TIFDA JOURNAL
Technology Information and Data Analytic
Journal Website : http://jeptira.id/index.php/tifda
e-ISSN: 3064-0660

Implementation of The Random Forest Algorithm for Early Detection Indications of Autism in Special Needs School (SLB) Students

Bagus Tri Mahardika[1*], Duha Nur Pambudi[2]
[1.2] Program Studi Teknologi Informasi Fakultas Teknik,Universitas DarmaPersada
[1,2] Jl. Taman Malaka Selatan No.8 Pd. Kelapa, Kec. Duren Sawit, Kota Jakarta Timur, Daerah Khusus Ibukota Jakarta, 13450, Indonesia
*bagusunsada @gmail.com

Abstract — This study aims to develop a system for early detection signs of autism in pupils at Special Needs Schools (SLB) by applying the Random Forest method. The problem addressed is how to provide an accurate and easily accessible tool for the early identification of signs of autism. The solution involves developing a Random Forest-based classification model using data from the Autism Spectrum Quotient (AQ-10) questionnaire, and then integrating it into a web application system built with a PHP frontend and a Flask backend. This system allows users to complete the questionnaire, upload data, and obtain prediction results automatically. Test results show that the model has an average accuracy of 99%, precision of 98%, recall of 100%, and an F1-score of 99%, as well as an AUC value above 0.98 in every fold. Consequently, this system is effective as a tool for initial screening to detect signs of autism in students at special schools in a practical and efficient manner.

Keywords – Autism, Random Forest, Prediction, Special Needs School, AQ-10 Questionnaire, Web-Based System, Machine Learning

I. INTRODUCTION

Autism, or Autism Spectrum Disorder (ASD), is a neurodevelopmental disorder that affects a child's communication skills and behaviour. This condition can be identified from an early age, and timely detection and appropriate management are crucial to ensure that pupils receive optimal educational support. However, the process of identifying autism in Special Schools (SLB) still relies heavily on manual observation by teachers or parents, which tends to be subjective and has the potential to cause delays in intervention [1], [2].

The urgency of this research stems from the fact that autism is a neurodevelopmental disorder that requires early intervention to ensure children receive appropriate educational support. In practice, the process of early identification of autism within Special Education Schools (SLB) still relies heavily on the subjective observations of teachers and parents, which can potentially lead to delays in intervention. The limited number of child psychology and psychiatry specialists also poses a challenge to the process of regular screening. Therefore, there is a need for an artificial intelligence-based early detection system capable of assisting in the identification of Autism Spectrum Disorder (ASD) indicators more quickly, objectively, and with greater accessibility, serving as a tool for initial decision-making prior to diagnosis by professionals.

Advances in information technology, particularly in the fields of data mining and machine learning, have opened up opportunities to develop more objective and efficient autism screening systems. The Random Forest algorithm [3] is widely used in classification research due to its ability to handle high-dimensional data, both categorical and numerical, whilst providing high prediction accuracy [4]. Recent studies have also

demonstrated the success of machine learning in supporting the process of identifying autism based on questionnaire data and the demographic characteristics of pupils [5]. For the predictive model to function optimally, the data preparation process is crucial; this includes data cleaning, attribute transformation, one-hot encoding for categorical features, and class balancing using the Synthetic Minority Over-sampling Technique (SMOTE) [4]. After the data has been thoroughly processed, a Random Forest-based prediction system can be created and included into a web application that parents and teachers can utilize directly in the special needs school setting. In light of this, the goal of this project is to create a system that uses the Random to identify early indicators of autism in students attending special schools. current data processing methods and the forest method. It is anticipated that this system will function as an impartial, efficient, and user-friendly preliminary screening instrument in the field of special education.

II. METHODOLOGY

This study applies the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework as the main framework for developing a system to predict signs of autism in special needs school pupils. CRISP-DM was chosen because it provides a systematic and proven workflow for data analysis projects. This model consists of six main stages, namely: business understanding, data understanding, data preparation, modelling, evaluation, and deployment. Each stage has been adapted to the needs of the research as follows:

1. Business Understanding

The process begins with an understanding of the need for early autism screening in special schools to support appropriate educational interventions. This study aims to develop an accurate and user-friendly system for EARLY DETECTION signs of autism to serve as a decision-making tool.

2. Data Understanding

The data used comprises the combined results of the Autism Spectrum Quotient (AQ-10) questionnaire completed by students at special schools, parents, support teachers, and publicly available data. At this stage, data exploration was carried out, key attributes were identified (AQ-10 scores, age, gender, ethnicity, etc.), and the distribution of ASD and non-ASD groups was analysed.

3. Data Preparation

At this stage, data cleaning was carried out (removal of duplicates, correction of spelling errors, and handling of incomplete data), ages were converted to positive integers, and a new feature total_score was added, calculated as the sum of the AQ-10 scores. Subsequently, one-hot encoding was applied to categorical attributes and data balancing was performed using SMOTE to address class imbalance.

4. Modeling

The pre-processed data was used to build a classification model using the Random Forest algorithm. Training was carried out using stratified k-fold cross-validation to ensure robust evaluation results with minimal bias. SMOTE was applied to the training data in each fold to maintain balanced class proportions.

5. Evaluation

Model performance was evaluated using accuracy, precision, recall, F1-score, and area under the curve (AUC). In addition, feature importance analysis and model interpretation were carried out using SHAP to ensure that the prediction results are understandable to users.

6. Deployment

The best model has been implemented in a web application built using PHP and Flask, so that the prediction results can be accessed directly by teachers and parents of students at special needs schools.

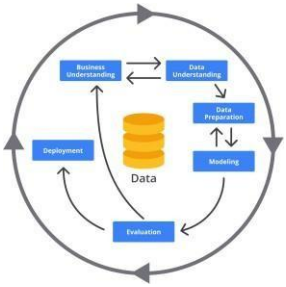


Fig 1. CRISP-DM Method

The dataset used in this study comprises a combination of data from the Autism Spectrum Quotient (AQ-10) questionnaire, participants' demographic data, and supporting health attributes. The variables used include:

AQ-10 scores (A1 to A10)

Student age, Gender, Ethnicity, Health history, Total AQ-10 score, ASD and Non-ASD class labels. The data then underwent data cleaning, attribute transformation, one-hot encoding for categorical data, and class balancing using the SMOTE method before being used in the Random Forest modelling stage.

The model was evaluated using the Stratified 5-Fold Cross-Validation method to ensure it possessed good generalisation ability and to minimise bias resulting from data partitioning. For each fold, several evaluation metrics were measured, namely: Accuracy, Precision, Recall, F1-Score, Area Under the Curve (AUC), Confusion Matrix. In addition, Feature Importance and SHAP (Shapley Additive

Explanations) analyses were performed to determine the contribution of each feature to the classification results, ensuring that the resulting model is not only accurate but also explainable.

III. THEORETICAL FRAMEWORK

This Literature Review section discusses the theories and concepts relevant to the research. It explains the underlying theory.

1. Autism Spectrum Disorder (ASD)

Autism, or Autism Spectrum Disorder, is a developmental disorder that affects an individual's communication skills, behaviour and social interaction. Early detection and intervention are crucial to ensure that children with ASD receive appropriate support and education tailored to their needs [2], [6].

2. Data Mining

Data mining is the process of automatically discovering patterns, insights or new information from large datasets. One of the key techniques in data mining is machine learning, which is the process of learning patterns from historical data for use in early detection or classifying new data [7], [8].

3. Random Forest

Random Forest is an ensemble learning algorithm consisting of many decision trees that operate in parallel. Each tree is constructed based on a data sample and a subset of features selected at random. For classification, the final prediction is determined by the majority vote across all the trees. Random Forest is known to be effective in

4. Synthetic Minority Over-sampling Technique (SMOTE)

SMOTE is a method for balancing class distributions in imbalanced datasets. This technique generates synthetic data for the minority class by interpolating values between existing data points, thereby making the data distribution more balanced and enabling the predictive model to learn optimally [4], [10].

5. One-Hot Encoding

One-hot encoding is a method of converting categorical data into a numerical format by creating a binary column (0 or 1) for each unique category within a feature. This technique is used to enable machine learning algorithms to process categorical data without assuming any order between categories [9], [11].

6. Model Evaluation

Classification models are typically evaluated using metrics such as accuracy, precision, recall, F1-score, and area under the curve (AUC). These metrics are used to assess how well the model distinguishes between target classes in both training and test data [4], [9].

7. CRISP-DM

CRISP-DM (Cross-Industry Standard Process for Data Mining) is a standard framework for data mining development comprising six stages: business understanding, data understanding, data preparation, modelling, evaluation, and implementation. This model helps researchers carry out the data analysis process in a structured and systematic manner [8].

IV. RESULTS AND DISCUSSION

This section discusses the results of the development and evaluation of an autism prediction system that has been implemented as a web application based on PHP and Flask. The performance of the prediction model was tested using data on special needs school pupils obtained from the AQ-10 questionnaire screening, demographic data, and health attributes, which were processed automatically via the web application. All stages, from application implementation and evaluation of the Random Forest model to the analysis of key features, are discussed in a structured manner to illustrate the reliability and benefits of the developed system.

1. Implementation of a web application system

An autism screening prediction system has been successfully implemented as a web application built using PHP and Flask. Users such as teachers and parents can access the application's home page to enter screening data digitally, including AQ-10 questionnaire scores, demographic data and pupils' health information.

2. Input Process and Output Prediction

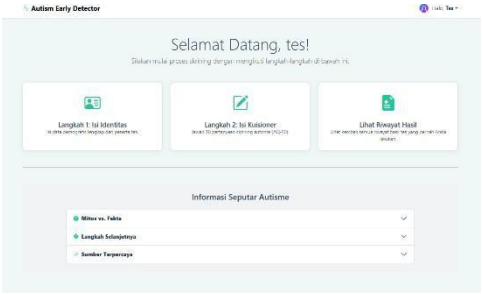


Fig 2. Dashboard User

Once the data has been entered into the web application form, the system automatically processes the data using a pre-trained Random Forest model. The results of the autism screening are displayed in real time within the application, making the screening process faster and more efficient. The application also stores a history of predictions for users' reference.

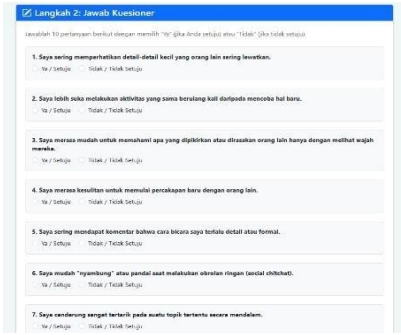


Fig 3. Questionnaire Input

Fig 4. Output Test

3. Model Evaluation with Cross-Validation

The Random Forest model was tested using stratified 5-fold cross-validation. The evaluation results for each fold showed accuracy ranging from 92% to 96%, with very high AUC values (0.98–0.99) across all folds. This is evident in the ROC curve, which demonstrates the model's excellent ability to distinguish between the ASD and non-ASD classes.

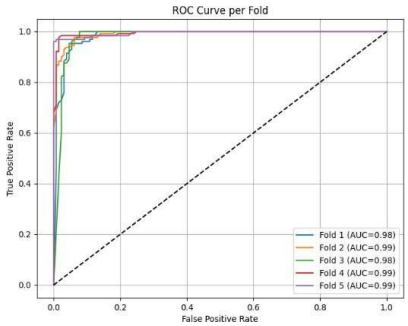
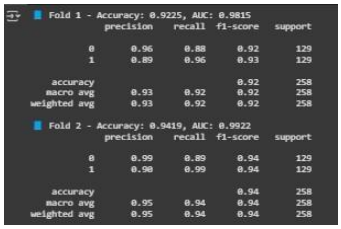


Fig 5. Roc Curve

4. Accuracy, Precision, Recall and F1-Score Results

A summary of the evaluation results for each fold can be seen in the metrics table, where the average precision, recall and F1-score are all above 0.90. This indicates that the model performs very consistently across various subsets of the data.



```
Fold 1 - Accuracy: 0.9225, AUC: 0.9815
              precision    recall  f1-score   support

           0       0.96      0.88      0.92       129
           1       0.89      0.96      0.93       129

    accuracy                           0.92       258
   macro avg       0.93      0.92      0.92       258
weighted avg       0.93      0.92      0.92       258

Fold 2 - Accuracy: 0.9419, AUC: 0.9922
              precision    recall  f1-score   support

           0       0.99      0.89      0.94       129
           1       0.90      0.99      0.94       129

    accuracy                           0.94       258
   macro avg       0.95      0.94      0.94       258
weighted avg       0.95      0.94      0.94       258
```

Fig 6. Fold Accuracy

5. Confusion Matrix and classification balance

Analysis of the confusion matrix shows that the number of incorrect predictions is very small; the model is able to classify the Non-ASD and ASD classes equally well, with near-perfect precision and recall values for both classes.

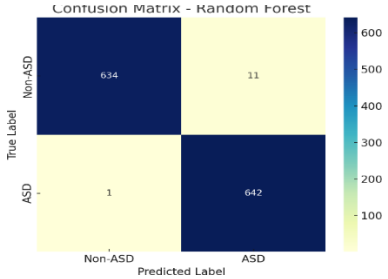


Fig 7. Confusion Matrix

6. Feature Importance Analyst

The results of the feature importance analysis in the Random Forest model show that the features most influential on the prediction are the AQ-10 score (particularly A6_Score), age, and certain demographic attributes. This supports the view that a combination of screening scores and individual learner factors is crucial in the identification of autism.

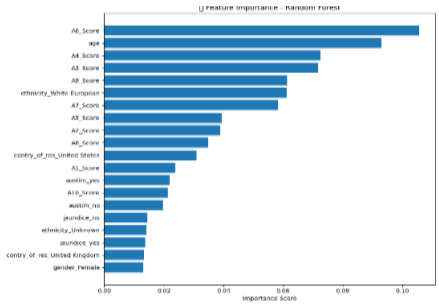


Fig 8. Feature Importance

7. Model Interpretation with SHAP

The model was interpreted using the SHAP (Shapley Additive Explanations) method to provide a deeper understanding of the influence of each feature on the prediction results. SHAP visualisations help users see the contribution of key features, such as the AQ-10 score and age, to the model's decisions, whilst ensuring the system's transparency. Consequently, the model is not only accurate but also interpretable to non-technical users.

Fig 9. Shap Explainer

IV. CONCLUSION

This study successfully developed a system for early detection signs of autism in special needs school pupils, based on a Random Forest algorithm integrated into a web application. The model built demonstrated excellent performance with high accuracy, precision, recall, F1-score, and AUC values across the entire test dataset. The data preparation process, which included data cleaning, one-hot encoding, feature engineering, and class balancing using SMOTE, was found to contribute significantly to the model's performance.

Feature importance analysis and model interpretation using SHAP confirmed that AQ-10 questionnaire scores and demographic factors such as age are the primary determinants in the prediction. The implementation of the web application facilitates teachers and parents in conducting initial autism screening objectively, quickly, and efficiently. Overall, the results of this study demonstrate that the application of the Random Forest method and data-driven machine learning can serve as an effective solution in supporting the early detection of autism within special education settings.

This research makes the following contributions: (1) Developing an early screening system for Autism Spectrum Disorder (ASD) in special needs school pupils, based on Random Forest and integrated into a web application. (2) Combining AQ-10 questionnaire data, demographic data, and health attributes to improve screening accuracy. (3) Applying the SMOTE technique to address class imbalance, thereby making the model more stable in distinguishing between ASD and non-ASD classes. (4) Implementing Explainable Artificial Intelligence (XAI) via the SHAP method so that screening results can be explained to teachers, parents, and educational staff. (5) Producing a web-based system that can be used as a practical and efficient early screening tool within the special education setting.

REFERENCES

[1] I. Sopiandi, R. Waliyudin Hidayat, and R. Nurahmat Damara, "Analisis Pemetaan Ilmiah tentang Perkembangan Explainable Artificial Intelelligence," EXPLORE, vol. 15, no. 2, pp. 2087–894, Jul. 2025, doi: 10.35200/ex.v15i2.166.

[2] S. Sukma, Memahami Autisme. DIVA Press, 2023.

[3] A. L. Puspanagara, "Penerapan Explainable AI untuk Prediksi Performa Akademik Mahasiswa Menggunakan Random Forest dan SHAP," Infoman's, vol. 19, no. 1, pp. 1–7, May 2025, doi: 10.13140/RG.2.2.27853.14565.

[4] M. H. Musyaffa, T. H. Saragih, D. T. Nugrahadi, D. Kartini, and A. Farmadi, "Effectiveness of SMOTE in Enhancing Adult Autism Spectrum Disorder Diagnosis Predictive Performance With Missforest Imputation And Random Forest," Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics, vol. 7, no. 2, pp. 270–280, Apr. 2025, doi: 10.35882/ijeeemi.v7i2.66.

[5] C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 3rd ed. Lulu.com, 2022. Accessed: Jul. 07, 2025. [Online].Available: https://christophm.github.io/interpretable-ml-book/

[6] Shinta Delfianti, Khalida Ayuni, Alifah Rizki, and Hijriati Hijriati, "Analisis Karakteristik Anak Berkebutuhan Khusus: Autisme Di Flexi School Banda Aceh," Ta'rim: Jurnal Pendidikan dan Anak Usia Dini, vol. 5, no. 2, pp. 97–106, May 2024, doi: 10.59059/tarim.v5i2.1244.

[7] Alex J. Smola & S.V.N. Vishwanathan, Introduction to Machine Learning. Cambridge University Press, 2022.

[8] A. Blum, J. Hopcroft, and R. Kannan, Foundations of Data Science. Cambridge University, 2022.

[9] A. Géron, Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow : concepts, tools, and techniques to build intelligent systems, 3rd ed. O'Reilly Media, Inc., 2022.

[10] A. Novianto and M. D. Anasanti, "Autism Spectrum Disorder (ASD) Identification Using Feature-Based Machine Learning Classification Model," IJCCS (Indonesian Journal of Computing and Cybernetics.