JURNAL INFOTEL
Informatics - Telecommunication - Electronics
Website Jurnal : http://ejournal.st3telkom.ac.id/index.php/infotel
ISSN : 2085-3688; e-ISSN : 2460-0997

PREDICTION OF STUDENT ACHIEVEMENT AT SMK TELKOM LAMPUNG USING ARTIFICIAL NEURAL NETWORK

Received, Revised, Accepted

Abstract — The utilization of artificial intelligence (AI) is crucial for analyzing student performance. In this context, artificial neural networks (ANNs) emerge as a relevant and effective method. ANNs, inspired by the structure and function of the human brain, consist of interconnected artificial neurons capable of learning from input data to generate complex predictions, including student delays. This research aims to predict student delays at SMK Telkom Lampung using the ANN method, comparing its performance with other techniques such as Support Vector Machine (SVM). Primary data collected from SMK Telkom Lampung comprises 4939 examples with 550 cases, 26 features, and 4 meta-attributes. Performance evaluation involves metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and coefficient of determination (R2). Comparative performance analysis of the models is also conducted using scatter plots and box plots. The research findings indicate that the Artificial Neural Network model slightly outperforms the Support Vector Machine model, with lower prediction error rates and better ability to explain data variability.

Keywords: Achievement, SMK Telkom, Coefficient Determination, Prediction

I. INTRODUCTION

Vocational High School (SMK) are vocational education institutions aimed at preparing individuals with the skills needed to work in the business/industry sector or to become entrepreneurs independently. One of the institutions that holds a large amount of data is the educational organization[1]. This organization utilizes data to obtain information, especially about the students. Student data has various attributes, allowing us to make predictions such as academic and non-academic achievements of students in school. Although in the independent curriculum, students are not evaluated with a ranking or rating system in academic achievements, as it is considered inaccurate to depict the potential and talents of students, academic performance is often used as a parameter for educational success. One indicator of this achievement is the academic performance of students expressed by the ranking. Achieving quality education is related to the roles of teachers, student motivation, student discipline, socioeconomic conditions of students, and past learning outcomes. Data mining is a process for discovering patterns using statistical techniques to unearth hidden information within a large database[2]. Data mining with the assistance of machine learning is very useful for solving various kinds of problems[3]. Machine Learning (ML) techniques are used to implement sentiment analysis methods on data[4]. By knowing students' performance earlier, schools can take necessary actions to ensure good academic and non-academic achievements for students. The ultimate hope is that all students from various backgrounds can maximize their academic and non-academic achievements.

In the field of education, research predicting the success rate of student studies is necessary to recommend improvements in future academic performance[5]. Many studies have been conducted on predicting students' learning achievements. For example, in research conducted by Sadimin and Handoyo Widi Nugroho entitled "Comparison of Data Mining Algorithm Performance for Student Graduation

Commented [A1]: Start the abstract with a problem. What is the problem? Why analyzing student performance is important?

Commented [A2]: These are not ideal keywords that describe the paper

Commented [A3]: Problem is not well introduced here.

Prediction" the attributes used in this study as indicators of student learning achievement are the Cumulative Grade Point Average (GPA) of students[1]. Another study was conducted by Yahia Baashar, Gamal Alkawsi, Abdulsalam Mustafa, Ammar Ahmed Alkahtani, Yazan A. Alsariera, Abdulrazzaq Qasem Ali, Wahidah Hasyim, and Sieh Kiong Tiong in a systematic literature review on "Towards Predicting Academic Performance of Students Using Artificial Neural Networks (ANN)" the systematic literature review concluded that artificial neural networks (ANN) have shown high accuracy in predicting academic performance outcomes, even though similar results were obtained with other data mining approaches[6]. Similar research was also conducted by Nalindren Naicker, Timotius Adeliy, and Jeanette Sayap titled "Linear Support Vector Machine for Predicting Student Performance in School-Based Education" in this study, experimental research was conducted using feature selection on a dataset available to the public consisting of 1,000 alphanumeric student records and the linear support vector machine, measured with ten categorical machine learning algorithms, showed superior performance in predicting student performance[7].

Based on the background and previous research studies, some of the identified research gaps include limitations in adjusting input variables, insufficient understanding of the prediction phenomenon and researchers only providing general performance comparisons.

In recent years, educational data mining has become one of the hottest topics in scientific research[8]. Artificial neural networks (ANN) are one of the machine learning and data mining algorithms used in various research literature and are claimed to provide superior and accurate results in predicting student performance[6]. This model uses a set of linear and non-linear mathematical equations that do not consider the physical processes at all. The most important aspect of this model is that the output produced approaches the actual results. ANN is also widely used because of its ability to recognize patterns and training[9]. In addition to ANN, some famous machine learning algorithms used in classification and prediction are Support Vector Machine (SVM)[10]. The advantage of SVM is its fast learning process, while its disadvantage lies in the difficulty of use when encountering problems, especially with large amounts of data[11].

Based on the above issues, this research aims to predict the performance of SMK Telkom Lampung students using the artificial neural network method. It is hoped that the research can utilize technological advancements by comparing the ANN and SVM methods to predict student delays at SMK Telkom Lampung using the orange software as an analysis tool to enable systematic and comprehensive comparison between the two methods.

Thus, this study not only explores the potential of artificial intelligence technology in the educational context but also provides insights into the comparison of the performance of ANN and SVM methods in predicting student delays at SMK Telkom Lampung. This allows for a deeper understanding of the effectiveness of both methods in the specific educational context of predicting student achievements.

II. RESEARCH METHOD

The stages of this research follow steps as outlined in the following flowchart:

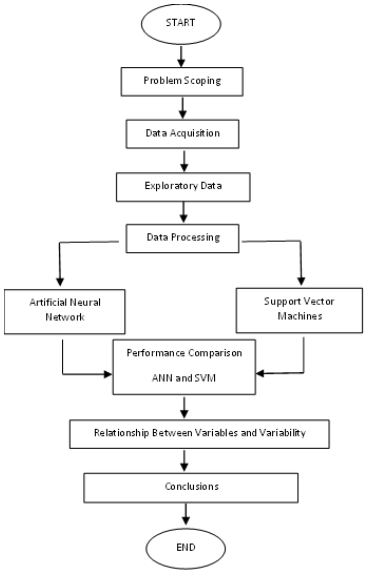

Figure 1. Research Stages

The explanation of the research stages above is as follows:

A. Problem Scoping

The problem scoping in this research includes focusing on Vocational High School (SMK) Telkom Lampung as the main research location and utilizing several core variables including absenteeism, delays, ranking, extracurricular activities, academic, non-academic student achievement data, etc. The focused analysis methods are Artificial Neural Network (ANN) and Support Vector Machine (SVM), with implementation and analysis conducted using the Orange software.

B. Data Acquisition

Data acquisition for this research involves several steps, as:

a) Data Collection

At this stage, it is determined which sources the data will be stored from, which data features will be used, and whether the collected data aligns with its objectives[12]. Instances refer to the number of records

Commented [A4]: You do not need to write all author names. Learn how to use "et al."

Commented [A6]: When it comes to the detail of the data, one should not use "etc." Mention all variables used in the research

Commented [A5]: What is student delay? Why should you predict the student delay? I think this matter should be explored more deeply as a background of this study.

Commented [A7]: Where does the data come from? Online learning platform such as Moodle?

in the dataset. The number of instances in the dataset is 550 samples.

b) Data Cleansing

This stage aims to ensure there are no duplicate data, identify inconsistent data, and rectify errors in the data, such as printing errors so that the data can be processed and used for data mining. During the data cleansing process, incorrect data, duplicate data, and inconsistent data were found. Thus, the authors are still in the process of cleaning the data to ensure good quality data is produced.

c) Data Transformation

This stage involves transforming the data into a format suitable for data mining processing. The application used for simulation is Orange Data Mining, an open-source data mining application that has been proven to assist researchers in analyzing data. When using the orange software, data preparation in CSV or Excel format is required to ensure smooth analysis processes and compliance with the software compatibility requirements.

d) Data Partitioning

The data will be divided into two main parts: training data and testing data. The training data will be used to train the prediction model, while the testing data will be used to test the performance of the trained model.

e) Data Validation

Before analysis is conducted, data validation will be performed to ensure that the data used aligns with the research objectives and does not contain significant errors.

C. Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an initial investigation process aimed at identifying patterns, discovering anomalies, testing hypotheses, and reviewing assumptions that ultimately reveal interesting insights from the data.

D. Prepocessing

In general, the steps of data preprocessing involve cleaning[13]. Data cleaning is performed to inspect missing values, duplicate data, or outliers[14]. The proper preprocessing process will yield clear information with good accuracy values[15]. The researcher's cleaning process uses the Microsoft Excel application to filter each column one by one and search for empty data. Columns such as date of birth, class, number of siblings, and distance from home contain many poorly recorded data. In the prediction column regarding tardiness, attendance, academic achievement, non-academic achievement and etc, they are grouped into 4 scores. A score of 0 represents invalid data, typically for students who have been expelled or transferred schools but their records still exist. A score of 1 is given for data with no influence, while a score of 2 is for moderate influence, and a score of 3 indicates significant influence.

E. Prediction

The term prediction can be equated with both prophecy and estimation[16]. The ability to make predictions is the ability to forecast future events or benefits based on observations, measurements, collected data, or research findings indicating certain trends in phenomena[17].

F. Artificial Neural Network (ANN)

ANN is one of the most widely used techniques in the field of Artificial Intelligence, where this algorithm can be learned independently during the data training process[18]. In general, each ANN can be characterized by (1) the arrangement of connections between neurons (referred to as topology), (2) its approach to obtaining strength or weights on connections (referred to as training or learning algorithms), and (3) its activation function[19].

G. Support Vector Machines (SVM)

Support Vector Machine (SVM) algorithm is one of the supervised machine learning algorithms based on statistical learning theory, this model selects from the training samples a set of characteristic subsets so that the classification of these characteristic subsets is equivalent to the division of the entire dataset[20].

H. Performance Comparison

The performance comparison between Artificial Neural Network (ANN) and Support Vector Machine (SVM) is conducted by comparing several evaluation metrics, such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R2).

Here are the formulas for each commonly used evaluation metric in prediction:

Mean Squared Error (MSE): MSE measures the average of the squared differences between predicted values (Y_pred) and actual values (Y_true).

MSE=$n1\sum i=1n$ (Ytrue$i-Y$pred$i$)2

Where:

n is the number of samples.

Ytrue$i$ is the actual value of sample i.

Ypred$i$ is the predicted value of sample i.

Root Mean Squared Error (RMSE): RMSE is the square root of MSE, providing a measure of the average deviation between predicted and actual values.

RMSE= $\sqrt{n_1 \Sigma_{i=ln(Y_{tr\,uei}-ypr\grave{e}d)^2}}$

Commented [A8]: Explain clearly steps performed in the data cleansing

Commented [A9]: What format used for data mining processing? How was the transformation performed?

Commented [A13]: ANN is a general name, what kind of ANN was used in this research? The detail including architecture, training method, etc are not clearly explained in this research.

Commented [A10]: What is the ratio for training vs testing data?

Commented [A14]: I believe that this research explores regression problem, not classification. Therefore, SVM cannot be used since SVM is applicable for classification, not regression. Authors should explore more about SVR (Support Vector Regressor)

Commented [A11]: How?

Commented [A12]: What kind of EDA performed in this research?

Commented [A15]: Always write an equation using Equation Editor in Word
Equation must be numbered

Mean Absolute Error (MAE): MAE measures the average of the absolute differences between predicted and actual values.

$$MAE=n1\sum i=1n|Ytruei-Ypredi|$$

Coefficient of Determination (R2 score): R2 score measures how well the variability in the dependent variable can be explained by the model.

$$R2=1-\sum i=1n\ (Ytruei-Ytrue)2\sum i=1n\ (Ytruei-Ypredi)2$$

Where $\bar{}Ytrue$ is the mean of the actual values true*Y*true.

III. RESULT

The results of predicting student achievement at SMK Telkom Lampung using ANN and SVM methods are as follows:

A. *Dataset*

The data used in this study is primary data obtained from SMK TELKOM Lampung. The total number of data is 4939, with 550 instances, 26 features, and 4Meta-Attributes.

Table 1. Data Exploration

No	Name Parameter	Status	Data Type
1	Name	Input	Nominal
2	NISN	Input	Numerik
3	Place of Birth	Input	Nominal
4	Date of Birth	Input	Numeric
5	Gender	Input	Nominal
6	Religion	Input	Nominal
7	Address	Input	Nominal
8	Subdistrict	Input	Nominal
9	Child Number	Input	Numeric
10	Number of Siblings	Input	Numeric
11	Type of Residence	Input	Nominal
12	Transportation	Input	Nominal
13	Current Class	Input	Nominal
14	Distance from Home to School (KM)	Input	Numeric
15	KPS Recipient Input	Input	Nominal
16	KIP Recipient	Input	Nominal
17	Eligible for PIP (proposed by the school)	Input	Nominal
18	Reason for Eligibility for PIP	Input	Nominal
19	Special Needs	Input	Nominal
20	Sum of Value Scores	Input	Numeric
21	Mean	Input	Numeric
22	Rank	Input	Numeric
23	Absenteeism	Input	Numeric
24	Delay Input	Input	Numerik
25	Extracurricular Activities	Input	Nominal
26	Academic Achievement	Input	Nominal
27	Non-Academic Achievement	Input	Nominal
28	Prediction	Output	Numeric

B. *Data Mining Process*

To select the best method with high accuracy, a comparison of several data mining methods is conducted by ANN and SVM as shown in Figure 2.

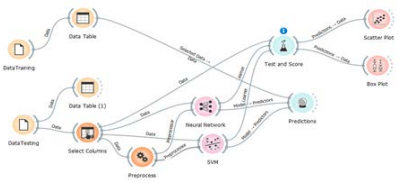

Figure 2. Orange Tools Widget Design

C. *Performance Comparison of ANN and SVM*

Based on the tested data, the calculation results of each model are depicted in Figure 3. Evaluating based on the metrics, the Neural Network model demonstrates superior performance compared to the Support Vector Machine (SVM) model. Specifically, the Neural Network model yields the following values: MSE: 0.352, RMSE: 0.593, MAE: 0.383, and R2: 0.003.

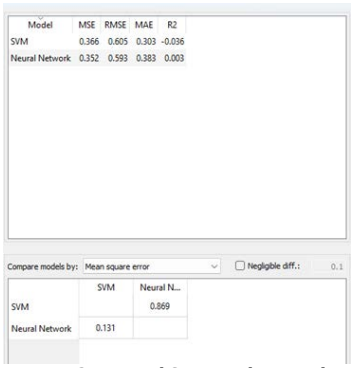

Figure 3. Test and Score Widget Results

On the other hand, although the SVM model has higher values for MSE: 0.366, RMSE: 0.605, MAE: 0.303, the R2 value of -0.036 indicates that the model is not suitable for the data. Therefore, despite the R2 value for the Neural Network model approaching zero, the evaluation concludes that the Neural Network model has better overall performance because of its ability to provide predictions closer to the actual values of the data.

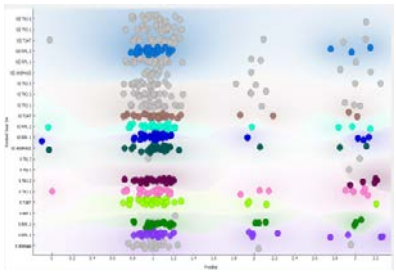

Figure 4. Scatter Plot of Prediction Data

The scatter plot of predictions only shows the visualization between predictions and classes.

Commented [A16]: Detail of each method is not well explained.

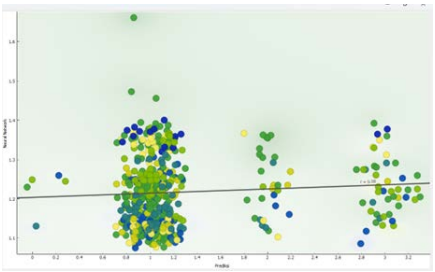

Figure 5. ANN Scatter Plot

The scatter plot of predictions with the Neural Network (NN) shows that the relationship between the model predictions and the actual values has a low correlation, with a Pearson correlation coefficient (r) of 0.08. This indicates that the predictions from the Neural Network model have a weak relationship with the actual values, resulting in a flat regression line.

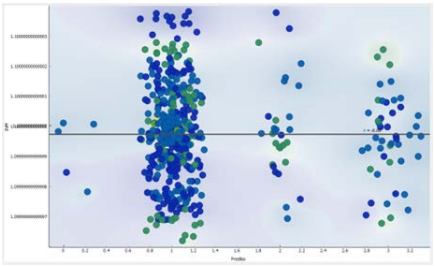

Figure 6. SVM Scatter Plot

The scatter plot of predictions with the Support Vector Machine (SVM) shows that there is no significant correlation between the model predictions and the actual values, with a Pearson correlation coefficient (r) close to zero, which is -0.00. This indicates that the predictions from the SVM model do not have a clear linear relationship with the actual values, so there is no identifiable regression line.

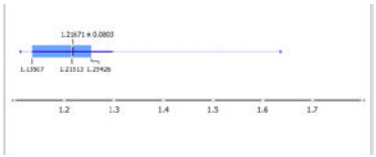

Figure 7. Neural Network Box Plot

The box plot results for the Neural Network (NN) show that the distribution of model predictions has a mean of 1.21671 with a standard deviation of 0.0803. This indicates that predictions from the NN model tend to have a slightly higher average value than 1.21, with a relatively low variation of around 0.0803.

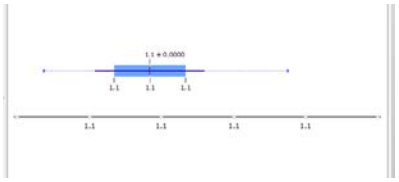

Figure 8. Support Vector Machine Box Plot

The box plot results for the Support Vector Machine (SVM) as depicted in Figure 8. show that the distribution of model predictions has a mean of 1.1 with no visible variation, and a standard deviation close to zero, which is 0.0000. This indicates that predictions from the SVM model tend to be stable around the value of 1.1 without significant variation.

IV. DISCUSSION

This research highlights the issue of model complexity, which may affect the interpretation and generalization of findings. While Neural Network (NN) demonstrates better performance than SVM, the complexity of the NN model can pose its challenges. The use of complex NN often requires a large number of parameters and intricate training processes. This can lead to difficulties in understanding the relative contributions of each variable or feature in making predictions, as well as exacerbating the problem of overfitting. Therefore, careful consideration is needed on how the complexity of NN models can impact the interpretation of results and their utility in real-world practice. Recognizing these issues, further research can focus on developing simpler approaches that are easier to interpret without sacrificing high predictive performance.

V. CONCLUSION

The research findings indicate that the Neural Network (NN) performs better in predicting students' academic performance at SMK Telkom Lampung compared to the Support Vector Machine (SVM). Despite the Neural Network's (NN) R2 value approaching zero, it provides more accurate predictions compared to SVM, which yields a negative R2 value. The analysis also shows that NN's predictions have a weak correlation with the actual values, although still better than SVM, which lacks significant correlation. Additionally, NN's prediction distribution exhibits slightly higher variability compared to the stable distribution of SVM. Therefore, for future research, it is recommended to consider the use of additional data, model adjustments, and further analysis of student performance to enhance prediction accuracy and deepen understanding of this phenomenon.

REFERENCES

[1] H. Sadimin , Widi Nugroho, "Performance Comparison Of Datamining Algorithm For Prediction Of Student Graduation," vol. 17, no. 2, pp. 512–520, 2023.

[2] S. Mukodimah and C. Fauzi, "Comparison of Tree Implementation, Regression Logistics, and Random Forest To Detect Iris Types," *J. TAM (Technology Accept. Model.*, vol. 12, no. 2, p. 149, 2021, doi: 10.56327/jurnaltam.v12i2.1074.

[3] A. Ishaq *et al.*, "Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques," *IEEE Access*, vol. 9, pp. 39707–39716, 2021, doi: 10.1109/ACCESS.2021.3064084.

[4] K. Baker, Mohammed Rashad, Najmaldeen Taher, Yalmaz, H. Jihad, "Prediction Of People Sentiments on Twitter Using Machine Learning Classifiers During Russian Aggression In Ukraine," *Jordanian J. Comput. Inf. Technol.*, vol. Vol. 09, N, no. 717, p. 2023, 2023.

[5] D. Kurniadi, E. Abdurachman, H. L. H. S. Warnars, and W. Suparta, "Predicting student performance with multi-level representation in an intelligent academic recommender system using backpropagation neural network," *ICIC Express Lett. Part B Appl.*, vol. 12, no. 10, pp. 883–890, 2021, doi: 10.24507/icicelb.12.10.883.

[6] Y. Baashar *et al.*, "Toward Predicting Student's Academic Performance Using Artificial Neural Networks (ANNs)," *Appl. Sci.*, vol. 12, no. 3, 2022, doi: 10.3390/app12031289.

[7] N. Naicker, T. Adeliyi, and J. Wing, "Linear Support Vector Machines for Prediction of Student Performance in School-Based Education," *Math. Probl. Eng.*, vol. 2020, 2020, doi: 10.1155/2020/4761468.

[8] D. Wang, D. Lian, Y. Xing, S. Dong, X. Sun, and J. Yu, "Analysis and Prediction of Influencing Factors of College Student Achievement Based on Machine Learning," *Front. Psychol.*, vol. 13, no. 1, 2022, doi: 10.3389/fpsyg.2022.881859.

[9] A. Çetinkaya and Ö. K. Baykan, "Prediction of middle school students' programming talent using artificial neural networks," *Eng. Sci. Technol. an Int. J.*, vol. 23, no. 6, pp. 1301–1307, 2020, doi: 10.1016/j.jestch.2020.07.005.

[10] K. M. O. Nahar, R. M. Al-khatib, M. A. Al-shannaq, and M. Malek, "An Efficient Holy Quran Recitation Recognizer Based on Svm Learning Model," vol. 06, no. 04, pp. 392–414, 2020.

[11] D. A. Anggoro and D. Novitaningrum, "Comparison of accuracy level of support vector machine (SVM) and artificial neural network (ANN) algorithms in predicting diabetes mellitus disease," *ICIC Express Lett.*, vol. 15, no. 1, pp. 9–18, 2021, doi: 10.24507/icicel.15.01.9.

[12] M. Yagcı, "Educational data mining: prediction of students' academic performance using machine learning algorithms," *Smart Learn. Environ.*, vol. 9, no. 1, 2022, doi: 10.1186/s40561-022-00192-z.

[13] S. Sandino Berutu, H. Budiati, Jatmika, and F. Gulo, "Data preprocessing approach for machine learning-based sentiment classification," *J. Infotel*, vol. 15, no. 4, pp. 317–325, 2023.

[14] A. Purwanto and H. Widi Nugroho, "Comparative Analysis Of The Performance Of The C4 . 5 Algorithm And The K-Nearest Neighbors Algorithm For Classification," vol. 17, no. 1, pp. 236–243, 2023.

[15] R. Toro and S. Lestari, "Comparison of Classification Algorithms for Determining Promotional Locations for New Student Admissions at IIB Darmajaya Lampung," vol. 22, no. 1, pp. 223–234, 2023.

[16] N. Maylani Ferdy , Sriyanto, "Implementation of Data Mining Methods to Predict the Color of Kittens in the Breeding Process Race Cats Using Algorithms Support Vector Machine (SVM)," no. August, pp. 114–125, 2021.

[17] C. H. Rosnani Ginting, "Application of Data Mining : Toyota Car Sales Prediction Using Artificial Neural Network on Orange Software TALENTA Conference Series ApplicationData Mining : Toyota Car Sales Prediction Using Artificial Neural NetworksonOrange Software," vol. 4, no. 1, 2021, doi: 10.32734/ee.v4i1.1228.

[18] M. Uzair and N. Jamil, "Effects of Hidden Layers on the Efficiency of Neural networks," *Proc. - 2020 23rd IEEE Int. Multi-Topic Conf. INMIC 2020*, pp. 1–6, 2020, doi: 10.1109/INMIC50486.2020.9318195.

[19] C. F. Rodríguez-Hernández, M. Musso, E. Kyndt, and E. Cascallar, "Artificial neural networks in academic performance prediction: Systematic implementation and predictor evaluation," *Comput. Educ. Artif. Intell.*, vol. 2, no. December 2020, 2021, doi: 10.1016/j.caeai.2021.100018.

[20] A. Mustafa Abdullah, Dakhaz,Mohsin Abdulazeez, "Machine Learning Applications based on SVM Classification: A Review," *Qubahan Acad. J.*, vol. 3, no. 4, pp. 206–218, 2023, doi: 10.48161/Issn.2709-8206.