# Classification of Autism in Children Based on the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) Using Support Vector Machine (SVM)

Avinda Nur Azizah[1]
[1.2]*Department of Informatics Engineering, Faculty of Science and Technology, State Islamic University Syarif Hidayatullah Jakarta*
West Tangerang, Indonesia
avinda.nurazizah21@mhs.uinjkt.ac.id[1]

Siti Ummi Masruroh[2]
[1.2]*Department of Informatics Engineering, Faculty of Science and Technology, State Islamic University Syarif Hidayatullah Jakarta*
West Tangerang, Indonesia
ummi.masruroh@uinjkt.ac.id[2]

Neneng Tati Sumiati[3]
[1.2]*Department of Informatics Engineering, Faculty of Science and Technology, State Islamic University Syarif Hidayatullah Jakarta*
West Tangerang, Indonesia
neneng.tati@uinjkt.ac.id[3]

***Abstract*— Autism is a developmental disorder that affects a child's ability to communicate and interact socially. Early diagnosis is crucial to prevent delays in treatment; however, there is a lack of parental awareness of early symptoms, and currently, no classification tool is available to identify the severity level of autism based on DSM-5 criteria. This study aims to classify autism in children based on DSM-5 criteria using the Support Vector Machine (SVM) algorithm with a CRISP-DM approach, and to address data imbalance using the SMOTE oversampling technique. The dataset consists of children's symptom data labeled as ASD (ASD 1, ASD 2, ASD 3) and Non-ASD. Evaluation results show that the SVM model without SMOTE achieved the best performance with an overall accuracy of 98.77%, along with high and balanced precision, recall, and F1-scores. Meanwhile, the application of SMOTE increased sensitivity for minority classes but reduced the accuracy to 90.12%. Therefore, the SVM model without SMOTE is considered the most effective. This model has potential as an early diagnostic aid but still requires further clinical validation.**

***Keywords*— *Autism, DSM-5, Support vector machine, SMOTE, Machine learning, and Classification.***

## I. INTRODUCTION

Autism is a neuropsychiatric developmental disorder that affects brain functions involved in social skills, communication, and behavior [1]. According to the World Health Organization (WHO), approximately 1 in 100 children worldwide are affected by autism [2]. In Indonesia, it is estimated that around 2.4 million children live with ASD [3]. The causes of autism involve various factors, both genetic and environmental. Children diagnosed with autism tend to have difficulties responding to others, face challenges in communication and language, and display unusual reactions to their surroundings [4].

The diagnostic process for mental disorders, including autism, refers to the guidelines outlined in the fifth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5), published by the American Psychiatric Association (APA). DSM-5 categorizes the severity of ASD into three levels: Level 1 (requiring support), Level 2 (requiring substantial support), and Level 3 (requiring very substantial support) [5]. Observations and interviews conducted at Therapy Centers and Hospitals indicate that the majority of children diagnosed with ASD fall into Level 3, the most severe category. Delayed diagnosis in autistic children has serious implications for their development, especially in communication, behavior, and social skills. These children tend to struggle with independence and require long-term assistance, which adds psychological and economic burdens to families.

This highlights that the early diagnosis process is still lacking and requires technological support. One of the technologies gaining significant attention is machine learning, which is rapidly advancing across various sectors, including healthcare [6]. Machine learning algorithms can identify patterns in autism symptoms from child data, offering potential for early and accurate diagnosis [7]. One widely used machine learning algorithm known for its accuracy, especially with small datasets, is the Support Vector Machine (SVM). SVM is a learning system that utilizes hypotheses in the form of linear functions on high-dimensional features and is trained using learning algorithms based on optimization theory principles [8].

In the classification process using machine learning, data quality has a significant impact on the results obtained. One common issue encountered in data processing is class distribution imbalance. In this case, the number of children with mild or moderate symptoms is significantly lower than those at the severe level. This imbalance can affect the model's performance in classification. There are two approaches to handling class imbalance: data-level and algorithm-level approaches [9]. In this study, the class imbalance is addressed using a data-level approach with the SMOTE oversampling technique. The Synthetic Minority Over-Sampling Technique (SMOTE) is an oversampling method used to address class distribution imbalance by generating synthetic samples for the minority class using its K nearest neighbors [10].

Previous research on autism classification was conducted by Hazarika et al. (2022), which concluded that many studies had not incorporated DSM-5 criteria more comprehensive than DSM-4 and among several algorithms, SVM demonstrated one of the best performances, thereby creating an opportunity to develop DSM-5-based models [11]. Another study by Supriyadi et al. (2022) used the Naïve Bayes and Support Vector Machine algorithms to predict autism using the CRISP-DM approach, resulting in fairly high performance [12]. Additionally, a study by Mulla et al. (2021) combined

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

PCA and SMOTE to enhance performance across several algorithms, with KNN achieving the highest accuracy [10].

Considering the importance of early diagnosis in managing autism in children and the differing conclusions from previous studies, this study is conducted by referring to DSM-5 criteria, using the CRISP-DM approach, the Support Vector Machine (SVM) algorithm, and the SMOTE oversampling technique. The aim of this research is to evaluate the performance of SVM in classifying autism based on DSM-5 and to assess the impact of SMOTE on classification performance.

## II. METHODS

### A. Data Collection

The data used in this study are primary data, consisting of medical records or assessments of children with ASD, including severity levels based on DSM-5, obtained from two institutions, a Therapy Center and a Hospital. The dataset includes symptoms experienced by children aged 2 to 15 years, categorized into ASD Level 1, ASD Level 2, ASD Level 3, and Non-ASD. All data have been anonymized and coded in accordance with ethical procedures. In addition, data collection was supported by observations and interviews to understand the context and validate the data used.

### B. Experimental Implementation Method

This study uses the CRISP-DM (Cross-Industry Standard Process for Data Mining) approach. In this research, the process is carried out up to the evaluation stage, as follows:

#### 1) Business Understanding

In this stage, the research objectives are defined and understood. This study aims to utilize machine learning for ASD classification as a tool for early detection, supporting faster intervention. The model is also expected to provide insights into significant features in the diagnosis process. Observations and literature reviews are used to determine an appropriate model that can improve classification accuracy and support the diagnostic process.

#### 2) Data Understanding

This study uses 81 child data entries obtained from a Therapy Center and a Hospital, including children diagnosed with ASD based on DSM-5 criteria and Non-ASD. The data were collected through observations, medical records (assessments), and with permission from the relevant parties. The dataset consists of 13 parameters reflecting aspects of communication, social interaction, repetitive behavior, and demographic factors in accordance with DSM-5. The details of these parameters are as follows:

TABLE I. PARAMETER DETAILS

| Parameter | Category & Score |
|---|---|
| Eye Contact | 0 = No, 1 = Good |
| Speaking Ability | 0 = Non-verbal, 1 = Limited, 2 = Fluent |
| Language Comprehension | 0 = Does not understand, 1 = Limited, 2 = Good |
| Social Interaction | 0 = None, 1 = Limited, 2 = Good |
| Stimming (Repetitive Movements) | 0 = No, 1 = Yes |
| Rigidity (Flexibility) | 0 = Flexible, 1 = Rigid |
| Echolalia (Repeating Words) | 0 = No, 1 = Yes |
| Dreaming (Withdrawn Behavior) | 0 = No, 1 = Yes |
| Social Awareness | 0 = Low, 1 = Moderate, 2 = High |
| Gender | M (Male), F (Female) |
| Age | 2–15 Years |
| CARS (Childhood Autism Rating Scale) | 0 = Normal (0–29), 1 = Mild (30–36), 2 = Moderate (37–40), 3 = Severe (40–60) |
| Diagnosis | 0 = ASD 1, 1 = ASD 2, 2 = ASD 3, 3 = Non-ASD |

The determination of ASD severity levels is based on the criteria established in the DSM-5. ASD classification is divided into three levels of severity: Level 1 (Requiring Support), Level 2 (Requiring Substantial Support), and Level 3 (Requiring Very Substantial Support).

#### 3) Data Preparation

This stage aims to prepare data from various sources so it can be optimally used in the autism classification process for children. The process was conducted in Google Colab using Python libraries such as pandas, numpy, matplotlib, and seaborn to facilitate data manipulation, analysis, and visualization.

*a) Data Selection*

The dataset consists of 81 records of children with 13 attributes, including 11 numerical and 2 categorical variables. The children's ages range from 2 to 15 years. The data is stored in CSV format for analysis purposes.

*b) Data Transformation*

Data transformation was performed using LabelEncoder to convert categorical data into numerical form, such as converting the gender column from "L" and "P" to 0 and 1. The diagnosis column was also converted, with ASD 1 labeled as 0, ASD 2 as 1, ASD 3 as 2, and Non-ASD as 3.

*c) Data Cleaning*

The purpose of data cleaning is to ensure the dataset is free from errors or missing values. In this study, checks were carried out for missing and duplicate data. As a result, no missing or duplicate entries were found in the dataset, so no further action was required. Afterward, the features (X) and the label (y) were separated. The diagnosis column was removed from the features (X) and used as the label (y) to be predicted by the model.

*d) Data Normalization*

Normalization was performed using StandardScaler, a method that transforms each feature value into a standard scale with a mean of 0 and a standard deviation of 1. The goal is to ensure all features are on the same scale, helping the model perform more optimally.

#### 4) Modelling

At this stage, after the dataset has gone through the preparation process, the next step is to perform modeling using the prepared dataset. This study uses one of the machine learning models, namely Support Vector Machine (SVM). The modeling process involves building two models: one SVM model without SMOTE, and one with SMOTE to handle data imbalance.

*a) Dataset Splitting*

The dataset in this study was split into a training set and a test set. The split was performed using an 80:20 ratio, where 80% of the data was used for training and the remaining 20% for testing [13]. This ratio is suitable for small datasets and provides an ideal balance between learning and evaluation data.

*b) Synthetic Minority Over-Sampling Technique (SMOTE)*

SMOTE is an oversampling technique used to address class imbalance in machine learning, which can affect a model's ability to recognize minority classes [13]. Unlike simply duplicating data, SMOTE generates new synthetic samples for the minority class, resulting in a more balanced data distribution before training. Specifically, SMOTE selects a data point $E_i$ and one of its neighbors $E_j$ at random. A new synthetic sample is then generated using the following formula, where $E_j$ is taken from the nearest neighbors set [10]:

$$Enew = E_i + (E_j - E_i)\,\delta \quad (1)$$

Where δ is a value randomly selected from the interval (0,1). The value of δ determines the distance of the generated synthetic sample from the original point $E_i$ [10]. SMOTE is applied only to the training data, as applying it to the test data could result in data leakage, making the model evaluation invalid [13]. In this study, SMOTE is used to handle class imbalance in the "diagnosis" category by building a pipeline that includes data standardization, oversampling using SMOTE, and training an SVM model with a linear kernel.

*c) Support Vector Machine Classification*

Support Vector Machine (SVM) is a powerful method for building classifiers that aims to create a decision boundary between two classes, allowing for the labeling of one or more feature vectors. The hyperplane or decision boundary is oriented to be as far as possible from the closest data points of each class—these closest points are called support vectors [14]. Identifying a reproducible hyperplane is part of the SVM decision function training process. This process maximizes the margin, or distance, between the support vectors of both class labels. Therefore, the optimal hyperplane is the one that maximizes the margin between the classes. After splitting the data, an SVM model with a linear kernel is trained on the training set to find the best hyperplane that optimally separates the classes [15]. The optimal hyperplane can be defined by the following equation:

$$wx^T + b = 0 \quad (2)$$

Where w is the weight vector, x is the input feature vector, and b is the bias. The values of w and b must satisfy the following inequalities for all elements in the training set:

$$wx_i^T + b \geq +1 \text{ if } y_i = 1 \quad (3)$$

$$wx_i^T + b \leq +1 \text{ if } y_i = -1 \quad (4)$$

The goal of training the SVM model is to find w and b such that the hyperplane separates the data while maximizing the margin, which is defined as $\frac{1}{\|w\|^2}$ [14].

*5) Evaluation*

At this stage, the model is evaluated to determine how well the SVM performs in classifying the data. The evaluation is conducted using K-Fold Cross-Validation with k=5, Confusion Matrix based on the applied cross-validation, and AUC-ROC. This evaluation provides a more comprehensive picture of the model's performance in ASD data classification.

*a) K-Fold Cross-Validation*

K-Fold Cross-Validation is a popular evaluation method in machine learning because it balances accuracy and computational efficiency. This technique divides the dataset into k equal-sized folds. The model is trained k times, with each iteration using one fold as the test set and the remaining folds as the training set. The results of all iterations are then averaged to obtain a more stable accuracy estimate. In this study, the value of k=5 is used, as it provides a good balance between bias and variance and yields a more accurate error estimate [16].

*b) Confusion Matrix*

Confusion Matrix is a 2x2 table that compares the model's predicted results with the actual data. It is used as an evaluation tool in classification tasks. The confusion matrix contains the following components [17]:

- True Positive (TP) – The data is actually positive and predicted as positive.
- False Positive (FP) – The data is actually negative but predicted as positive.
- False Negative (FN) – The data is actually positive but predicted as negative.
- True Negative (TN) – The data is actually negative and predicted as negative.

The confusion matrix also enables the calculation of other performance metrics such as precision, recall, and F1-scores.

*c) Accuracy*

Accuracy indicates the percentage of predictions the model correctly makes, both for positive and negative cases. The formula for calculating accuracy is as follows [18]:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

*d) Precision*

Precision measures how accurate the model is when predicting something as positive. The formula for calculating precision is [18]:

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

*e) Recall*

Recall, also known as sensitivity, focuses on how well the model identifies all actual positive cases. The formula for calculating recall is [18]:

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

*f) F1-Scores*

F1-Scores is a metric that combines precision and recall into a single number using the harmonic mean. F1-Scores consider the balance between how accurately the model classifies (precision) and how completely it captures all positive cases (recall). The formula for calculating F1-Scores is as follows [18]:

$$F1 - Scores = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (8)$$

*g) Receiver Operating Characteristic*

Receiver Operating Characteristic (ROC) is a curve or graph that illustrates the performance of a classification model. In this graph, the x-axis represents the False Positive Rate (FPR), while the y-axis represents the True Positive Rate (TPR) or sensitivity [13].

*h) Area Under the Curve*

Area Under the Curve (AUC) is a quantitative measure of the area under the ROC curve. The AUC value ranges between 0.5 and 1.0. A value of 0.5 indicates that the model performs no better than random guessing, while a value of 1.0 indicates that the model is capable of performing perfect classification [13].

## III. RESULTS

### A. Data Collection

Data collection was carried out through observations and interviews to obtain primary data in the form of medical records or assessments of children with and without ASD. A total of 81 child data entries were used, consisting of symptoms of children diagnosed with ASD based on the DSM-5 criteria.

### B. Minority Class Distribution Analysis

In this study, the label used is "diagnosis," with an imbalanced number of data samples. The following shows the number of samples in each diagnosis class:

TABLE II. DIAGNOSIS DISTRIBUTION

| Category | Number |
|---|---|
| ASD 1 | 8 |
| ASD 2 | 21 |
| ASD 3 | 33 |
| Non-ASD | 19 |

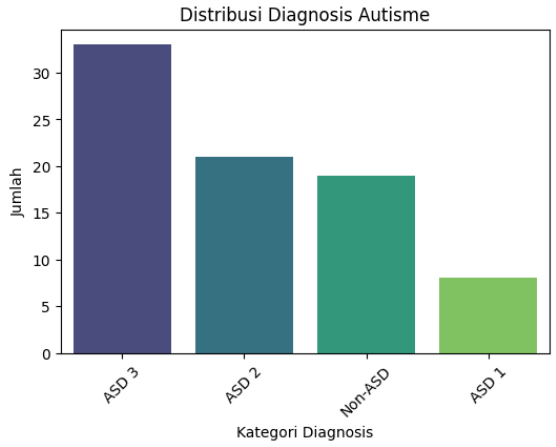


Fig. 1. Autism Diagnosis Distribution

In the distribution of ASD diagnoses, the majority of children in this dataset fall under the ASD Level 3 category, while ASD Level 1 has the fewest data entries compared to the other categories. This imbalance will be addressed by applying an oversampling method—specifically, the Synthetic Minority Over-Sampling Technique (SMOTE)—to increase the number of samples in the minority classes. This approach helps ensure that the classification model can accurately identify ASD cases without being biased toward the majority class.

### C. Handling Imbalanced Dataset

The Synthetic Minority Over-Sampling Technique (SMOTE) is an oversampling method used in machine learning to address class imbalance in training data. The table below shows the number of samples in each diagnosis category within the training dataset:

TABLE III. DIAGNOSIS DISTRIBUTION

| Category | Number |
|---|---|
| ASD 1 | 7 |
| ASD 2 | 15 |
| ASD 3 | 27 |
| Non-ASD | 15 |

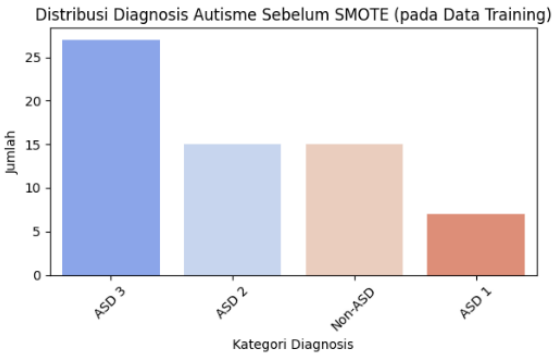


Fig. 2. Diagnosis Distribution in Training Data

A distribution like this may cause the model to become biased toward the majority class. Therefore, SMOTE is applied to increase the number of samples in the minority classes. SMOTE with k_neighbors=1 means that it uses the single nearest neighbor to generate synthetic samples. After training the model using a pipeline that combines SVM with SMOTE, the number of samples in each diagnosis category within the training data becomes more balanced, as shown below:

TABLE IV. TRAINING DATA DISTRIBUTION AFTER SMOTE

| Category | Number |
|---|---|
| ASD 1 | 27 |
| ASD 2 | 27 |
| ASD 3 | 27 |
| Non-ASD | 27 |

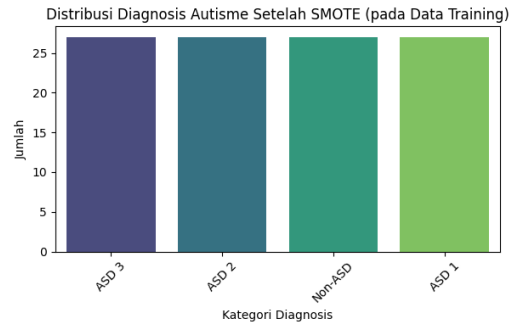


Fig. 3. Diagnosis Distribution After SMOTE

As seen in Table 4 and Fig. 3, the diagnosis distribution in the training data becomes balanced, allowing the model to learn more effectively without bias toward the majority class.

### *D. Evaluation of SVM Model Without SMOTE*

The evaluation was carried out on the SVM model without applying the SMOTE oversampling technique. A 5-fold cross-validation was performed to maximize the use of a small dataset and to prevent overfitting.

TABLE V. 5-FOLD CROSS-VALIDATION RESULTS

| Fold | Accuracy |
|---|---|
| Fold 1 | 100% |
| Fold 2 | 100% |
| Fold 3 | 100% |
| Fold 4 | 93.75% |
| Fold 5 | 100% |
| **Average (Mean)** | **98.75%** |

The 5-fold cross-validation produced results where 3 folds achieved 100% accuracy, and 1 fold achieved 93.75%, resulting in an average accuracy of 98.75%. These results indicate that the model is stable, accurate, and capable of generalizing well. The evaluation was further conducted using a confusion matrix based on the cross-validation results.

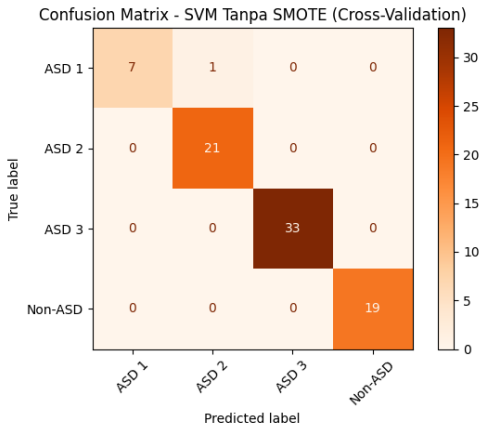


Fig. 4. Confusion Matrix of SVM Without SMOTE

The model shows a high True Positive (TP) rate for ASD 2, ASD 3, and Non-ASD classes, indicating strong classification capability. However, for the ASD 1 class, there are still some False Positives (FP) and False Negatives (FN), suggesting challenges in distinguishing between closely related categories. Despite this, the overall performance of the SVM model remains strong, with a relatively high level of accuracy. Based on the confusion matrix, the precision, recall, F1-scores, and accuracy metrics were calculated as follows:

TABLE VI. CLASSIFICATION REPORT

| Class | Precision | Recall | F1-scores | ROC-AUC |
|---|---|---|---|---|
| ASD 1 | 1.00 | 0.88 | 0.93 | 1.00 |
| ASD 2 | 0.95 | 1.00 | 0.98 | 1.00 |
| ASD 3 | 1.00 | 1,00 | 1.00 | 1.00 |
| Non-ASD | 1.00 | 1.00 | 1.00 | 1.00 |
| **Accuracy** | **0.9877** | | | |

SVM model without SMOTE achieved an accuracy of 98.77%, which aligns with the cross-validation result of 98.75%. The highest precision was obtained in the ASD 3 and Non-ASD classes (100%), while recall was generally high across all classes. However, recall for ASD 1 was slightly lower (88%) due to some misclassifications into other classes. The high F1-scores demonstrate a good balance between precision and recall. Overall, the model demonstrates excellent classification performance, despite a slight drop in performance for the minority class (ASD 1) due to class imbalance.

### *E. Evaluation of SVM Model with SMOTE*

The evaluation was conducted on the SVM model by applying the SMOTE oversampling technique. 5-fold cross-validation was performed on the SMOTE + SVM pipeline.

TABLE VII. 5-FOLD CROSS-VALIDATION RESULTS

| Fold | Accuracy |
|---|---|
| Fold 1 | 100% |
| Fold 2 | 100% |
| Fold 3 | 93.75% |
| Fold 4 | 87.50% |
| Fold 5 | 68.75% |
| **Average (Mean)** | **90%** |

Table 7 shows high accuracy in most folds, with Fold 1 and Fold 2 reaching 100%. However, Fold 5 experienced a significant drop to 68.75%, possibly due to the dominance of minority class samples or data variance. The overall average accuracy was 90%. Further evaluation was conducted using a confusion matrix based on the cross-validation results.

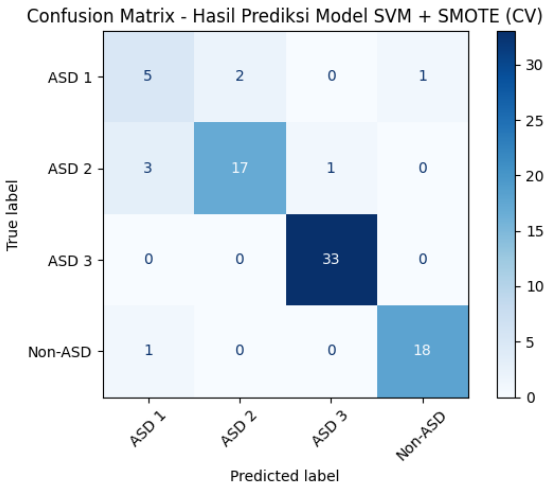


Fig. 5. Confusion Matrix of SVM with SMOTE

The model demonstrates a high True Positive (TP) rate for the ASD 2, ASD 3, and Non-ASD classes, indicating good classification performance. However, in the ASD 1 class, there are still some False Positives (FP) and False Negatives (FN), particularly involving misclassifications between ASD 2 and Non-ASD. This suggests difficulties in distinguishing between closely related categories. After applying SMOTE, the prediction performance became more balanced, especially for minority classes such as ASD 1 and ASD 2. However, there was a slight increase in FP and FN errors in these classes. Overall, the SVM model with SMOTE still demonstrated strong classification performance and maintained a high accuracy. Based on the previous confusion matrix, the precision, recall, F1-scores, and AUC-ROC values were calculated as follows:

TABLE VIII. CLASSIFICATION REPORT

| Class | Precision | Recall | F1-scores | AUC-ROC |
|---|---|---|---|---|
| ASD 1 | 0.56 | 0.62 | 0.59 | 0.98 |
| ASD 2 | 0.89 | 0.81 | 0.85 | 0.99 |
| ASD 3 | 0.97 | 1,00 | 0.99 | 1.00 |
| Non-ASD | 0.95 | 0.95 | 0.95 | 0.99 |
| **Accuracy** | **0.9012** | | | |

The SVM model with SMOTE achieved an accuracy of 90.12%, consistent with the cross-validation result of 90%. The highest precision and recall scores were observed in the ASD 3 and Non-ASD classes, both exceeding 95%, indicating the model's strong ability to recognize majority classes. However, for the ASD 1 class, both precision and recall were relatively low, with an F1-score of only 59%, reflecting the model's difficulty in distinguishing this class. Although SMOTE improved the representation and sensitivity of minority classes, the overall performance of the model slightly decreased compared to the SVM model without SMOTE. Nonetheless, the general accuracy remains high.

### *F. Comparison of Evaluation Results*

After evaluating the SVM models with and without SMOTE using cross-validation, confusion matrix, classification report, and ROC-AUC, a performance comparison was conducted to determine which model demonstrates better accuracy and overall performance.

TABLE IX. COMPARISON OF MODEL EVALUATION RESULTS

| Method | CV Accuracy (%) | Test Accuracy (%) | Precision (%) | Recall (%) | F1-Scores (%) | AUC-ROC |
|---|---|---|---|---|---|---|
| SVM without SMOTE | 98,75% | 98,77% | 99% | 97% | 98% | 1.00 |
| SVM with SMOTE | 90% | 90,12% | 84% | 85% | 84% | 0.99 |

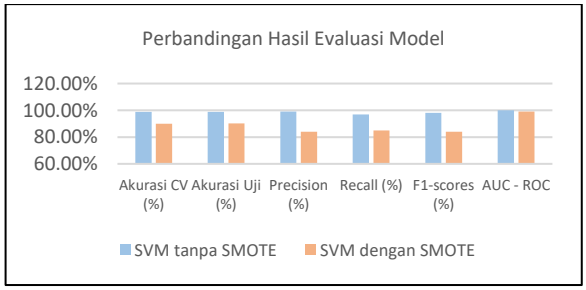


Fig. 6. Comparison of Model Evaluation Results

Based on Table 8 and Fig. 6, the SVM model without SMOTE achieved the highest cross-validation accuracy of 98.75%, while the accuracy dropped to 90.00% with SMOTE. This result is consistent with the test accuracy, which was 98.77% without SMOTE and 90.12% with SMOTE. Precision, recall, and F1-scores were also higher in the model without SMOTE. Although SMOTE reduced the overall accuracy, it helped improve sensitivity to minority classes such as ASD 1. Overall, the SVM model without SMOTE is considered the best-performing model, as it delivers higher accuracy and more balanced classification performance.

### *G. Feature Importance*

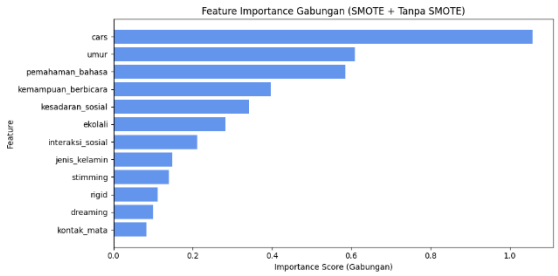


Fig. 7. Feature Importance Chart

Based on the permutation feature importance method, Fig. 7 shows that the most influential feature in autism classification is CARS, with the highest score of 1.0572, indicating its critical role in predicting ASD. Other significantly impactful features include speech ability, social awareness, and language comprehension. These results suggest that the model places greater emphasis on core clinical aspects such as CARS and communication skills, rather than minor behavioral features, in the ASD classification process.

## IV. DISCUSSION

The results of this study demonstrate that the Support Vector Machine (SVM) is capable of classifying autism effectively based on DSM-5 criteria. The SVM model without SMOTE achieved higher performance as it learned from the original data with a more natural distribution. While SMOTE helps balance the class distribution, applying it to a small dataset can introduce noise. In clinical practice, classification accuracy is crucial, as incorrect classification may lead to inappropriate interventions. This finding aligns with previous studies, which have shown that although SMOTE can improve sensitivity to minority classes, it may reduce the overall accuracy of the model [19]. This study has several limitations, including a small dataset size, the decreased accuracy due to SMOTE, and the scope limited only to the evaluation phase. Future research could involve larger and more diverse datasets, explore alternative sampling techniques, and consider developing the model into a mobile or web-based application.

## V. CONCLUSION

Support Vector Machine (SVM) proved to be effective in classifying autism based on DSM-5 criteria, with the SVM model without SMOTE achieving the highest accuracy and

the most balanced performance, making it the best choice in this study. Although SMOTE was able to improve sensitivity, it resulted in an overall decrease in accuracy, with a difference of 8.65% compared to the SVM model without SMOTE considered a significant gap, especially in the context of medical classification [20]. This study highlights the potential of SVM as a tool for early autism screening. However, its implementation must still undergo further clinical validation to ensure its reliability in real-world applications.

## REFERENCES


[1] Dewi, S., & Morawati, S. (2024). Gangguan Autis pada Anak. Scientific Journal, 3(6), 418–431.

[2] autism. (2023). Retrieved from World Health organization: https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders

[3] Marini, S.KM, D. (2024, November 13). Kajian Epidemiologis, Anak dengan Autisme di Indonesia. Retrieved from Orang Tua Hebat: https://www.orangtuahebat.id/kajian-epidemiologis-anak-autisme/

[4] Eko Cahyo, A., & Nilogiri, A. (2022). Klasifikasi Gangguan Autisme Pada Anak Menggunakan Algoritma C4.5 Dengan Teknik Random forest. National Multidisciplinary Sciences, 1(6), 846–851. https://doi.org/10.32528/nms.v1i6.241

[5] APA. (2013). Diagnostic And Statistical Manual of Mental Disorders Fifth Edition DSM-5. In United States.

[6] Zaenuddin, I., & Riyan, A. B. (2024). Perkembangan Kecerdasan Buatan (AI) Dan Dampaknya Pada Dunia Teknologi. Jurnal Informatika Utama, 2(2), 128–153.

[7] Sugara, B., & Subekti, A. (2019). Penerapan Support vector machine (Svm) Pada Small Dataset Untuk Deteksi Dini Gangguan Autisme. Jurnal Pilar Nusa Mandiri, 15(2), 177–182. https://doi.org/10.33480/pilar.v15i2.649

[8] Raj, S., & Masood, S. (2020). Analysis and Detection of Autism Spectrum Disorder Using Machine learning Techniques. Procedia Computer Science, 167(2019), 994–1004. https://doi.org/10.1016/j.procs.2020.03.399.

[9] Widjiyati, N. (2021). Implementasi Algoritme Random forest Pada Klasifikasi Dataset Credit Approval. Jurnal Janitra Informatika Dan Sistem Informasi, 1(1), 1–7. https://doi.org/10.25008/janitra.v1i1.118

[10] Mulla, G. A. A., Demİr, Y., & Hassan, M. M. (2021). Combination of PCA with SMOTE Oversampling for Classification of High-Dimensional Imbalanced Data. BEU Journal of Science, 10(3), 858–869.

[11] Hazarika, A., Kumar, C. J., & Das, P. R. (2022). Autism Spectrum Disorder diagnosis and machine learning: a review. International Journal of Medical Engineering and Informatics, 14(6), 512. https://doi.org/10.1504/ijmei.2022.10050777

[12] Supriyadi, R., Maulidah, N., Fauzi, A., Nalatissifa, H., & Diantika, S. (2022). Penerapan Algoritma Naive Bayes Dan Support Vector Machine dalam Memprediksi Autisme. Swabumi, 10(1), 55–59. https://doi.org/10.31294/swabumi.v10i1.12294

[13] Nigrou, N. (2024). Predicting Autism Spectrum Disorder Using Machine learning Classifiers. January. https://doi.org/10.13140/RG.2.2.35833.44646

[14] Huang, S., Nianguang, C. A. I., Penzuti Pacheco, P., Narandes, S., Wang, Y., & Wayne, X. U. (2018). Applications of support vector machine (SVM) learning in cancer genomics. Cancer Genomics and Proteomics, 15(1), 41–51. https://doi.org/10.21873/cgp.20063

[15] Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. In Machine Learning: Methods and Applications to Brain Disorders. Elsevier Inc. https://doi.org/10.1016/B978-0-12-815739-8.00006-7

[16] Nti, I. K., Nyarko-Boateng, O., & Aning, J. (2021). Performance of Machine Learning Algorithms with Different K Values in K-fold Cross-Validation. International Journal of Information Technology and Computer Science, 13(6), 61–71. https://doi.org/10.5815/ijitcs.2021.06.05.

[17] Zeng, G. (2020). On the confusion matrix in credit scoring and its analytical properties. Communications in Statistics - Theory and Methods, 49(9), 2080–2093. https://doi.org/10.1080/03610926.2019.1568485

[18] Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics, 21(1), 1–13. https://doi.org/10.1186/s12864-019-6413-7

[19] Putri, V. M., Masjkur, M., & Suhaeni, C. (2021). Performance of SMOTE in a random forest and naive Bayes classifier for imbalanced Hepatitis-B vaccination status. Journal of Physics: Conference Series, 1863(1). https://doi.org/10.1088/1742-6596/1863/1/012073.

[20] Sowjanya, A. M., & Mrudula, O. (2023). Effective treatment of imbalanced datasets in health care using modified SMOTE coupled with stacked deep learning algorithms. Applied Nanoscience (Switzerland), 13(3), 1829–1840. https://doi.org/10.1007/s13204-021-02063-4