Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics e-ISSN: 2656-8624
Homepage: https://ijeeemi.org/; Vol. 7, No. 2, pp. 270-280, April 2025

RESEARCH PAPER

OPEN ACCES

Effectiveness of SMOTE in Enhancing Adult Autism Spectrum Disorder Diagnosis Predictive Performance With Missforest Imputation And Random Forest

Muhammad Hafizh Musyaffa, Triando Hamonangan Saragih, Dodon Turianto Nugrahadi, Dwi Kartini, Andi Farmadi

Department of Computer Science, Lambung Mangkurat University, Kalimantan Selatan,Indonesia

ABSTRACT

Autism Spectrum Disorder (ASD), originally described by Leo Kanner in 1943, is a complex developmental condition that manifests through social, emotional, and behavioral challenges, often including speech delays and difficulties in interpersonal interactions. Despite significant advancements in diagnostic criteria over the years, accurate diagnosis of ASD in adults remains challenging due to limited access to comprehensive datasets and inherent methodological constraints. The Autism Screening Adult dataset used in this study exemplifies these issues, as it contains missing values and exhibits a marked class imbalance, both of which can adversely affect model performance. To address these challenges, we proposed a framework that integrates Random Forest classification with MissForest imputation and the Synthetic Minority Over-sampling Technique (SMOTE). MissForest effectively imputes missing data by employing an iterative random forest approach that preserves the underlying structure of the data without relying on strict parametric assumptions. Meanwhile, SMOTE generates synthetic samples for the minority class, thereby balancing the dataset and reducing prediction bias. Experimental evaluation through 10-Fold Cross Validation demonstrated that the application of SMOTE significantly enhanced model performance. Notably, the overall accuracy improved from 70.17% to 79.32%, and the AUC-ROC increased from 47.13% to 85.84%, indicating a robust improvement in the model's ability to distinguish between positive and negative cases. These results underscore the critical importance of addressing data imbalance and missing values in predictive modeling for ASD. The promising outcomes of this study provide a solid foundation for developing more reliable diagnostic tools for adult ASD, and future research may further refine feature selection and incorporate additional data sources to optimize performance even further.

PAPER HISTORY

Received April 02, 2024
Revised May 31, 2024
Accepted May 31, 2024

KEYWORDS

Autism Spectrum Disorder;
Imbalanced Data;
Missing Values;
MissForest;
Random Forest;
SMOTE;

AUTHOR EMAIL

hafizhktb@gmail.com
Triando.saragih@ulm.ac.id
dodonturianto@ulm.ac.id
dwikartini@ulm.ac.id
andifarmadi@ulm.ac.id

1. INTRODUCTION

In 1943, Leo Kanner first introduced the term "autism" as a diagnostic label to define a specific syndrome that is characterized by early onset, characteristic symptomatology, and disrupted social and emotional relationships in young children [1]. The neurodevelopmental abnormalities known as autism spectrum disorders (ASD) are multifaceted, widespread, and intricate. The diagnosis is based on the observation of abnormal conduct, with criteria centered on limited, repetitive patterns of behavior, interests, or hobbies as well as deficits in social communication and engagement [2]. Because of the intricate underlying pathomechanisms that are activated by a variety of events, ASD is highly heterogeneous. While some people and children with ASD need significant assistance to carry out fundamental tasks, others are completely capable of doing all activities of daily living [3]. Additionally, individuals on the autism spectrum may exhibit various neurological, psychiatric, and medical co-morbidities that are important to take into account when planning interventions. They may also exhibit various problem behaviors (such as self-harm, hetero-aggression, compulsivity, hyperactivity, etc.) that are concerning to educators, therapists, and caregivers and frequently lead to social stigma [4]. Autism Spectrum Disorder (ASD) is most commonly diagnosed during early childhood because its characteristic symptoms often emerge in infancy [5]. The diagnostic process involves a careful evaluation of an individual's behaviors and developmental patterns against established criteria. These criteria serve as a framework that outlines the specific symptoms required to confirm an ASD diagnosis while also identifying any factors that might rule it out. By systematically assessing both the presence and absence

Corresponding author: Triando Hamonangan Saragih, Triando.saragih@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Jalan A. Yani Km 36, Banjarbaru 70714, Kalimantan Selatan, Indonesia.
DOI: https://doi.org/10.35882/ijeeemi.v7i2.66

of these key symptoms, clinicians can ensure that the diagnosis is both accurate and appropriate for the individual's unique profile [6]. The Autism Spectrum Quotient (AQ), the Childhood Autism Rating Scale (CARS-2), and the Screening Tool for Autism in Toddlers and Young Children (STAT) are a few examples of screening techniques [7].

Even while most research on ASD has traditionally focused on children, there has been an increase in interest in evaluating ASD presentations in adults in recent years. Since milder types of ASD often go undetected in childhood, several researchers have placed special emphasis on the importance of detecting them in adult populations that do not include intellectual disability. This is particularly crucial because adult outcomes have been subpar, with many people finding it difficult to get and keep a job, succeed in higher education, and/or develop the skills necessary for independent life. Therefore, more study aimed at adults is required [8][9].

The diagnosis of ASD, particularly in adults, poses unique challenges due to the heterogeneity of symptoms and the lack of standardized diagnostic tools tailored for this population. Early diagnosis (between the ages of two and five) can open doors for therapy that could help a young child grow in particular areas, like social interaction, physical skills, and communication [10]. It is recommended that to increase the effectiveness and accessibility of the screening procedure and to speed up the diagnosis [11]. Now a days, machine learning has been applied to detect various diseases including ASD [12]. Machine learning (ML) presents a promising approach for early screening and diagnosis of ASD, enabling the identification of risk factors for prevention while offering efficient, accurate, and objective predictive capabilities [13][14].

Building on these advancements, one promising approach is to implement the Random Forest algorithm for ASD diagnosis. Random Forest classifier is a supervised classifier based on an ensemble of decision trees. The ensemble of trees is training by the bagging method [15]. Random Forest has been used in various fields, for instance, it had been used in banking for predicting customer response, for predicting the direction of stock market prices, in the medicine/pharmaceutical industry, e-commerce, etc [16].

Imputation, the process of substituting expected values for missing or anomalous ones, is a useful strategy for dealing with comparable missing values and associated data quality issues in many applications [17]. In this research, the MissForest imputation method is utilized. MissForest imputation uses random forest models to repeatedly impute nonresponses variable by variable in increasing order of the number of nonresponses after first imputation by the mean[18]. Unlike simple mean/mode substitution or k-nearest neighbors, which can oversimplify the data and fail to capture complex, non-linear relationships, MissForest leverages the power of RF to iteratively predict missing values, thereby preserving the intrinsic structure and variability of the data. This approach is particularly advantageous for our dataset, as it ensures that the underlying behavioral and demographic relationships remain intact, ultimately enhancing the reliability of subsequent analyses and predictions in Autism Spectrum Disorder diagnosis.

Another significant challenge in this research is the data imbalance, which can lead to biased predictions favoring the majority class. To address this issue, this research uses SMOTE. The SMOTE method generates synthetic data by applying linear interpolation between a minority class point and one of its K nearest neighbors [19].

In recent studies, Ramya and Arokiaraj (2024) employed a hybrid approach combining CNN and Random Forest, which yielded an accuracy of 98.75% in predicting an autism dataset. In order to increase the accuracy of predicting the severity of autism, this study presented a hybrid model that included Random Forest (RF) and Convolutional Neural Networks (CNNs). In addition to overcoming the drawbacks of independent models, the CNN and RF combination produced a more comprehensive and precise forecast. The CNN-RF model consistently outperformed the CNN and RF models separately in a variety of K-fold cross-validation scenarios, as evidenced by increased accuracy, precision, recall, and Kappa statistic. This finding confirmed the widely held data science observation that combining different models frequently produced better results than using them separately [20]. Novianto and Anasanti (2023) confirmed the effectiveness of the MissForest imputation method in achieving high-precision predictions for Autism Spectrum Disorder. In their study, MissForest proved instrumental in handling missing and anomalous values by iteratively refining the imputation process, thereby preserving the integrity of the underlying data. By integrating MissForest with advanced feature selection (SpFSR) and classification techniques (SVM), their hybrid model achieved an exceptional accuracy of 100%, surpassing previous studies. Despite the limited dimensionality of the dataset, the success of MissForest in this context underscored its potential to enhance predictive performance, reduce computational complexity, and lower data collection costs. Future research was suggested to further explore its application on larger and more complex datasets, along with more robust validation strategies, to confirm these promising results [21]. Ismail et al. (2023) demonstrated the effectiveness of a hybrid Stacking-SMOTE model for predicting autism spectrum disorder (ASD)-causing genes. By integrating the MissForest imputation method to address missing and anomalous data with SMOTE to overcome class imbalance, their approach leveraged robust feature selection (SpFSR) and advanced classification techniques, including ensemble methods like gradient boosting on Random Forest (GBBRF) within a stacking framework. This combination resulted in a significant improvement in predictive performance,

Corresponding author: Triando Hamonangan Saragih, Triando.saragih@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Jalan A. Yani Km 36, Banjarbaru 70714, Kalimantan Selatan, Indonesia.
DOI: https://doi.org/10.35882/ijeeemi.v7i2.66

achieving an accuracy of 95.5%—a notable advancement over previous studies. Furthermore, the incorporation of a hybrid gene similarity function based on GO annotations proved effective in enhancing the model's ability to capture functional similarities between genes, although the limited availability of GO annotations for some genes remained a challenge. Future research was expected to further integrate additional data sources, such as gene expression profiles and protein–protein interaction networks, to improve the model's accuracy and generalizability [22].

2. MATERIALS AND METHOD

In this study, we employed a comprehensive methodology that integrates several advanced techniques. Specifically, we utilized Random Forest as our primary machine learning model due to its robustness and effectiveness in classification tasks. To address the challenge of missing data, we applied the MissForest imputation method, which iteratively refines missing value estimates while preserving data integrity. Additionally, we implemented SMOTE to balance the dataset, ensuring that minority classes were adequately represented. Lastly, to thoroughly assess our model's performance across various data subsets, we used k-fold cross-validation throughout data processing. Fig. 1 shows the workflow for the research.

A. Data Collection

The Autism Screening Adult dataset, which was obtained from this link https://archive.ics.uci.edu/dataset/426/autism+screening+adult, consists of responses from adult participants who underwent an autism screening test designed to detect behavioral indicators associated with Autism Spectrum Disorder (ASD). The Autism Screening Adult dataset consists of 704 instances and 20 features that capture a range of behavioral responses and demographic information, making it a valuable resource for analyzing autism-related traits in adults. Each record in the dataset includes demographic attributes (e.g., age, gender, ethnicity), responses to specific screening questions, and an overall classification or label indicating whether the individual screened positive for ASD. Because this dataset is publicly accessible and has been curated for research purposes, it is well-suited for developing and evaluating machine learning models aimed at early ASD detection and intervention planning.

B. MissForest Imputation

Data imputation is a crucial process in machine learning and statistical analysis, as it helps to address missing or incomplete values within datasets. Without proper imputation, missing data can lead to biased results, reduced model performance, and a loss of valuable information. MissForest (MF), an iterative technique for imputation that uses random forests (RF), sets itself apart from conventional imputation techniques by not assuming normalcy or requiring modeling parameter requirements [23].

Missing values in a dataset are initially filled with a default value typically the mean or mode of the available data for each variable, although median or custom values can also be used. The missForest imputation method then employs an iterative process that uses Random Forest models to predict and update the missing values for each variable. At each iteration, the algorithm builds a Random Forest on the observed cases for a given variable and uses it to impute the missing entries. This process continues until a convergence criterion is met or the maximum number of iterations is reached. Throughout this iterative process, the initial imputed values and the trained Random Forest models for each variable are stored, allowing the same sequence of operations to be applied to new data. As a non-parametric imputation method, missForest is versatile, effectively handling both categorical and numerical data by leveraging the predictive power of Random Forests to maximize imputation accuracy [24][25]. Algorithm 1 gives a representation of the missForest method.

Algorithm 1. missForest

Require: X an $n \times p$ matrix, stopping criterion γ
1: Sort X by amount of missing values of stations descend;
2: Make an initial guess for missing values using another method;
3: **while** not γ **do**
4: $\mathrm{X}_{\mathrm{old}}^{\mathrm{imp}} \leftarrow$ store previously imputed matrix;
5: **for** s in $1 \cdots p$ **do**
6: Fit a random forest: $y_{obs}^{(s)} \sim x_{obs}^{(s)}$;
7: Predict $y_{mis}^{(s)}$ using $x_{mis}^{(s)}$;
8: $\mathrm{X}_{\mathrm{old}}^{\mathrm{imp}} \leftarrow$update impute matrix, using predicted $y_{obs}^{(s)}$;
9: update γ;
10: **return** the imputed matrix X^{imp}

C. Synthetic Minority Over-sampling Technique (SMOTE)

SMOTE, introduced in 2002 by Chawla et al., was developed to address the challenges posed by imbalanced classification. Its main purpose is to create synthetic samples for the minority class by leveraging k-nearest neighbors and Euclidean distance [26]. The SMOTE algorithm (Algorithm 2) operates in a straightforward and understandable manner. It chooses the minority class instances' k-nearest neighbors for every instance of the minority class. Then, along the line segments that link the chosen instance to one of its k-nearest neighbors, it generates synthetic instances. These artificial examples expand the minority class's representation and add fresh data points to the feature

Corresponding author: Triando Hamonangan Saragih, Triando.saragih@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Jalan A. Yani Km 36, Banjarbaru 70714, Kalimantan Selatan, Indonesia.
DOI: https://doi.org/10.35882/ijeeemi.v7i2.66

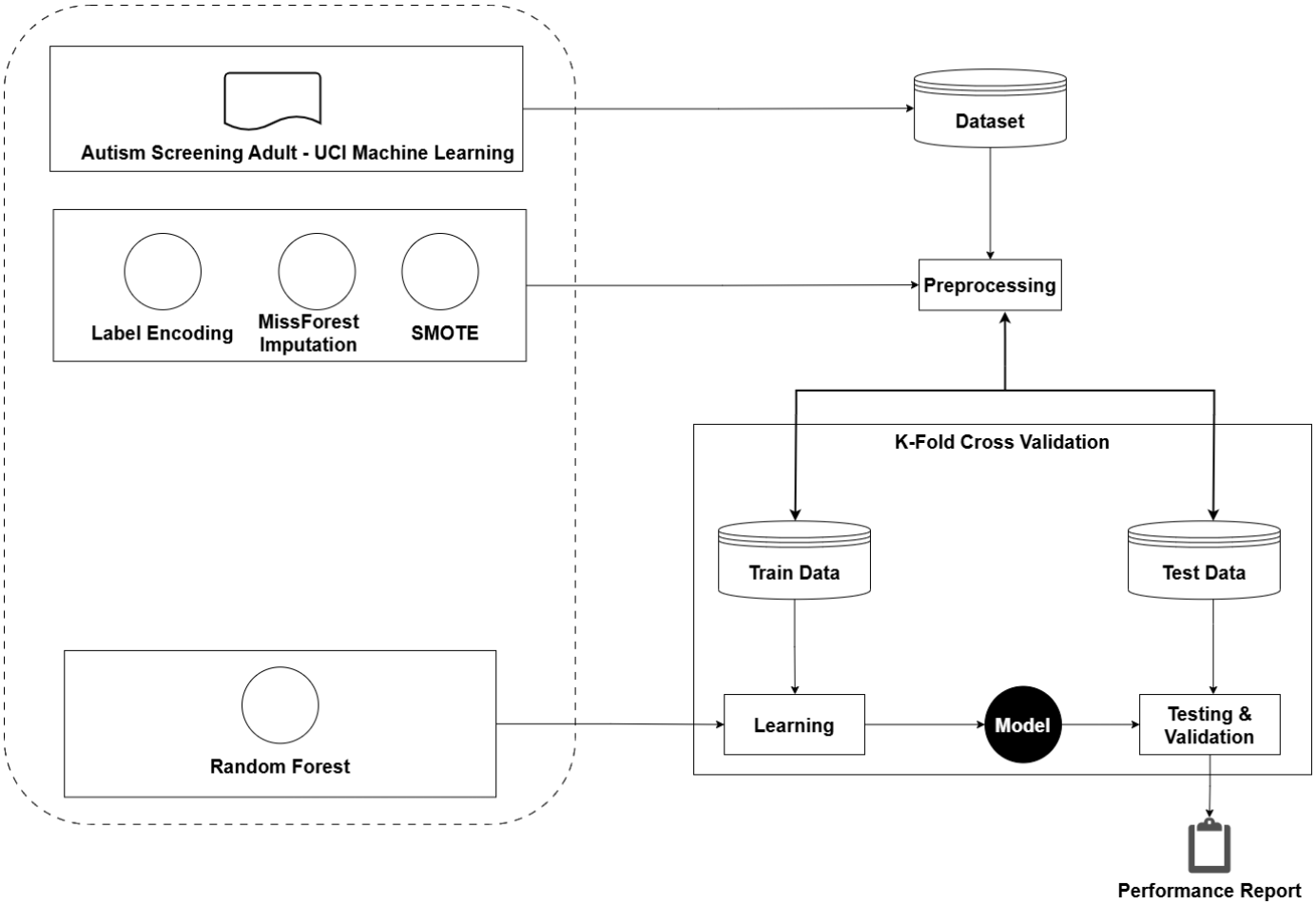

Fig 1. Data preprocessing, feature selection, modeling, evaluation, and final results interpretation.

space [27]. It typically outperforms simple oversampling and is commonly employed in various applications. The SMOTE method generates a synthetic sample by linearly combining two samples from the minority class (X_i and X_j) as follows in Eq. (1):

$$X_{new} = X_i + (X_j - X_i) * \alpha \quad (1)$$

For the new artificial instance X_{new} of the minority class, a sample X_i chosen randomly. Then, using the Euclidean distance, X_i is selected at random from the minority class's five closest neighbors [28]. A random float value between 0 and 1 is entered for the parameter α [29]. In our study, we employed SMOTE with the default setting of k=5, meaning that for each minority class instance, the 5 nearest neighbors were considered for generating synthetic samples (Algorithm 2). Using 5 neighbors helps to ensure that the new synthetic data points accurately reflect the underlying distribution of the minority class while minimizing the risk of overfitting or introducing noise.

Algorithm 2. SMOTE

Require: $X \in \mathbb{R}^{nxp}$ the features
Require: $Y \in \{0,1\}^n$ the binary class label outputs
Require: $k \in$ N the number of neighbors to select for the k-Nearest Neighbors.
Ensure: Generated data $X_{new} \in \mathbb{R}^{qxp}$ and $Y_{new} \in \{0,1\}^q$ with q points created.

1: Denote by S_1 the number of points labelled as the minority class and $S_0$the number of points labelled as the majority class.
2: Initialize $X_{new}$and Y_{new} as empty vectors.
3: **while** $S_1 < S_0$ **do**
4: Filter $\mathcal{D} = X_i | Y_i = 1$, the set of points labelled as minority class 1
5: Randomly choose $r \in \mathcal{D}$ and find the indices of its k nearest neighbors
6: Randomly choose an index r_2 among these neighbors.
7: $x^{new} \leftarrow \alpha \times x_{r1} + (1-\alpha) \times x_{r2}$ with $\alpha \in$ [0,1] randomly chosen
8: $y^{new} \leftarrow 1$
9: $S_1 = S_1 + 1$
10: Append x^{new} to X_{new}, append y^{new} to Y_{new}
11: **end while**
12: **return** X_{new} , Y_{new}

Corresponding author: Triando Hamonangan Saragih, Triando.saragih@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Jalan A. Yani Km 36, Banjarbaru 70714, Kalimantan Selatan, Indonesia.
DOI: https://doi.org/10.35882/ijeeemi.v7i2.66

D. Random Forest

Random Forest is an ensemble learning algorithm that generates multiple decision trees and aggregates their predictions, typically by averaging the outputs. This approach mitigates the risk of overfitting that individual decision trees may exhibit, thereby enhancing the model's generalization and predictive accuracy [30].

Ensemble approaches typically rely on either bagging or boosting. Random Forest adopts bagging by training each model on random subsets of the dataset in parallel. In ensemble frameworks, the final outcome is determined by combining all individual model outputs, such as by averaging in regression tasks [31]. Random Forest created a random vector Θ_k for every k^{th} tree, all vectors $\Theta_1, \ldots, \Theta_{k-1}$ are independent but have the same distribution; each tree builds a regression model $h(x, \Theta_k)$ where x is an input vector. The result is a set of regression values, these are evaluated by the mean and this becomes the Random Forest result.

$$E = -\sum_{i=1}^{c} pi \times \log(pi) \quad (2)$$

The Eq. (2) provides the entropy at each internal node of the decision tree, where pi is the prior probability of each class and c is the number of distinct classes. To obtain the greatest information at each decision tree split, this value is maximized. The mean squared error at each internal node is a frequently used splitting criterion for regression issues [32]. Algorithm 3 gives a representation of the missForest method.

Algorithm 3. Random Forest

Precondition: A training set ***S*** := $(x_1, y_1), \ldots, (x_n, y_n)$, features ***F,*** and number of trees in forest ***B.***

function RANDOMFOREST (***S***, ***F***)
  $\boldsymbol{H} \leftarrow \emptyset$
  for $i \in 1,\ldots,$ ***B*** **do**
    $S^{(i)} \leftarrow$ A bootsrap sample from ***S***
    $h_i \leftarrow$ RANDOMIZEDTREELEARN ($\boldsymbol{S^{(i)}}, \boldsymbol{F}$)
    $\boldsymbol{H} \leftarrow \boldsymbol{H} \cup \{h_i\}$
  end for
  return H
end function
function RANDOMIZEDTREELEARN (***S***, ***F***)
  At each node :
    $f \leftarrow$ very small subset of ***F***
    Split on best feature in f
  return the learned tree
end function

E. K-Fold Cross Validation

The K-fold CV is simple to build, adaptable, and rarely depends on model structure [33]. K-fold cross-validation divides the dataset into k subgroups of equal size.For each iteration, one subset is used for validation while the remaining k-1 subsets are used for training. This process is repeated k times, ensuring each subset is used exactly once for validation. The mean performance across all folds then serves as an estimate of the model's overall effectiveness [34]. We chose 10 folds because this number is widely recognized for offering a good balance between bias and variance using 10 folds generally results in a lower bias in performance estimation compared to methods like leave-one-out cross-validation, while still maintaining reasonable computational efficiency and a lower variance than using fewer folds.

$$k = \frac{N}{\alpha} \quad (3)$$

$$t = N - t \quad (4)$$

In Eq. (3), the variable "k" represents the ratio of the total number of data points (N) to the size of each fold. Thus, the total number of data points is calculated by multiplying the fold size by the number of folds. Eq. (4) further indicates that the training set size is derived by subtracting the test set size from the total number of data points [35]. Algorithm 4 gives a representation of the K-Fold Cross Validation method.

Algorithm 4. K-Fold Cross Validation

Divide data into k equal folds
for k in range(0,K)
  $V \leftarrow Fold_k$ in data
  $T \leftarrow data \backslash V$
  Train T
  $Acc_k \leftarrow$ evaluate V with trained model
end for
$Acc \leftarrow \frac{1}{K}\sum_{k=1}^{K} Acc_k$

F. Performance Metrics

The performance evaluation begins with the construction of a confusion matrix, which tabulates the true positives, true negatives, false positives, and false negatives observed in the model's predictions. This matrix provides a comprehensive overview of the classifier's decision outcomes, as expressed in Table (1).

Table 1. Predicted versus actual outcomes delineate classifier performance and error

Actual Class	Predicted Class: True	Predicted Class: False
True	True Positive (TP)	False Negative (FN)
False	False Positive (FP)	True Negative (TN)

Corresponding author: Triando Hamonangan Saragih, Triando.saragih@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Jalan A. Yani Km 36, Banjarbaru 70714, Kalimantan Selatan, Indonesia.
DOI: https://doi.org/10.35882/ijeeemi.v7i2.66

Derived from the confusion matrix, key performance metrics namely accuracy, precision, recall, and F1 score are computed. Accuracy is defined as the proportion of correct predictions out of all predictions made, providing an overall measure of model performance. Precision measures the percentage of instances classified as ASD that are truly ASD, indicating the model's reliability in its positive predictions. Recall (or sensitivity) assesses the percentage of actual ASD cases that were correctly identified, which is particularly important for ensuring that critical cases are not missed. The F1-score, the harmonic mean of precision and recall, offers a balanced metric that accounts for both false positives and false negatives. Detailed information is given in Table 2.

Table 2. Precision, recall, F1-score, and accuracy summarize model classification performance.

Performance Metrics	Formula
Accuracy	$\frac{TP+TN}{TP+FN+FP+TN}$
Recall	$\frac{TP}{TP+FN}$
Precision	$\frac{TP}{TP+FP}$
F1-Score	$\frac{2\times precison\times recall}{precision+recall}$

To further assess the model's discriminative capability, the Receiver Operating Characteristic (ROC) curve is generated, and the area under this curve (AUC) is calculated. A higher AUC indicates superior performance, with values ranging from 0 to 1, where 1 denotes perfect classification. This metric reflects the model's ability to differentiate between positive and negative cases across all possible classification thresholds. The ROC AUC is particularly significant for ASD classification as it encapsulates the trade-off between sensitivity (recall) and specificity, providing a single measure that is robust to class imbalance. This measure is mathematically defined as expressed in Eq. (5).

$$AUC = \frac{\left(\frac{TP}{TP+FN}\right) \times \left(\frac{TN}{TN+FP}\right)}{2} \quad (5)$$

Moreover, the interpretation of the AUC value offers valuable insights into a model's ability to distinguish between positive and negative classes. Additionally, AUC serves as a useful tool for model comparison and selection, allowing practitioners to assess the relative effectiveness of different classifiers [36]. To see the value of classification quality based on the AUC value can be seen at Table 3 [37].

Table 3. AUC-based accuracy score evaluates model performance in classification tasks

AUC Scores	Category
0.90 – 1.00	Excellent
0.80 – 0.90	Good
0.70 – 0.80	Fair
0.60 – 0.70	Poor
0.50 – 0.60	Failure

3. RESULTS

Before SMOTE was applied, our dataset exhibited a marked class imbalance, with the majority class containing 515 instances and the minority class only 189 instances,as shown in Figure 2. This disparity, illustrated by a 73:26 ratio, raised concerns about biased model training, as classifiers tend to favor the majority class, which could result in diminished predictive accuracy for the underrepresented group.

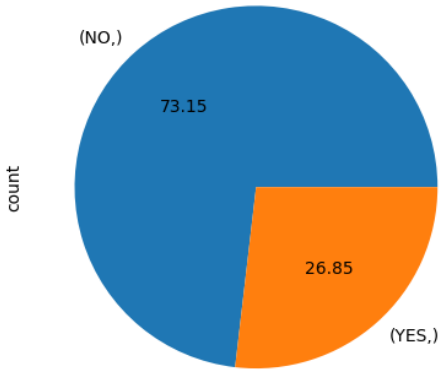

Fig 2. Visualization of class distribution before SMOTE was applied to the dataset.

After SMOTE was implemented, the minority class was synthetically oversampled to achieve a more balanced distribution of samples. The application of SMOTE generated additional minority class instances, effectively equalizing the class representation. This rebalancing was expected to enhance the classifier's performance by improving its ability to detect and correctly classify instances from both classes, thereby positively impacting key performance metrics such as precision, recall, F1-score, and AUC-ROC.

This section focuses on evaluating the Random Forest algorithm for predicting Autism Spectrum Disorder (ASD) outcomes. To ensure data quality, we incorporated imputation techniques using MissForest Imputation. The primary aim was to evaluate the efficacy of the classification model in forecasting complications related to ASD. To achieve this, a variety of performance metrics, including precision, recall, accuracy, F1-score, and the Area Under the Curve (AUC) of the ROC curve, were employed. Additionally, we compared the performance of the model before and after applying SMOTE to address

Corresponding author: Triando Hamonangan Saragih, Triando.saragih@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Jalan A. Yani Km 36, Banjarbaru 70714, Kalimantan Selatan, Indonesia.
DOI: https://doi.org/10.35882/ijeeemi.v7i2.66

data imbalance, thereby highlighting the impact of oversampling on enhancing prediction accuracy. Table 4 presents the performance metrics, illustrating the improvements achieved through SMOTE application. The table above shows the performance of the Random Forest classifier before and after applying the Synthetic Minority Oversampling Technique (SMOTE). Notably, the accuracy increases from 70.17% to 79.32%, and precision jumps from 58.47% to 79.55%. Similarly, recall improves from 70% to 79.32%, leading to a corresponding rise in the F1-score from 62% to 79.28%. The most significant gain is observed in the AUC-ROC, which surges from 47.13% to 85.84%. Overall, these findings indicate that using SMOTE to address class imbalance substantially enhances the model's predictive performance in classifying Autism Spectrum Disorder (ASD) cases. Fig. 3 shows the chart comparing performance metrics between SMOTE and non-SMOTE models, further illustrating the impact of oversampling on classification effectiveness.The model using SMOTE showed an increase from 47.13% to 85.84%, indicating a much-enhanced capacity to differentiate between positive and negative classes, highlighting the significance of AUC-ROC as well.

Table 4. Comparison of performance metrics between models using SMOTE and non-SMOTE approaches.

Evaluation	NO SMOTE	SMOTE
Accuracy	70.17%	79.32%
Precision	58.47%	79.55%
Recall	70.17%	79.32%
F1 - Score	61.98%	79.28%
AUC - ROC	47.13%	85.84%

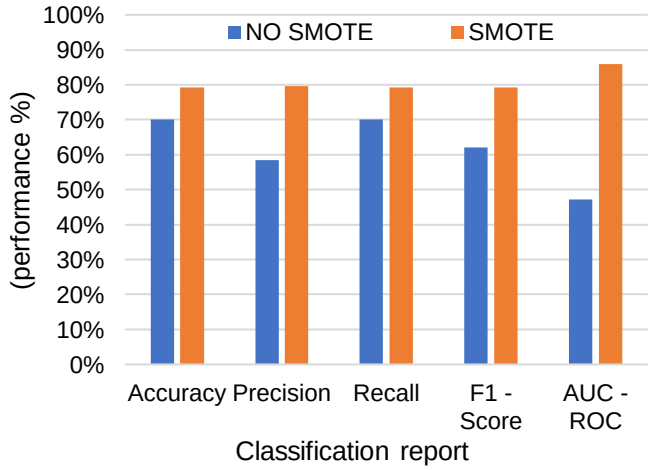

Fig 3. Visualization of performance metrics comparison between SMOTE and non-SMOTE models.

Consequently, the significant increase in AUC-ROC highlights that the SMOTE-based strategy is more successful in this medical diagnostic setting, even though certain trade-offs may be seen in other metrics. Fig. 4 shows the visualization of ROC-AUC for SMOTE and non-SMOTE models.

4. DISCUSSION

This study investigated the efficacy of a Random Forest classifier in predicting Autism Spectrum Disorder (ASD) outcomes, with a particular focus on the impact of addressing class imbalance through SMOTE. The model was evaluated using stratified 10-fold cross-validation, and several performance metrics were analyzed, including accuracy, precision, recall, F1-score, and AUC-ROC. Our results indicated that applying SMOTE led to a notable improvement in most performance metrics. Specifically, accuracy increased from 70.17% to 79.32%, and precision rose from 58.47% to 79.55%. These enhancements suggested that the classifier became more reliable in correctly predicting positive instances when the minority class was oversampled. The improvement in recall, which increased from 70.17% to 79.32%, further indicated that the model with SMOTE was better at identifying positive cases overall. Consequently, the F1-score, which balanced precision and recall, increased from 61.98% to 79.28%.

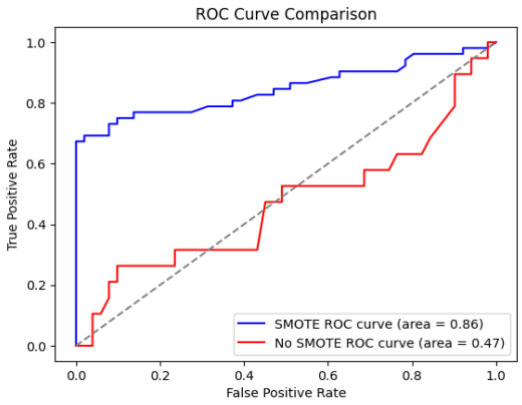

Fig 4. ROC Curve Comparison Between SMOTE and Non-SMOTE Model Performance Visualization

One of the most significant findings was reflected in the AUC-ROC metric. The model without SMOTE achieved an AUC-ROC of 47.13%, while the use of SMOTE boosted this value to 85.84%. This substantial increase in AUC-ROC highlighted the enhanced discriminative ability of the model in distinguishing between positive and negative classes, a critical factor in medical diagnostics where misclassifications could have had significant consequences.

In comparison with previous studies, our approach utilizing Random Forest with SMOTE and MissForest imputation achieved an accuracy of 79.32%, with precision, recall, and F1-scores around 79%. By contrast, Ramya and Arokiaraj [20] reported substantially higher

Corresponding author: Triando Hamonangan Saragih, Triando.saragih@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Jalan A. Yani Km 36, Banjarbaru 70714, Kalimantan Selatan, Indonesia.
DOI: https://doi.org/10.35882/ijeeemi.v7i2.66

Table 5. Comparison Between Past Research and Current Model Performance Results

Study	Model & Method	Accuracy (%)	Precision (%)	Recall (%)	F1 – Score (%)
This Study (2025)	Random Forest + SMOTE, MissForest	79.32	79.55	79.32	79.28
Ramya and Arokiaraj [21]	CNN + Random Forest	98.75	97.89	97.18	2.68
Novianto & Anasanti [22]	MissForest + SVM	100	100	100	100
Ismail et al. [23]	Random Forest + SMOTE	90.7	90.0	90.7	88.8

performance using a hybrid CNN and Random Forest model, with an accuracy of 98.75% and precision and recall exceeding 97%. Moreover, Novianto and Anasanti [21] obtained perfect classification (100% across all metrics) employing MissForest imputation combined with SVM, while Ismail et al. [22] achieved an accuracy of 90.7% using Random Forest with SMOTE. These variations may be attributed to differences in dataset characteristics, preprocessing methods, and model architectures, highlighting the trade-offs and challenges in optimizing classification performance for ASD prediction. Table 5 presents the performance metrics comparison between previous studies and the current research, illustrating the differences in accuracy, precision, recall, and F1-score across various methodologies.

The improvements observed with SMOTE underscore the importance of addressing data imbalance in predictive modeling. While some trade-offs in certain metrics (such as a potential decrease in recall in some contexts) might be observed with oversampling techniques, the overall enhancement in performance metrics especially the dramatic rise in AUC-ROC demonstrates that SMOTE is highly effective in this application.

In conclusion, this study provides compelling evidence that integrating SMOTE into the data preprocessing pipeline significantly enhances the predictive performance of the Random Forest classifier in the context of ASD diagnosis. These findings not only highlight the critical role of addressing class imbalance but also pave the way for future research that can explore the integration of additional data sources and alternative oversampling techniques to further optimize model performance.

5. CONCLUSION

This study aimed to predict Autism Spectrum Disorder (ASD) outcomes using a Random Forest classifier in combination with SMOTE for addressing class imbalance and MissForest for imputing missing values. The proposed approach achieved an accuracy of 79.32%, with precision, recall, and F1-scores of 79.55%, 79.32%, and 79.28%, respectively.

Future research should focus on refining the feature selection process, exploring alternative oversampling strategies, and validating the model on larger, more diverse datasets. Additionally, incorporating boosting techniques, such as Gradient Boosting or XGBoost, into the ensemble framework could further enhance the predictive performance and reliability of ASD diagnosis models.

These findings underscore the critical role of oversampling techniques in managing imbalanced datasets. Consequently, this study provides deeper insights into the effectiveness of different data imputation approaches and machine learning algorithms. Looking ahead, future work could focus on advanced feature selection strategies and the integration of multiple classification models to further enhance predictive performance.

REFERENCES

[1] A. Genovese and M. G. Butler, “Clinical assessment, genetics, and treatment approaches in autism spectrum disorder (ASD),” Jul. 01, 2020, *MDPI AG*. doi: 10.3390/ijms21134726.

[2] A. Sheik Abdullah, K. V, S. Geetha, and U. Mishra, “Leveraging deep learning for enhanced diagnosis of autism spectrum disorder using resting-state functional magnetic resonance imaging and clinical data,” *Results in Engineering*, vol. 25, p. 104444, Mar. 2025, doi: 10.1016/j.rineng.2025.104444.

[3] A. K. Sauer, J. E. Stanton, S. Hans, and A. M. Grabrucker, “Autism Spectrum Disorders: Etiology and Pathology,” 2021, doi: 10.36255/exonpublications.

[4] G. M. Guazzo, “Self-Injurios Behaviour in an Adult with Autism Spectrum Desorder,” *European Journal of Medical and Health Research*, vol. 2, no. 5, pp. 80–86, Sep. 2024, doi: 10.59324/ejmhr.2024.2(5).09.

[5] B. M. Lupindo, A. Maw, and N. Shabalala, “Late diagnosis of autism: exploring experiences of males diagnosed with autism in adulthood,” *Current Psychology*, vol. 42, no. 28, pp. 24181–24197, Oct. 2023, doi: 10.1007/s12144-022-03514-z.

[6] P. Joon, A. Kumar, and M. Parle, “What is autism?,” Oct. 01, 2021, *Springer Science and*

Corresponding author: Triando Hamonangan Saragih, Triando.saragih@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Jalan A. Yani Km 36, Banjarbaru 70714, Kalimantan Selatan, Indonesia.
DOI: https://doi.org/10.35882/ijeeemi.v7i2.66

Business Media Deutschland GmbH. doi: 10.1007/s43440-021-00244-0.

[7] F. Thabtah and D. Peebles, "A new machine learning model based on induction of rules for autism detection," *Health Informatics J*, vol. 26, no. 1, pp. 264–286, Mar. 2020, doi: 10.1177/1460458218824711.

[8] A. Hillier, A. Buckingham, and D. Schena, "Physical Activity Among Adults With Autism: Participation, Attitudes, and Barriers," *Percept Mot Skills*, vol. 127, no. 5, pp. 874–890, Oct. 2020, doi: 10.1177/0031512520927560.

[9] B. Carpita *et al.*, "Platelet Levels of Brain-Derived Neurotrophic Factor in Adults with Autism Spectrum Disorder: Is There a Specific Association with Autism Spectrum Psychopathology?," *Biomedicines*, vol. 12, no. 7, Jul. 2024, doi: 10.3390/biomedicines12071529.

[10] C. Okoye *et al.*, "Early Diagnosis of Autism Spectrum Disorder: A Review and Analysis of the Risks and Benefits," *Cureus*, Aug. 2023, doi: 10.7759/cureus.43226.

[11] A. Kusumaningsih, C. V. Angkoso, and A. K. Nugroho, "Autism Screening Prediction Based on Multi-kernel Support Vector Machine," in *Proceeding - IEEE 9th Information Technology International Seminar, ITIS 2023*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ITIS59651.2023.10420224.

[12] R. Nurfalah, S. Rahayu, and M. F. Akbar, "The Analysis of Adult Autism Spectrum Disorders Screening Using Neural Network," *SinkrOn*, vol. 4, no. 1, p. 196, Oct. 2019, doi: 10.33395/sinkron.v4i1.10148.

[13] O. Khan Durrani, "Innovative Autism Spectrum Disorder Prediction Using Machine Learning," 2024. [Online]. Available: www.ijsdr.org

[14] A. F. Pina, M. J. Meneses, I. Sousa-Lima, R. Henriques, J. F. Raposo, and M. P. Macedo, "Big data and machine learning to tackle diabetes management," Jan. 01, 2023, *John Wiley and Sons Inc*. doi: 10.1111/eci.13890.

[15] A. Basri and M. Arif, "Classification of Seizure Types Using Random Forest Classifier," *Advances in Science and Technology Research Journal*, vol. 15, no. 3, pp. 167–178, 2021, doi: 10.12913/22998624/140542.

[16] F. Farooq *et al.*, "A comparative study of random forest and genetic engineering programming for the prediction of compressive strength of high strength concrete (HSC)," *Applied Sciences (Switzerland)*, vol. 10, no. 20, pp. 1–18, Oct. 2020, doi: 10.3390/app10207330.

[17] B. Cho *et al.*, "Effective Missing Value Imputation Methods for Building Monitoring Data," in *Proceedings - 2020 IEEE International Conference on Big Data, Big Data 2020*, Institute of Electrical and Electronics Engineers Inc., Dec. 2020, pp. 2866–2875. doi: 10.1109/BigData50022.2020.9378230.

[18] A. Tripet, E. Eustache, and Y. Tillé, "Improving Donor Imputation Using the Prediction Power of Random Forests: a Combination of SwissCheese and missForest," *J Surv Stat Methodol*, Nov. 2023, doi: 10.1093/jssam/smad040.

[19] D. Elreedy, A. F. Atiya, and F. Kamalov, "A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning," *Mach Learn*, vol. 113, no. 7, pp. 4903–4923, Jul. 2024, doi: 10.1007/s10994-022-06296-4.

[20] R. Ramya and S. P. Arokiaraj, "ENHANCED AUTISM SEVERITY PREDICTION: A FUSION OF CONVOLUTIONAL NEURAL NETWORKS AND RANDOM FOREST MODEL," 2024.

[21] 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 Systems)*, vol. 17, no. 3, p. 259, Jul. 2023, doi: 10.22146/ijccs.83585.

[22] E. Ismail, W. Gad, and M. Hashem, "A hybrid Stacking-SMOTE model for optimizing the prediction of autistic genes," *BMC Bioinformatics*, vol. 24, no. 1, Dec. 2023, doi: 10.1186/s12859-023-05501-y.

[23] Y. H. Hu, R. Y. Wu, Y. C. Lin, and T. Y. Lin, "A novel MissForest-based missing values imputation approach with recursive feature elimination in medical applications," *BMC Med Res Methodol*, vol. 24, no. 1, p. 269, Dec. 2024, doi: 10.1186/s12874-024-02392-2.

[24] F. Domebale Maale, O. O. Awe, and G. A. Okeyere, "Effects of Imputations Techniques on Predictive Performance of Supervised Machine Learning Algorithms: Empirical Insights from Health Data Classification." [Online]. Available: https://ssrn.com/abstract=4437047

[25] E. Albu, S. Gao, L. Wynants, and B. Van Calster, "missForestPredict -- Missing data imputation for prediction settings," Jul. 2024, [Online]. Available: http://arxiv.org/abs/2407.03379

[26] N. Mohanty, B. K. Behera, C. Ferrie, and P. Dash, "A Quantum Approach to Synthetic Minority Oversampling Technique (SMOTE)," Feb. 2024, [Online]. Available: http://arxiv.org/abs/2402.17398

[27] S. Matharaarachchi, M. Domaratzki, and S. Muthukumarana, "Enhancing SMOTE for imbalanced data with abnormal minority instances," *Machine Learning with Applications*, vol. 18, p. 100597, Dec. 2024, doi: 10.1016/j.mlwa.2024.100597.

Corresponding author: Triando Hamonangan Saragih, Triando.saragih@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Jalan A. Yani Km 36, Banjarbaru 70714, Kalimantan Selatan, Indonesia.
DOI: https://doi.org/10.35882/ijeeemi.v7i2.66

[28] M. Kivrak, U. Avci, H. Uzun, and C. Ardic, "The Impact of the SMOTE Method on Machine Learning and Ensemble Learning Performance Results in Addressing Class Imbalance in Data Used for Predicting Total Testosterone Deficiency in Type 2 Diabetes Patients," *Diagnostics*, vol. 14, no. 23, Dec. 2024, doi: 10.3390/diagnostics14232634.

[29] A. J. Mohammed, "Improving Classification Performance for a Novel Imbalanced Medical Dataset using SMOTE Method," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 3, pp. 3161–3172, Jun. 2020, doi: 10.30534/ijatcse/2020/104932020.

[30] B. Grillone, S. Danov, A. Sumper, J. Cipriano, and G. Mor, "A review of deterministic and data-driven methods to quantify energy efficiency savings and to predict retrofitting scenarios in buildings," Oct. 01, 2020, *Elsevier Ltd*. doi: 10.1016/j.rser.2020.110027.

[31] B. R. Ramos Collin *et al.*, "Random forest regressor applied in prediction of percentages of calibers in mango production," *Information Processing in Agriculture*, 2024, doi: 10.1016/j.inpa.2024.12.002.

[32] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," *Stata Journal*, vol. 20, no. 1, pp. 3–29, Mar. 2020, doi: 10.1177/1536867X20909688.

[33] A. M. Peco Chacón, I. Segovia Ramírez, and F. P. García Márquez, "K-nearest neighbour and K-fold cross-validation used in wind turbines for false alarm detection," *Sustainable Futures*, vol. 6, Dec. 2023, doi: 10.1016/j.sftr.2023.100132.

[34] M. Alruqi, P. Sharma, and Ü. Ağbulut, "Investigations on biomass gasification derived producer gas and algal biodiesel to power a dual-fuel engines: Application of neural networks optimized with Bayesian approach and K-cross fold," *Energy*, vol. 282, Nov. 2023, doi: 10.1016/j.energy.2023.128336.

[35] S. Ünalan, O. Günay, I. Akkurt, K. Gunoglu, and H. O. Tekin, "A comparative study on breast cancer classification with stratified shuffle split and K-fold cross validation via ensembled machine learning," *J Radiat Res Appl Sci*, vol. 17, no. 4, p. 101080, Dec. 2024, doi: 10.1016/j.jrras.2024.101080.

[36] A. M. Akbar, R. Herteno, S. W. Saputro, M. R. Faisal, and R. A. Nugroho, "Enhancing Software Defect Prediction through Hybrid Optimization for Feature Selection and Gradient Boosting Classification," *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 6, no. 2, pp. 169–181, Apr. 2024, doi: 10.35882/jeeemi.v6i2.388.

[37] A. Tajali, H. Saragih, M. I. Mazdadi, I. Budiman, and A. Farmadi, "ShareAlike 4.0 International License (CC BY-SA 4.0). The Impactness of SMOTE as Imbalance Class Handling for Myocardial Infarction Complication Classification using Machine Learning Approach with Data Imputation and Hyperparameter," vol. 6, no. 4, pp. 227–239, 2024, doi: 10.35882/ijeeemi.v6i4.13.

AUTHOR BIOGRAPHY

Muhammad Hafizh Musyaffa is a dedicated student at Lambung Mangkurat University. He has been pursuing his studies in the Department of Computer Science since 2021, focusing on Data Science. His research primarily targets data analysis, machine learning, and predictive modeling, with an emphasis on solving complex real-world problems using innovative, data-driven solutions. He is extremely passionate about leveraging his advanced technical and analytical skills to drive meaningful advancements in technology, ultimately benefiting society at large. For further details or collaboration inquiries, please contact him at hafizhktb@gmail.com. He continually strives to expand his knowledge and contribute innovative research solutions for excellence.

Triando Hamonangan Saragih is a lecturer in the Department of Computer Science at Lambung Mangkurat University, deeply engaged in the field of Data Science. He completed his bachelor's degree in Informatics at Brawijaya University in Malang in 2016, and earned his master's degree in Computer Science from the same institution in 2018. His research is primarily focused on Data Science, contributing to the advancement of knowledge in this field. His innovative approach and scholarly commitment have significantly contributed to academic research and practical applications in Data Science. For further information or collaborative inquiries, please contact him at triando.saragih@ulm.ac.id (ORCID: 0000-0003-4346-3323).

Dodon Turianto Nugrahadi is a lecturer in the Department of Computer Science at Lambung Mangkurat University, specializing in Data Science and Computer Networking. He earned his bachelor's degree in Informatics Engineering from Petra University, Surabaya, in 2004, and later pursued a master's degree in Information Engineering at Gajah Mada University, Yogyakarta, in 2009. Currently, his research interests encompass

Corresponding author: Triando Hamonangan Saragih, Triando.saragih@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Jalan A. Yani Km 36, Banjarbaru 70714, Kalimantan Selatan, Indonesia.
DOI: https://doi.org/10.35882/ijeeemi.v7i2.66

network studies, data science, the Internet of Things (IoT), and network Quality of Service (QoS). His work contributes to advancements in technology and academic knowledge, fostering innovation and excellence in his field. He actively inspires peers and students with innovative, impactful research.application of neural networks and deep learning application for power system.

Dwi Kartini earned her Bachelor's and Master's degrees in Computer Science from the Faculty of Computer Science at Putra Indonesia "YPTK" in Padang, Indonesia. She is also a lecturer in the Department of Computer Science, where she teaches a variety of subjects including linear algebra, discrete mathematics, research methods, and others. Her research interests focus on the applications of Artificial Intelligence and Data Mining. Currently, she serves as an assistant professor in the Department of Computer Science, Faculty of Mathematics and Natural Sciences at Lambung Mangkurat University in Banjarbaru, Indonesia, and is the head of the Computer Science study program.

Andi Farmadi, a senior lecturer in the Computer Science program at Lambung Mangkurat University, has been teaching since 2008 and currently serves as the Head of the Data Science Lab since 2018. He completed his undergraduate studies at Hasanuddin University and his graduate studies at Bandung Institute of Technology. His research area, up to the present, focuses on Data Science. One of his research projects, along with other researchers, was published in the International Conference of Computer and Informatics Engineering (IC2IE), is titled "Hyperparameter tuning using GridsearchCV on the comparison of the activation function of the ELM method to the classification of pneumonia in toddlers," and this research was published in 2021.Email: andifarmadi@ulm.ac.id. Orcid ID: 0009-0009-0926-8082

Corresponding author: Triando Hamonangan Saragih, Triando.saragih@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Jalan A. Yani Km 36, Banjarbaru 70714, Kalimantan Selatan, Indonesia.
DOI: https://doi.org/10.35882/ijeeemi.v7i2.66