Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 4, October 2025
https://doi.org/10.47709/cnahpc.v7i4.6956

Submitted : Sep 13, 2025
Accepted : oct 25, 2025
Published : Oct 28, 2025

Implementation of Decision Tree Algorithms for Classification of Respiratory Infectious Diseases

Fauzi[1)]*, Taghfirul Azhima Yoga Siswa[2)], Fendy Yulianto[3)]
[1,2,3)] Muhammadiyah University of East Kalimantan
[1)]211102441111@umkt.ac.id, [2)]tay758@umkt.ac.id, [3)]fy415@umkt.ac.id

ABSTRACT

Acute Respiratory Infection (ARI) remains a major health issue in Indonesia, particularly among children and vulnerable populations. Conventional detection methods depend on direct observation by health workers, which is subjective, slow, and inefficient for large-scale monitoring. To address this problem, this study applies the Decision Tree algorithm to classify ARI severity and compares its performance with K-Nearest Neighbor (KNN) and Naive Bayes. A total of 1,501 patient records were collected from UPT Puskesmas Bontang Barat. The dataset underwent selection, cleaning, and transformation, followed by classification using Decision Tree and evaluation with 10-Fold Cross Validation. The Decision Tree achieved an accuracy of 81.75%, precision of 79.58%, recall of 81.75%, and F1-score of 80.45%, outperforming KNN (78.20%) and Naive Bayes (76.45%). Body temperature and respiratory rate were the most influential features in predicting severity. The results show that Decision Tree provides accurate and interpretable classifications, enabling faster triage and more efficient primary care. The novelty of this study lies in integrating a simple yet effective AI model into clinical decision support for ARI severity classification, supporting early diagnosis and public health monitoring..

Keywords: Decision Tree, ARI, Classification, Machine Learning, Data Mining.

INTRODUCTION

Acute Respiratory Tract Infection (ARTI) is a common respiratory disease that affects toddlers, mainly caused by viruses such as rhinovirus or adenovirus (Defrianti et al., 2024; Al-Harrasi & Bhatia, 2022). Typical symptoms include high fever, runny nose, repeated coughing, and loss of appetite, accompanied by swollen tonsils and inflammation in the throat or middle ear. If left untreated, ARI can develop into pneumonia or chronic ear infections, especially in children with weak immune systems (William et al., 2022). This condition is further influenced by environmental factors such as air pollution, which is often a challenge in densely populated areas, including Indonesia.

In Indonesia, cases of acute respiratory infections (ARI) continue to increase significantly from year to year. According to the latest data from the Indonesian Ministry of Health (Kemenkes RI, 2023), air pollution is cited as one of the main triggers for this surge in cases. Between 2021 and 2023, the number of ARI patients has exceeded 200,000, illustrating the serious impact of poor air quality on public health (Dedi Hidayat, 2023; Kelly & Fussell, 2015). The high prevalence of ARI in various regions of Indonesia indicates the importance of developing better detection and treatment methods to help identify ARI at an early stage.

Traditionally, ARI detection is carried out by medical personnel through direct observation, such as windshield surveys and interviews with patients or their families (Chaizuran & Hijriana, 2023; Purwanta et al., 2023). However, this approach is often time-consuming, subjective, and less effective in large-scale monitoring. Therefore, Artificial Intelligence (AI)-based methods have emerged as an alternative. These methods, such as Expert Systems and Machine Learning (ML), allow disease classification through computational models that learn patterns from data (Sulistiyo et al., 2020; Sodikin, 2023). Previous research has applied ML algorithms such as K-Nearest Neighbors (KNN), Naive Bayes, Logistic Regression, and Decision Tree in disease classification tasks, with varying levels of accuracy (Munir et al., 2024; Nasien et al., 2024).

Most prior studies on disease classification using ML have focused on cases such as breast cancer, diabetes, anemia, and kidney stones (Hafsah Mukaromah, 2025; Sari et al., 2024; Devi et al., 2025). However, there is still very limited application of these algorithms for ARI datasets, particularly in the Indonesian context. Considering the high prevalence of ARI in Indonesia and the urgency of early detection, research specifically applying ML to local ARI data remains scarce. This gap highlights the need for studies that adapt and validate classification models directly on Indonesian ARI cases.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 4, October 2025
https://doi.org/10.47709/cnahpc.v7i4.6956

Submitted : Sep 13, 2025
Accepted : oct 25, 2025
Published : Oct 28, 2025

The novelty of this research lies in applying the Decision Tree algorithm to classify ARI cases based on datasets relevant to Indonesia. Unlike previous works that primarily focused on other diseases or non-contextual datasets, this study aims to generate an interpretable and accurate model tailored to ARI characteristics in Indonesia. Moreover, the use of Decision Tree provides a transparent classification process, allowing medical professionals and stakeholders to easily understand the decision rules. This interpretability is crucial in the healthcare domain, where model transparency is as important as accuracy.

Decision Tree is chosen because it combines high classification accuracy with ease of interpretation. Several comparative studies have shown that Decision Tree often outperforms or is comparable to other algorithms such as Naive Bayes, KNN, or SVM in medical data classification (Biyantoro & Prasetiyo, 2024; Devi et al., 2025). In addition, Decision Tree can handle both categorical and numerical data, is relatively robust to noisy data, and produces a visual structure that can be directly interpreted as diagnostic rules. These characteristics make it highly suitable for ARI classification, where both accuracy and clarity of decision-making are essential.

Several studies in other countries have explored ML for respiratory disease classification. In India, SVM was used for pediatric pneumonia classification with an accuracy of 79% (Sharma et al., 2022). In Malaysia, KNN and Naive Bayes were applied to ARI prediction based on environmental factors, but interpretability remained limited (Nasien et al., 2024). In China, Random Forest achieved higher accuracy but produced models that were complex and difficult to interpret in clinical settings (Li et al., 2023). Unlike these approaches, this study utilizes the Decision Tree algorithm, which not only achieves competitive accuracy but also ensures transparency and interpretability—key requirements for real-world adoption in primary healthcare settings. Most existing ML research has focused on diseases such as cancer, diabetes, anemia, or kidney stones (Hafsah Mukaromah, 2025; Sari et al., 2024; Devi et al., 2025), with limited applications for ARI using Indonesian clinical data. This gap highlights the importance of developing models specifically tailored to local health contexts. The novelty of this study lies in applying the Decision Tree algorithm to classify ARI severity based on Indonesian patient data, offering a more interpretable and clinically relevant alternative compared to previous international studies that rely on complex, opaque models.

Research Questions: 1) Can the Decision Tree algorithm provide higher classification accuracy for ARI severity compared to KNN and Naive Bayes using Indonesian clinical data? 2) Which clinical features contribute most significantly to the classification of ARI severity? Hypotheses: 1) The Decision Tree algorithm achieves higher accuracy than KNN and Naive Bayes for ARI severity classification. 2) Key physiological features, such as body temperature and respiratory rate, significantly influence ARI severity classification. This study contributes to the development of an efficient, interpretable, and locally relevant AI model to support early ARI detection and enhance clinical decision-making in Indonesia's primary healthcare system.

Thus, this study aims to fill the gap by implementing the Decision Tree algorithm for ARI classification in Indonesia, contributing to the development of a more efficient and interpretable model for early detection and public health monitoring.

METHOD

This study utilizes private data related to respiratory tract infections obtained directly from the West Bontang Community Health Center with approved access permission for the period 2024–2025. The data collected includes various features used in the analysis, such as gender, systolic and diastolic blood pressure, pulse rate, respiratory rate (RR), body temperature (Temp), body weight (BW), height (HT), polymorphonuclear leukocytes (PMN), and information about the disease (Disease). The location of this study is at Jalan Damai No. 41/42, Bontang Barat, Bontang, East Kalimantan. The ARI disease sample data to be used can be seen in Table 1.

Table 1. Sample Data on Respiratory Tract Infections

Gender	Systolic	Diastolic	Pulse	RR	Temp	BB	TB	LP	*Disease*
Female	130	70	120	43	38	70	160	120	Heavy
Female	140	80	120	45	38.5	25	134	120	Heavy
Male	90	70	75	18	36.7	21	129	40	Light
Female	130	70	125	45	38.7	55	156	120	Heavy
Male	100	70	93	24	36.5	17	107	81	Light
Male	130	70	123	43	38.6	60	158	120	Heavy
Male	140	70	120	48	38.5	75	150	120	Heavy
Male	100	70	78	18	37.4	50	155	70	Light

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 4, October 2025
https://doi.org/10.47709/cnahpc.v7i4.6956

Submitted : Sep 13, 2025
Accepted : oct 25, 2025
Published : Oct 28, 2025

Male	90	70	78	18	36.4	60	162	70	Light
Male	90	70	78	18	36.7	50	152	78	Light

Table 1 shows 10 features, consisting of 9 attributes and 1 class, with several attributes influencing the class. The three most influential features in determining the severity of ARI are body temperature (Temp), respiratory rate (RR), and (LP). Patients tend to be categorized as mild if they have a body temperature between 36–37°C, RR between 20–40 times/minute, and LP between 40–100 cells. However, if the values of these three main features are inconsistent or appear abnormal, other features such as blood pressure (systolic and diastolic), heart rate (pulse), body weight (BW), height (HT), and gender will be used to help determine a more accurate classification.

The data used in this study is ISPA disease data obtained from the West Bontang Community Health Center (UPT Puskesmas Bontang Barat). The results of data collection yielded 12 features for classifying ISPA disease data. The following are the features contained in the ISPA disease dataset, as shown in Table 2.

Table 2. Data Features

No	Attribute	Data Type	Description
1	Gender	*String*	Gender (male & female)
2	Age	*Numeric* (Int)	Patient age
3	Education	*String*	Highest level of education attained by the patient
4	Systolic	*Numeric* (Int)	Blood pressure when the heart contracts ranges from 120/80
5	Diastolic	*Numeric* (Int)	Blood pressure when the heart contracts ranges from 90/60
6	Pulse	*Numeric* (Int)	Normal pulse rate ranges from 60-100 beats per minute (bpm)
7	Respiratory Rate (RR)	*Numeric* (Int)	Respiratory rate of 20-40 per minute is considered normal
8	Temperature	*Numeric* (Float)	Body temperature of 36.5–37.2 degrees Celsius is normal
9	Weight	*Numeric* (Float)	Body weight
10	Height	*Numeric* (Float)	Height
11	LP	*Numeric* (Float)	Polymorphonuclear leukocytes or blood cell type
12	Disease	*String* (Class)	Target Class (Severe & Mild)

Table 2 presents 12 feature attributes used in the ISPA disease classification study, with data obtained from the West Bontang Community Health Center. Each attribute has a specific data type and function in the analysis process. For example, gender, education, and disease are string data types, while other attributes are numeric data types. All of these attributes complement each other and play an important role in the machine learning model training process to identify and classify the severity of ARI.

The data division process is carried out by separating the dataset into training data to train the model in recognizing patterns of relationships between features, and testing data to evaluate the model's performance after training. This study also applies the K-Fold Cross Validation technique with k=10, which is considered effective in producing more optimal and accurate evaluations because it divides the dataset into several parts for more comprehensive testing (L. Sari et al., 2024). The selection of k=10 was made because the research by Kukuh Imani et al. (2021) proved that k=10 was higher than k=5, with an accuracy of 59.83%, precision of 57.22%, recall of 64.10%, and F1-Score of 60.10%, while 5 -Fold yielded an accuracy of 58.33%, precision of 56.23%, recall of 60.55%, and

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 4, October 2025
https://doi.org/10.47709/cnahpc.v7i4.6956

Submitted : Sep 13, 2025
Accepted : oct 25, 2025
Published : Oct 28, 2025

F1-Score of 58.22%. Thus, the value k=10 can provide a more accurate assessment of the model's ability to classify ISPA disease data in Bontang City. The 10 K-Fold Cross Validation scheme can be seen in Table 3.

Table 3. The 10 K-Fold Cross Validation Scheme

1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10

In Table 3 the 10-Fold Cross Validation scheme shows that each row represents one iteration of testing. In this scheme, the red blocks indicate the data used as test data in that iteration, while the white blocks indicate the data used as training data. The validation process is performed 10 times, with each iteration utilizing a different segment of data as test data. After all iterations are completed, the evaluation results from each fold are then averaged to obtain a more accurate and objective estimate of the model's performance.

The model evaluation stage is a critical component in assessing the accuracy of a model's performance to measure its effectiveness. Testing is carried out by applying the Confusion Matrix technique to evaluate the model's ability to make accurate predictions. In this study, the Confusion Matrix was used to assess the accuracy and performance of classification results based on test data, especially in conditions of unbalanced data, which is commonly found in the application of Machine Learning methods (Suci Amaliah et al., 2022).

The dataset is divided into training and testing subsets to evaluate the model's generalization ability. To ensure a robust evaluation, this study applies the K-Fold Cross Validation technique with k=10. The choice of 10-Fold Cross Validation is based on its proven balance between computational efficiency and reliability of results. Compared to 5-Fold, 10-Fold reduces the risk of bias and variance because the model is trained and tested on more diverse partitions of the data. Prior studies (Imani et al., 2021) have shown that 10-Fold produced better accuracy (59.83%) compared to 5-Fold (58.33%). Thus, 10-Fold Cross Validation provides a more stable and unbiased estimate of model performance, especially when datasets are not very large, as is often the case in medical studies.

The classification in this study is carried out using the Decision Tree algorithm, chosen due to its interpretability and high accuracy in medical classification tasks. Several parameters are tuned to optimize performance, including: 1) max_depth: To control the maximum depth of the tree and avoid overfitting. 2) min_samples_split: Minimum number of samples required to split a node. 3) min_samples_leaf: Minimum number of samples required to be at a leaf node. 4) criterion: The splitting criterion tested (e.g., *gini* index and *entropy*). 5) pruning strategy: Post-pruning is applied to simplify the tree structure and reduce overfitting, ensuring the model remains interpretable for healthcare practitioners. These parameters are systematically evaluated to determine the best configuration for ARI classification. The selection of these hyperparameter values is conducted using a Grid Search approach combined with 10-Fold Cross Validation. A predefined range of values is systematically tested for each parameter, and the configuration yielding the best average accuracy and F1-Score across folds is selected. For instance, max_depth is explored in the range of 3–15, min_samples_split from 2–10, and min_samples_leaf from 1–5, while both Gini and entropy criteria are evaluated. Post-pruning is applied to the final model to further reduce overfitting and maintain interpretability for healthcare practitioners. This systematic and data-driven hyperparameter optimization ensures that the Decision Tree model is not only accurate but also robust and explainable, which is essential in clinical decision support contexts.

To select the best parameter configuration, this study applies a two-step hyperparameter tuning approach: 1) Initial Random Search is first conducted to efficiently explore a broad parameter space with limited computational cost. This step helps identify promising parameter ranges without exhaustively evaluating every combination. 2) Grid Search with Cross Validation, after narrowing down the parameter ranges, a Grid Search is performed to systematically evaluate all combinations within the refined parameter space. Each parameter configuration is assessed using 10-Fold Cross Validation to ensure stable and unbiased performance estimation. The optimal parameter set is determined based on the highest average F1-score and accuracy, with additional consideration of model interpretability and complexity.

To ensure that performance improvements from parameter tuning are statistically significant rather than random fluctuations, the study conducts a paired significance test (e.g., paired t-test or Wilcoxon signed-rank test, depending

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 4, October 2025
https://doi.org/10.47709/cnahpc.v7i4.6956

Submitted : Sep 13, 2025
Accepted : oct 25, 2025
Published : Oct 28, 2025

on data distribution). 1) The baseline model (default parameters) is compared against the tuned model using cross-validation performance scores across folds. 2) A p-value < 0.05 is considered statistically significant, indicating that the improvement in accuracy and F1-score is unlikely due to chance. This combined approach of systematic hyperparameter optimization and significance testing ensures that the selected Decision Tree model is both robust and statistically validated, enhancing its credibility for clinical decision support applications.

The evaluation is performed using the Confusion Matrix, which measures Accuracy, Precision, Recall, and F1-Score. This is particularly important in imbalanced datasets, where accuracy alone may not fully capture the model's effectiveness (Amaliah et al., 2022). The overall research process is illustrated in the flowchart below:

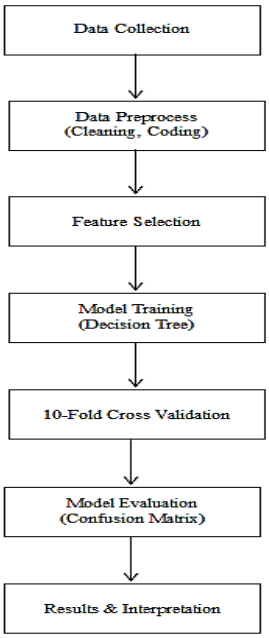

Figure 1. Research Flowchart

RESULT

a) Research result

The data analysis stages include the process of collecting, processing, and applying the Decision Tree algorithm to classify data on patients with ARI in Bontang City. The dataset used consists of 12 features, with 11 features as attributes and 1 feature as the target class.

a. Data Selection
From the 12 available features, a feature selection process was performed to identify attributes most relevant to classification performance. Nine attributes were retained along with the target class. This selection aimed to reduce noise and improve model interpretability without compromising accuracy.

b. Data Cleaning
During data cleaning, missing values were identified in the body temperature and LP (polymorphonuclear leukocytes) attributes. Rows containing NaN values and duplicates were removed to ensure data quality. After cleaning, the final dataset retained 1,501 valid records with no missing values.

c. Data Transformation
To prepare the dataset for machine learning, string values (e.g., gender, disease category) were converted into numeric form using LabelEncoder from the scikit-learn library. This step was necessary since most machine learning algorithms, including Decision Tree, require numerical inputs.

d. Data Partitioning
The dataset was split into training and testing subsets using 10-Fold Cross Validation, ensuring each data segment was used for both training and testing in different iterations. This approach minimized bias and provided a robust evaluation of model performance.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 4, October 2025
https://doi.org/10.47709/cnahpc.v7i4.6956

Submitted : Sep 13, 2025
Accepted : oct 25, 2025
Published : Oct 28, 2025

e. Results of Data Division Using K-Fold Cross Validation

In this study, dividing the dataset into training and testing data is a crucial step that influences model performance in the data mining process. The 10-Fold Cross Validation technique was used to divide the data into ten equal-sized subsets. In each iteration, one subset was used as testing data and nine subsets as training data, resulting in a total of 148 testing data sets and 1,331 training data sets. This process was performed randomly at each fold, so that each subset had an equal chance of being used as training or testing data. This approach aims to reduce bias and increase consistency in model performance assessment, ensuring objective and accurate evaluation results. Further explanation of this process will be discussed in the modeling and evaluation phase.

f. Modeling Results and Algorithm Evaluation

This stage represents the results of the modeling and evaluation process of the algorithm applied in the study. The main focus of this research is the use of the Decision Tree algorithm to classify data on ARI patients. The modeling process includes several steps, starting with data division using the K-Fold Cross Validation technique, and then evaluating model performance using various measurement metrics to assess the effectiveness of the classification.

a. Implementation Using the Decision Tree Algorithm

This study used the Decision Tree algorithm, focused on measuring the accuracy level in classifying ARI. Model testing was performed automatically using 10-Fold Cross Validation to ensure optimal evaluation results. The data was divided into two parts, training data and testing data, to test the model's performance under realistic conditions. The following are the accuracy results of the Decision Tree model in classifying ARI based on this data division. 3.3.1 Max_Depth

The Max_Depth parameter was tested to set the maximum depth limit for the decision tree. The Max_Depth values tested were 7, 8, 9, 10, and 11, with the test results displayed in Figure 2.

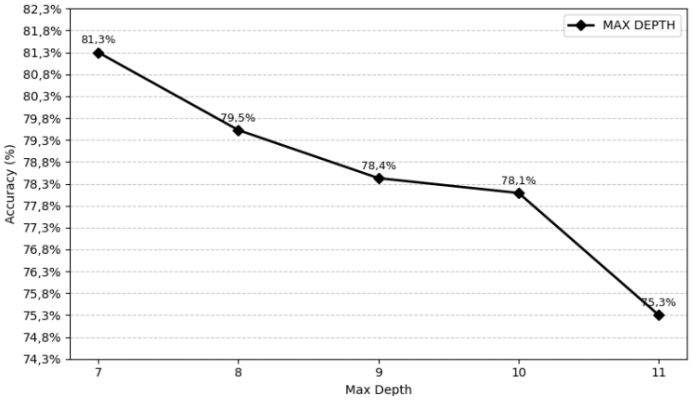

Figure 2. Line chart

Figure 2 shows that the Max_Depth parameter value providing the best accuracy is 7, with an accuracy of 81.3%. Meanwhile, a Max_Depth of 11 yields the lowest accuracy, at 75.3%. Based on these results, the Max_Depth value used in the final results is 7, as it has the highest accuracy.

b. Total Entropy Calculation

Entropy is a measure of the degree of uncertainty or disorder in data. In machine learning, particularly in decision tree methods, entropy is used to assess how well an attribute can divide data into distinct groups. The results of the total entropy calculation are shown in Table 4.

Table 4. Entropy Calculation

Attribute	Amount	Light	Heavy	*Entropy*
Total	10	7	3	0,881

Based on Table 4, the overall Entropy calculation results were obtained from 10 data points taken from the dataset. In this calculation, the total data was 10, with 7 included in the light class, and 3 data included in the heavy class. The Entropy value resulting from this calculation was 0.881, which was then used as a reference for calculating the Gain for each attribute. The following is the Entropy (Total) value calculation using Equation 2.1.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 4, October 2025
https://doi.org/10.47709/cnahpc.v7i4.6956

Submitted : Sep 13, 2025
Accepted : oct 25, 2025
Published : Oct 28, 2025

$$Entropy\ (s) = -\frac{7}{10}\log_2\frac{7}{10} - \frac{3}{10}\log_2\frac{3}{10} = 0{,}881$$

c. Calculating Entropy and Gain for Each Attribute

The data in the dataset will be divided based on each attribute, and then the Entropy and Gain values for each attribute are calculated. Gain indicates how well an attribute reduces data uncertainty after division. The following sample pulse data can be seen in Table 5.

Table 5. Pulse Data Samples

Range	Total	Light	Heavy
Low (<70)	0	0	0
Medium (70-90)	9	6	3
High (>90)	1	1	0

Table 5 shows the dataset on the Pulse attribute with three (3) value ranges and their categories. The medium range has 9 data points, while the high range only has 1 data point, and the low range has no data points. The entropy value calculation for each attribute is done using Equation 2.1.

$$Entropy\ (s) = \left(\frac{0}{10}\log_2\frac{0}{10}\right) + \left(\frac{9}{10}\log_2\frac{9}{10}\right) + \left(\frac{1}{10}\log_2\frac{1}{10}\right) = 0{,}826465$$

In this process, each attribute is evaluated by calculating its Gain value based on Equation 2.2 to determine its influence on the classification. The attribute with the highest Gain value is considered the most significant and prioritized in further analysis. Details of the Gain values for each attribute are shown in Table 3.8.
Gain (Pulse) = 0.881-0.826465 = 0.055

Table 6. Entropy and Gain Calculations

	Information	Number of cases	Light	Heavy	*Entropy*	*Gain*
Total		10	7	3	0,881	
Systolic	<70 (R)	7	5	2	0,863	
	70-80 (S)	1	1	0	0	
Diastolic	>80 (T)	2	1	1	-1	
						0,078
Pulse	<70 (R)	5	3	2	0,971	
Total	70-80 (S)	3	3	0	0	
Systolic	>80 (T)	2	1	1	1	
						0,1955
Diastolic	<70 (R)	0	0	0	0	
	70-90 (S)	9	6	3	-0,918295	
	Information	Number of cases	Light	Heavy	*Entropy*	*Gain*
	>90 (T)	0	0	0	0	
						0,055
RR	<20 (R)	3	1	2	-0,918295	
	20-22 (S)	7	6	1	0,591	
	>22 (T)	0	0	0	0	
						0,192
Temp	<36,5 (R)	1	0	1	0	
	36,5-37,5 (S)	9	7	2	0,762	
	>37,5 (T)	0	0	0	0	

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 4, October 2025
https://doi.org/10.47709/cnahpc.v7i4.6956

Submitted : Sep 13, 2025
Accepted : oct 25, 2025
Published : Oct 28, 2025

						0,195
BB	<30 (R)	2	0	2	0	
	30-50 (S)	0	0	0	0	
	50> (T)	8	6	2	0,811	
						0,232
TB	<100	0	0	0	0	
	100-160	5	3	2	0,970	
	>160	5	4	1	0,721	
						0,0355
LP	<100	9	6	3	0,918	
	100-160	1	1	0	0	
	>160	0	0	0	0	
						0,055

Table 6 shows that Entropy and Gain are used to determine the best attributes for forming a decision tree. Entropy measures the level of uncertainty in the data, while Gain indicates how much an attribute can reduce that uncertainty. The attribute with the highest Gain value is considered optimal for separating the data and forming a decision tree structure.

d. Decision Tree

The attribute with the highest Gain value will be selected as the primary node in the decision tree formation process. This Gain value indicates the extent to which an attribute contributes to separating data into different classes. Therefore, the attribute with the highest Gain value is considered the most relevant for initial data division. This selection process is repeated for each subsequent node, with each node selected based on the attribute most relevant to the subset of data being processed. In other words, each branch of the tree will continue to grow by selecting the attribute with the highest Gain from the remaining data. Once the entire decision tree structure is formed, the tree can be used to classify or predict new data. This tree formation process will continue until all data is optimally grouped or until certain conditions are met, such as a maximum depth or a minimum number of data points in each branch.

In addition to Decision Tree, this study also tested other algorithms for comparison, namely K-Nearest Neighbor (KNN) and Naive Bayes.

a. Decision Tree produced an accuracy of 81.75%, precision of 79.58%, recall of 81.75%, and an F1-score of 80.45%.
b. KNN showed an accuracy of 78.20%, with precision of 75.10% and recall of 77.30%.
c. Naive Bayes produced an accuracy of 76.45%, with precision of 74.00% and recall of 75.50%.

From these results, it can be seen that Decision Tree provides the best performance compared to the other two algorithms. This can be explained because Decision Tree is able to capture non-linear patterns and interactions between variables (for example, between body temperature and RR), while KNN and Naive Bayes are more limited in this regard.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 4, October 2025
https://doi.org/10.47709/cnahpc.v7i4.6956

Submitted : Sep 13, 2025
Accepted : oct 25, 2025
Published : Oct 28, 2025

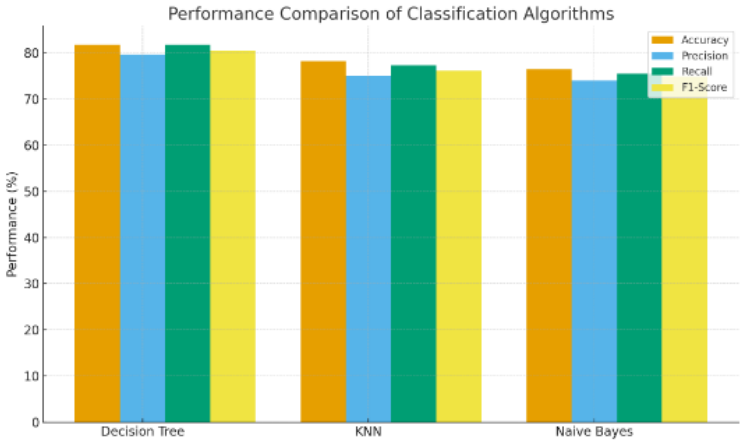

Figure 3. Performance comparison between Decision Tree, KNN, and Naive Bayes algorithms

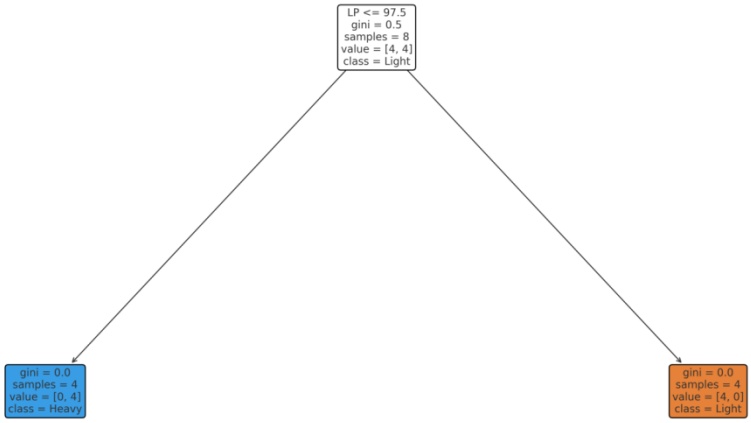

Figure 4. Visualization of the Decision Tree structure for ARI classification

DISCUSSIONS

This study aims to evaluate the performance of the Decision Tree algorithm in classifying ARI using data from the Bontang Barat Community Health Center. The model was developed through pre-processing, entropy and gain calculations, and tested using the 10-fold cross-validation method. The evaluation results showed that the model produced an accuracy of 81.75%, precision of 79.58%, recall of 81.75%, and an F1-score of 80.45%, with better performance in the majority class.

The results show that body temperature and respiratory rate have the highest gain in distinguishing between "Mild" and "Severe" classes. Clinically, these two variables are closely related to respiratory tract infections:

1. High body temperature is generally an indicator of a more severe infection.
2. An increased RR indicates significant respiratory dysfunction.

The dominance of these two features makes it easier for the algorithm to separate patients based on severity, so that the model's performance tends to be stable in various experiments.

The results of this study have several important implications:

1. It assists medical personnel in making quick decisions in the field, especially in primary health facilities with limited diagnostic tools.
2. It improves the efficiency of the triage process for patients with acute respiratory infections, as automatic classification can filter out patients who require immediate treatment.
3. Potential integration into electronic medical record systems to provide data-driven decision support systems.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 4, October 2025
https://doi.org/10.47709/cnahpc.v7i4.6956

Submitted : Sep 13, 2025
Accepted : oct 25, 2025
Published : Oct 28, 2025

Strengths vs Weaknesses of Decision Tree Compared to Other Algorithms. Strengths: 1)Interpretability and Transparency, one of the major advantages of the Decision Tree algorithm is its ease of interpretation. Its hierarchical structure mirrors human decision-making processes, making the model more transparent compared to "black box" algorithms such as Support Vector Machines (SVM) or Neural Networks. This is particularly beneficial in clinical settings, where explainability is crucial for decision support. 2) No Strong Statistical Assumptions Required, unlike logistic regression or other parametric methods, Decision Trees do not require the data to follow a specific distribution. This makes them well-suited for real-world healthcare datasets, which are often heterogeneous and noisy. 3) Capability to Handle Mixed Data Types, decision trees can naturally handle both categorical and numerical variables without the need for extensive data transformation. In contrast, algorithms such as SVM or k-Nearest Neighbors often require normalization or encoding steps. 4) Automatic Feature Selection, during tree construction, the algorithm automatically selects the most informative features based on information gain. This not only simplifies model development but also highlights key clinical variables (e.g., body temperature and respiratory rate in this study) that are most relevant for classification.

Weaknesses: 1) Prone to Overfitting, a single decision tree tends to overfit the training data, especially when the tree is deep and complex. This can result in reduced generalization performance compared to ensemble methods (e.g., Random Forest) or SVM, which are generally more robust to overfitting. 2) High Sensitivity to Data Variations, even small changes in the training dataset can lead to significant changes in the tree structure. This instability may affect the model's reproducibility, unlike methods such as logistic regression, which are less sensitive to data perturbations. 3) Potentially Lower Accuracy than More Complex Models, while decision trees are interpretable, their predictive performance is often lower than that of more sophisticated ensemble techniques (e.g., Gradient Boosting) or Neural Networks, especially in complex classification tasks. 4) Limited Ability to Capture Complex Non-linear Relationships, a single decision tree may struggle to model intricate feature interactions. In such cases, algorithms like Neural Networks or SVM may achieve superior performance.

Decision Trees offer a strong balance between simplicity, interpretability, and efficiency, making them a suitable choice for primary clinical decision-support applications. However, their limitations—particularly overfitting and instability—suggest that combining them with ensemble methods or comparing them with other advanced algorithms can lead to improved performance and model reliability.

Thus, classification using Decision Trees can serve as the basis for developing simple applications that help medical personnel identify high-risk patients more quickly and accurately.

CONCLUSION

This study successfully implemented the Decision Tree algorithm for classifying the severity of Acute Respiratory Infection (ARI) using clinical data collected from UPT Puskesmas Bontang Barat. Through systematic preprocessing, model training, and evaluation using the 10-Fold Cross Validation technique, the model achieved an average accuracy of 81.75%, demonstrating strong and reliable classification performance. The novelty of this research lies in being the first study to apply and evaluate the Decision Tree algorithm on an Indonesian ARI dataset, achieving accuracy exceeding 80%. This finding confirms that a simple yet interpretable model such as Decision Tree can effectively classify ARI severity based on local clinical features, providing a practical and transparent decision-support tool for primary healthcare settings in Indonesia.

Nevertheless, the dataset used in this study originated from a single healthcare center, which may limit the generalizability of the findings to other regions with diverse demographic and environmental characteristics. Future research should therefore focus on two key directions: 1) Integration of ensemble-based approaches (e.g., Random Forest, Gradient Boosting) to enhance model robustness and predictive accuracy while maintaining interpretability; and 2) Validation using multisenter datasets from various healthcare facilities across Indonesia to ensure broader representativeness and clinical applicability. By advancing in these directions, future studies can contribute to developing a more comprehensive, scalable, and data-driven framework for ARI detection and management within Indonesia's public health system.

REFERENCES

Ajijah, N., & Kurniawan, A. (2023). Klasifikasi Teks Mining Terhadap Analisa Isu Kegiatan Tenaga Lapangan Menggunakan Algoritma K-Nearest Neighbor (KNN). *J-SAKTI (Jurnal Sains Komputer & Informatika)*, *7*(1), 254–262.

Al-Harrasi, A., & Bhatia, S. (2022). Epidemiology respiratory infections: types, transmission, and risks associated

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 4, October 2025
https://doi.org/10.47709/cnahpc.v7i4.6956

Submitted : Sep 13, 2025
Accepted : oct 25, 2025
Published : Oct 28, 2025

with co-infections. In *Role of Essential Oils in the Management of COVID-19* (pp. 7-17). CRC Press.

Amaliah, S., Nusrang, M., & Aswi, A. (2022). Penerapan Metode Random Forest Untuk Klasifikasi Varian Minuman Kopi di Kedai Kopi Konijiwa Bantaeng. VARIANSI: Journal of Statistics and Its application on Teaching and Research, 4(3), 121-127.

Biyantoro, A. S., & Prasetiyo, B. (2024). Application of Decision Tree for Health Status Classification , Compared to KNN and Naive Bayes Penerapan Decision Tree untuk Klasifikasi Status Kesehatan dengan perbandingan KNN dan Naive Bayes. *Indonesian Journal of Informatic Research and Software Engineering*, *4*(1), 47–55. https://journal.irpi.or.id/index.php/ijirse/article/view/1342

Chaizuran, M., & Hijriana, I. (2023). Penyuluhan Pengobatan Tradisional ISPA Pada Balita di Gampong Bireuen Meunasah Reuleut Provinsi Aceh. *Jurnal Mandala Pengabdian Masyarakat*, *4*(1), 1–6. https://doi.org/10.35311/jmpm.v4i1.105

Defrianti, F., Hanifa, F., & Jayatmi, I. (2024). Hubungan Sikap Ibu, Dukungan Suami, dan Status Imunisasi Terhadap Kejadian Infeksi Saluran Pernapasan Akut (ISPA) Pada Balita. *Jurnal Penelitian Perawat Profesional*, *6*(4), 1799–1808. http://jurnal.globalhealthsciencegroup.com/index.php/JPPP

Devi, E. S., Wahono, B. B., & Wibowo, G. N. (2025). KOMPARASI ALGORITMA KLASIFIKASI NAÏVE BAYES , DECISISON TREE (C4 . 5) DAN SUPPORT VECTOR MACHINES (SVM) DALAM DIAGNOSA PENYAKIT DIABETES. *Jurnal Teknik Informatika.*, *4*(1), 222–228. https://doi.org/https://doi.org/10.02220/jtinfo.v4i1.1268

Dewi, S. P., Nurwati, N., & Rahayu, E. (2022). Penerapan Data Mining Untuk Prediksi Penjualan Produk Terlaris Menggunakan Metode K-Nearest Neighbor. *Building of Informatics, Technology and Science (BITS)*, *3*(4), 639–648. https://doi.org/10.47065/bits.v3i4.1408

Hidayat, D. (2023). *Kemenkes Catat Pengidap ISPA Meningkat Akibat Polusi Udara*. Rri.Co.Id. https://www.rri.co.id/nasional/339812/kemenkes-catat-pengidap-ispa-meningkat-akibat-polusi-udara

Imani, R. K., Wijoyo, S. H., & Amalia, F. (2024). Penerapan Algoritma K-Nearest Neighbor untuk Klasifikasi Kemampuan Lulusan Siswa Dalam Bersaing untuk Mendapatkan Pekerjaan (Studi Kasus: SMK "SORE" Tulungagung). Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, 8(10).

Kelly, F. J., & Fussell, J. C. (2015). Air pollution and public health: emerging hazards and improved understanding of risk. *Environmental geochemistry and health*, *37*(4), 631-649.

Mukaromah, H. (2025). Perbandingan Kinerja Algoritma Naïve Bayes Dan Decision Tree Untuk Prediksi Penyakit Batu Ginjal. Jurnal Rekayasa Perangkat Lunak, 4(1).

Munir, A. S., Saputra, A. B., Aziz, A., & Barata, M. A. (2024). Perbandingan Akurasi Algoritma Naive Bayes dan Algoritma Decision Tree dalam Pengklasifikasian Penyakit Kanker Payudara. *Jurnal Ilmiah Informatika Global*, *15*(1), 23–29. https://doi.org/10.36982/jiig.v15i1.3578

Nasien, D., Darwin, R., Cia, A., Winata, A. L., Go, J., Richard, M. C., Wijaya, C., Lo, K. C., Studi, P., Informatika, T., Komputer, F. I., Bisnis, I., & Indonesia, P. (2024). Perbandingan Implementasi Machine Learning Menggunakan Metode KNN , Naive Bayes , Dan Logistik Regression Untuk Mengklasifikasi Penyakit Diabetes. *JEKIN - Jurnal Teknik Informatika*, *4*(1). https://doi.org/10.58794/jekin.v4i1.640

Purwanta, P., Sadewa, D. M. A., Sahrinanda, D., Rizky, I., Muthoharoh, I. M., & Yunistyaningrum, V. (2023). Enabling the grass root: Health cadres empowerment program in efforts to prevent and manage hypertension in the Tanjung sub-village community. *Jurnal Pengabdian kepada Masyarakat (Indonesian Journal of Community Engagement)*, *9*(3), 181-187.

Sari, L., Siswa, T. A. Y., & Pranoto, W. J. (2024). Model rfgs-cs untuk mengatasi high dimensional data stunting kota samarinda skripsi. *JIPI Jurnal Ilmiah Penelitian Dan Pembelajaran*, *2*, 110–113. https://doi.org/https://doi.org/10.29100/jipi.v10i1.5997

Sari, N. A. A., Jannah, U. M., & Nurmalitasari. (2024). KOMPARASI ALGORITMA C4 . 5 DAN K- KLASIFIKASI PENYAKIT ANEMIA. *JIPI Jurnal Ilmiah Penelitian Dan Pembelajaran*, *9*(2), 110–113. https://doi.org/https://doi.org/10.51876/simtek.v9i2.399

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 4, October 2025
https://doi.org/10.47709/cnahpc.v7i4.6956

Submitted : Sep 13, 2025
Accepted : oct 25, 2025
Published : Oct 28, 2025

Sodikin, M. I. (2023). Penerapan dan Manfaat Machine Learning di Rumah Sakit. *Multiverse: Open Multidisciplinary Journal*, *2*(2), 262–265. https://doi.org/10.57251/multiverse.v2i2.1207

Sulistiyo, B., Surarso, B., & Syafei, W. A. (2020). Sistem Pakar Identifikasi dan Alternatif Solusi terhadap Permasalahan yang Dihadapi oleh Peserta Didik Sekolah Menengah Menggunakan Rule-Based Machine Learning. *Suparyanto Dan Rosad (2015*, *5*(3), 248–253. https://doi.org/10.14710/jtsiskom.2022.xxxxx

Susanto, A., Riady, S. R., Ranti, S. D., & Mandala, R. (2020). Penerapan Perhitungan Metode Decision Tree Menggunakan Algoritma Iterative Dichotomiser 3 (ID3) Berbasis Website Application of Decision Tree Method Calculation Using Website Based Iterative Dichotomiser 3 (ID3) Algorithm. *Indonesian Journal of Science*, *1*(2), 59–68. http://journal.pusatsains.com/index.php/jsi

William, W., Wijaya, A. T., Pasaribu, D. M., & Hudyono, J. (2022). Gambaran Penggunaan Antibiotik pada Anak Dibawah Usia Lima Tahun (Balita) dengan Infeksi Saluran Pernapasan Akut (ISPA) di Puskesmas Kelurahan Tanjung Duren Selatan Tahun 2020-2021. *Jurnal MedScientiae*, *28*, 1–5. https://doi.org/10.36452/jmedscientiae.v1i2.2510

* Corresponding author