**Journal of Electronics, Electromedical Engineering, and Medical Informatics**
**Multidisciplinary: Rapid Review: Open Access Journal** **Vol. 6, No. 3, July 2024, pp: 302-311; eISSN: 2656-8632**

**RESEARCH ARTICLE** **OPEN ACCESS**

Manuscript received May 2, 2024; revised May 23, 2024; accepted May 27, 2024; date of publication July 8, 2024
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v6i3.470

How to cite: Difa Fitria, Triando Hamonangan Saragih, Muliadi, Dwi Kartini, and Fatma Indriani, "Classification of Appendicitis in Children Using SVM with KNN Imputation and SMOTE Approach to Improve Prediction Quality", Journal of Electronics, Electromedical Engineering, and Medical Informatics, vol. 6, no. 3, pp. 302-311, July 2024.

# Classification of Appendicitis in Children Using SVM with KNN Imputation and SMOTE Approach to Improve Prediction Quality

**Difa Fitria, Triando Hamonangan Saragih, Muliadi, Dwi Kartini, and Fatma Indriani**

Computer Science Department, Lambung Mangkurat University, Banjarbaru, South Kalimantan, Indonesia
Corresponding author: Triando Hamonangan Saragih (e-mail: triando.saragih@ulm.ac.id)

**ABSTRACT** This study evaluates the effect of using the Synthetic Minority Over-sampling Technique (SMOTE) on the performance of Support Vector Machine (SVM) classification models in the diagnosis of appendicitis in children. Class imbalance in medical data is often a significant challenge that can reduce the accuracy of predictive models. To address this, K-Nearest Neighbors (KNN) imputation is used to handle missing data in the dataset. An SVM model with a polynomial kernel was chosen for its ability to capture the non-linear relationship between clinical features and diagnosis. The polynomial kernel parameters were set with d = 3 and c = 1 to balance the model complexity and risk of overfitting and the use of SmOTE to address class imbalance. The results showed that the use of SMOTE increased the model precision from 87.00% to 98.65% and AUC-ROC from 85.96% to 88.04%. However, there was a decrease in recall from 92.55% to 77.66% and F1-Score from 89.69% to 86.90%. This suggests a trade-off between the model's increased ability to distinguish between positive and negative classes and its decreased ability to detect all positive instances. This study makes an essential contribution to medical informatics by showing that while SMOTE can improve some of the model's performance metrics, there are significant trade-offs that must be considered. These findings can assist medical professionals in making better decisions based on more accurate and representative data analysis, particularly in the diagnosis of appendicitis in children.

**INDEX TERMS** Appendicitis, KNN Imputation, Support Vector Machine, SMOTE.

## I. INTRODUCTION

Appendicitis is one of the indications for emergency abdominal surgery in children. If the appendix is perforated, morbidity and mortality will increase. Therefore, surgeons strive to make an accurate diagnosis as quickly as possible. This is not always possible in children. The history may be confusing or unrememberable, and the clinical presentation is also confusing in young children. Dr. Kottmeier found that at the time of appendectomy, 1% to 32% of patients did not have appendicitis. in general, a more accurate diagnosis is more difficult to make when the child is younger [1].

Lack of fiber and carbohydrate intake is a major factor in appendicitis in Asian, Indian, and African countries. The incidence of appendicitis is less than 10% compared to European countries, where low fiber intake has a higher risk of appendicitis. A US case study found that children in the 50th percentile and above who ate a diet high in fiber had a 30% lower risk of appendicitis compared to children in the lowest percentile [2].

The clinical presentation of appendicitis is fever, anorexia, nausea, and pain from the navel to the lower right. However, in many studies, the symptoms of appendicitis are relatively mild [3]. It is still difficult to diagnose experienced surgeons [4], [5]. An initial examination is relied upon to diagnose appendicitis, making it easier for the doctor or surgeon to make a quick decision [6]. Diagnosis is essential to avoid complications of appendicitis [7].

Machine learning algorithms are being studied and applied to various clinical workflow tasks, including disease prognosis and diagnosis, medical treatment, and patient care plan creation[8]. The use of technology, such as classification, is an effective solution that can improve the accuracy of diagnosis and provide better care to patients. [9]. Classification is one of the models in data mining. Classification models are data

prediction techniques that make predictions of the value of data that is derived from different data. Classification is one of the main tasks in machine learning and data mining, and it belongs to the supervised learning type.[10].

A classification technique that allows medical professionals to sort through and identify patterns of relevant information, enabling them to diagnose appendicitis patients based on specific features associated with the condition [3]. Technology is an effective solution that can improve diagnosis accuracy and provide better patient care. One classification method that can be used is the SVM (Support Vector Machine) classification method, which has proven successful in various medical applications, including diagnostic, prognostic, and clinical decision-making [11].

SVM kernels for classification and regression in recent years into the data mining, pattern recognition, and machine learning communities have been attracted to the tremendous generalization and discriminative power of SVMs. SVM has been used to solve practical binary classification problems. SVM has been proven to be better than other supervised learning methods [12]. SVM was developed in 1995 by Cortes and Vapnik. It is a supervised learning algorithm used in classification and regression analysis. SVM is also used for applications such as pattern recognition, data mining, and machine learning[13]. SVM is a classification technique used to create a decision boundary between two classes and allows the prediction of labels for one or more features [14].

When building SVM classifiers, it is necessary to specify a particular kernel function, such as a polynomial or radial basis function (RBF), which is an important learning parameter. However, little research focuses on assessing the predictive performance of SVM classifiers built using different kernel functions. Moreover, it is known that combining multiple classifiers or an ensemble of classifiers, another active research area in pattern classification, often provides better performance than a single classifier [15].

The dataset used in this study has a problem, namely the presence of missing values. The missing value problem can be solved by using data mining techniques. Missing or missing imputation replaces the value of the missing data with a reasonable one, although there is no standard rule regarding the terrible percentage of missing data [16]. Data mining is extracting information from data sets stored in data warehouses. Data mining is a series of processes that get patterns from data sets [17], [18]. One way to handle missing values is to fill them with possible values based on the information available in the data, commonly referred to as imputation techniques.

One method can be used in classification to handle missing value problems, namely K-Nearest Neighbor Imputation (KNNI). The K-Nearest Neighbor Imputation algorithm is a system that uses a supervised learning algorithm and aims to find new data patterns by connecting existing data patterns with new data. KNNI is an approach used to identify objects based on specific information that is the closest distance to the object. KNN imputation is a model-free method, although the majority rule is straightforward [19].

Previous research [20] The SMOTE method has been used to handle data imbalance, followed by hybrid feature selection and SVM optimization using genetic algorithms. The results showed an accuracy of 81.02%, sensitivity of 82.89%, and specificity of 79.23%, which outperformed various other methods such as LACE score, logistic regression, naïve Bayes, decision tree, and advanced neural network in identifying patients at risk of readmission. A study [21] found that among various machine learning models for detecting complicated appendicitis, Gradient Boosting (GB) had the highest validity, with AUC and accuracy values around 0.8 or more, both before and after applying SMOTE to balance the data. Another study [22] found that among various machine learning models for lung nodule detection, a combination of SVM with random undersampling (RU) and SMOTE yielded the highest classification accuracy, with an average value of more than 92.94% on various sizes of training datasets.

To solve the class imbalance problem, SMOTE (Synthetic Minority Oversampling Technique) can be used. SMOTE is an oversampling method that aims to balance the class distribution by generating synthetic samples of minority classes. By creating new data similar to the existing minority data, SMOTE helps improve the representation of minority classes in the dataset [23].

The findings of this study are anticipated to contribute in the following ways:

a. Increase the understanding of how classification techniques can be effectively applied to patient medical record data, particularly in the context of class imbalance.
b. Provide insight into the impact of using SMOTE in improving classification performance, including an increase in precision and AUC-ROC despite a decrease in recall and F1-Score.
c. Support medical professionals in improving their decision-making process through more sophisticated data analysis, especially in the diagnosis of pediatric appendicitis.

## II. METHOD

This research compares results from two different classification methods: SVM with KNNI without SMOTE and SVM with KNNI utilizing SMOTE. These two methods are evaluated using different parameters, where SVM uses polynomial parameters to fit its kernel, while KNNI uses weighting parameters to handle missing values. This study aims to understand how resampling methods and parameter tuning affect classification performance on the pediatric appendicitis dataset used as a research subject. This study has several sequential stages: appendicitis dataset collection, data preprocessing, KNNI imputation, SVM classification with and without smote, and evaluation. The flow carried out in this study can be seen in FIGURE 1 :

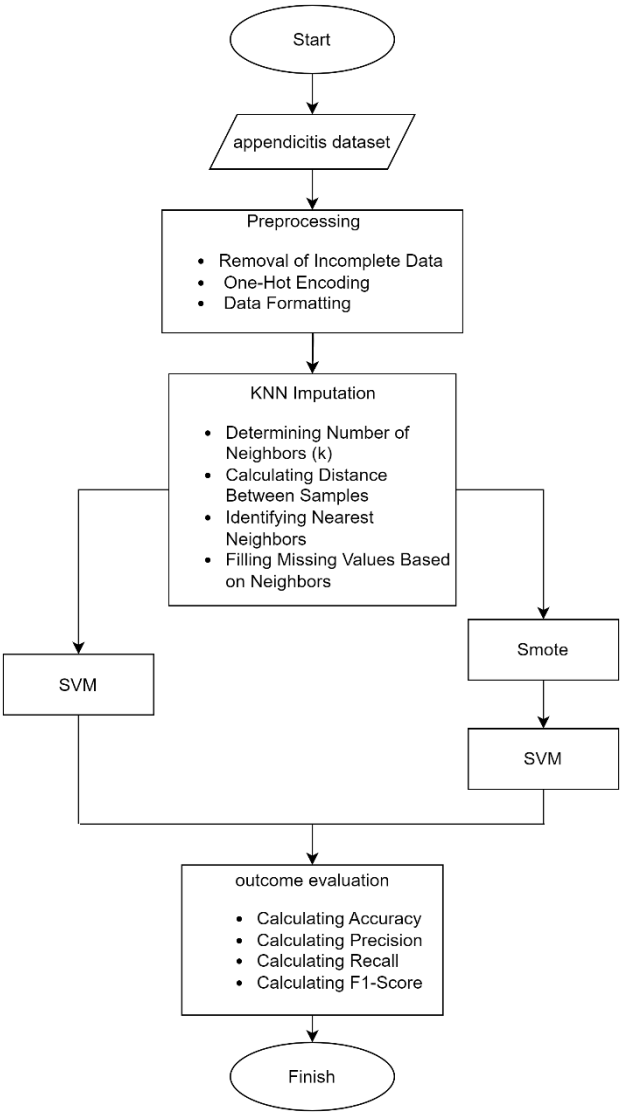


FIGURE 1. **Research Flow Chart**

### A. DATASET

The dataset used in this study is appendicitis in children. The dataset taken from the UCI Machine Learning Repository site can be seen at https://archive.ics.uci.edu/dataset/938/regensburg+pediatric+appendicitis. The data used consists of 783 records and 58 features. The other feature is the target class, which consists of 2 classes: diagnosed appendicitis and not diagnosed appendicitis. TABLE 1, Which shows samples from the Regensburg Pediatric Appendicitis dataset [24].

TABLE 1
**Sample Data Appendicitis in children**

| Age | BMI | Sex | Height | …. | Enteritis | Gynecological Findings |
|---|---|---|---|---|---|---|
| 12.68 | 16.90 | female | 148.00 | …. | | |
| 14.10 | 31.90 | male | 147.00 | …. | | |
| 14.14 | 23.30 | female | 163.00 | …. | yes | |
| 16.37 | 20.60 | female | 165.00 | …. | yes | |
| 11.08 | 16.90 | female | 163.00 | …. | yes | |
| … | … | …. | … | …. | …. | … |
| 11.40 | 18.80 | male | 149.56 1628 | …. | no | |
| 11.40 | 18.80 | male | 149.59 9601 | …. | | yes |
| 11.40 | 18.80 | male | 149.59 9601 | …. | | yes |

### B. PREPROCESSING

Data preprocessing is essential for data mining because it affects the generalization performance of machine learning algorithms. This is obvious since machine learning methods are popular and commonly used in data mining [25].

1. ONE-HOT ENCODING.

One-hot coding is a process used when the dataset contains categorical data. In this coding, each categorical feature is replaced by a set of binary features, where each can only take the value 0 or 1. The number of these binary features equals the number of possible categories (k>2) of the original features. Within each set of binary features, one feature is "hot" with a value of 1, while the others have a value of 0. Hence, this method is known as "one-hot encoding." [26]. feature samples that implement the encoding algorithm in TABLE 2.

TABLE 2
**Features that Implement One-Hot Encoding**

| BMI | Appendicitis | No Appendicitis |
|---|---|---|
| 17.50 | 1 | 0 |
| 33.10 | 1 | 0 |

2. KNN IMPUTATION.

Missing data can occur in all fields and can cause problems such as biased results, reduced statistical accuracy, and invalid conclusions [27][28]. The most common machine learning technique for handling missing values is imputation. Many imputation methods have been proposed to solve this problem [29][30].

One classification method that can solve the missing value problem is the K-Nearest Neighbor Imputation (K-NNI) algorithm. This algorithm determines the value of the missing attribute based on the similarity between the new case and the old case on the corresponding feature [31]. In the appendicitis dataset, missing values are indicated by empty columns, as can be seen in TABLE 1.

This study uses KNN imputation with distance weighting parameters, which can handle binary, categorical, ordered, continuous, and semi-continuous distance variables. The distance between two values is a weighted average of each variable's contribution, where the weights should represent the variable's importance[32].

The equation of KNN imputation can be seen. Based on the dataset, the K-NN imputation method can be applied to overcome missing data. The steps for the K-NN method are as follows [33]:

1) Determining the K parameter. There is no specific method for determining the value of k in the KNN imputation method. If the value of k is too small or too large, it can cause noise, cause natural errors to limit the value taken, and indirectly affect the level of accuracy in classification[33][34].
2) Calculate the Euclidian distance between the missing examples and the complete data with (Equation(1))[35]:

$$d_{(x,y)} = \sqrt{\sum_{j=1}^{s}(x_j - y_j)^2} \quad (1)$$

Where $d_{(x,y)}$ is the euclidian distance, j is the attribute data with j = 1,2,3, ... . s, s is the data domain, $x_{aj}$ is the value of the jth attribute containing missing data and $y_{bj}$ is the value of the j-th attribute containing complete data.
3) Based on the information from the obtained distances, the minimum Euclidian distance is used as an estimated value for the missing data based on a predetermined parameter k. The imputed value is then calculated using weighted average estimation according to (Equation(2))[36]:

$$d_{i,j} = \frac{\sum_{k=1}^{p} w_k \delta_{i,j,k}}{\sum_{k=1}^{p} w_k} \quad (2)$$

where $x_j$ is the estimated average weight, K is the number of parameters k used with k = 1,2,3 . . . K, $w_k$ is the value of K nearest neighbor observations, $v_k$ is the value of complete data on attributes containing missing data based on parameter k. An analysis was conducted based on the accuracy levels achieved by various classification methods to assess the quality of the estimated values obtained through the imputation method. Thus, the weight of $w_k$ was determined through the inverse of the square of the Euclidean distance, which was then used in the calculation of the imputed values through weighted mean estimation. The accuracy of the imputation results was then evaluated using the classification method to determine the effectiveness of the imputation method used [37].

### C. SMOTE

After the preprocessing stage, the SMOTE algorithm, which uses a random oversampling approach, is used to handle the imbalance of data distribution between the majority and minority groups. The SMOTE procedure is based on interpolating between close minority class cases [38], [39]. The goal is to increase the number of minority class instances to balance the dataset by adding artificially created minority class instances to their nearest neighbors [40]. Here are the steps of the SMOTE algorithm [41]:

1. For each sample x in the training set, calculate their Eucliden Distance to each minority class sample xi and get the k nearest neighbors of each sample from the minority class.
2. Based on the degree of sample imbalance, randomly assign a sampling ratio N to xi and then randomly select N samples from its k nearest neighbors denoted as xh.
3. Based on equation (a), construct new samples based on xi and xh until the classes become balanced, denoted as x new.

$$x_{new} = x_i + rand(0,1) * (x_h - x_i) \quad (3)$$

SMOTE also affects variable selection. For example, the p-values obtained from comparing two classes by t-test after SMOTE show a more minor increase in data than those obtained with the original data. This happens because SMOTE reduces the data obtained with the original data, which increases the sample size and reduces the variance. In contrast, the difference between the sample means is not significant [42]. Comparison after and before smote implementation can be seen in FIGURE 2.

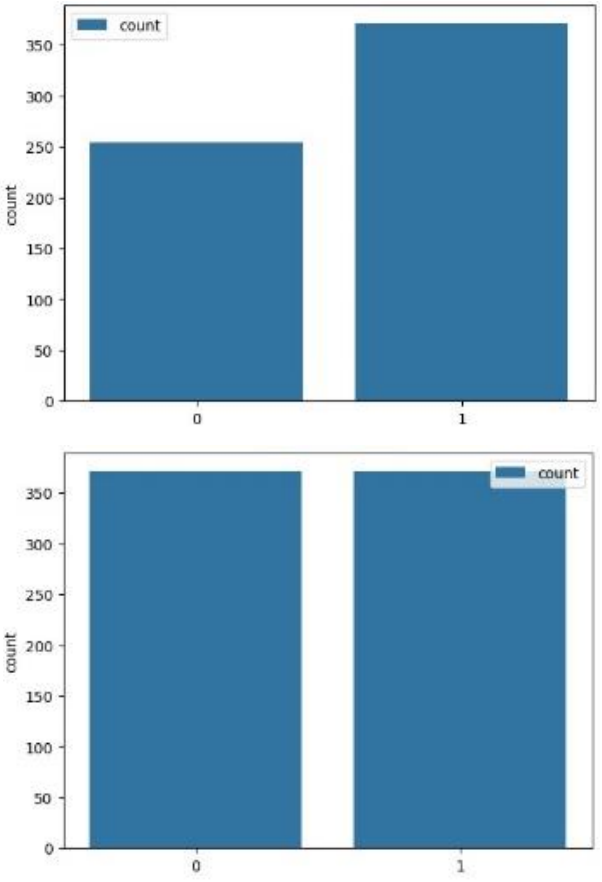


**FIGURE 2. Before and After SMOTE for Appendicitis Dataset**

### D. (SUPPORT VECTOR MACHINE) SVM

A Support Vector Machine (SVM) is a machine learning algorithm used to classify a set of training data with a label [43]. The best decision boundary is the one with the most significant distance and margin from both data classes. SVM finds the best hyperplane to separate the data [44], [15] (FIGURE 3).

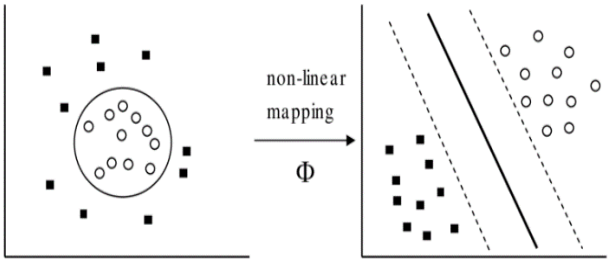


FIGURE 3. SVM Model Generation

To separate the data into two linearly distinct classes, SVM searches for the optimal hyperplane by maximizing the distance or margin between the hyperplane and the closest data sample from each class [45]. Kernels are methods applied to data that are not linearly separable. The basis of the kernel is to map the data into a higher dimensional space using functions of $\theta(x)$. Many data mining or machine learning techniques are developed by assuming that the relationship between data is linear, so the resulting algorithms are limited to the linear case (Equation(4))[46][47]:

$$(\theta(x_i).\theta(x_j)) = K(x_i, x_j) \quad (4)$$

where $x$ is an individual data point, and $\phi$ is a function that maps the data to a higher dimensional space. Using the mapping function $\theta(x)$, each multiplication of $x_i . x_i$will be calculated by $K(x_i, x_j)$. Furthermore, $xi$ will be mapped into a higher dimensional space. In this research, a polynomial kernel is used for Multiclass SVM classification[47].

In the context of a Support Vector Machine (SVM) with a polynomial kernel, $Kij$ are the elements in the i-th row and j-th column of the kernel matrix. The vectors $xi$ and $xj$ are the feature vectors of the i-th and j-th data. The following is the equation for SVM classification and polynomial parameters. The polynomial kernel is calculated using the formula(Equation (5)) [48][49]:

$$K(x_i, x_j) = (x_i, x_j + c)^d \quad (5)$$

where $c$ is a constant added to control bias, and is the polynomial degree that determines the complexity of the kernel function. This kernel function allows the SVM to operate in a high-dimensional feature space, measuring the similarity between data points without explicitly computing the coordinates in that space, leveraging what is known as the kernel trick.

### *E. EVALUATION*

#### 1. CONFUSION MATRIX

In machine learning, the performance of the integrated model in classification performance is commonly done through the utilization of confusion matrices. Confusion matrices are a more efficient way to present the results of problems in classification [50]. This matrix provides information about the real and predicted classification results [51].

False Negative (FN), False Positive (FP), True Negative (TN), and True Positive (TP) are terms used in the confusion matrix. The terms are defined in TABLE 3.

TABLE 3
**Confusion matrix**

| Actual class | Predicted Class | |
|---|---|---|
| | True | False |
| True | True Positive (TP) | False Negative (FN) |
| False | False Positive (FP) | True Negative (TN) |

Here is the evaluation matrix that has been considered, the confusion matrix parameters have been used to measure each parameter evaluation can be seen in (Equation (6), Equation (7), Equation (8), and Equation (9)) [52].

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \quad (6)$$

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

$$Precision = \frac{TP}{TP+FP} \quad (8)$$

$$F1 = \frac{2*precision*Recall}{precision+Recal} \quad (9)$$

#### 2. AREA UNDER THE ROC CURVE (AUC-ROC)

Calculation using a formula that integrates the curves. AUC can be interpreted as the probability that the classification model will correctly distinguish between positive and negative samples [53]. The categorization approach implies that when randomly selected, positive examples will have higher scores than negative examples. Therefore, a higher AUC indicates a better classification model's ability to distinguish between positive and negative classes effectively. Maximizing the AUC value is the main goal in developing an effective classification model [54].

AUC ranges between 0 and 1, with a higher AUC indicating better model performance. AUC can be modeled mathematically in (Equation (10))[55].

$$AUC = \frac{\left(\frac{TP}{TP+FN}\right)x\left(\frac{TN}{TN+FP}\right)}{2} \quad (8)$$

In addition, the interpretation of the AUC value indicates the ability of the model to distinguish between positive and negative classes. In addition, AUC is a valuable tool for selecting and comparing models, which allows practitioners to evaluate the relative effectiveness of various classifiers. To see the value of classification quality based on the AUC value can be seen at TABLE 4 [55].

TABLE 4
**Categories of results from classification based on AUC values**

| AUC Values | 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 |

## III. RESULT

This section shows the performance of the SVM classification algorithm with and without using the SMOTE technique. The aim is to see how well SVM can classify unbalanced data. This analysis uses various evaluation metrics such as accuracy, precision, recall, F1-score, and AUC of the ROC curve. The evaluation is done to compare the performance of SVM with and without SMOTE and to understand whether the use of SMOTE can improve SVM's ability to distinguish between positive and negative classes.

### *A. SUPPORT VECTOR MACHINE(SVM)PERFORMANCE*

This section reveals the experimental findings obtained from the Support Vector machine (SVM) classification model with and without SMOTE.

**TABLE 5**
**SVM accuracy with and without SMOTE**

| Evaluation | SMOTE | NO SMOTE |
|---|---|---|
| Accuracy | 85.99% | 87.26% |
| Precision | 98.65% | 87.00% |
| Recall | 77.66% | 92.55% |
| F1-Score | 86.90% | 89.69% |
| AUC-ROC | 88.04% | 85.96% |

The results of evaluating classification models using Support Vector Machine (SVM), both with and without using Synthetic Minority Over-sampling Technique (SMOTE), show significant differences in performance metrics. When using SMOTE, there was a notable increase in precision, reaching 98.65% from 87.00% without SMOTE. This indicates that most of the optimistic predictions made by the model with SMOTE are correct. However, the use of SMOTE also resulted in a decrease in recall, which decreased from 92.55% to 77.66%. This means that the model with SMOTE is more likely to miss some positive instances. In addition, F1-Score, which is the harmonic mean of precision and recall, also decreased from 89.69% to 86.90% when using SMOTE.

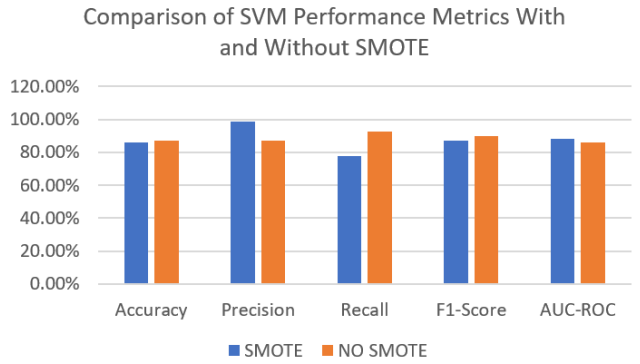


**FIGURE 4. SVM Classification Performance: With and Without SMOTE**

However, it is worth noting that the accuracy of the model only experienced a slight decrease, from 87.26% to 85.99%, and there was a significant increase in AUC-ROC from 85.96% to 88.04%. This emphasis on improving AUC-ROC suggests that the use of SMOTE can improve the model's ability to distinguish between positive and negative classes, which is often a critical factor in evaluating the performance of classification models. While there is some compromise to precision, recall, and F1-Score, it is essential to note that AUC-ROC is a more critical matrix. For more details, it can be seen the chart in FIGURE 4 comparison of the performance results of the suppert vector mechine algorithm with and without smote. The following are the evaluation results of the classification model using a Support Vector Machine (SVM) for the diagnosis of appendicitis in children, both with and without using the Synthetic Minority Oversampling Technique (SMOTE). The confusion matrix can be seen in TABLE 6 below using the smote:

**TABLE 6**
**Confusion matrix with SMOTE**

| | Predicted class | |
|---|---|---|
| | Positive | Negative |
| Actual Positive | 62 | 1 |
| Actual Negative | 21 | 73 |

Then for the confision matrix that does not use smote can be seen in TABLE 7:

**TABLE 7**
**Confusion matrix without SMOTE**

| | Predicted class | |
|---|---|---|
| | Positive | Negative |
| Actual Positive | 50 | 13 |
| Actual Negative | 7 | 87 |

The performance evaluation of the model with and without SMOTE shows a significant difference. The model without SMOTE has 50 True Positives and 13 False Positives, while the model with SMOTE has 62 True Positives and 1 False Positives. This indicates that the model with SMOTE is better at reducing False Positives, which means that most of the positive predictions made by the model with SMOTE are correct. However, it should be noted that the model with SMOTE experienced an increase in False Negatives from 7 to 21, indicating that it is more likely to miss some positive cases. In contrast, True Negatives decreased from 87 to 73 with the use of SMOTE, indicating that the model with SMOTE is less effective in identifying negative cases.

The importance of AUC-ROC was also seen, where the model with SMOTE showed an increase from 85.96% to 88.04%, indicating a better ability to distinguish between positive and negative classes. Therefore, despite some compromises in other performance metrics, the improvement in AUC-ROC indicates that the model with SMOTE is more effective in this context of medical diagnosis.

## IV. DISCUSSION

The selection of polynomial kernel in the analysis of appendicitis diagnosis is based on several considerations, including the complexity of the non-linear relationship between clinical features and diagnosis, the complex distribution between "positive" and "negative" classes in the dataset, and the polynomial interaction between features. The

use of a polynomial kernel in SVM is considered strategic as it can overcome the structural challenges in the dataset and improve the model's ability to understand and predict appendicitis diagnoses. Then, the parameter selection in the polynomial kernel (d) and regularization parameter (c) in SVM affect the complexity and generalization of the model. Choosing d = 3 provides a good balance between model complexity and the ability to capture patterns in the data, while c = 1 reflects a balanced trade-off between model flexibility and the risk of overfitting. Thus, using these values can result in a model that is flexible enough to handle the complexity of the data without losing important generalizations.

Following the use of SMOTE to handle a class imbalance in the dataset, this study used a value of K = 15 to control the number of neighbors considered when creating synthetic samples for minority classes. The value of K = 15 was chosen as it resulted in a significant improvement in AUC-ROC compared to other K values. This suggests that by considering the 15 nearest neighbors in the formation of synthetic samples, the model can obtain more representative information from the variation in the minority data, improving the model's performance in handling class imbalance. And Results show that the AUC rises when using SVM with SMOTE, indicating an improvement in the model's ability to distinguish between positive and negative classes.

evaluation results show that using SMOTE affects the classification model's performance in several important aspects. The model's accuracy slightly decreased with SMOTE (from 87.26% to 85.99%). The increase in precision (from 87.00% to 98.65%) indicates that the model is more precise in its optimistic predictions when using SMOTE. However, the significant decrease in recall (from 92.55% to 77.66%) indicates that the model is less effective in recognizing all positive instances in the test data after using SMOTE. F1-Score, which combines precision and recall, was also higher in the model without SMOTE (89.69%) compared to the model with SMOTE (86.90%). Meanwhile, AUC-ROC was slightly higher with SMOTE (88.04%) than without SMOTE (85.96%), indicating that SMOTE helps distinguish between positive and negative classes. Overall, although SMOTE can increase precision and AUC-ROC, the decrease in recall and F1-Score suggests that the use of SMOTE needs to be carefully considered based on the specific needs of the classification application

The drawback of this study is that the decrease in recall using SMOTE causes a significant decrease in recall, which means that the model becomes less effective in detecting positive instances after the data is balanced with SMOTE. The size of the dataset used may not be large enough or representative, so the results of this study may not be generalizable to more extensive or different datasets. This study may have yet to explore various SMOTE parameter settings and classification models that could have yielded better results. For example, the value of k_neighbors in SMOTE and the hyperparameters in the SVM model may be better tuned. The use of SMOTE increases the complexity of the model, which can lead to overfitting, especially if not accompanied by adequate cross-validation.

This study aims to evaluate the impact of using the Synthetic Minority Over-sampling Technique (SMOTE) on the performance of Support Vector Machine (SVM) classification models on class-imbalanced datasets. The results show that the use of SMOTE can improve the model's ability to distinguish between positive and negative classes, especially seen from the increase in AUC-ROC values. However, this improvement is also accompanied by a decrease in several other performance metrics, such as recall and F1-Score, indicating a trade-off between the precision and sensitivity of the model. This suggests that the use of SMOTE may improve the model's ability to address class imbalance, but it should be noted that there are trade-offs that must be taken into account in terms of overall model performance.

This research highlights several limitations that need to be addressed in future research. Firstly, the restriction to using SVM and SMOTE in the classification of certain medical data raises the potential for further exploration of other classification algorithms and class balancing techniques. Furthermore, this study underscores the importance of expanding the sample size to represent a wider variety of class imbalances in clinical settings. In addition, integration with other classification methods can provide a more comprehensive insight into the best approach to handling class imbalance in medical data. Finally, improvements to the evaluation methodology, such as the implementation of a more detailed cross-validation scheme, are expected to yield more accurate information on the model's performance under various data conditions. With these limitations in mind, future research in this area can further explore such aspects to expand our understanding of the performance of classification models in the face of class imbalance in medical data.

## V. CONCLUSION

The conclusion of this study confirms that the use of the Synthetic Minority Over-sampling Technique (SMOTE) in combination with the Support Vector Machine (SVM) classification algorithm has a significant impact in dealing with class imbalance in medical datasets, particularly in the context of diagnosing pediatric appendicitis. The main findings show a significant improvement in the model's ability to distinguish between the positive and negative classes, which is reflected by an increase in the AUC-ROC value. However, this improvement was also offset by a decrease in several other performance metrics, such as recall and F1-Score.

The implication of this study is that the use of SMOTE can be an effective strategy to improve model performance in handling class imbalance in medical datasets. However, the decision to apply SMOTE should be made carefully, considering side effects such as decreased recall. These results have practical implications in the development of medical decision support systems, which can improve accuracy and efficiency in diagnosing cases such as pediatric appendicitis.

As such, this research makes an important contribution to developing the field of medical informatics by providing better insight into the use of class-balancing techniques in medical data classification. In addition, these findings can serve as a basis for the development of more advanced methods in diagnosing and managing complex medical conditions, strengthening the role of medical informatics in improving overall healthcare.

In future research, there is potential to further explore different SMOTE parameter settings or combinations of other clustering and class-balancing techniques to improve model performance. In addition, it is important to conduct further studies to understand the effects of using SMOTE on different types of datasets and other classification algorithms, as well as to develop more holistic evaluation methods to evaluate model performance in the context of class imbalance. Thus, future research in this area can provide deeper insights and more effective solutions in addressing the class imbalance problem in data classification. This can contribute to the development of more advanced classification techniques in the field of medical informatics, which in turn can improve healthcare quality and patient outcomes

## REFERENCES

[1] D. Frush, C. Beam, and E. L. Effman, "Acute appendicitis in children: An evaluation with ultrasound," *Invest. Radiol.*, vol. 27, no. 6, pp. 489–490, 1992, doi: 10.1097/00004424-199206000-00017.

[2] J. Pagane and S. G. Rothrock, "Acute appendicitis in children: emergency department diagnosis and management," *Ann. Emerg. Med.*, vol. 36, no. 1, pp. 39–51, 2000.

[3] C. C. Glass and S. J. Rangel, "Overview and diagnosis of acute appendicitis in children," *Semin. Pediatr. Surg.*, vol. 25, no. 4, pp. 198–203, 2016, doi: 10.1053/j.sempedsurg.2016.05.001.

[4] S. Y. Park and S. M. Kim, "Acute appendicitis diagnosis using artificial neural networks," *Technol. Heal. Care*, vol. 23, pp. S559–S565, 2015, doi: 10.3233/THC-150994.

[5] S. G. Rothrock, G. Skeoch, J. J. Rush, and N. E. Johnson, "Clinical features of misdiagnosed appendicitis in children," *Ann. Emerg. Med.*, vol. 20, no. 1, pp. 45–50, 1991, doi: 10.1016/S0196-0644(05)81117-5.

[6] S. Di Saverio *et al.*, "Diagnosis and treatment of acute appendicitis: 2020 update of the WSES Jerusalem guidelines," *World J. Emerg. Surg.*, vol. 15, no. 1, pp. 1–42, 2020, doi: 10.1186/s13017-020-00306-3.

[7] A. B. Goldin, P. Khanna, M. Thapa, J. A. McBroom, M. M. Garrison, and M. T. Parisi, "Revised ultrasound criteria for appendicitis in children improve diagnostic accuracy," *Pediatr. Radiol.*, vol. 41, no. 8, pp. 993–999, 2011, doi: 10.1007/s00247-011-2018-2.

[8] S. Mishra, "Artificial Intelligence: A Review of Progress and Prospects in Medicine and Healthcare," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 4, no. 1, pp. 1–23, 2022, doi: 10.35882/jeeemi.v4i1.1.

[9] N. Diabetes and D. Group, "Guide to diagnosis and classification of diabetes mellitus and other categories of glucose intolerance," *Diabetes Care*, vol. 20, no. 1 SUPPL., pp. 1039–1057, 1997, doi: 10.2337/diacare.20.1.s21.

[10] J. Ha, M. Kambe, and J. Pe, *Data Mining: Concepts and Techniques*. 2011. doi: 10.1016/C2009-0-61819-5.

[11] A. Kalantari, A. Kamsin, S. Shamshirband, A. Gani, H. Alinejad-Rokny, and A. T. Chronopoulos, "Computational intelligence approaches for classification of medical data: State-of-the-art, future challenges and research directions," *Neurocomputing*, vol. 276, pp. 2–22, 2018, doi: 10.1016/j.neucom.2017.01.126.

[12] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," *Neurocomputing*, vol. 408, no. xxxx, pp. 189–215, 2020, doi: 10.1016/j.neucom.2019.10.118.

[13] M. S. Uzer, N. Yilmaz, and O. Inan, "Feature selection method based on artificial bee colony algorithm and support vector machines for medical datasets classification," *Sci. World J.*, vol. 2013, 2013, doi: 10.1155/2013/419187.

[14] S. Huang, C. A. I. Nianguang, P. Penzuti Pacheco, S. Narandes, Y. Wang, and X. U. Wayne, "Applications of support vector machine (SVM) learning in cancer genomics," *Cancer Genomics and Proteomics*, vol. 15, no. 1, pp. 41–51, 2018, doi: 10.21873/cgp.20063.

[15] M. W. Huang, C. W. Chen, W. C. Lin, S. W. Ke, and C. F. Tsai, "SVM and SVM ensembles in breast cancer prediction," *PLoS One*, vol. 12, no. 1, pp. 1–14, 2017, doi: 10.1371/journal.pone.0161501.

[16] A. Jadhav, D. Pramod, and K. Ramanathan, "Comparison of Performance of Data Imputation Methods for Numeric Dataset," *Appl. Artif. Intell.*, vol. 33, no. 10, pp. 913–933, 2019, doi: 10.1080/08839514.2019.1637138.

[17] S. M. Weiss and N. Indurkhya, "Decision-rule solutions for data mining with missing values," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 1952 LNAI, pp. 1–10, 2000, doi: 10.1007/3-540-44399-1_1.

[18] X. Wang, A. Li, Z. Jiang, and H. Feng, "Missing value estimation for DNA microarray gene expression data by Support Vector Regression imputation and orthogonal coding scheme," *BMC Bioinformatics*, vol. 7, pp. 1–10, 2006, doi: 10.1186/1471-2105-7-32.

[19] S. Zhang, "Nearest neighbor selection for iteratively kNN imputation," *J. Syst. Softw.*, vol. 85, no. 11, pp. 2541–2552, 2012, doi: 10.1016/j.jss.2012.05.073.

[20] S. Cui, D. Wang, Y. Wang, P. W. Yu, and Y. Jin, "An improved support vector machine-based diabetic readmission prediction," *Comput. Methods Programs Biomed.*, vol. 166, pp. 123–135, 2018, doi: 10.1016/j.cmpb.2018.10.012.

[21] T. A. Phan-Mai, T. T. Thai, T. Q. Mai, K. A. Vu, C. C. Mai, and D. A. Nguyen, "Validity of Machine Learning in Detecting Complicated Appendicitis in a Resource-Limited Setting: Findings from Vietnam," *Biomed Res. Int.*, vol. 2023, 2023, doi: 10.1155/2023/5013812.

[22] Y. Sui, Y. Wei, and D. Zhao, "Computer-aided lung nodule recognition by SVM classifier based on combination of random undersampling and SMOTE," *Comput. Math. Methods Med.*, vol. 2015, 2015, doi: 10.1155/2015/368674.

[23] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009, doi: 10.1109/TKDE.2008.239.

[24] P. N. Srinivasu, U. Sirisha, K. Sandeep, S. P. Praveen, L. P. Maguluri, and T. Bikku, "An Interpretable Approach with Explainable AI for Heart Stroke Prediction," *Diagnostics*, vol. 14, no. 2, 2024, doi: 10.3390/diagnostics14020128.

[25] S. A. N. Alexandropoulos, S. B. Kotsiantis, and M. N. Vrahatis, *Data preprocessing in predictive data mining*, vol. 34. 2019. doi: 10.1017/S026988891800036X.

[26] M. Alruily, S. A. El-Ghany, A. M. Mostafa, M. Ezz, and A. A. A. El-Aziz, "A-Tuning Ensemble Machine Learning Technique for Cerebral Stroke Prediction," *Appl. Sci.*, vol. 13, no. 8, 2023, doi: 10.3390/app13085047.

[27] F. Sciences, "Working With Missing Values," *J. Marriage Fam.*, vol. 67, no. November, pp. 1012–1028, 2005.

[28] A. Farhangfar, L. Kurgan, and J. Dy, "Impact of imputation of missing values on classification error for discrete data," *Pattern Recognit.*, vol. 41, no. 12, pp. 3692–3705, 2008, doi: 10.1016/j.patcog.2008.05.019.

[29] A. R. T. Donders, G. J. M. G. van der Heijden, T. Stijnen, and K. G. M. Moons, "Review: A gentle introduction to imputation of missing values," *J. Clin. Epidemiol.*, vol. 59, no. 10, pp. 1087–1091, 2006, doi: 10.1016/j.jclinepi.2006.01.014.

[30] R. Pan, T. Yang, J. Cao, K. Lu, and Z. Zhang, "Missing data imputation by K nearest neighbours based on grey relational structure and mutual information," *Appl. Intell.*, vol. 43, no. 3, pp. 614–632,

2015, doi: 10.1007/s10489-015-0666-x.
[31] L. Muflikhah, N. Hidayat, and D. J. Hariyanto, "Prediction of hypertention drug therapy response using K-NN imputation and SVM algorithm," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 15, no. 1, pp. 460–467, 2019, doi: 10.11591/ijeecs.v15.i1.pp460-467.
[32] A. Kowarik and M. Templ, "Imputation with the R package VIM," *J. Stat. Softw.*, vol. 74, no. 7, 2016, doi: 10.18637/jss.v074.i07.
[33] D. M. P. Murti, U. Pujianto, A. P. Wibawa, and M. I. Akbar, "K-Nearest Neighbor (K-NN) based Missing Data Imputation," *Proceeding - 2019 5th Int. Conf. Sci. Inf. Technol. Embrac. Ind. 4.0 Towar. Innov. Cyber Phys. Syst. ICSITech 2019*, pp. 83–88, 2019, doi: 10.1109/ICSITech46713.2019.8987530.
[34] S. Zhang, D. Cheng, Z. Deng, M. Zong, and X. Deng, "A novel kNN algorithm with data-driven k parameter computation," *Pattern Recognit. Lett.*, vol. 109, pp. 44–54, 2018, doi: 10.1016/j.patrec.2017.09.036.
[35] M. Bazmara and S. Jafari, "K Nearest Neighbor Algorithm for Finding Soccer Talent," *J. Basic. Appl. Sci. Res*, vol. 3, no. 4, pp. 981–986, 2013, [Online]. Available: www.textroad.com
[36] L. Wang and D. M. Fu, "Estimation of missing values using a weighted k-nearest neighbors algorithm," *Proc. - 2009 Int. Conf. Environ. Sci. Inf. Appl. Technol. ESIAT 2009*, vol. 3, no. 2, pp. 660–663, 2009, doi: 10.1109/ESIAT.2009.206.
[37] J. Luengo, S. García, and F. Herrera, *On the choice of the best imputation methods for missing values considering three groups of classification methods*, vol. 32, no. 1. 2012. doi: 10.1007/s10115-011-0424-2.
[38] A. Fernández, S. García, F. Herrera, and N. V. Chawla, "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary," *J. Artif. Intell. Res.*, vol. 61, pp. 863–905, 2018, doi: 10.1613/jair.1.11192.
[39] G. Douzas, F. Bacao, and F. Last, "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE," *Inf. Sci. (Ny).*, vol. 465, pp. 1–20, 2018, doi: 10.1016/j.ins.2018.06.056.
[40] M. K. Suryadi, R. Herteno, S. W. Saputro, M. R. Faisal, and R. A. Nugroho, "A Comparative Study of Various Hyperparameter Tuning on Random Forest Classification with SMOTE and Feature Selection Using Genetic Algorithm in Software Defect Prediction," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 6, no. 2, pp. 137–147, 2024, doi: 10.35882/jeeemi.v6i2.375.
[41] F. Duan, S. Zhang, Y. Yan, and Z. Cai, "An Oversampling Method of Unbalanced Data for Mechanical Fault Diagnosis Based on MeanRadius-SMOTE," *Sensors*, vol. 22, no. 14, 2022, doi: 10.3390/s22145166.
[42] R. Blagus and L. Lusa, "SMOTE for high-dimensional class-imbalanced data," *BMC Bioinformatics*, vol. 14, 2013, doi: 10.1186/1471-2105-14-106.
[43] D. X. Zhou and K. Jetter, "Approximation with polynomial kernels and SVM classifiers," *Adv. Comput. Math.*, vol. 25, no. 1–3, pp. 323–344, 2006, doi: 10.1007/s10444-004-7206-2.
[44] M. Yalsavar, P. Karimaghaee, A. Sheikh-Akbari, M. H. Khooban, J. Dehmeshki, and S. Al-Majeed, "Kernel Parameter Optimization for Support Vector Machine Based on Sliding Mode Control," *IEEE Access*, vol. 10, pp. 17003–17017, 2022, doi: 10.1109/ACCESS.2022.3150001.
[45] S. Ding, X. Hua, and J. Yu, "An overview on nonparallel hyperplane support vector machine algorithms," *Neural Comput. Appl.*, vol. 25, no. 5, pp. 975–982, 2014, doi: 10.1007/s00521-013-1524-6.
[46] M. Alida and M. Mustikasari, "Rupiah Exchange Prediction of US Dollar Using Linear, Polynomial, and Radial Basis Function Kernel in Support Vector Regression," *J. Online Inform.*, vol. 5, no. 1, pp. 53–60, 2020, doi: 10.15575/join.
[47] D. A. Adyanti, D. C. R. Novitasari, and A. Fanani, "Support Vector Machine Multiclass using Polynomial Kernel for Osteoporosis Detection," no. ICMIs 2018, pp. 384–390, 2020, doi: 10.5220/0008522303840390.
[48] R. Damaševičius, "Optimization of SVM parameters for recognition of regulatory DNA sequences," *Top*, vol. 18, no. 2, pp. 339–353, 2010, doi: 10.1007/s11750-010-0152-x.
[49] X. Song *et al.*, "Bayesian-Optimized Hybrid Kernel SVM for Rolling Bearing Fault Diagnosis," *Sensors*, vol. 23, no. 11, 2023, doi: 10.3390/s23115137.
[50] M. Jena and S. Dehuri, "An Integrated Novel Framework for Coping Missing Values Imputation and Classification," *IEEE Access*, vol. 10, no. May, pp. 69373–69387, 2022, doi: 10.1109/ACCESS.2022.3187412.
[51] Siti Napi'ah, Triando Hamonangan Saragih, Dodon Turianto Nugrahadi, Dwi Kartini, and Friska Abadi, "Implementation of Monarch Butterfly Optimization for Feature Selection in Coronary Artery Disease Classification Using Gradient Boosting Decision Tree," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 5, no. 4, pp. 314–323, 2023, doi: 10.35882/jeeemi.v5i4.331.
[52] J. Wei, Z. Lu, K. Qiu, P. Li, and H. Sun, "Predicting drug risk level from adverse drug reactions using smote and machine learning approaches," *IEEE Access*, vol. 8, pp. 185761–185775, 2020, doi: 10.1109/ACCESS.2020.3029446.
[53] W. Islam *et al.*, "A Neoteric Feature Extraction Technique to Predict the Survival of Gastric Cancer Patients," *Diagnostics*, vol. 14, no. 9, 2024, doi: 10.3390/diagnostics14090954.
[54] Shalehah, Muhammad Itqan Mazdadi, Andi Farmadi, Dwi Kartini, and Muliadi, "Implementation of Particle Swarm Optimization Feature Selection on Naïve Bayes for Thoracic Surgery Classification," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 5, no. 3, pp. 150–158, 2023, doi: 10.35882/jeemi.v5i3.305.
[55] Angga Maulana Akbar, R. Herteno, S. W. Saputro, M. R. Faisal, and R. A. Nugroho, "Optimizing Software Defect Prediction Models: Integrating Hybrid Grey Wolf and Particle Swarm Optimization for Enhanced Feature Selection with Popular Gradient Boosting Algorithm," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 6, no. 2, pp. 169–181, 2024, doi: 10.35882/jeeemi.v6i2.388.

## AUTHORS BIOGRAPHY

**Difa Fitria** is from Banjarbaru, South Kalimantan, has been studying Computer Science at Lambung Mangkurat University since 2020. Her current research focus is on data classification using machine learning algorithms. This study program provides her with the opportunity to delve into her interest in data science. She chose this particular interest due to her fascination with data science and her deep passion for the field. Additionally, her final project revolves around conducting research centered on the classification of Appendicitis Disease in children.

**Triando Hamonangan Saragih** is a lecturer in the Department of Computer Science, Lambung Mangkurat University. His research interest is centered on Data Science and Computer Networking. He completed his bachelor's degree in Informatics at Brawijaya University, Malang in 2016. After that, he pursued a master's degree in Computer Science at Brawijaya University, Malang in 2018. The research field he is involved in is Data Science.

**Muliadi** is a lecturer in the Department of Computer Science at Lambung Mangkurat University, where he specializes in Artificial Intelligence, Decision Support Systems, and Data Science. His academic journey began with a bachelor's degree in Informatics Engineering from STMIK Akakom in 2004, followed by the attainment of a master's degree in Computer Science from Gadjah Mada University in 2009. With expertise in Data Science, he also brings valuable skills in Start-up Business Development, Digital Entrepreneurship, and Data Management Staff.

**Dwi Kartini** received her Bachelor's and master's degrees in computer science from the Faculty of Computer Science, Putra Indonesia "YPTK" Padang, Indonesia. She is a lecturer too in the Department of Computer Science. She instructs various subjects such as linear algebra, discrete mathematics, research methods and others Her research interests include the applications of Artificial Intelligence and Data Mining. She is an assistant professor in the Department of Computer Science, Faculty of Mathematics and Natural Sciences, Lambung Mangkurat University in Banjarbaru, Indonesia.

**Fatma Indriani** is a lecturer in the Department of Computer Science at Lambung Mangkurat University. Her research interests include data science and artificial intelligence. She earned her Bachelor's Degree in Informatics Engineering from Institut Teknologi Bandung. Subsequently, she completed her Master's studies at Monash University, Australia, and her PhD studies at Kanazawa University, Japan.