Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.6, December 2025, pp. 3419~3430
e-ISSN: 2548-6861

Optimization of Early Diagnosis Prediction Models for Acute Respiratory Infections (ARI) in Children Using Decision Tree, Random Forest, and Resampling Techniques

Caesario Gumilang Firdaus [1]*, Asih Rohmani [2]*, Suharnawi [3]*
* Faculty of Computer Science, Dian Nuswantoro University, Semarang
112202206910@mhs.dinus.ac.id [1], asecharsoyo@dsn.dinus.ac.id [2], nt@dosen.dinus.ac.id [3]

Article Info

Article history:

Received 2025-10-23
Revised 2025-11-21
Accepted 2025-11-26

Keyword:

Decision Tree,
Data Imbalance,
Pediatric ARI,
Random Forest,
SMOTE-ENN.

ABSTRACT

Acute Respiratory Tract Infections (ARI) are the leading cause of childhood morbidity in Indonesia, with challenges in early detection due to limited medical personnel and diagnostic data imbalance, where LRTI cases are far fewer than URTI cases. This study developed and optimized an ARI classification prediction model (URTI and LRTI) based on machine learning with resampling techniques to address imbalance. An explanatory quantitative design was used with secondary data from the Mijen Community Health Center, Semarang (2020–2025, 12.177 valid data), with preprocessing including outlier handling (Winsorizing, IQR), stratified split (70:30), and RobustScaler on the training data. Three resampling techniques (SMOTE, ADASYN, SMOTE-ENN) were applied, then tested using Decision Tree and Random Forest with GridSearchCV and 5-fold cross-validation, focusing on Recall and AUC-PR evaluation for minority classes. The results showed that Random Forest with SMOTE-ENN provided the best performance, increasing the LRTI recall from 0.02 to 0.37 and F1-macro to 0.54, while Decision Tree with SMOTE-ENN produced the highest AUC-PR of 0.31. Despite this significant improvement, a recall of 0.37 is still low for clinical applications because the risk of false negatives remains high, potentially delaying patient treatment Future implementation requires the integration of clinical symptom data (e.g., respiratory rate) to achieve clinically acceptable sensitivity. These findings confirm that resampling can improve model capabilities, but additional feature exploration is needed to achieve adequate diagnostic sensitivity in the context of healthcare analytics.

I. INTRODUCTION

Acute respiratory infections (ARI) are one of the infectious diseases that remain a major cause of morbidity and mortality in children worldwide, especially in developing countries [1]. The World Health Organization (WHO) notes that ARI accounts for a significant proportion of deaths in children under five years of age, with the burden of disease showing no downward trend globally [2]. Based on the results of the 2023 Indonesian Health Survey, the national prevalence of ARI was recorded at 2.2%, with the rate in Central Java Province slightly higher at 2.5%. Among children under five years of age, the prevalence reached 4.8% nationally and 6.9% in Central Java [3]. This situation shows that ARI remains a public health issue that requires serious attention at both the national and regional levels.

Respiratory tract infections can be classified into upper respiratory tract infections (URTI) and lower respiratory tract infections (LRTI) based on the location of the infection, which have different clinical manifestations and risks of complications [2]. URTI are usually mild to moderate, while LRTI are more severe and require more intensive clinical attention [4]. This classification approach has been used in clinical practice to support more specific diagnoses and appropriate interventions in children. 2 Clinically, ARI has a significant impact on the health system because it has the

potential to cause serious complications such as pneumonia, bronchitis, and bronchiolitis, especially in children with low immunity [5]. Delayed treatment or inaccurate diagnosis can worsen the patient's condition and increase the risk of disease spread [6]. A study conducted at the Srondol Community Health Center and Batur Village, Getasan District, showed a relationship between poor nutritional status and the incidence of ARI in toddlers, indicating that biological and social factors contribute to the risk of infection [7].

Community health centers have abundant medical record data, such as demographic, physical, and administrative data, which can be used for early prediction. However, community health center data often varies in quality, has missing values, and contains noise. Frequent staff turnover, seasonal workloads, and changes in medical record storage operational standards can result in incomplete or inconsistent data. This phenomenon is common in Puskesmas and has been reported in several studies highlighting the variation in the quality of medical record inputs and the risk of administrative bias in primary health facilities in Indonesia [8]. Therefore, a rigorous preprocessing process is necessary to minimize this potential bias and ensure that the prediction model accurately describes clinical conditions.

The clinical and demographic variables used in this study are medically relevant in the assessment of ARI. Gender plays a role as a biological factor because several studies have shown that boys have a higher risk of developing ARI due to differences in the maturation of the immune system and respiratory tract [9]. The age of the child remains an important component because younger age groups, including toddlers, have a higher proportion of ARI cases than older age groups[10]. Anthropometric parameters such as weight, height, and Body Mass Index (BMI) also play an important role because nutritional status has been shown to affect ARI susceptibility and severity [11]. Blood pressure, although not a primary indicator of ARI, can reflect the body's physiological response to acute infection [12]. In addition, administrative information such as place of residence and type of health insurance can describe the socioeconomic conditions of families, which are important determinants in the incidence and management of ARI [13]. Thus, the selection of features in this study has a strong clinical and epidemiological basis to support comprehensive ARI diagnosis prediction.

Previous studies have attempted to apply machine learning (ML) methods to detect ARI and pneumonia. One study developed an Artificial Neural Network-based model to detect pneumonia in children, without applying data balancing techniques [14]. Another study proved that the application of resampling methods such as SMOTE and ADASYN significantly improved the performance of classification models on maternal health data, especially in minority classes [15]. Another study [16] confirmed the effectiveness of SMOTE-ENN in improving classification performance, especially in cases of imbalanced data, while study [17] found that Random Forest with resampling produced higher accuracy and recall in heart failure patients compared to models without resampling. These findings indicate that machine learning has great potential in supporting early diagnosis of diseases in children, although most of the previous literature has not integrated URTI/LRTI classification and demographic, biometric, and administrative variables simultaneously. Most studies have focused on toddlers and have not covered the population of children aged 1-10 years in primary health care services such as community health centers [14]. In addition, few studies have compared the performance of Decision Tree (DT) and Random Forest (RF) algorithms, and rarely integrated resampling techniques such as SMOTE, ADASYN, and SMOTE-ENN to address data imbalance [15], [16]. Previous studies have also not combined demographic, biometric, and administrative variables in a single predictive model, so the potential for comprehensive medical record analysis has not been maximized [18]. This gap is the basis for the need for this study to expand understanding in the application of machine learning techniques in the field of child health. Several studies have applied boosting algorithms such as XGBoost and LightGBM to predict ARI in children by balancing the data using a cost-sensitive approach, where the minority class is given a higher weight [19]. This approach shows that boosting models can be used as a strong baseline for tabular health data. However, these studies did not apply synthetic oversampling techniques such as SMOTE, ADASYN, or SMOTE-ENN, nor did they examine the classification of URTI and LRTI as two separate clinical categories. Therefore, this study continues to focus on the use of Decision Tree and Random Forest as core models that are more suitable for clinical interpretability in primary care, while still recognizing XGBoost as a common baseline in similar domains.

In addition to these findings, the selection of the SMOTE, ADASYN, and SMOTE-ENN methods in this study was based on the characteristics of the data distribution, which showed extreme imbalance between the URTI and LRTI classes, with a relatively small proportion of minority classes and the dominance of numerical features such as age, BMI, and blood pressure. This condition requires an oversampling approach that is capable of forming representative synthetic samples without losing important information, so simple undersampling methods were not selected. SMOTE and ADASYN were used because both generate new minority samples based on feature proximity, while ADASYN adaptively adds more samples to areas of features that are difficult to learn. On the other hand, SMOTE-ENN was combined to handle noise through a sample cleaning process, resulting in a more stable data distribution suitable for outlier-sensitive models such as Random Forest and Decision Tree.

Furthermore, although the dataset used covers the 2020-2025 service period (from January 1 to August 26) medical records do not provide temporal information such as visit dates, months, or seasons. This condition prevents the analysis of seasonal patterns and epidemiological trends. Since the medical record data does not contain consistent visit

timestamps, this study focuses on a pure classification approach based on patient features, rather than time-series or seasonal analysis.

This study aims to develop and optimize a model for predicting early diagnosis of ARI in children using Decision Tree and Random Forest algorithms by applying resampling techniques (SMOTE, ADASYN, and SMOTE-ENN) to address the imbalance in medical record data from the Mijen Community Health Center, Semarang City. The contributions of this study are both theoretical and practical. Theoretically, this study enriches the literature on the application of machine learning in pediatric health with a focus on data imbalance issues, while practically producing a medical record-based screening tool to support early detection of ARI in primary health care [16]. The novelty of this research lies in the integration of DT and RF with three resampling methods in the context of URTI and LRTI classification in primary health facilities, an approach that has not been widely explored in Indonesia.

II. METHOD

The research methodology included several interrelated stages, starting from data collection and preprocessing, dataset division, application of class balancing techniques, classification model development, to the evaluation stage using various model performance metrics.

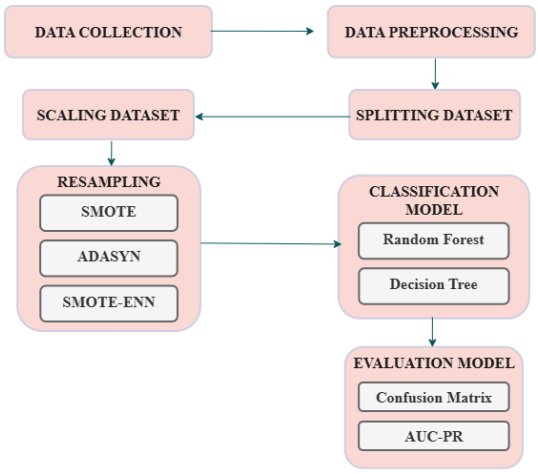

Figure 1. Research Flow

The stages begin with data collection, followed by preprocessing to ensure the quality of the dataset. Next, the data is divided into training data and test data in the dataset splitting stage. Two approaches are applied to the training data, namely without resampling (baseline) and with resampling using SMOTE, ADASYN, and SMOTE-ENN. Both approaches then enter the modeling stage with the Decision Tree (DT) and Random Forest (RF) algorithms, and end with model evaluation using accuracy, precision, recall, F1-score, and confusion matrix metrics. These stages are illustrated in Figure 1.

A. Data Collection

This study used secondary data from the medical records of pediatric patients diagnosed with ARI (ICD-10 J00-J22) in the Mijen Community Health Center Information System (SIMPUS), Semarang City, for the period 2020-2025. This dataset consists of 13.156 rows and 12 attributes covering demographic, clinical, and administrative variables. Demographic variables include gender, age in months, and patient type, while clinical variables include weight, height, body mass index (BMI), systolic blood pressure, and diastolic blood pressure. Administrative variables include polyclinic, payment type, and region of origin (subdistrict)

The target variable of the study is disease diagnosis based on the ICD-10 code, which is then mapped into two broad categories, namely Upper Respiratory Tract Infection (URTI) and Lower Respiratory Tract Infection (LRTI). The URTI category includes codes J00-J06, which cover acute nasopharyngitis, sinusitis, pharyngitis, laryngitis, and acute tracheitis. Meanwhile, the LRTI category includes codes J09-J22, which cover influenza, pneumonia, bronchitis, bronchiolitis, and undefined lower respiratory tract infections [2].

This mapping is a simplified form of analysis that refers to classification practices in international medical literature, which generally distinguishes ARIs based on the location of infection, namely the upper and lower respiratory tract [1], [4].

Although the data range is from 2020 to 2025, the dataset does not include consistent visit timestamps, such as the date or month of examination. The absence of this temporal information makes it impossible to analyze seasonal patterns or annual trends. Therefore, this study focuses on a pure feature-based classification approach that includes clinical, biometric, and demographic variables, without incorporating time series or seasonal analysis that requires complete and structured time attributes.

B. Data PreProcessing

The data preprocessing stage was carried out to ensure the quality and suitability of the data before it was used in the modeling process. The first step was label mapping from the ICD-10 code into two target classes, namely Upper Respiratory Tract Infection (URTI) and Lower Respiratory Tract Infection (LRTI) [2].

$$Diagnosis = \begin{cases} URTI, if\ ICD\text{-}10 \in [J00 - J06] \\ LRTI, if\ ICD\text{-}10 \in [J08 - J22] \end{cases}$$

This formula is used to simplify the variety of diagnoses into two main categories based on the anatomy of the respiratory tract [4].

Missing values were handled using a simple imputation method for numerical and categorical attributes, so that no important information was lost from the dataset. Next, duplicate data was removed to avoid repetition of the same patient entries, which could affect the model training results.

The next step is to remove constant attributes, which are features that have identical values across all data rows and do not contribute to the classification process. After that, outliers are detected and handled using two statistical approaches, namely Interquartile Range (IQR) and Winsorizing.

The IQR method is used to detect and remove extreme values based on the distance between the first quartile (Q1) and the third quartile (Q3)

$$Lower\ Bound = Q1 - (1.5 \times IQR)$$
$$Upper\ Bound = Q3 + (1.5 \times IQR)$$

The Winsorizing method is used not to remove outliers, but to limit extreme values so that they do not affect the data distribution.

$$x_i' = \begin{cases} P_\alpha, & if\, x_i < P_\alpha \\ x_i, & if\, P_\alpha \le x_i \le P_{1-\alpha} \\ P_{1-\alpha}, & if\, x_i > P_{1-\alpha} \end{cases}$$

This approach aims to identify extreme values that deviate significantly from the normal distribution of data without eliminating natural variations that are still clinically relevant [20].

Next, categorical variables are converted to numerical form using Label Encoding, so that they can be processed by machine learning algorithms without losing their original categorical meaning [21].

C. Splitting Dataset

This stage is carried out after the data cleaning process to separate the dataset into two main parts, namely 70% training data (training set) and 30% test data (testing set). The division is done using stratified sampling techniques so that the distribution of the URTI and LRTI target classes remains balanced in both subsets. This approach is important to maintain the representativeness of patterns in each class and prevent model bias towards the majority class [22]. The stratification method has also been widely used in medical research to ensure that model performance remains stable even though the dataset is limited in size [23].

D. Scaling Dataset

After the dataset is divided, normalization is performed only on the training set using RobustScaler, then the parameters obtained from the training data are used to transform the test data. This approach prevents data leakage, where information from the test data should not influence the model training process. RobustScaler was chosen because it effectively suppresses the influence of outliers without changing the original distribution of numerical data such as body temperature, respiratory rate, and blood pressure [24]. Normalization is performed so that each feature has a uniform scale, allowing the Decision Tree and Random Forest algorithms to learn patterns more efficiently and consistently.

E. Resampling

Class imbalance (imbalanced data) is a significant challenge in disease modeling, where the number of URTI cases is much higher than LRTI cases. This condition can cause model bias towards the majority class and reduce detection capabilities in the minority class. To overcome this, three resampling techniques were applied, namely SMOTE, ADASYN, and SMOTE-ENN.

The selection of these three resampling methods was based on the characteristics of the dataset, which consisted mostly of continuous numerical features resulting from clinical measurements, allowing for a more accurate synthetic sample generation approach. In addition, the very small proportion of LRTI classes compared to URTI required a method capable of improving the representation of minority classes without removing the original data, so the undersampling technique was not used. In this context, SMOTE and ADASYN were chosen because they are capable of safely enriching the pattern variation in the minority class, while SMOTE-ENN was used to overcome potential noise that arises after the oversampling process so that the data distribution becomes cleaner and more stable when used for model training.

SMOTE (Synthetic Minority Oversampling Technique) generates synthetic data based on the proximity between minority vectors [25]. ADASYN (Adaptive Synthetic Sampling) multiplies synthetic data in areas that are difficult to classify [15]. Meanwhile, SMOTE-ENN combines oversampling and data cleaning using Edited Nearest Neighbor to reduce noise [26].

All of these techniques are applied only to the training data to prevent data leakage to the test data, thereby improving the stability and sensitivity of the model to minority classes without sacrificing overall accuracy.

In the GridSearchCV process with 5-fold cross validation, the resampling technique and fitting scaler process are applied only to the subset that acts as training data in each fold, while the subset that acts as validation data per fold does not undergo resampling or refitting. Validation data is only transformed using scaling parameters obtained from the training fold. Meanwhile, the test data, which is separated from the outset, is not involved in the training, resampling, scaling, or tuning processes, and is only used once in the final evaluation stage. With this procedure, all stages of the analysis are free from data leakage.

F. Classification Model

The modeling stage aims to build a classification model capable of predicting the initial diagnosis of Acute Respiratory Infection (ARI) in children based on medical record data. Two algorithms used are Decision Tree (DT) and Random Forest (RF), as both are known to be effective for medical tabular data and have high interpretability. The modeling process was carried out using the Scikit-learn library, with a series of training, validation, and

hyperparameter tuning procedures to obtain optimal performance from each algorithm.

During the model tuning process, GridSearchCV with a 5-fold cross validation scheme was used to obtain the best parameter combination. During this process, all training stages such as scaling and resampling were applied exclusively to the training data in each fold, while the validation data was kept in its original condition without modification. Furthermore, the test data obtained from the initial division was never involved in the resampling process, additional scaling, or hyperparameter tuning. This approach ensures that no data leakage occurs during the training and cross-validation processes, so that the final evaluation truly reflects the model's ability to generalize.

1) Decision Tree (DT)

The Decision Tree algorithm works by building a decision tree structure based on the principle of information gain or Gini index as the attribute separation criterion [27]. conditions, and the leaves contain the final class labels. The hyperparameter tuning process is carried out to adjust parameters such as max_depth, min_samples_split, and criterion, in order to avoid overfitting and improve the model's generalization to new data. DT was chosen because it produces a model that is easy for medical personnel to interpret and is suitable for clinical decision support systems [28].

2) Random Forest (RF)

Random Forest is an ensemble learning algorithm that combines a number of decision trees through a bagging mechanism and random feature selection [29]. This approach improves prediction stability and reduces the variance that often occurs in single tree models. Hyperparameter tuning is performed on parameters such as n_estimators, max_features, and max_depth to achieve a balance between accuracy and computational efficiency. RF is widely used in the healthcare domain due to its ability to detect complex patterns and its resilience to data noise [28].

G. Evaluation Model

Model evaluation is performed to assess the performance of classification algorithms in predicting the initial diagnosis of ARI based on clinical data. Five main metrics are used, namely Accuracy, Precision, Recall, F1-Score, and Confusion Matrix [30]. Each metric has a specific function to describe the overall performance of the model.

Accuracy measures the proportion of correct predictions out of all test data and is used to assess the overall performance of the model [31].

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$$

Precision describes the accuracy of the model in identifying positive cases from all positive predictions generated [32].

$$Precision = \frac{TP}{TP + FP}$$

Recall (Sensitivity) shows the model's ability to detect all positive cases that actually exist in the data [33].

$$Recall = \frac{TP}{TP + FN}$$

The F1-Score is the harmonic mean between precision and recall, providing a balance between the sensitivity and specificity of the model [31].

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

In addition, the Confusion Matrix is used to visualize the distribution of classification results into four main categories [34] : *True Positive (TP)*, *True Negative (TN)*, *False Positive (FP)*, and *False Negative (FN)*.

TABLE I
CONFUSION MATRIX

	Positive	Negative
Positive	**TP**	**FN**
Negative	**FP**	**TN**

In addition to understanding the distribution of predictions through the Confusion Matrix, it is necessary to select evaluation metrics that can accurately describe the performance of the model in conditions of unbalanced class distribution. The selection of metrics in this study focused on Recall, F1-macro, and AUC-PR because accuracy tends to be biased in unbalanced data [35]. Recall is prioritized to minimize the risk of false negatives in LRTI cases, F1-macro provides a balanced assessment between classes, and AUC-PR is more informative than ROC-AUC in imbalanced conditions. The combination of these three metrics is considered most relevant for evaluating model performance in the early diagnosis of ARI.

III. RESULT AND DISCUSSION

A. Dataset Description

The dataset used in this study consists of medical records of pediatric ARI patients sourced from the Mijen Community Health Center, Semarang City. The initial dataset consists of 13.156 rows of data. The feature variables analyzed included demographic data and basic physical examination data, such as Gender, age in months (Age_months), Weight, Height, body mass index (BMI), blood pressure (Systolic, Diastolic), and administrative data such as Poly, Patient, Payment, and Subdistrict. The raw target variable is Diagnosis, which uses the ICD-10 code (J00 to J22). Examples of some features

from the dataset before the preprocessing stage are shown in Table 2.

TABLE II
DATASET BEFORE PREPROCESSING

No.	Gender	Age_ months	…	Subdistrict	Diagnosis
1	M	84	…	Ngadirjo	J02
2	F	84	…	Mijen	J02
3	M	72	…	Tambangan	J06
4	F	96	…	Wonolopo	J18
5	F	96	…	Jatibarang	J02
…	…	…	…	…	…
…	…	…	…	…	…
…	…	…	…	…	…
13154	M	9	…	Out of area	J06.9
13155	M	4	…	Cangkiran	J06.9
13156	M	5	…	Out of area	J00

B. Preprocessing result

The preprocessing process is carried out to ensure data quality before it is used in the ARI classification model training stage. The preprocessing stages range from label mapping to normalization and encoding of categorical variables.

1) Label Mapping

The first step is to map the diagnosis labels from the ICD-10 codes J00-J22 into two main target categories. Based on the location of the respiratory tract infection, codes J00-J06 are grouped as Upper Respiratory Tract Infections (URTI), and codes J09-J22 are grouped as Lower Respiratory Tract Infections (LRTI). This mapping aims to simplify the variety of diagnoses into two main classes based on the anatomy of the respiratory tract.

The mapping process and examples of labeling results are presented in Table 3. The results of this mapping produced two target classes with an initial unbalanced distribution, namely URTI with 11.481 data (majority) and LRTI with 1.675 data (minority). This distribution is presented in Figure 2.

TABLE III
LABEL MAPPING

No.	Gender	Age_ months	…	Subdistrict	Diagnosis
1	M	84	…	Ngadirjo	URTI
2	F	84	…	Mijen	URTI
3	M	72	…	Tambangan	URTI
4	F	96	…	Wonolopo	LRTI
5	F	96	…	Jatibarang	URTI

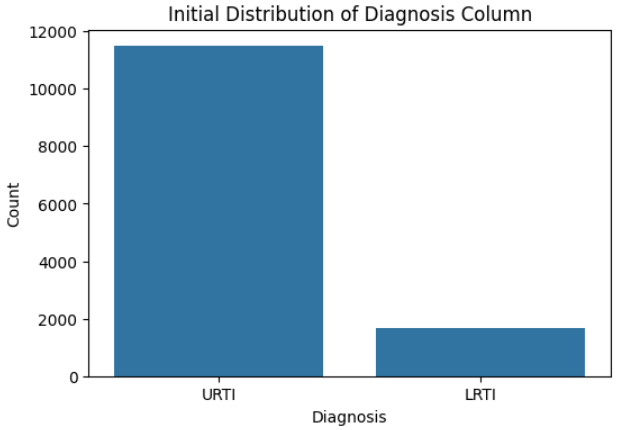

Figure 2. Distribution class diagnosis

2) Handling Missing Value

Missing values were handled using the median imputation method to maintain the stability of numerical data. The attributes Age_months, Weight, Height, Systolic, and Diastolic had empty values. All empty values were replaced with the median value as shown inTable 4.

TABLE IV
HANDLING MISSING VALUE

Attribute	Missing Value	After Imputation	Method
Age_months	1	0	Median
Weight	2	0	Median
Height	10	0	Median
Systolic	432	0	Median
Diastolic	427	0	Median

3) Handling Duplicate

The duplicate checking stage was carried out after the label mapping process to ensure that the data used was truly unique at the diagnosis category level (URTI and LRTI). Of the total 13.156 initial records, 979 duplicate data were found and then deleted, leaving 12.177 clean data as shown in Table 5.

TABLE V
HANDLING DUPLLICATE

Description	Number of Records
Total Initial Data	13.156
Duplicate Data	979
Final Data Count	12.177

4) Handling Variable Constant

The next step is to remove constant attributes, which are variables that do not vary in value and do not contribute information to the model learning process. In this study, the Poly attribute was removed because all entries had the same value, namely "MTBS". Variables such as this are considered uninformative features, so their removal is necessary to

prevent the model from being burdened by irrelevant features and to improve data processing efficiency.

5) Handling Outlier

Outlier handling was performed to maintain the consistency of numerical values in attributes such as height, weight, BMI, and blood pressure. The Interquartile Range (IQR) method was used to detect and adjust extreme values in the Weight, Height, BMI, and Diastolic variables by replacing values outside the lower and upper limits to remain within the physiological range for children. Additionally, the Systolic attribute is handled using a winsorizing approach at the 1st and 99th percentiles to reduce the influence of extreme values without deleting data. This combination strategy of IQR and winsorizing maintains the stability of the data distribution before the normalization and modeling stages.

6) Encoding Features

subdistrict, and diagnosis were converted into numerical values using LabelEncoder so that they could be processed by the algorithm. The results are shown in Table 6.

TABLE VI
ENCODING FEATURES

Attribute	UniqueValue	Encoding
Gender	['M' 'F']	[0 1]
Patient	['OLD' 'NEW']	[0 1]
Payment	['BPJS NON PBI' 'FREE' 'BPJS PBI' 'GENERAL']	[0 1 2 3]
Subdistrict	['Ngadirgo' 'Mijen' 'Tambangan' …'Cangkiran']	[0 1 3 … 146]
Diagnosis	['URTI' 'LRTI']	[0 1]

C. Splitting Dataset

The dataset was divided into two parts using a stratified train-test split approach to maintain a balanced proportion between the URTI and LRTI classes. A total of 70% of the data (8.523 rows) was used as training data, while the remaining 30% (3.654 rows) was used for testing. Stratified division was performed so that the model would obtain a balanced data representation and be able to generalize well.

In addition, the division results show that both subsets have 10 independent features and one target diagnosis variable, with a consistent data structure. This stage is important to ensure that the model validation process runs objectively and reduces the risk of overfitting.

D. Scaling Dataset

The standardization process was performed using RobustScaler to reduce the influence of outliers on numerical features. This technique was chosen because it is more stable against data distributions with extreme values than conventional normalization methods. The fitting process was performed only on the training data, while the test data was transformed using the same parameters to prevent data leakage.

E. Resampling

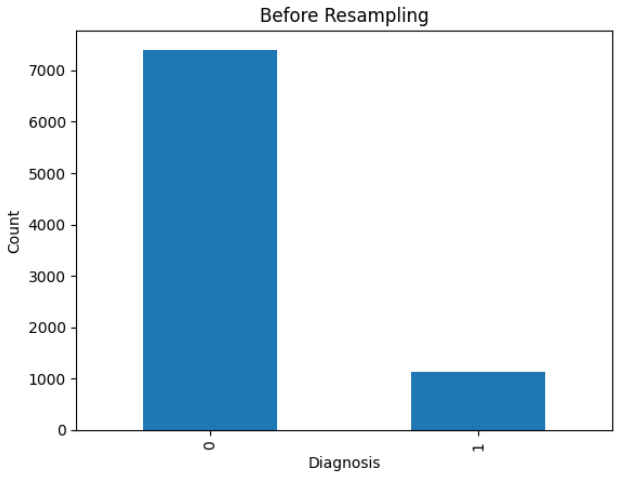

Figure 3. Initial Distribution

Figure 3 illustrates the significant class imbalance present in the initial dataset. Class 0 (URTI) comprises 7.400 samples, while Class 1 (LRTI) contains only 1.123 samples in the training data. This substantial disparity had the potential to impair the model's ability to accurately recognize the minority class. To address this issue, three data balancing techniques were subsequently applied: SMOTE, ADASYN, and SMOTE-ENN.

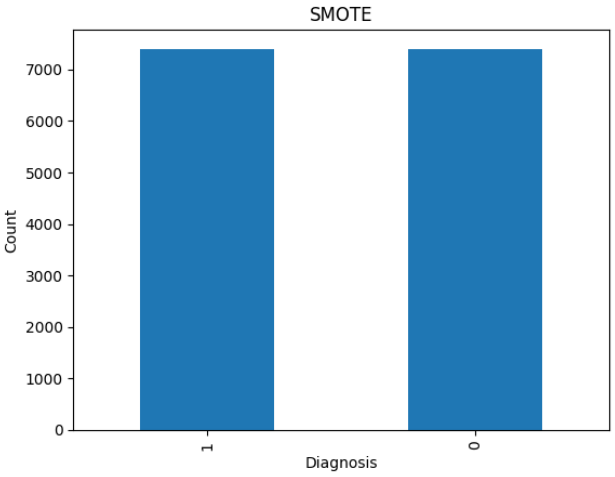

Figure 4. Distribution after SMOTE

Figure 4 demonstrates the outcome of employing the SMOTE technique. This method successfully achieved perfect class equilibrium, wherein both classes now possess an identical number of samples: 7.400 samples for Class 0 and 7.400 samples for Class 1.

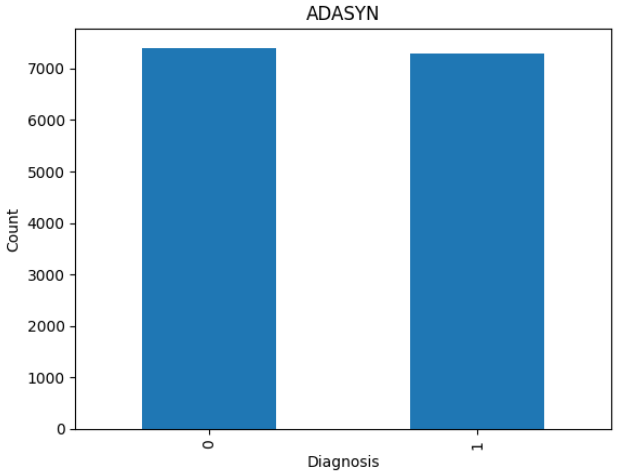

Figure 5. Distribution after ADASYN

Figure 5 displays the sample distribution following the application of the ADASYN technique. This approach yielded a distribution that is in near-parity, with Class 0 remaining at 7.400 samples and Class 1 increasing to 7.290 samples.

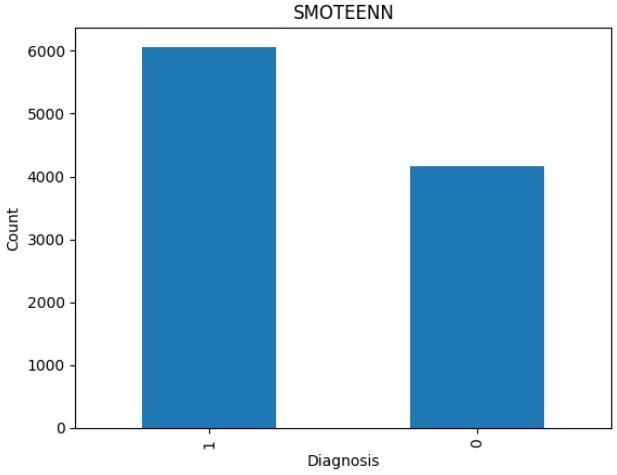

Figure 6. Distribution after SMOTEENN

Figure 6 illustrates the sample distribution after utilizing the hybrid SMOTE-ENN technique. This method not only augmented the minority class samples but also cleaned the dataset by reducing potentially noisy instances in the majority class. Consequently, the resulting distribution shows a decrease in Class 0 to 4.170 samples and an increase in Class 1 to 6.060 samples, marking a significant reduction in class disparity compared to the initial data condition.

F. Classification Model Result

The classification models used in this study consisted of Decision Tree (DT) and Random Forest (RF), which were applied to the ARI patient dataset after undergoing preprocessing and resampling using SMOTE, ADASYN, and SMOTE-ENN. Before tuning, both models showed relatively low initial performance in the minor class (LRTI). For Random Forest, the accuracy on the original data reached 84.9%, but the F1-score for the minor class was only 0.03, while on SMOTE, ADASYN, and SMOTE-ENN data, the accuracy ranged from 70-82% with a low minor F1-score (0.11-0.26). Decision Tree showed a similar pattern, with accuracy before tuning between 68-76% and minor class F1-score between 0.16-0.23. This confirms the challenges arising from unbalanced class distribution.

After tuning using GridSearchCV with 5-fold cross-validation and the parameter class_weight = 'balanced', the performance of both models improved, especially in predicting minor classes. Random Forest on SMOTEENN data produced an accuracy of 71.6% and an F1-macro of 0.54, with minor class recall increasing to 0.37. Other data (original data, SMOTE, ADASYN) showed similar improvements in minor class recall, although the overall accuracy remained relatively stable. The post-tuning Decision Tree SMOTEENN also showed an increase in minor class recall to 0.37, with an F1-macro of 0.54, while the original data and ADASYN showed more moderate improvements.

Overall, Random Forest tends to provide a more stable and consistent prediction distribution compared to Decision Tree, although both models show comparable capabilities in effectively classifying ARI diagnoses. The results before and after tuning confirm that parameter optimization is very important for improving prediction performance, especially for minor classes that are more difficult to predict.

G. Evaluation Model

The evaluation of classification model performance in this study focused on the ability of Random Forest (RF) and Decision Tree (DT) to detect majority (0) and minority (1) classes in the ARI dataset. The analysis was conducted on models before and after parameter tuning, as well as for each resampling method, namely Original Data, SMOTE, ADASYN, and SMOTEENN.

TABLE VII
CONFUSION MATRIX RANDOM FOREST

Method	Tuning	TN	FP	FN	TP
Original Data	Before	3096	473	77	8
SMOTE	Before	2946	441	227	40
ADASYN	Before	2934	439	239	42
SMOTEENN	Before	2396	294	777	187
Original Data	After	2566	329	607	152
SMOTE	After	2945	435	228	46
ADASYN	After	2949	436	224	45
SMOTEENN	After	2438	304	735	177

Based on the results of the Random Forest confusion matrix model in Table 7, Random Forest on the original data before tuning shows a strong bias towards the majority class. The model tends to prioritize dominant patterns in the dataset, resulting in limited ability to detect minority classes. This can be seen from the high True Negative performance but low True Positive performance.

After applying resampling methods such as SMOTE and ADASYN, there was an increase in True Positive, but this increase was still limited because these methods only added synthetic samples without optimizing class boundaries. SMOTEENN provided more significant improvements because the combination of oversampling and instance cleaning made the model more sensitive to the minority class, although on the other hand it reduced negative predictions in the majority class. This pattern remains visible after tuning, with the highest increase in True Positive in SMOTEENN, but there is still a decrease in the accuracy of negative predictions.

TABLE VIII
CONFUSION MATRIX DECISION TREE

Methode	Tuning	TN	FP	FN	TP
Original Data	Before	2714	402	459	79
SMOTE	Before	2632	390	541	91
ADASYN	Before	2564	376	609	105
SMOTEENN	Before	2307	304	866	177
Original Data	After	2079	202	1094	279
SMOTE	After	2716	394	457	87
ADASYN	After	2621	381	552	100
SMOTEENN	After	2438	301	742	180

Based on the results of the confusion matrix in Decision Tree Table 8, Decision Tree is more sensitive to class distribution variations than Random Forest because trees directly form separations based on data. In the original data, the model's ability to recognize minority classes is still limited, but resampling, especially SMOTEENN, can significantly improve True Positive thanks to instance cleaning that clarifies class boundaries.

After tuning, the Decision Tree on the original data showed a drastic increase in True Positive, although it was followed by a decrease in True Negative. This reflects the tendency of this single model to overfit on unbalanced datasets. SMOTEENN still provides the highest True Positive after tuning, but the increase in errors in the majority class shows the limitations of the model in balancing Precision and Recall.

TABLE IX
EVALUATION METRICS RANDOM FOREST

Method	Tune	Acc	Prec	Recall	F1
Original Data	Before	0.849	0.765	0.849	0.801
SMOTE	Before	0.817	0.775	0.817	0.794
ADASYN	Before	0.814	0.775	0.814	0.792
SMOTEENN	Before	0.706	0.798	0.706	0.743
Original Data	After	0.743	0.796	0.743	0.766
SMOTE	After	0.818	0.778	0.818	0.796
ADASYN	After	0.819	0.778	0.819	0.796
SMOTEENN	After	0.715	0.797	0.715	0.749

TABLE X
EVALUATION METRICS DECISION TREE

Method	Tune	Acc	Prec	Recall	F1
Original Data	Before	0.764	0.775	0.764	0.769
SMOTE	Before	0.745	0.775	0.745	0.759
ADASYN	Before	0.730	0.776	0.730	0.751
SMOTEENN	Before	0.679	0.789	0.679	0.723
Original Data	After	0.645	0.818	0.645	0.701
SMOTE	After	0.767	0.779	0.767	0.773
ADASYN	After	0.744	0.778	0.744	0.760
SMOTEENN	After	0.714	0.798	0.714	0.748

Additional evaluation metrics (Accuracy, Precision Recall, and F1-Score) presented in Table 9 and Table 10. Random Forest shows better performance stability than Decision Tree because the voting mechanism in the ensemble is able to reduce prediction variability between trees, so that changes in data distribution due to resampling and tuning do not drastically affect performance. Conversely, Decision Tree, which is a single model, is more easily affected by changes in data structure, so that its metrics are more volatile.

The evaluation metric results (Accuracy, Precision, Recall, and F1-Score) support this insight, with Random Forest's average F1-Score being more consistent across all resampling and tuning methods. This shows that the stability of the ensemble model is not only evident technically, but also reflected empirically in the performance of the evaluation metrics.

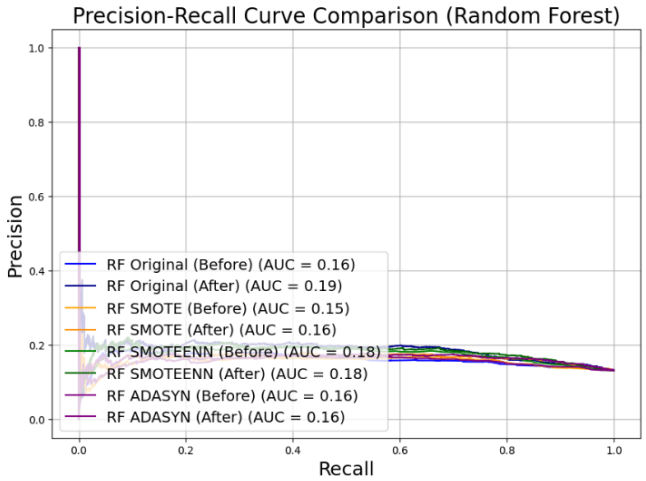

Figure 7. Precision Recall Curve Random Forest

Precision-Recall (PR) curve analysis provides crucial insight into the model's ability to address class imbalance. Based on the test results on Random Forest, which are visualized in Figure 7, in Random Forest, the highest AUC-PR value was achieved by the original model after tuning (AUC = 0.19), followed by SMOTEENN (AUC = 0.18). This low AUC value indicates that although Random Forest is stable in terms of accuracy and F1-score, the model as a whole has difficulty balancing Precision and Recall in the minority

class (LRTI). This confirms that the ensemble mechanism prioritizes majority class patterns, and tuning only provides limited improvement in minority sensitivity.

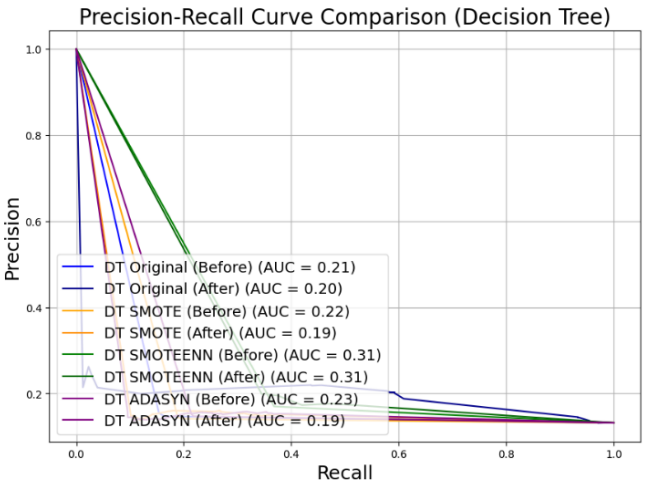

Figure 8. Precision Recall Curve Decision Tree

This contrasts with the Decision Tree model, as shown in Figure 8. Decision Tree shows a better response to resampling techniques, especially SMOTEENN. The highest AUC-PR value was achieved by DT SMOTEENN before and after tuning (AUC = 0.31). This improvement indicates that Decision Trees, with aggressive resampling combinations, are more effective at increasing sensitivity to minority classes without sacrificing Precision to an extreme degree. The tuning process on Decision Tree SMOTEENN did not produce significant changes, confirming that the quality of the resampled data distribution is more decisive for AUC-PR performance than model hyperparameters.

From the results of the confusion matrix analysis and evaluation metrics, the SMOTE-ENN method proved to be the most effective in improving recall in the minority class (LRTI) in both algorithms, although it was always accompanied by a trade-off in the form of a decrease in Precision in the majority class (URTI). Specifically, the Random Forest algorithm showed more stable and balanced performance (high macro F1-Score), while Decision Tree showed the best Recall potential (the highest achieved by Decision Tree Original After with 279 TP) but with a higher risk of False Positives in the majority class. These findings confirm the fundamental trade-off between overall accuracy and minority class detection capability, which is a major challenge in classifying medical data with an imbalanced distribution.

However, the recall value of 0.37 in both algorithms is still relatively low for clinical applications due to the high potential for false negatives, which means that some cases of LRTI remain undetected. The increase from the very low baseline, which was 0.02 in Random Forest and 0.16 in Decision Tree, shows that the balancing technique provides significant progress, but the sensitivity of the model is still inadequate for healthcare purposes. These findings indicate that LRTI screening cannot rely solely on routine administrative and anthropometric data from community health centers. Resampling techniques help improve performance, but the addition of clinical features such as symptoms, immunization history, or environmental exposure is an absolute necessity for the model to achieve higher sensitivity. Thus, low recall results are not merely a model failure, but an important scientific finding that highlights the intrinsic limitations of Puskesmas data in supporting early detection of LRTI without feature enrichment.

These findings reinforce the theory of data balancing in the health domain, particularly in disease prediction with imbalanced class distributions. The effectiveness of SMOTE-ENN in improving recall and F1-macro is in line with studies [15] and [17], which show that resampling techniques can improve model sensitivity to minority classes in clinical data. These results are also consistent with [16], which shows that the combination of oversampling and instance cleaning can improve data structure so that the model is more stable in distinguishing classes. However, the highest AUC–PR achievement in the SMOTE-ENN Decision Tree shows characteristics that have not been widely discussed in previous research. Thus, this study makes an important contribution that SMOTE-ENN is not only effective in ensemble models such as Random Forest, but also optimally improves the performance of interpretable models such as Decision Trees in the context of pediatric ISPA diagnosis.

IV. CONCLUSION

Based on the results of analysis using the Decision Tree and Random Forest algorithms with the application of resampling techniques (SMOTE, ADASYN, and SMOTE-ENN), this study shows that balancing plays a significant role in improving the model's ability to detect minority classes in ISPA data for children. The Random Forest model with SMOTE-ENN provided the most balanced performance, as indicated by an increase in the recall of the LRTI class to approximately 0.37 and an F1-macro value of 0.54. Meanwhile, the Decision Tree with SMOTE-ENN produced the highest AUC–PR value of 0.31, indicating that aggressive resampling techniques can improve the sensitivity of simple models to imbalanced data distributions. These findings support the research hypothesis that resampling methods can improve prediction accuracy in medically imbalanced data.

Although the recall of the minority class increased to approximately 0.37 in both algorithms, this value is still not ideal for the clinical context because the risk of false negatives remains high. However, the improvement from a very low baseline indicates that balancing techniques such as SMOTE-ENN are an important first step in improving model sensitivity.

Theoretically, this study reinforces the concept in the literature that balancing, particularly through SMOTE-ENN, not only increases the number of minority samples but also

improves the data structure so that the model becomes more capable of distinguishing classes with unequal distributions. These results are in line with previous studies on the effectiveness of resampling in the health domain, but provide a new contribution by demonstrating that SMOTE-ENN is capable of optimizing the performance of two types of algorithms simultaneously, both ensemble models such as Random Forest and interpretable models such as Decision Tree in the specific context of pediatric ISPA diagnosis. Thus, this study adds empirical evidence that the combination of oversampling and instance cleaning can be a consistent approach to improving classification performance on imbalanced clinical data.

In practical terms, this study has important implications for the development of early detection systems in primary health care facilities. Community health centers and health departments are advised to adopt machine learning-based prediction models with resampling techniques such as SMOTE-ENN as tools for the early identification of LRTI cases that are at risk of being missed in routine clinical diagnosis. For future research, it is recommended to expand the scope of the dataset, include additional clinical variables such as symptoms, immunization, and environmental factors, and evaluate advanced ensemble methods or deep learning to improve the generalization and external validity of the model. Overall, these results show that this study not only provides empirical contributions but also relevant theoretical and practical contributions to the development of data-driven health policies.

REFERENCES

[1] V. History, "Describing, characterising and predicting winter respiratory accident and emergency attendances, hospital and intensive care unit admissions and deaths in Scotland Version History," hal. 1–8, 2023.

[2] M. Del Riccio *et al.*, "Burden of Respiratory Syncytial Virus in the European Union: Estimation of RSV-Associated hospitalizations in children under 5 years," *J. Infect. Dis.*, vol. 228, no. 11, hal. 1528–1538, 2023, doi: 10.1093/infdis/jiad188.

[3] Survey kesehatan indonesia (Ski), "Survei Kesehatan Indonesia 2023 (SKI)," *Kemenkes*, hal. 235, 2023.

[4] R. Boracchini *et al.*, "A silent strain: the unseen burden of acute respiratory infections in children," *Ital. J. Pediatr.*, vol. 50, no. 1, hal. 2–5, 2024, doi: 10.1186/s13052-024-01754-2.

[5] S. Azis, H. Jusuf, dan L. Kadir, "Risiko Kejadian Penyakit Infeksi Saluran Pernapasan Akut pada Balita di Puskesmas Momunu Kabupaten Buol," *Myjurnal.Poltekkes-Kdi.Ac.Id*, vol. 14, no. 2, hal. 2087–2122, 2023.

[6] M. T. Hidayat, E. A. Jayadipraja, M. Asrullah, dan K. W. Astawa, "Analisis Time Trend Kualitas Udara Ambien dan Peningkatan ISPA di Kota Kendari," *Miracle J. Public Heal.*, vol. 7, no. 1, hal. 53–65, 2024, doi: 10.36566/mjph/Vol7.Iss1/365.

[7] K. P. A. Nugroho, B. P. S. Adi, dan R. Angelina, "Gambaran Status Gizi Kurang Dan Kejadian Penyakit Ispa Pada Balita Di Desa Batur, Kecamatan Getasan, Kabupaten Semarang," *J. Kesehat. Kusuma Husada*, hal. 233–242, 2018, doi: 10.34035/jk.v9i2.285.

[8] A. Information, "Evaluasi Dan Kontrol Kualitas Kelengkapan Berkas," vol. 2, hal. 627–634, 2024.

[9] K. Ritonga dan K. Kunci, "Hubungan Faktor Risiko Dengan Kejadian Ispa Pada Anak Di Wilayah Kerja Puskesmas Tanjung Beringin Kabupaten Serdang," vol. IV, no. Ii, hal. 108–114, 2021.

[10] S. Billa, N. Suhada, C. Novianus, dan I. R. Wilti, "Faktor-Faktor yang Berhubungan dengan Kejadian Ispa pada Balita di Puskesmas Cikuya Kabupaten Tangerang Tahun 2022," vol. 3, no. 2, hal. 115–124, 2023.

[11] E. R. Molenaar, F. Hans, dan M. Mawo, "Hubungan Status Gizi Dengan Kejadian Ispa Pada Balita Di Klinik Julia Likupang," vol. 6, hal. 6823–6830, 2025.

[12] I. D. Lubis, K. A. Khalil, R. Nurmalinda, dan N. I. A, "Artikel Pengabdian Masyarakat Edukasi Memahami Secara Umum Penyakit Hipertensi Dan Penyakit Infeksi Saluran Pernafasan Atas," vol. 6, no. 3, hal. 61–67, 2025.

[13] J. Ilmiah dan W. Pendidikan, "Gambaran Faktor-Faktor Yang Memengaruhi Kejadian Infeksi Saluran Pernapasan Akut (ISPA) Pada Balita Di Wilayah Kerja Puskesmas Sokaraja I Hima," vol. 11, no. September, hal. 51–58, 2025.

[14] A. I. Harahap, R. D. Priyatna, H. P. Figna, dan N. Rambe, "Aplikasi Cerdas Terintegrasi dalam Mendiagnosa Penyakit ISPA Pneumonia Pada Balita Menggunakan Algoritma Neural Network Backprogation di Kabupaten Langkat," *G-Tech J. Teknol. Terap.*, vol. 7, no. 4, hal. 1703–1712, 2023, doi: 10.33379/gtech.v7i4.3343.

[15] N. Mauliza, A. S. Iedwan, Y. Pristyanto, A. D. Hartanto, dan A. N. Rohman, "The Effect of Resampling Techniques on Model Performance Classification of Maternal Health Risks," *J. RESTI*, vol. 8, no. 4, hal. 496–505, 2024, doi: 10.29207/resti.v8i4.5934.

[16] A. Mukherjee *et al.*, "SMOTE-ENN resampling technique with Bayesian optimization for multi-class classification of dry bean varieties," *Appl. Soft Comput.*, vol. 181, no. June, hal. 113467, 2025, doi: 10.1016/j.asoc.2025.113467.

[17] R. Arisandi, "Perbandingan Model Klasifikasi Random Forest Dengan Resampling Dan Tanpa Resampling Pada Pasien Penderita Gagal Jantung," *J. Gaussian*, vol. 12, no. 1, hal. 136–145, 2023, doi: 10.14710/j.gauss.12.1.136-145.

[18] A. Zolanda, M. Raharjo, dan O. Setiani, "Faktor Risiko Kejadian Infeksi Saluran Pernafasan Akut Pada Balita Di Indonesia," *Link*, vol. 17, no. 1, hal. 73–80, 2021, doi: 10.31983/link.v17i1.6828.

[19] Z. B. Tadese *et al.*, "Interpretable prediction of acute respiratory infection disease among under- five children in Ethiopia using ensemble machine learning and Shapley additive explanations (SHAP)," 2024, doi: 10.1177/20552076241272739.

[20] P. R. Sihombing, S. Suryadiningrat, D. A. Sunarjo, dan Y. P. A. C. Yuda, "Identifikasi Data Outlier (Pencilan) dan Kenormalan Data Pada Data Univariat serta Alternatif Penyelesaiannya," *J. Ekon. Dan Stat. Indones.*, vol. 2, no. 3, hal. 307–316, 2023, doi: 10.11594/jesi.02.03.07.

[21] F. Septian, "Optimasi Klusterisasi pada Lama Tempo Pekerjaan Berbasis Gradient Boost Algorithm," *IJITECH Indones. J. Inf. Technol.*, vol. 2, no. 1, hal. 1–5, 2024, doi: 10.71155/vpny7m62.

[22] R. F. Ramadhan dan W. M. Ashari, "Performance Comparison of Random Forest and Decision Tree Algorithms for Anomaly Detection in Networks," *J. Appl. Informatics Comput.*, vol. 8, no. 2, hal. 367–375, 2024, doi: 10.30871/jaic.v8i2.8492.

[23] Fashihullisan, Dodi Vionanda, Yenni Kurniawati, dan Fadhilah Fitri, "Comparing Classification and Regression Tree and Logistic Regression Algorithms Using 5×2cv Combined F-Test on Diabetes Mellitus Dataset," *UNP J. Stat. Data Sci.*, vol. 1, no. 4, hal. 344–352, 2023, doi: 10.24036/ujsds/vol1-iss4/84.

[24] E. Virantika, K. Kusnawi, dan J. Ipmawati, "Evaluasi Hasil Pengujian Tingkat Clusterisasi Penerapan Metode K-Means Dalam Menentukan Tingkat Penyebaran Covid-19 di Indonesia," *J. Media Inform. Budidarma*, vol. 6, no. 3, hal. 1657, 2022, doi: 10.30865/mib.v6i3.4325.

[25] R. Ridwan, E. H. Hermaliani, dan M. Ernawati, "Penerapan: Penerapan Metode SMOTE Untuk Mengatasi Imbalanced Data Pada Klasifikasi Ujaran Kebencian," *Comput. Sci.*, vol. 4, no. 1, hal. 80–88, 2024, [Daring]. Tersedia pada: https://jurnal.bsi.ac.id/index.php/co-science/article/view/2990

[26] F. Gurcan dan A. Soylu, "Learning from Imbalanced Data: Integration of Advanced Resampling Techniques and Machine Learning Models for Enhanced Cancer Diagnosis and Prognosis," *Cancers (Basel).*, vol. 16, no. 19, 2024, doi:

10.3390/cancers16193417.

[27] A. H. Putra dan A. Salam, "A Comparative Performance of SMOTE, ADASYN and Random Oversampling in Machine Learning Models on Prostate Cancer Dataset," *J. Appl. Informatics Comput.*, vol. 9, no. 3, hal. 603–610, 2025, doi: 10.30871/jaic.v9i3.9308.

[28] Putri Ayu Firnanda, Litasya Shofwatillah, Fauziah Rahma, dan Fatkhurokhman Fauzi, "Analisis Perbandingan Decision Tree dan Random Forest dalam Klasifikasi Penjualan Produk pada Supermarket," *Emerg. Stat. Data Sci. J.*, vol. 3, no. 1, hal. 445–461, 2025, doi: 10.20885/esds.vol3.iss.1.art2.

[29] S. Riyanto, I. S. Sitanggang, T. Djatna, dan T. D. Atikah, "Comparative Analysis using Various Performance Metrics in Imbalanced Data for Multi-class Text Classification," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 6, hal. 1082–1090, 2023, doi: 10.14569/IJACSA.2023.01406116.

[30] E. R. Susanto, M. R. Inzaghi, A. Amarudin, dan N. Neneng, "Evaluasi Kinerja Model Random Forest Dalam Memprediksi Diabetes Berdasarkan Dataset Kesehatan di Indonesia," *J. Pendidik. dan Teknol. Indones.*, vol. 5, no. 7, hal. 1857–1866, 2025, doi: 10.52436/1.jpti.871.

[31] Rizky Fauzan, Anik Vega Vitianingsih, Dwi Cahyono, Anastasia Lidya Maukar, dan Yoyon Arie Budi Suprio, "Application of Classification Algorithms in Machine Learning for Phishing Detection - Penerapan Algoritma Klasifikasi pada Machine Learning untuk Deteksi Phishing," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 5, no. 2, hal. 531–540, 2025.

[32] A. R. Hanum *et al.*, "Mendeteksi Berita Hoaks Performance Analysis of the Bert Text Classification Algorithm," *J. Teknol. dan Ilmu Komput.*, vol. 11, no. 3, hal. 537–546, 2024, doi: 10.25126/jtiik2024118093.

[33] T. H. Pinem dan Z. P. Putra, "Evaluasi Kinerja Algoritma Klasifikasi Deep Learning dalam Prediksi Diabetes," *J. Ilm. FIFO*, vol. 17, no. 1, hal. 17, 2025, doi: 10.22441/fifo.2025.v17i1.003.

[34] J. Haviar Saviola dan N. Deny Hendrawan, "Implementasi Klasifikasi Kualitas Susu Menggunakan Algoritma Decision Tree, K-Nearest Neighbors Dan Naive Bayes," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 9, no. 5, hal. 8953–8960, 2025, doi: 10.36040/jati.v9i5.15260.

[35] D. Chicco dan G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," hal. 1–13, 2020.