J. Innovative Eng. Nat. Sci., vol. 5, no. 2, pp. 490-506, 2025.
http://doi.org/10.61112/jiens.1592608

Journal of Innovative Engineering and Natural Science
(Yenilikçi Mühendislik ve Doğa Bilimleri Dergisi)
https://dergipark.org.tr/en/pub/jiens

Original Paper

Leveraging machine learning for improved outcomes in pediatric appendicitis diagnosis and management

Zeynep Özer[a]

[a]*Department of Management Information Systems, Bandırma Onyedi Eylul University, Balıkesir 10200, Turkey.*

ARTICLE INFO

Article history:
Received 28 November 2024
Received in revised form 26 January 2025
Accepted 23 February 2025

Available online

Keywords:
Pediatric appendicitis
Clinical decision support
Machine learning

ABSTRACT

Pediatric appendicitis, as a critical condition, represents clinical challenges in both diagnostic and treatment management due to the variability in its presentation and the absence of a specific biomarker for both diagnosis and outcome prediction. Leveraging Machine Learning (ML) algorithms, this study aims to improve diagnostic accuracy and treatment strategies utilizing a robust dataset from the Children's Hospital St. Hedwig in Regensburg, Germany, containing extensive clinical data and a broad spectrum of patient demographics. We evaluated the efficiency of three ML techniques, including Multilayer Neural Networks (MLNN), Support Vector Machines (SVM), and Linear Discriminant Analysis (LDA), using 10-fold cross-validation to assess the diagnosis, management, and severity of pediatric appendicitis. The findings reveal SVM's consistently strong performance across all metrics, achieving highly accurate classification results, followed by the competitive performance of MLNN. Conversely, LDA demonstrated limitations due to its linear nature, proving insufficient for handling the intricate and nonlinear relationships present in the complex dataset. The study highlights the potential of using ML-powered clinical decision support systems, providing a holistic approach to the treatment management of pediatric appendicitis.

I. INTRODUCTION

Pediatric appendicitis, a leading cause of hospital admissions for abdominal pain in children, presents substantial diagnostic and therapeutic management challenges in pediatric healthcare [1]. With a lifetime risk estimated between 6 to 9% and a peak incidence between the interval of 10 to 19-year age group, appendicitis not only constitutes a common health concern but also poses a higher risk of perforation in preschool children compared to older age groups [2]. This amplifies the need for accurate diagnosis and effective treatment to prevent severe complications. Traditionally, the diagnosis of appendicitis has relied heavily on clinical assessment supported by laboratory data and imaging techniques such as abdominal ultrasonography [3]. However, despite its routine use, the lack of a specific biomarker for appendicitis in clinical practice leaves room for diagnostic uncertainty. This is further complicated by the variability in the effectiveness of commonly used diagnostic tools and scoring systems, such as the Alvarado and Pediatric Appendicitis Scores (AS and PAS, respectively), which are not consistently applied across clinical settings [4]. The management of acute appendicitis in children lacks standardized international guidelines, oscillating between surgical intervention and conservative therapy with antibiotics [5, 6]. This variability in treatment approaches underscores the necessity for a more precise diagnostic and management strategy. Moreover, despite well-established prediction models in determining and assessing the diagnosis and severity of acute appendicitis in children, applying traditional statistical modeling techniques [7–9], these models have yet to achieve widespread clinical acceptability and applicability [10].

*Corresponding author. Tel.: +90-266-717-4024; e-mail: zozer@bandirma.edu.tr

The recent advancements in machine learning (ML) algorithms, a subset of Artificial Intelligence (AI), are being increasingly applied for effective clinical prediction [2] and decision-making processes [11] to enhance the early detection, treatment management, and severity assessment of pediatric appendicitis. As a data-driven approach, ML delves into the discovery and application of advanced algorithms that analyze data to forecast outcomes and inform decision-making processes [12], preventing unnecessary operations and decreasing the burden of appendicitis for patients and health systems [2, 10, 11, 13]. With that in mind, the aim is to advance the development of an ML-powered virtual assistant that provides real-time information and feedback to physicians for diagnosing appendicitis upon presentation, assessing its severity, and determining optimal treatment strategies. So, we evaluated three ML methods on a robust dataset obtained from the Children's Hospital St. Hedwig in Regensburg, Germany [14]. These ML methods utilized are MLNN, SVM, and LDA. We analyzed three critical measures from the dataset, including diagnosis, guiding management (conservative vs. operative), and risk stratifying severity, to enhance the clinical decision-making processes.

To this end, we described each ML method based on how each differentiates from the others, then explained the obtained data and its post-processing, specifically applied missing data analysis. This paper details our methodology and findings, highlighting the potential of SVM in enhancing diagnosis, management, and severity assessment of pediatric appendicitis through ML and paving the way for future advancements in pediatric healthcare technology.

II. EXPERIMENTAL METHOD / TEORETICAL METHOD

2.1 Linear Discriminant Analysis

LDA is a projection technique classifier designed to minimize the dimensionality of the data. It is based on Kernel Fisher Discriminant Analysis [15]. The main goal of LDA is to find the ideal balance between maximizing the variation between classes and reducing the variance within classes [12]. This property is particularly useful when dealing with datasets with different frequencies within classes and when assessing accuracy on randomly generated test data [12]. In cases where classes are labeled C1 and C2, LDA undertakes the task of identifying a projection direction (w) that ensures maximum separability in the spatial model [16]. The chosen direction is strategically oriented to enhance discrimination between classes [17]. The formulation of LDA's mathematical underpinnings is encapsulated in Eqs. 1-3 [12, 16]:

$$z = w^T x \tag{1}$$

where x samples are projected onto w. If the training sample is $X = \{x^t, r^t\}$, then

$$X\{t\} = \begin{cases} r^t = 1, & x^t \in C1 \\ r^t = 0, & x^t \in C2 \end{cases} \tag{2}$$

$$J(w) = \frac{w^T S_B w}{w^T S_w w} = \frac{|w^T(m1 - m2)|^2}{w^T S_w w} \tag{3}$$

where x represents the input, r represents the output in the training sample pairs, S_w is the total within-class scatter, and S_B indicates the between-class scatter matrix. Figure 1 depicts the LDA projection technique [16].

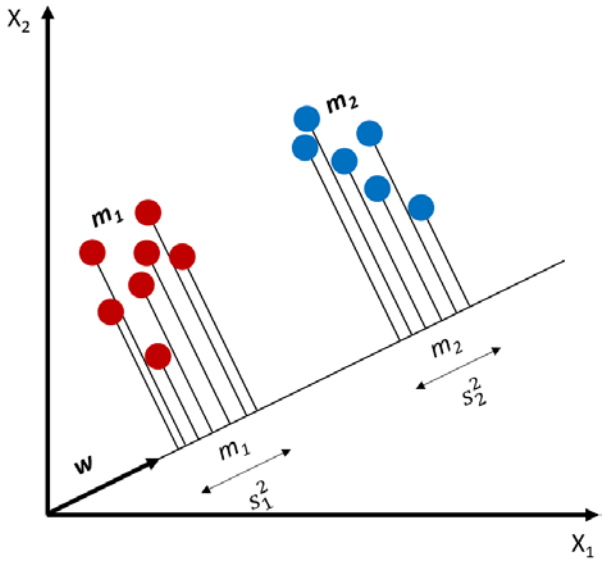

Figure 1. The projection method of LDA [16]

where, $\boldsymbol{m_1} \in \boldsymbol{\Re^d}$ and $m_1 \in \Re$ are the means of $C1$ samples before and after projection, respectively. The same holds true for $\boldsymbol{m_2}$ and m_2. S_1^2 and S_2^2 are the scatter of samples from $C1$ and $C2$ [12,18].

2.2 Support Vector Machine

Corinna Cortes and Vladimir Vapnik [19] introduced the SVM as a machine learning algorithm with considerable impact. Widely acknowledged in the literature, SVM stands out as an effective tool in the realm of ML, emphasizing the statistical learning principle for both classification and regression analysis [12]. Specifically designed as a kernel-based classification algorithm, SVM finds frequent applications in the classification of bio-signal patterns.

In the SVM framework, the weight vector is calculated after training, and the support vectors are the instances from the training data that are closest to the decision boundary (hyperplane) [20]. These support vectors play a crucial role as they provide critical insight into ambiguous and erroneous states. As a result, the hyperplane becomes a central entity that delineates the decision space in the classification process. The margin, the distance from the hyperplane to the closest support vector on either side, is a critical measure in this context. For optimal determination of hyperplanes, the focus is on maximizing the margins, represented by dashed lines defining class boundaries in Figure 2 [21]. This approach enhances the discriminative power of the SVM and ensures robust decision-making in the classification process.

The circled instances in Figure 2 show support vectors for a two-class problem, with examples of classes represented by a square pattern and a dot pattern. For generalization, SVM only works with examples close to the boundary, neglecting those in the center. SVM is calculated using the following formulas, as defined by Eqs. (4-6) [21]:

$$X\{t\} = \begin{cases} r^t = +1, & x^t \in C1 \\ r^t = -1, & x^t \in C2 \end{cases} \quad (4)$$

$$g\{x\} = \begin{cases} w^T x^t + w_0 \geq +1, & x^t \in C1 \\ w^T x^t + w_0 \leq -1, & x^t \in C2 \end{cases} \quad (5)$$

$$r^t (w^T x^t + w_0) \geq +1 \quad (6)$$

The input space is represented as $X = \{x^t, r^t\}$, with $C1$ and $C2$ as different classes and +1/-1 as labels, the hyperplane is defined by $g\{x\}$, and w_0 specifies the localized hyperplane. The SVM operates without heuristic parameters such as learning rate, initialization, or convergence control. However, this doesn't negate the importance of the hyperparameters. The kernel, a crucial SVM hyperparameter, significantly influences the performance of the algorithm. Common choices include the linear kernel, radial kernel, or polynomial kernel; each serves a different purpose [12].

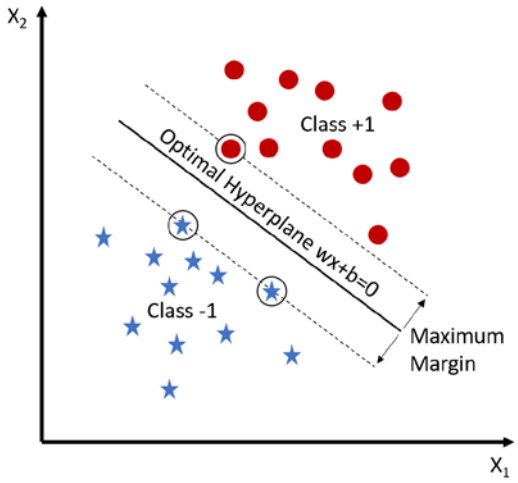

Figure 2. The basic concepts of the SVM structure [21]

2.3 Multilayer Neural Network

A MLNN is an artificial neural network (ANN) that consists of many layers of interconnected nodes or neurons. Unlike simpler models such as the perceptron, which consists of a single layer of neurons, MLNNs are designed to capture complex patterns and correlations within data [22]. The MLNN is capable of learning and generalizing the pattern. These algorithms can adapt their mathematical model to new problems because of their trainable structure [23, 24]. MLNN training is defined as the act of updating the weights to achieve better convergence to the desired results. The structure of the MLNN consists of an input layer that receives the initial input data, hidden layers where complex calculations are performed between input and output, and an output layer that produces the prediction or classification of the network [25]. Figure 3 depicts an example of an MLNN structure consisting of two hidden layers [22].

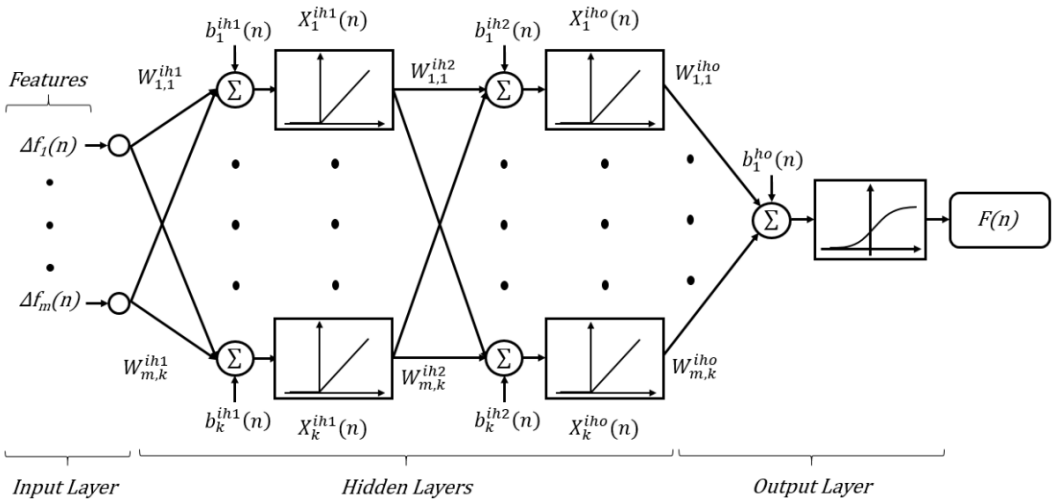

Figure 3. MLNN structure with input-hidden-and-output layers [22]

A neuron's output is expressed as the sum of the weighted inputs and the biased value. Next, an activation function is applied to this weighted summation to produce an output. Activation functions play a crucial role in MLNNs by introducing non-linearities into the model. These non-linearities are essential for the network to learn and represent complex patterns in the data. ReLU, a computationally efficient and easy-to-implement activation function, is used explicitly in the hidden layers of MLNNs. The function allows positive values to pass through unchanged. Negative values are set to zero. Although ReLU is a linear function for positive inputs, its overall behavior is nonlinear. This non-linearity is crucial for the network to learn complex patterns. Softmax, widely used in the output layer of MLNNs for multi-class classification problems, was used in the output layer. Softmax assigns probabilities to each class, and the class with the highest probability is selected as the predicted class. The formulae for the ReLU and the Softmax are given in Eq. 7 and Eq. 8, respectively [26–28].

$$ReLU(x) = max(0, x) \tag{7}$$

$$\text{Softmax}(x_i) = \frac{e^{x_i}}{\sum_{j-1}^{K} e^{x_j}} \tag{8}$$

where x represents the input to the activation function, and K is the number of classes. In MLNN architecture, a critical challenge is to define the optimal configuration of hidden layers and neurons. This process lacks strict rules, and determining the correct structure involves finding a balance. The goal is to identify the minimum number of hidden layers necessary to perform the task effectively while maximizing the network's generalizability [23].

III. ANALYTICAL PERSPECTIVE

3.1 Regensburg Pediatric Appendicitis Dataset

The dataset from the retrospective study at Children's Hospital St. Hedwig in Regensburg, Germany [14] encapsulates a diverse and comprehensive collection of clinical data from a cohort of pediatric patients admitted

with abdominal pain. This dataset is particularly distinguished by its inclusion of a multitude of abdominal B-mode ultrasound images for most patients. These images, ranging from 1 to 15 views per patient, encompass various abdominal regions such as the right lower quadrant, appendix, intestines, lymph nodes, and reproductive organs. The variability in ultrasound views underscores the depth of the imaging approach, offering a rich visual insight into the patient's abdominal conditions.

Beyond these detailed ultrasonographic images, the dataset includes an extensive range of clinical data, encompassing laboratory test results, physical examination outcomes, and scores from clinical assessment tools, notably the AS and PAS. This blend of data provides a multifaceted perspective on each patient's health, enriching the dataset's analytical depth. The dataset utilized in this study encompasses a total of 782 entries. For this research, we excluded three records due to issues with label information, resulting in 779 records being used for classification. It is important to note that for this study, the classification was performed based on the non-imaging data.

A key feature of this dataset is the classification of subjects concerning three critical target variables: diagnosis, management, and severity. The diagnosis category differentiates between cases of appendicitis and those without. Management is categorized as either primary surgical, secondary surgical, or conservative, while severity is classified into complicated, uncomplicated, or no appendicitis. This tripartite classification not only elucidates the immediate clinical decisions and outcomes but also serves as an invaluable resource for broader medical research. It enables the analysis of patterns and correlations between the clinical, laboratory, and imaging data against health outcomes, thus offering insights into effective diagnostic and treatment strategies for pediatric abdominal pain.

In terms of management, the dataset categorizes subjects into three classes: conservative, primary surgical, and secondary surgical, with 483, 270, and 26 records in each class, respectively. The categories of severity and diagnosis are both binary. Under severity, there are two subclasses: uncomplicated and complicated, containing 660 and 119 records, respectively. In the diagnosis category, the dataset includes 463 records labeled as appendicitis and 316 records labeled as no appendicitis. The comprehensive nature of the dataset, melding detailed imaging data with a broad spectrum of clinical information, renders it an asset in medical research. It offers a unique platform to study the interplay of various diagnostic tools and develop models to enhance the accuracy and efficiency of medical diagnoses and interventions in pediatric care.

The summary statistics shed light on patient demographics, clinical characteristics, and treatment outcomes. The average age of subjects is around 11.35 years, indicating a moderate age range among the patients. The Body Mass Index (BMI) average suggests a normal weight range for this age group, with a standard deviation signifying some variability. The length of hospital stays, averaging about 4.28 days, shows a range of hospitalization durations. The Alvarado Score (AS), a key metric in appendicitis assessment, displays a wide range of severity in appendicitis symptoms among the patients. The appendix diameter, averaging around 7.76 mm, highlights the variability in appendicitis cases. Other clinical parameters like body temperature and blood parameters exhibit average values with notable variations, reflecting the diverse health statuses of the patient cohort.

Visual analyses, including histograms (see Figure 4) and count plots (see Figure 5), further elucidate these findings. Histograms for continuous variables such as age, BMI, length of stay, and AS illustrate their respective distributions. In contrast, count plots for categorical variables like management, severity, and diagnosis visually

represent their frequency distribution within the dataset. These plots are instrumental in revealing the prevalence of different management strategies, the levels of severity encountered, and the types of diagnoses made.

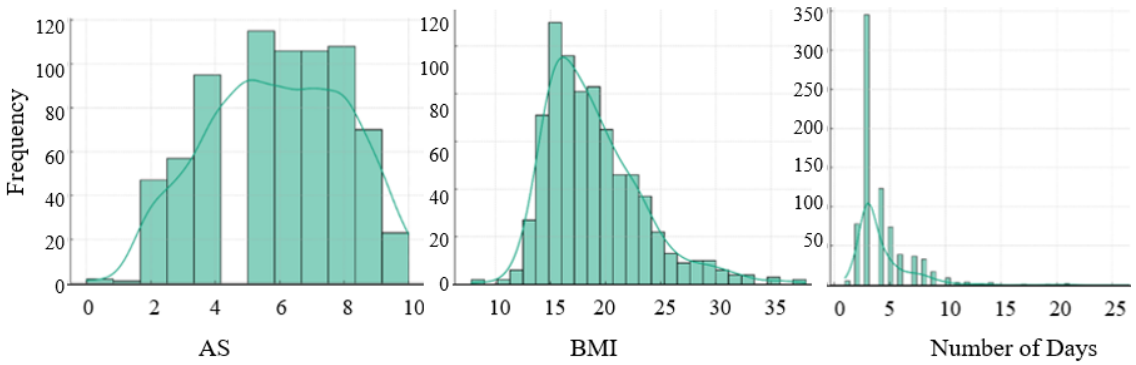

Figure 4. Statistical distribution of some features across the three measures

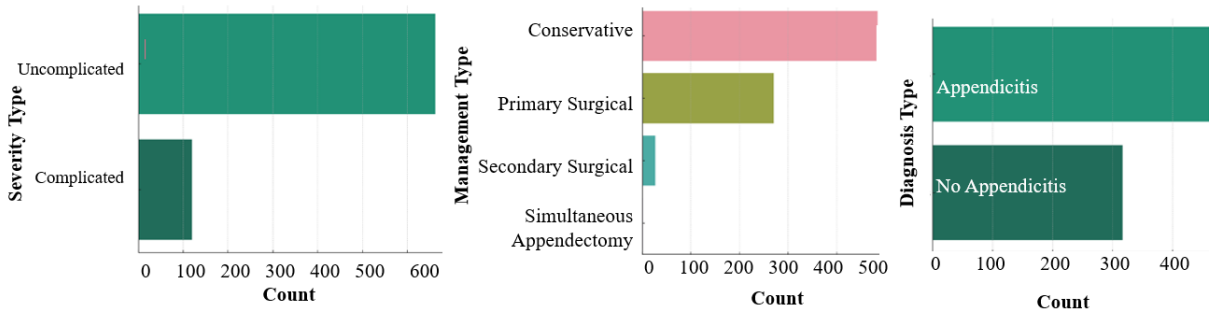

Figure 5. Number of class information across the three measures

The histogram analysis of the dataset reveals different distribution patterns for BMI, Alvarado Score, and Length of Stay. The BMI histogram appears roughly symmetric and centers around a moderate range, suggesting that most patients have a BMI within a normal or slightly overweight category, with fewer instances at the extreme ends of underweight and obese categories. This distribution approximates a normal distribution, indicating that BMI values among the patient population do not vary widely.

In contrast, the Alvarado Score histogram shows a distribution that, while not perfectly symmetric, is spread across a range of scores from low to high. This distribution isn't strictly normal as it exhibits multiple peaks, reflecting varied likelihoods of appendicitis among the patients. This suggests a diverse patient group in terms of symptoms and signs associated with appendicitis.

Lastly, the histogram for Length of Stay is right-skewed, with a majority of the data clustered at shorter stay durations and fewer instances extending toward longer stays. This indicates that while most patients experience shorter hospital stays, a minority have prolonged stays, possibly due to complications or more severe manifestations of their conditions. This positive skew highlights the presence of outliers or exceptional cases within the dataset.

In conclusion, this dataset offers a holistic view of patient demographics, clinical characteristics, and outcomes, essential for identifying underlying patterns and informing decision-making in healthcare and research. The

combination of statistical and visual analyses enhances our understanding of patient profiles and is pivotal in advancing patient care and medical research. In the realm of data science and ML, data preprocessing is a critical phase where raw data is transformed into a format more conducive to modeling. This phase involves a series of pivotal steps, each tailored to enhance the overall quality of the data and ensure its compatibility with the chosen analytical model.

3.2 Preprocessing Pipeline

In the realm of data science and ML, the preprocessing of data is a critical phase, where raw data is transformed into a format that is more conducive to modeling. This phase involves a series of pivotal steps, each tailored to enhance the overall quality of the data and ensure its compatibility with the chosen analytical model.

The preprocessing pipeline initiates with the identification and segregation of the data based on its type. The dataset is first analyzed to differentiate between numerical and categorical data. This distinction is crucial as it dictates the subsequent preprocessing techniques that will be applied to each data type. Numerical columns are identified, typically encompassing data types such as ‘float64’ and ‘int64’, while categorical columns are recognized as those containing non-numeric data, typically of the ‘object’ type.

Following the identification of data types, the first substantive step in the preprocessing pipeline is the imputation of missing values. The presence of missing data can significantly impair the integrity of statistical analyses and the effectiveness of predictive modeling. For numerical data, missing values are commonly replaced with the median of the respective column. This approach, known as ‘median imputation’, is chosen for its robustness, particularly in datasets where outliers may skew the mean. This technique ensures a more accurate representation of the central tendency of the data.

After addressing missing numerical values, the pipeline normalizes the numerical data. This normalization, often achieved through techniques such as Standard Scaling, ensures that each numerical feature contributes equally to the model. It involves adjusting the scale of the data so that its distribution has a mean of zero and a standard deviation of one, thereby mitigating the potential bias that can arise from features with larger scales.

Simultaneously, the pipeline deals with missing values in categorical data by employing the ‘most frequent imputation’ method, where missing values are substituted with the mode of the respective feature. This technique preserves the dataset’s underlying distribution and minimizes the introduction of bias.

After the imputation of missing values, categorical variables within the dataset undergo transformation through label encoding. Categorical data, which often presents in a non-numeric form, must be converted into a numerical format to be processed effectively by ML algorithms. Label encoding assigns a unique integer to each category within the feature, differing from one-hot encoding, which uses binary vectors.

A custom transformer, the Multi Column Label Encoder, has been implemented to handle the challenges of processing multiple categorical columns. This transformer extends the standard functionality of label encoders, enabling them to process either multiple columns in a Pandas Data Frame or feature indices in a NumPy array. This versatility ensures seamless integration into a wide range of data processing pipelines.

The steps of imputing missing values, normalizing numerical data, and encoding categorical data are then integrated into a cohesive pipeline using Scikit-Learn’s Pipeline tool. This approach encapsulates the sequential

application of data transformation processes, promoting modularity and ease of replication. It streamlines the preprocessing workflow and enhances the reproducibility and scalability of the model development process. Figure 6 illustrates a block diagram representing the preprocessing steps, as outlined in the preceding sections. Initially, records containing problematic label information are excluded. Subsequently, any missing values in the Neutrophils feature are replaced with 0. This step is followed by the implementation of the preprocessing procedures described earlier.

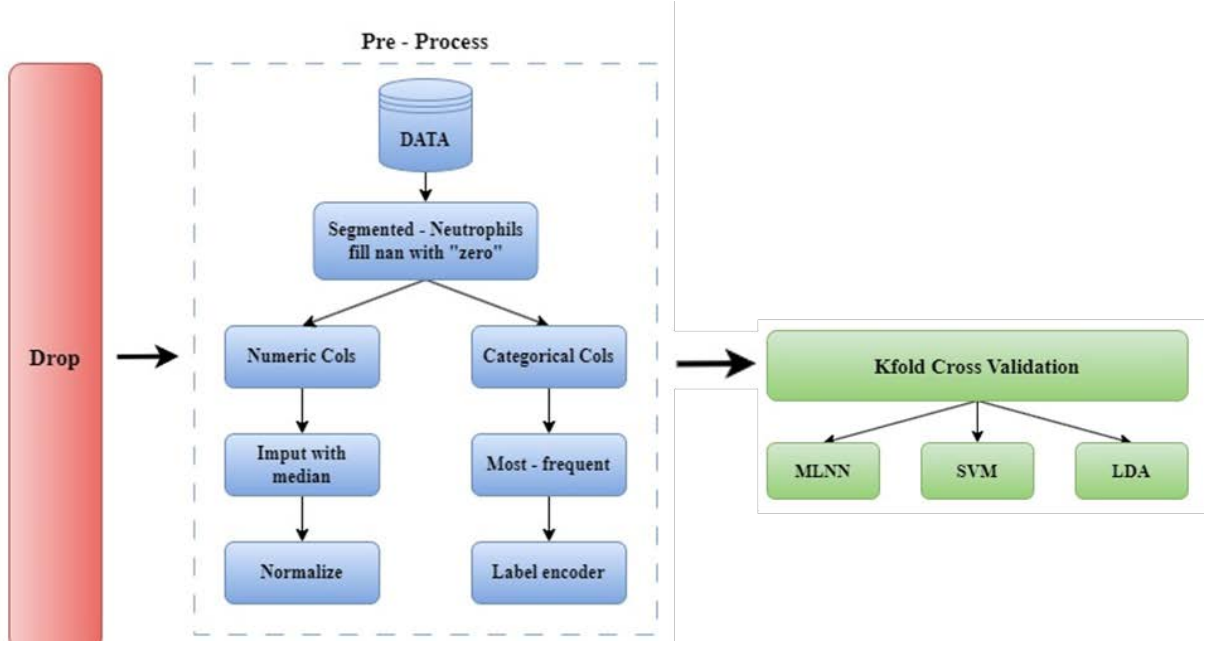

Figure 6. Block diagram of the proposed model

In conclusion, the preprocessing pipeline, encompassing the identification of data types, imputation of missing values, normalization of numerical data, and encoding of categorical data, plays a vital role in preparing raw data for sophisticated analyses in ML. By meticulously designing these steps, the input data is rendered into a format compatible with algorithmic requirements and reflective of the dataset's inherent structure. The systematic and efficient data preparation exemplified in this pipeline is fundamental in the broader context of data-driven research and analysis.

Figure 7 presents the correlation matrix derived from the data following preprocessing. Examination of the matrix reveals a pronounced negative correlation between the "peritonitis" variable and the "management" label, more so than with other variables. In contrast, a stronger positive correlation is evident among "weight," "length of stay," "pediatric appendicitis score," "appendix diameter," and "body temperature" variables with the "management" label, surpassing correlations with other variables. Furthermore, "weight," "appendix diameter," and "body temperature" exhibit a notable negative correlation with the "severity" label, distinct from other characteristics. A positive correlation is also observed between the "peritonitis" feature and the "severity" label. Conversely, "weight," "length of stay," "Alvarado score," "pediatric appendicitis score," "appendix diameter," and "neutrophil percentage" demonstrate a negative correlation with the "diagnosis" label. Meanwhile, a positive correlation between the "diagnosis" label and "peritonitis" is also noted. It seems that the weight feature has a relationship

with all label titles. For this reason, the relationship of the weight feature with these labels was also evaluated using a violin plot (Figure 8).

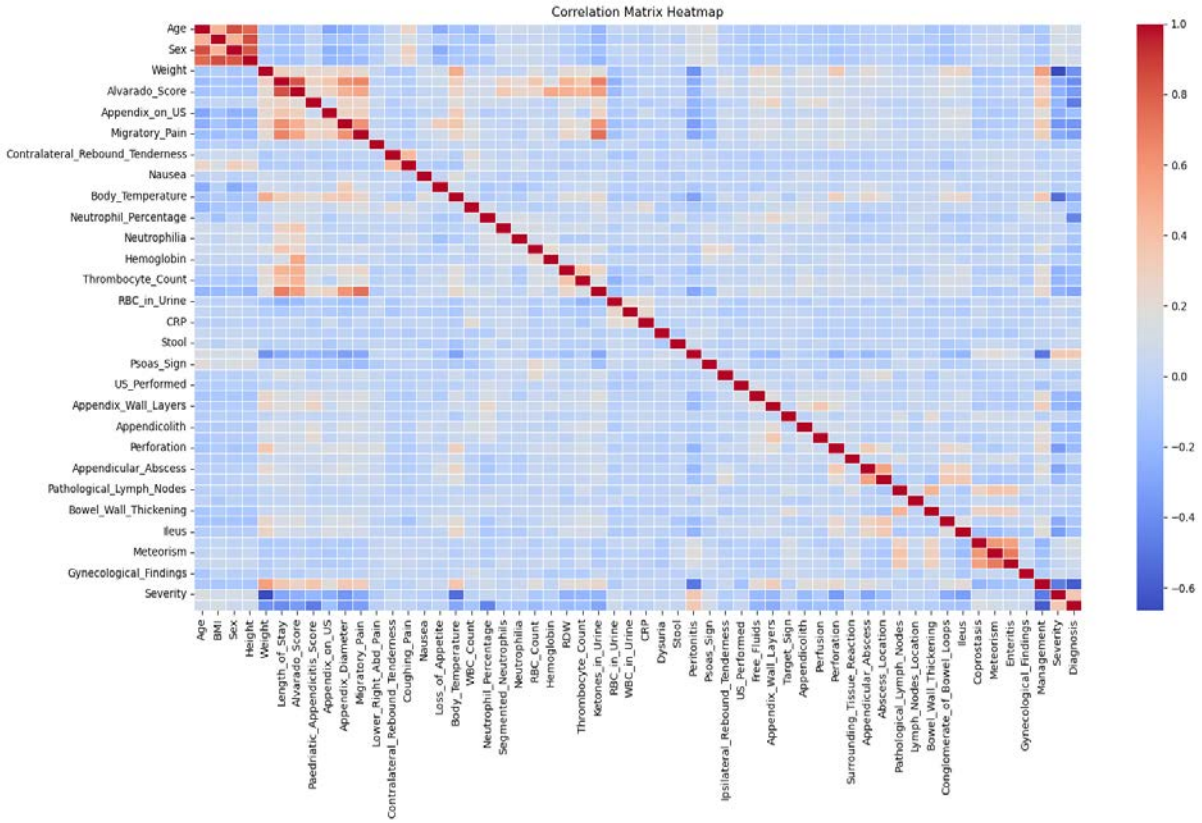

Figure 7. Correlation matrix of transformed data

According to the evaluation, Figure 8 comprises three violin plots, each illustrating the distribution of the standardized “Weight” variable across various stratifications of “Management,” “Severity,” and “Diagnosis.” These plots integrate features of box plots with kernel density estimations, thus providing a nuanced view of the data distribution.

The violin plot delineating “Weight” by “Management” reveals three distinct categories labeled conservative, primary surgical, and secondary surgical. Conservative exhibits a highly concentrated distribution, with negligible variability indicated by the slim profile of the violin. This suggests homogeneity in the weight measurements within this management group. Primary and secondary surgical, in contrast, display broader distributions, indicative of higher variability in patient weight. Notably, primary surgical shows a pronounced bulge around the median, suggesting a higher density of observations in that region compared to conservative and secondary surgical.

When observing the “Weight” distribution by “Severity,” two categories are evident. Complicated distribution closely mirrors that observed in conservative management, which is characterized by limited variability. Uncomplicated contrasts starkly, showcasing a flattened, extended distribution that spans a wider range of weight values, thus indicating considerable heterogeneity in patient weight within this severity level.

Finally, the “Weight” distribution by “Diagnosis” presents a similar bimodal distribution to “Severity.” Appendicitis presents a constrained distribution, suggesting a clustering of weight measurements around the median. Conversely, no appendicitis exhibits a more dispersed distribution, albeit less pronounced than observed in the severity uncomplicated.

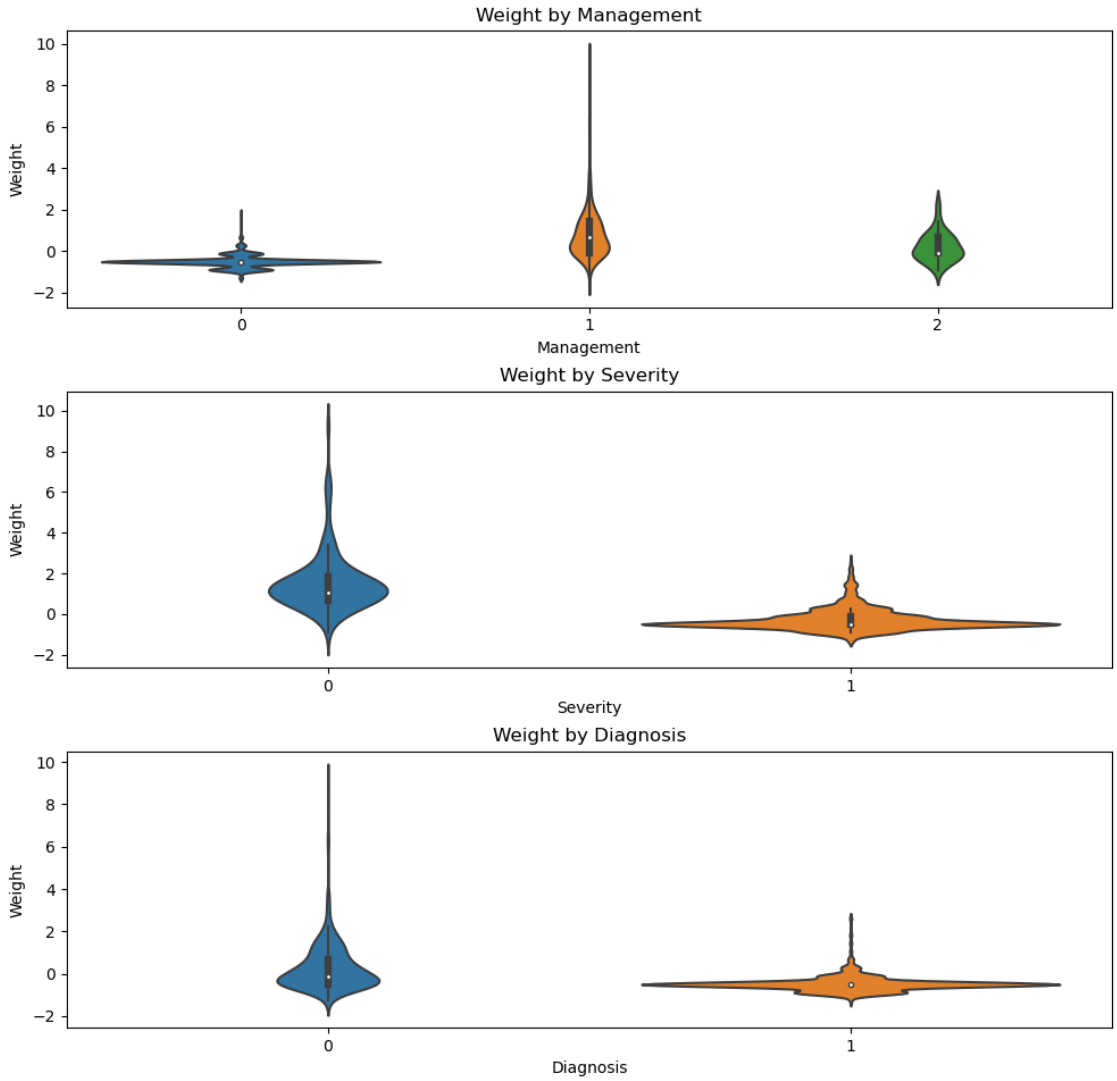

Figure 8. Violin plot representation for weight feature (Management: 0- Conservative, 1-Primary Surgical, 2-Secondary Surgical, Severity: 0-Complicate, 1-Uncomplicated, Diagnosis: 0- Appendicitis, 1- No appendicitis)

The variability in weight differs markedly across the categorical stratifications, possibly reflecting the diversity in patient characteristics or disease manifestations within each category. The plots do not convey the actual count of observations; instead, the width correlates with the density of data at scores. These visual patterns could imply potential associations between the standardized weight of patients and their respective management strategies, the severity of the condition, or diagnostic categories. Figure 9 shows the scatter plot of the weight and appendix diameter features. It is seen that the distributions for each label and class information differ in these two features.

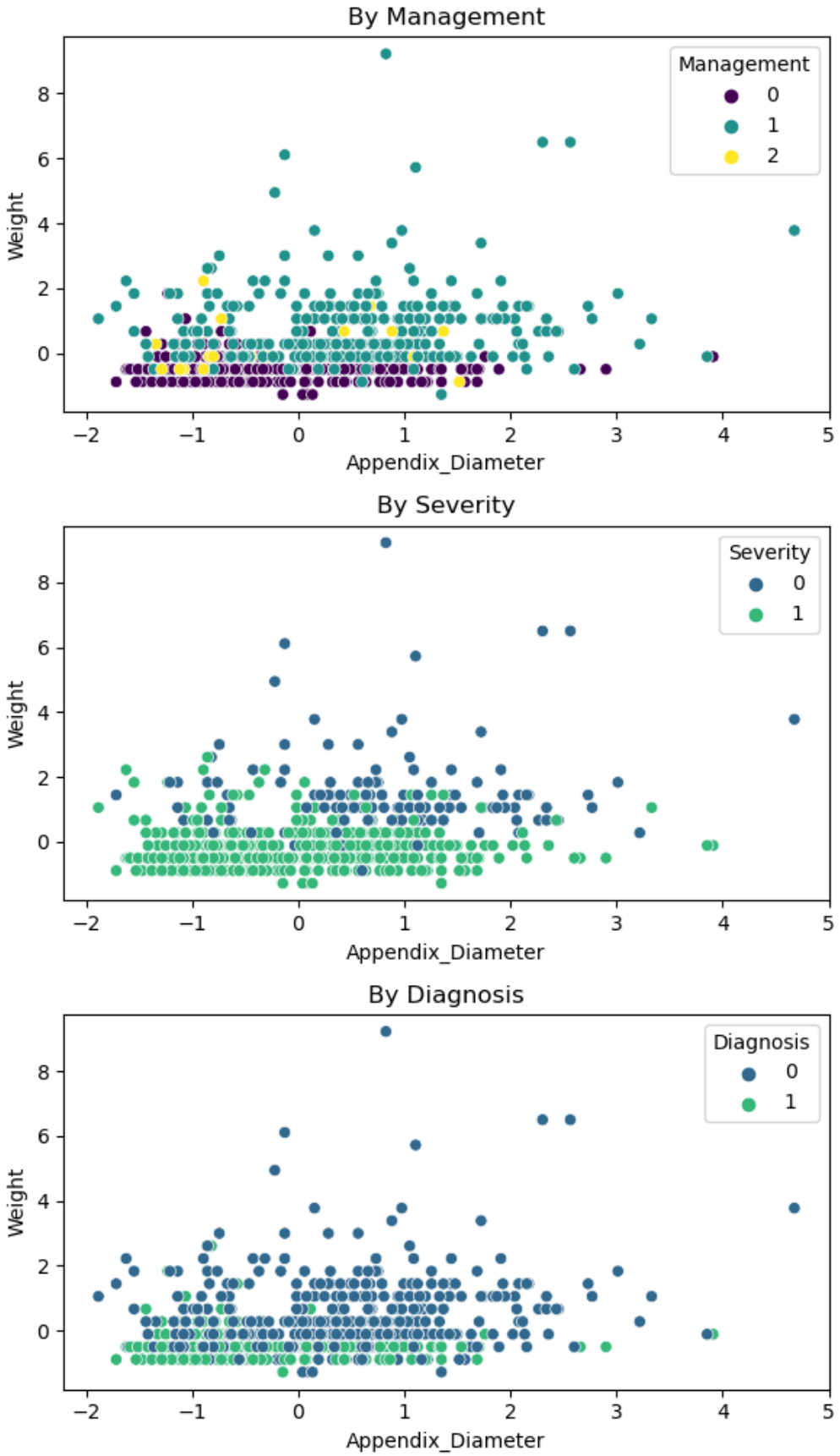

Figure 9. Scatter plot representation for weight and appendix diameter feature (Management: 0- Conservative, 1-Primary Surgical, 2-Secondary Surgical, Severity: 0-Complicate, 1-Uncomplicated, Diagnosis: 0- Appendicitis, 1- No appendicitis)

IV. RESULTS AND DISCUSSION

This study's classification results for the 'Management' category, derived from a 10-fold cross-validation approach, are profoundly illustrative of the comparative performance of three distinct ML algorithms: MLNN, SVM, and LDA.

The classification result of management is depicted in Table 1. Accordingly, the SVM algorithm demonstrated remarkable performance, achieving a perfect score across all metrics, which include Accuracy (Acc), Sensitivity (Sens), Specificity (Spec), Precision (Prec), and the F1 Score (F1). These results suggest that the SVM model is highly effective in this context, perfectly classifying the management strategies as conservative, primary surgical, or secondary surgical. The SVM's inherent ability to construct an optimal hyperplane in a high-dimensional space may contribute to this exemplary outcome, as it effectively maximizes the margin between different classes.

The classification result of management is depicted in Table 1. Accordingly, the SVM algorithm demonstrated exceptional performance, achieving an accuracy of 99.87%, sensitivity of 98.72%, specificity of 99.93%, precision of 99.88%, and an F1 score of 99.28%. These results suggest that the SVM model is highly effective in this context, nearly perfectly classifying the management strategies as conservative, primary surgical, or secondary surgical. The SVM's inherent ability to construct an optimal hyperplane in a high-dimensional space may contribute to this outstanding outcome, as it effectively maximizes the margin between different classes.

In comparison, the MLNN exhibited robust performance with an accuracy of 98.07%, sensitivity of 85.51%, specificity of 98.92%, precision of 98.32%, and an F1 score of 90.01%, see Table 1. While slightly less than perfect, these results are commendable and indicate that MLNNs can capture complex nonlinear relationships within the data. However, the discrepancy in sensitivity suggests that MLNN may be less adept at identifying true positives across all classes when compared to SVM.

The LDA algorithm, on the other hand, showed a marked divergence in performance, with Accuracy at 89.22%, Sensitivity at 63.80%, Specificity at 93.41%, Precision at 64.65%, and F1 Score at 64.18%, see Table 1. These findings indicate a substantial reduction in the LDA model's ability to correctly identify and classify management strategies. The lower Sensitivity and Precision suggest that the LDA model struggles with class imbalance and may incorrectly classify more cases into the dominant class.

Table 1. Classification result of management

	Acc	Sens	Spec	Prec	F1
MLNN	98.07	85.51	98.92	98.32	90.01
SVM	99.87	98.72	99.93	99.88	99.28
LDA	89.22	63.80	93.41	64.65	64.18

The stark contrast in the performance of SVM and LDA, especially, opens a dialogue on the suitability of linear methods versus those capable of capturing higher-order interactions in datasets with complex class boundaries. Furthermore, the perfect scores achieved by SVM raise questions regarding the potential overfitting of the model. While overfitting is typically identified during the validation phase, the 10-fold cross-validation used in this study is generally robust against this issue. Nonetheless, further investigation into the SVM model's performance on an independent test set would be prudent to confirm its generalizability.

Considering these findings, the selection of an appropriate ML model for the classification of management strategies in pediatric appendicitis appears to be a decisive factor in predictive performance. The choice between MLNN and SVM may be guided by considerations of dataset size, feature space complexity, and the computational resources at hand. Meanwhile, the results suggest that LDA may not be suitable for datasets where the classes are not linearly separable or when the prediction task requires a nuanced distinction between classes.

The 10-fold cross-validation classification results for the 'Severity' and 'Diagnosis' categories further elucidate the relative performances of the ML algorithms under consideration: MLNN, SVM, and LDA. For the 'Severity' classification, the SVM algorithm demonstrates exceptional proficiency, attaining an accuracy of 98.85%, sensitivity of 95.83%, specificity of 99.39%, precision of 96.64%, and an F1 score of 96.23%, depicted in Table 2. This strong performance indicates that SVM is exceptionally adept at distinguishing between uncomplicated and complicated cases of appendicitis, a distinction that is critical in clinical decision-making.

According to the results, shown in Table 2, MLNN also performed admirably in the 'Severity' classification, achieving 98.59% Accuracy, 95.80% Sensitivity, and a 95.40% F1 Score. These results indicate a high degree of model reliability, although slightly inferior to SVM, perhaps due to MLNN's sensitivity to the distribution of data and its inherent complexity in capturing the intricate patterns within. On the other hand, LDA showed a decrease in performance, with 92.17% Accuracy and notably lower Sensitivity and F1 Score of 64.71% and 71.63%, respectively. This suggests that LDA may not be as effective in handling the nuances of the severity classification, which can be crucial in assessing the degree of intervention required for pediatric appendicitis.

Table 2. Classification result of severity

	Acc	Sens	Spec	Prec	F1
MLNN	98.59	95.80	99.09	95.00	95.40
SVM	98.85	95.83	99.39	96.64	96.23
LDA	92.17	64.71	97.12	80.21	71.63

Turning to the 'Diagnosis' category, the results mirror the trend observed in 'Management' and 'Severity', with SVM achieving perfect scores in all metrics, underscoring its potential as a highly reliable classifier for diagnosing pediatric appendicitis (see Table 3). The MLNN model also exhibits stellar performance, with 99.61% Accuracy and an F1 Score of 99.68%. These results suggest that MLNN is almost as effective as SVM in diagnosing appendicitis, making it a valuable alternative when considering the trade-offs between computational complexity and predictive performance.

Conversely, LDA demonstrates a starkly contrasting performance in the 'Diagnosis' category, with a mere 17.18% Accuracy and an F1 Score of 29.33% (see Table 3). The 100% Sensitivity accompanied by 0% Specificity indicates a significant classification imbalance, where LDA may be over-predicting the majority class. This severe underperformance indicates that LDA is unsuitable for the diagnosis task within this dataset, likely due to its inability to manage the complex and nonlinear decision boundaries that the data presents.

Table 3. Classification result of diagnosis

	Acc	Sens	Spec	Prec	F1
MLNN	99.61	100.00	99.05	99.36	99.68
SVM	100.00	100.00	100.00	100.00	100.00
LDA	17.18	100.00	0.00	17.18	29.33

The consistency of SVM's perfect scores across all categories and metrics raises a critical discussion about the model's capacity for overfitting despite the robustness of 10-fold cross-validation against such risks. The results advocate for additional validation using external datasets to verify SVM's generalizability.

In MLNN's case, the slight discrepancies in performance across categories suggest a need for further fine-tuning of the model's architecture or parameters, particularly to improve its sensitivity in the 'Management' category.

The discussion must also address the feasibility of deploying these models in real-world clinical settings. While SVM shows superior performance, the interpretability of ML models is essential in clinical applications. Therefore, the trade-off between performance and interpretability must be carefully considered.

The article used the same dataset [14] and employed ML classifiers to predict pediatric appendicitis's diagnosis, management, and severity. The study utilized logistic regression, random forests, and gradient boosting machines, achieving areas under the precision-recall curve of 0.94, 0.92, and 0.70, respectively, for these three targets. These models outperformed conventional scores like Alvarado and Pediatric Appendicitis Score.

In comparison, the results of our study, using MLNN, SVM, and LDA, indicate that SVM outperforms the other models across all metrics in both the 'Management' and 'Severity' categories, achieving perfect scores. MLNN also shows excellent performance, but with slightly lower sensitivity in 'Management' and slightly lower metrics in 'Severity' compared to SVM. LDA displays significantly lower performance across all metrics.

When comparing the two sets of results, it is notable that SVM in our study demonstrates strong performance, surpassing the models used in the other study in most metrics for the 'Management' and 'Severity' categories. However, it is important to remain cautious about the potential risk of overfitting, as indicated by the consistently high performance of SVM, which might not generalize as effectively to unseen data. The results of the referenced article and your study suggest that while traditional scores are valuable, ML models, particularly SVM in this case, offer a substantial improvement in predictive accuracy for pediatric appendicitis.

In conclusion, the SVM's remarkable performance across the board suggests it is a frontrunner for classification tasks in pediatric appendicitis. However, MLNN's near-competitive results are a compelling alternative, especially in scenarios where model transparency and interpretability are important. The underperformance of LDA underscores the necessity of selecting models congruent with the data's complexity, emphasizing that the choice of model in medical diagnostics should be based on a combination of performance metrics and practical considerations such as explainability, computational demands and ease of integration into clinical workflows.

V. CONCLUSION

The utilization of MLNN and SVM in the prediction of pediatric appendicitis has demonstrated a strong ability to discern complex patterns in clinical data, thereby aiding in diagnosis, management, and severity assessment. SVM's robust performance, with near-perfect scores, positions it as a promising tool for clinical decision support, while the competitive results from MLNN provide a viable alternative. The underperformance of LDA underscores the importance of careful model selection that aligns with data complexity and task requirements. Although the consistently high performance of SVM raises some concerns about potential overfitting, the study's findings highlight the effectiveness of advanced ML algorithms in outperforming traditional diagnostic scores and

supporting clinical decision-making. Further external validation is encouraged to ensure the generalizability of the SVM model. Overall, this research represents a meaningful step toward integrating ML into pediatric healthcare, offering opportunities to improve patient outcomes and enhance healthcare efficiency.

REFERENCES

1. Andersson RE (2007) The Natural History and Traditional Management of Appendicitis Revisited: Spontaneous Resolution and Predominance of Prehospital Perforations Imply That a Correct Diagnosis is More Important Than an Early Diagnosis. World Journal of Surgery 31(1):86–92. https://doi.org/10.1007/s00268-006-0056-y
2. Marcinkevics R, Reis Wolfertstetter P, Wellmann S et al (2021) Using Machine Learning to Predict the Diagnosis, Management and Severity of Pediatric Appendicitis. Frontiers in Pediatrics 9:1–12. https://doi.org/10.3389/fped.2021.662183
3. Acharya A, Markar SR, Ni M, Hanna GB (2017) Biomarkers of acute appendicitis: systematic review and cost–benefit trade-off analysis. Surgical Endoscopy 31(3):1022–1031. https://doi.org/10.1007/s00464-016-5109-1
4. Shommu NS, Jenne CN, Blackwood J et al (2018) The Use of Metabolomics and Inflammatory Mediator Profiling Provides a Novel Approach to Identifying Pediatric Appendicitis in the Emergency Department. Scientific Reports 8(1):4083. https://doi.org/10.1038/s41598-018-22338-1
5. Svensson J, Hall N, Eaton S et al (2012) A Review of Conservative Treatment of Acute Appendicitis. European Journal of Pediatric Surgery 22(03):185–194. https://doi.org/10.1055/s-0032-1320014
6. Svensson JF, Patkova B, Almström M et al (2015) Nonoperative Treatment With Antibiotics Versus Surgery for Acute Nonperforated Appendicitis in Children. Annals of Surgery 261(1):67–71. https://doi.org/10.1097/SLA.0000000000000835
7. Samuel M (2002) Pediatric appendicitis score. Journal of Pediatric Surgery 37(6):877–881. https://doi.org/10.1053/jpsu.2002.32893
8. Alvarado A (1986) A practical score for the early diagnosis of acute appendicitis. Annals of Emergency Medicine 15(5):557–564. https://doi.org/10.1016/S0196-0644(86)80993-3
9. Andersson M, Andersson RE (2008) The Appendicitis Inflammatory Response Score: A Tool for the Diagnosis of Acute Appendicitis that Outperforms the Alvarado Score. World Journal of Surgery 32(8):1843–1849. https://doi.org/10.1007/s00268-008-9649-y
10. Lam A, Squires E, Tan S et al (2023) Artificial intelligence for predicting acute appendicitis: a systematic review. ANZ Journal of Surgery 93(9):2070–2078. https://doi.org/10.1111/ans.18610
11. Aydin E, Türkmen İU, Namli G et al (2020) A novel and simple machine learning algorithm for preoperative diagnosis of acute appendicitis in children. Pediatric Surgery International 36(6):735–742. https://doi.org/10.1007/s00383-020-04655-7
12. Alpaydın E (2010) Introduction to Machine Learning. MIT Press, London
13. Mani S, Ozdas A, Aliferis C et al (2014) Medical decision support using machine learning for early detection of late-onset neonatal sepsis. Journal of the American Medical Informatics Association 21(2):326–336. https://doi.org/10.1136/amiajnl-2013-001854
14. Pediatric Appendicitis Dataset (2024) Children's Hospital St. Hedwig in Regensburg, Germany
15. Muller K-R, Mika S, Ratsch G et al (2001) An introduction to kernel-based learning algorithms. IEEE Transactions on Neural Networks 12(2):181–201. https://doi.org/10.1109/72.914517
16. Cetin O (2023) Accent Recognition Using a Spectrogram Image Feature-Based Convolutional Neural Network. Arabian Journal for Science and Engineering 48 (2):1973–1990. https://doi.org/10.1007/s13369-022-07086-9
17. Gorur K, Bozkurt MR, Bascil MS, Temurtas F (2020) Comparative Evaluation for PCA and ICA on Tongue-Machine Interface Using Glossokinetic Potential Responses. Celal Bayar University Journal of Science 16(1):35–46. https://doi.org/10.18466/cbayarfbe.571994
18. Temurtas F, Gorur K, Cetin O, Ozer I (2023) Machine learning for thyroid cancer diagnosis. In: Comput. Intell. Cancer Diagnosis. Elsevier. pp 117–145.
19. Cortes C, Vapnik V (1995) Support-vector networks. Machine Learning 20(3):273–297. https://doi.org/10.1007/BF00994018
20. Liu J, Song S, Sun G, Fu Y (2019) Classification of ECG Arrhythmia Using CNN, SVM and LDA. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 11633 LNCS (2016):191–201. https://doi.org/10.1007/978-3-030-24265-7_17

21. Ozer I, Cetin O, Gorur K, Temurtas F (2021) Improved machine learning performances with transfer learning to predicting need for hospitalization in arboviral infections against the small dataset. Neural Computing and Applications 33(21):14975–14989. https://doi.org/10.1007/S00521-021-06133-0/TABLES/7
22. Cetin O, Temurtas F (2021) A comparative study on classification of magnetoencephalography signals using probabilistic neural network and multilayer neural network. Soft Computing 25(3):2267–2275. https://doi.org/10.1007/S00500-020-05296-7/FIGURES/4
23. Gorur K, Bozkurt MR, Bascil MS, Temurtas F (2018) Glossokinetic potential based tongue–machine interface for 1-D extraction using neural networks. Biocybernetics and Biomedical Engineering 38(3):745–759. https://doi.org/10.1016/j.bbe.2018.06.004
24. Shafi I, Ahmad J, Shah SI, Kashif FM (2006) Impact of Varying Neurons and Hidden Layers in Neural Network Architecture for a Time Frequency Application. In: 2006 IEEE Int. Multitopic Conf. IEEE. pp 188–193.
25. Temurtas H, Yumusak N, Temurtas F (2009) A comparative study on diabetes disease diagnosis using neural networks. Expert Systems with Applications 36(4):8610–8615. https://doi.org/10.1016/j.eswa.2008.10.032
26. Kiliçarslan S, Közkurt C, Baş S, Elen A (2023) Detection and classification of pneumonia using novel Superior Exponential (SupEx) activation function in convolutional neural networks. Expert Systems with Applications 217, 119503. https://doi.org/10.1016/j.eswa.2023.119503
27. Közkurt C, Kiliçarslan S, Baş S, Elen A (2023) α-SechSig and α-TanhSig: two novel non-monotonic activation functions. Soft Computing 27(24):18451–18467. https://doi.org/10.1007/s00500-023-09279-2
28. Zhu D, Lu S, Wang M et al. (2020) Efficient Precision-Adjustable Architecture for Softmax Function in Deep Learning. IEEE Transactions on Circuits and Systems II: Express Briefs 67(12):3382–3386. https://doi.org/10.1109/TCSII.2020.3002564