near-perfect classification for this category. Meanwhile, the Very Satisfactory category shows good performance, with a precision of 0.91, indicating that most predictions for this category are correct, and a recall of 0.97, meaning that most actual data in this category is correctly predicted. The F1-score for this category is 0.94, demonstrating strong performance, although not as high as the Satisfactory category. **Figure 2** is a confusion matrix illustrating the prediction results of study period against actual data. This matrix shows the distribution of the model's predictions across each study period category.

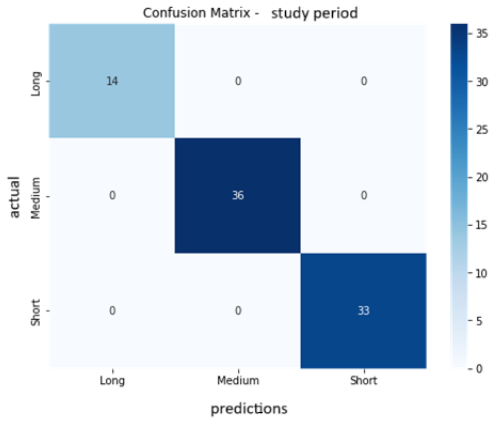

Figure 2. Confusion Matrix - Classification of Study Period

The confusion matrix illustrates the performance of a classification model in predicting study period categories, namely Long, Medium, and Short. The diagonal elements represent the correctly classified instances, with the model accurately predicting 14 instances of the Long category, 36 instances of the Medium category, and 33 instances of the Short category. Notably, all off-diagonal elements are zero, indicating that there were no misclassifications. This means the model perfectly categorized all instances into their respective study period classes, achieving 100% accuracy. Such results suggest that the model is highly effective in distinguishing between the categories. However, it is essential to consider that perfect accuracy might be influenced by factors such as a small or homogeneous dataset, and further evaluation on larger or more diverse datasets is recommended to confirm the model's robustness and generalizability.

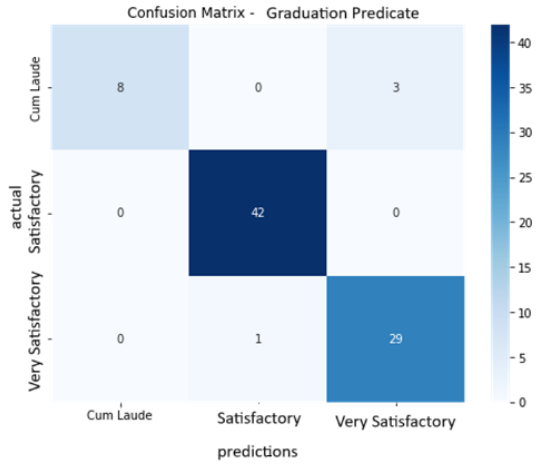

Figure 3. Confusion Matrix - Classification of Graduation Predicate

The confusion matrix shown in **Figure 3** represents the performance of a classification model for predicting graduation predicates. The matrix contains three categories: "Cum Laude," "Satisfactory," and "Very Satisfactory." Each cell indicates the number of predictions made by the model compared to the actual labels. For the "Cum Laude" category, the model correctly predicted 8 instances, but it misclassified 3 instances as "Very Satisfactory." No "Cum Laude" instances were mistakenly classified as "Satisfactory." For the "Satisfactory" category, the model performed flawlessly, correctly predicting all 42 instances without any misclassification. Meanwhile, for the "Very Satisfactory" category, the model correctly classified 29 instances but misclassified 1 instance as "Satisfactory." The high number of correct predictions along the diagonal highlights the strong overall performance of the model, especially for the "Satisfactory" and "Very Satisfactory" categories. However, the model demonstrates some difficulty distinguishing between "Cum Laude" and "Very Satisfactory," as indicated by the three misclassifications. This suggests potential overlap in features or thresholds used by the model for these categories, which could be addressed through further refinement of the model or feature engineering.

Discussion

The results confirm the efficacy of SVM models in handling classification tasks involving academic data, achieving strong accuracies in both study period and graduation predicate predictions. However, several limitations and challenges warrant further exploration. One significant issue is data imbalance, particularly in the 'Long' study period category and the 'Cum Laude' predicate, which adversely affected classification accuracy. Addressing this issue requires strategies such as data augmentation to increase representation of minority categories and employing synthetic sampling techniques like SMOTE (Synthetic Minority Oversampling Technique) to balance the dataset.

In terms of feature importance, GPA consistently emerged as the dominant feature in both tasks, while other features, such as age and semester, contributed minimally. Future research should consider exploring additional features, such as course grades, extracurricular activities, or attendance records, and applying feature selection algorithms to refine the input space for better model performance. Despite the SVM's linear kernel performing well, it may not fully capture complex patterns in the data. Advanced methods, such as ensemble learning techniques (e.g., Random Forest or Gradient Boosting) or neural networks, could provide better classification results, particularly for nuanced categories. Misclassifications in the 'Cum Laude' category highlighted overlapping GPA ranges, suggesting that stricter GPA thresholds for category demarcation or the inclusion of qualitative factors, such as research achievements, could improve accuracy.

While the model achieved high accuracy on small, homogeneous datasets, concerns about generalizability remain. Testing the model on larger, more diverse datasets is essential to ensure its robustness and applicability across institutions and demographic variations. Overall, the study demonstrates the potential of SVM models in academic data classification but emphasizes the need for future work focusing on data enrichment, advanced modeling techniques, and cross-institutional validations to address these challenges effectively.

4. Conclusion

This study demonstrates that the SVM-based classification model can be effectively utilized to predict the categories of study period and graduation predicates of students based on the available features. The evaluation results indicate that the model for study period outperforms the model for graduation predicates, which may be attributed to the complexity and variability within the graduation predicate data.

Acknowledgments

The authors express their gratitude to the Universitas Teknologi Digital Indonesia for providing funding support through the research budget for the 2023–2024 Academic Year.

References:

[1] R. A. Raj, "Classification and Prediction of Incipient Faults in Transformer Oil by Supervised Machine Learning using Decision Tree," *2023 3rd International Conference on Artificial Intelligence and Signal Processing, AISP 2023*. 2023, doi: 10.1109/AISP57993.2023.10134566.

[2] A. Singh and V. Gutte, "Classification of Breast Tumor Using Ensemble Learning," in *Mobile Computing and Sustainable Informatics*, 2022, pp. 491–507.

[3] L. S. Van Velzen, "Classification of suicidal thoughts and behaviour in children: results from penalised logistic regression analyses in the Adolescent Brain Cognitive Development study," *Br. J. Psychiatry*, vol. 220, no. 4, pp. 210–218, 2022, doi: 10.1192/bjp.2022.7.

[4] R. Hazra, "Machine Learning for Breast Cancer Classification with ANN and Decision Tree," *11th Annual IEEE Information Technology, Electronics and Mobile Communication Conference, IEMCON 2020*. pp. 522–527, 2020, doi: 10.1109/IEMCON51383.2020.9284936.

[5] Y. Wang, "Rice diseases detection and classification using attention based neural network and bayesian optimization," *Expert Syst. Appl.*, vol. 178, 2021, doi: 10.1016/j.eswa.2021.114770.

[6] B. H. Reddy, "Classification of Fire and Smoke Images using Decision Tree Algorithm in Comparison with Logistic Regression to Measure Accuracy, Precision, Recall, F-score," *14th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics, MACS 2022*. 2022, doi: 10.1109/MACS56771.2022.10022449.

[7] S. Markkandan, "SVM-based compliance discrepancies detection using remote sensing for organic farms," *Arab. J. Geosci.*, vol. 14, no. 14, 2021, doi: 10.1007/s12517-021-07700-4.

[8] Q. Tang, "Classification of Complex Power Quality Disturbances Using Optimized S-Transform and Kernel SVM," *IEEE Trans. Ind. Electron.*, vol. 67, no. 11, pp. 9715–9723, 2020, doi: 10.1109/TIE.2019.2952823.

[9] P. Purnawansyah, A. P. Wibawa, and ..., "Comparative Study of Herbal Leaves Classification using Hybrid of GLCM-SVM and GLCM-CNN," *Ilk. J. ...*, 2023, doi: 10.33096/ilkom.v15i2.1759.382-389.

[10] W. Wu, "An Intelligent Diagnosis Method of Brain MRI Tumor Segmentation Using Deep Convolutional Neural Network and SVM Algorithm," *Comput. Math. Methods Med.*, vol. 2020, 2020, doi: 10.1155/2020/6789306.

[11] P. Sharma, "Performance analysis of deep learning CNN models for disease detection in plants using image segmentation," *Inf. Process. Agric.*, vol. 7, no. 4, pp. 566–574, 2020, doi: 10.1016/j.inpa.2019.11.001.

[12] A. A. Ewees, "Performance analysis of Chaotic Multi-Verse Harris Hawks Optimization: A case study on solving engineering problems," *Eng. Appl. Artif. Intell.*, vol. 88, 2020, doi: 10.1016/j.engappai.2019.103370.

[13] A. Tuppad and S. D. Patil, "Data Pre-processing Issues in Medical Data Classification," *2023 Int. Conf. ...*, 2023, doi: 10.1109/NMITCON58196.2023.10275855.

[14] K. N. Myint and Y. Y. Hlaing, "Predictive Analytics System for Stock Data: methodology, data pre-processing and case studies," *2023 IEEE Conf. Comput. ...*, 2023 doi: 10.1109/ICCA51723.2023.10182047.

[15] S. Buriya, "Malware Detection Using 1d Convolution with Batch Normalization and L2 Regularization for Android," *2023 International Conference on System, Computation, Automation and Networking, ICSCAN 2023*. 2023, doi: 10.1109/ICSCAN58655.2023.10395472.

[16] L. Peng, "Dual-Structure Elements Morphological Filtering and Local Z-Score Normalization for Infrared Small Target Detection against Heavy Clouds," *Remote Sens.*, vol. 16, no. 13, 2024, doi: 10.3390/rs16132343.

[17] M. Sholeh, "Comparison of Z-score, min-max, and no normalization methods using support vector machine algorithm to predict student's timely graduation," *AIP Conference Proceedings*, vol. 3077, no. 1. 2024, doi: 10.1063/5.0202505.

[18] N. Mamat, "Enhancement of water quality index prediction using support vector machine with sensitivity analysis," *Front. Environ. Sci.*, vol. 10, 2023, doi: 10.3389/fenvs.2022.1061835.

[19] X. Ning, "Classification of Sulfadimidine and Sulfapyridine in Duck Meat by Surface Enhanced Raman Spectroscopy Combined with Principal Component Analysis and Support Vector Machine," *Anal. Lett.*, vol. 53, no. 10, pp. 1513–1524, 2020, doi: 10.1080/00032719.2019.1710524.

[20] M. Stusek, "Accuracy Assessment and Cross-Validation of LPWAN Propagation Models in Urban Scenarios," *IEEE Access*, vol. 8, pp. 154625–154636, 2020, doi: 10.1109/ACCESS.2020.3016042.

[21] S. Clark and R. Wicentowski, "SwatCS: Combining simple classifiers with estimated accuracy," 2013.

[22] P. A. Flach and M. Kull, "Precision-Recall-Gain Curves: PR Analysis Done Right.".

[23] R. Kiruthika, "Predictive Modeling of Poultry Growth Optimization in IoT-Connected Farms Using SVM for Sustainable Farming Practices," Proceedings of the 2024 10th International Conference on Communication and Signal Processing, ICCSP 2024. pp. 677–682, 2024, doi: 10.1109/ICCSP60870.2024.10543463.

[24] W. Zhuo and A. Ahmad, "HCRF: an improved random forest algorithm based on hierarchical clustering," Indones. J. Electr. Eng. Comput. Sci., vol. 38, no. 1, p. 578, Apr. 2025, doi: 10.11591/ijeecs.v38.i1.pp578-586.

[25] J. T. Hasić, "Breast Cancer Classification Using Support Vector Machines (SVM)," Lecture Notes in Networks and Systems, vol. 644. pp. 195–205, 2023, doi: 10.1007/978-3-031-43056-5_16.