Classification of Medical Complaints: Comparative Analysis of Machine Learning Algorithms with Determination of Dominant Factors Using Information Gain

Catherine Santoso Prasetya[1], I Gede Wiarta Sena[*2], Matthew Austen Fernando[3]

[1-2]Information System, Institut Informatika Indonesia

[3]Game Development Major, Dongseo University

E-mail: catherine@student.ikado.ac.id[1], dedek@ikado.ac.id[2], austenfernando25@gmail.com[3]

Abstract. This research compares three machine learning algorithms: Random Forest (RF), Decision Tree (DT), and K-Nearest Neighbors (KNN) for classifying illnesses influenced by climate, patient history, and clinical indicators. The dataset obtained from Kaggle contains 5,200 records combining meteorological and symptom data. Two pre-processing scenarios were tested to examine their impact on model performance: (1) normalization using Min-Max, and (2) normalization followed by balancing with the Synthetic Minority Over-sampling Technique (SMOTE). Results show that normalization significantly improves KNN's performance, increasing its accuracy from 0.324 on raw data to 0.968. In the first scenario, Random Forest achieved the highest accuracy of 0.985, followed by Decision Tree with 0.974 and KNN with 0.968. After applying SMOTE, Random Forest maintained stable accuracy at 0.985, while Decision Tree and KNN slightly decreased to 0.964. These findings indicate that Random Forest is the most robust and consistent algorithm for this classification task. Furthermore, the study reveals that SMOTE does not always enhance accuracy and must be applied selectively. Information gain analysis identifies symptom features as the strongest predictors. Overall, this research provides guidance in selecting the optimal algorithm and pre-processing strategy for building effective weather-related disease classification systems.

Keywords: Classification of Diseases, Decision Tree, K-Nearest Neighbors, Random Forest, SMOTE

1. Introduction

Disease classification systems play a crucial role in modern healthcare by supporting early diagnosis, treatment planning, and efficient healthcare management. Accurate disease identification often requires the integration of multiple risk factors, including patient demographics, environmental conditions, medical history, and clinical symptoms. Previous studies have shown that factors such as weather patterns, age, and comorbidities can significantly influence the occurrence and progression of various diseases, including respiratory and cardiovascular conditions. With the increasing volume and complexity of healthcare data, machine learning techniques have emerged as powerful tools for uncovering complex patterns and improving diagnostic accuracy [1][2][3].

Despite their potential, selecting the right machine learning algorithm for disease classification remains a challenging task. Different algorithms use different learning mechanisms, leading to varying performance even when applied to the same dataset. Tree-based models such as RF and DT construct hierarchical decision rules, while distance-based models like KNN classify instances based on similarity measures [4][5][6]. Consequently, algorithm performance is highly dependent on data characteristics and preprocessing strategies.

Data preprocessing techniques, particularly feature scaling and class imbalance management, are known to have a substantial impact on classification results. Normalization methods such as Min–Max scaling are commonly applied to ensure consistent feature ranges, which is crucial for distance-based models. Meanwhile, oversampling techniques such as the SMOTE are widely used to address class imbalance [7][8][9]. However, existing studies report inconsistent findings regarding the effectiveness of SMOTE, as synthetic data generation may not always improve model generalization and can sometimes introduce noise. Furthermore, the interaction between normalization and SMOTE, particularly the effect of the order in which they are applied, has not been thoroughly studied in the context of disease classification.

To address this gap, this study adopted a comparative experimental approach using a publicly available dataset of 5,200 patient records obtained from Kaggle. This dataset integrates demographic attributes (age and gender), meteorological factors (temperature, humidity, and wind speed), clinical symptoms, and medical history variables. Three widely used machine learning algorithms Random Forest, Decision Tree, and K-Nearest Neighbors were evaluated for their contrasting learning characteristics and broad applicability across various analytical domains [10][11][12]. However, their comparative performance in weather-related disease classification scenarios remains underexplored.

The primary objective of this study is to systematically evaluate the performance of RF, DT, and KNN under different preprocessing configurations, focusing on the role of normalization and SMOTE. In addition to algorithm comparison, this study aims to assess how preprocessing choices affect classification accuracy and class balance. Furthermore, a feature importance analysis using information gain was performed to identify the most influential feature groups. By addressing these objectives, this study seeks to provide empirically based guidance for selecting preprocessing strategies and machine learning models in developing robust disease classification systems [1][13][14].

2. Literature Review

Research related to data classification and class imbalance handling has been conducted extensively in various fields, ranging from health and finance to text analysis. This section discusses several relevant previous studies that form the basis for the design and development of models in this study.

Sediatmoko et al. [7], in their research named Sentiment Analysis of Customer Review Using Classification Algorithms and SMOTE for Handling Imbalanced Class, evaluate how the use of SMOTE enhances the performance of customer review text classification models on the Ralali.com platform. They evaluated three primary algorithms: Naïve Bayes, Support Vector Machine (SVM), and KNN. The findings indicated that SMOTE successfully enhanced recall and F-measure for minority classes while not greatly compromising performance in majority classes. However, in some cases, overall accuracy actually decreased slightly due to overfitting. This study is an important reference because it emphasizes that the use of SMOTE needs to be evaluated based on dataset characteristics, a principle that was also tested in this study.

Supangat et al. [1] in their study *Implementasi* Decision Tree C4.5 *Untuk Menentukan Status Berat Badan dan Kebutuhan Energi Pada Anak Usia* 7-12 *Tahun* applied the Decision Tree C4.5 algorithm to classify children's nutritional status based on age, weight, height, BMI, and BMR. With a dataset comprising 360 children aged 7 to 12 years, the findings indicated a significant accuracy in categorizing weight status into three groups: underweight, normal, and overweight. These findings demonstrate the superiority of tree-based models in handling non-linear multidimensional data, while also proving the interpretability of Decision Trees, which was the reason for choosing this algorithm as one of the comparison models in this study.

W Sena and Emanuel [4] in their study Mobile Legend Game Prediction Using Machine Learning Regression Method compared the performance of Artificial Neural Network (ANN) and RF in predicting Mobile Legends match results. Using a dataset of 852 match records, features such as average gold, average level, and first blood were the main determinants of prediction results. The findings indicated that ANN reached an accuracy of 82%, whereas RF attained 80%, suggesting that Random Forest exhibits strong stability despite not being the most accurate model. This research supports empirical evidence of RF's capability to manage intricate data and numerical factors, which underpins its choice for disease classification in this research.

Rahmawati et al. [15] in their work Decision Support System for Determining Final Project Supervisors Using Fuzzy and Simple Additive Weighting Based on Android: Case Study of IKADO Surabaya (*Sistem Pendukung Keputusan Penentuan Dosen Pembimbing Tugas Akhir Menggunakan* Fuzzy *dan* Simple Additive Weighting *Berbasis* Android: *Studi Kasus* IKADO Surabaya) designed a decision support system based on Fuzzy Logic and SAW to automatically determine final project advisors. This system assesses several criteria such as field of expertise, experience, and supervision load. Although the focus of the research is not on health, the weighting and ranking approach applied is relevant to the score-based classification process used in this study. This study highlights the importance of artificial intelligence-based systems to assist in complex decision-making processes.

Nanda et al. [10] studied the integration of SMOTE and Random Forest in bank credit risk assessment through their study Implementation of SMOTE to Improve the Performance of Random Forest Classification in Credit Risk Assessment in Banking. The dataset used was sourced from Kaggle with 27,591 transaction data, covering nine main variables such as loan amount, income, and credit history. The findings indicated an improvement in accuracy from 91.54% to 94.41% following the implementation of SMOTE. This study is one of the strongest empirical evidences that SMOTE can improve the performance of ensemble models such as Random Forest, especially on imbalanced data relevant to the RF performance analysis in this study.

Suresh et al. [16] in their publication A Hybrid Approach to Medical Decision-Making: Diagnosis of Heart Disease with Machine-Learning Model presented a blended method that integrates Random Forest and Support Vector Machine for diagnosing heart disease using various clinical factors. The experiment's results demonstrated an accuracy improvement of as much as 98.3% in contrast to employing just one model. This research validates the capability of ensemble algorithms like Random Forest in enhancing the consistency of predictive outcomes in the healthcare sector. These findings strengthen the theoretical basis for the use of RF in this study for the classification of diseases influenced by weather factors and clinical symptoms.

Barkah et al. [17] in their article Impact of Data Balancing and Feature Selection on Machine Learning-based Network Intrusion Detection emphasize the significance of feature selection and data balancing methods in enhancing the performance of medical classification. They compare the effects of normalization methods (Z-score and Min–Max) and data balancing (SMOTE and ADASYN) on RF, DT, and KNN algorithms. The results show that the combination of Min–Max normalization and Random Forest provides the best performance, while SMOTE does not always have a positive impact on accuracy. These findings are highly relevant to recent studies that also assess the effects of normalization and SMOTE on these three algorithms.

3. Methodology

This research intends to assess the effectiveness of three well-known machine learning algorithms, specifically RF, DT, and KNN, in categorizing illnesses affected by weather conditions, medical background, and clinical signs. The following is the research flow of two scenarios designed by the author:

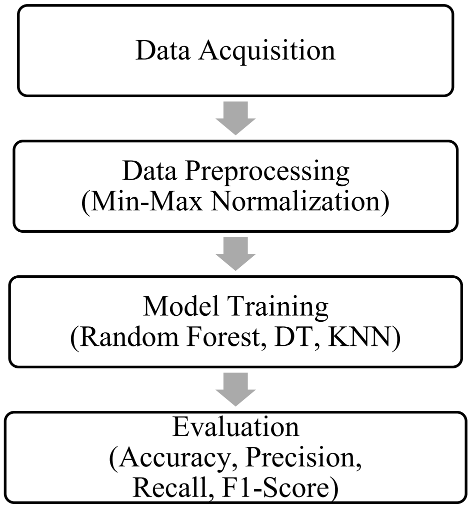

Figure 1. Research Flow Scenario 1 – Normalization

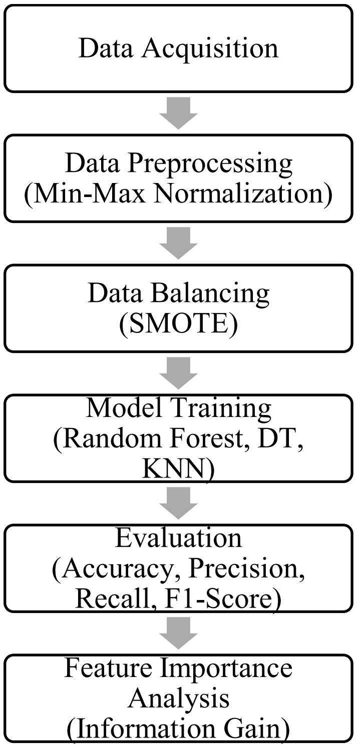

Figure 2. Research Flow Scenario 2 – Normalization, SMOTE, and Information Gain

The study was carried out using two situations to assess the effect of data pre-processing methods on model efficacy. In scenario 1 (Figure 1), the research process started with data collection, which was succeeded by pre-processing that solely included Min–Max normalization. The processed data was then used for model training and evaluation.

In scenario 2 (Figure 2), the research process was extended by incorporating a data balancing phase using SMOTE following the normalization step. Moreover, a stage for feature importance analysis utilizing information gain was included at the conclusion to pinpoint the most significant attributes in the model.

3.1. Dataset

The dataset used in this study was obtained from the Kaggle platform and consists of 5,200 patient medical records that integrate environmental conditions, demographic attributes, clinical symptoms, and medical history information [18]. To facilitate structured analysis and feature importance evaluation, all input variables were explicitly organized into three main feature groups based on their clinical and contextual relevance, as described below.

1. Environmental & Demographic Features

This feature group represents external and individual background factors that can influence the occurrence and severity of disease. These include patient demographic attributes and meteorological conditions at the time of diagnosis, namely age, gender, temperature, humidity, and wind speed. These variables capture environmental exposures and personal characteristics commonly associated with weather-related diseases.

2. Medical History Features

The medical history feature group contains pre-existing health conditions and chronic comorbidities that can increase susceptibility to certain diseases or worsen symptoms. This group includes a history of asthma, obesity, high cholesterol, HIV/AIDS, high blood pressure, diabetes, and nasal polyps. These features represent a patient's long-term health profile that can influence disease risk and progression.

3. Symptom Features

The symptom feature group consists of observable clinical manifestations self-reported by the patient during the consultation. This group includes 37 symptom variables, such as fever, cough, nausea, fatigue, shortness of breath, chest pain, and headache. These features reflect the patient's current clinical condition and are important indicators for disease classification.

The three characteristic groups and their corresponding variables are summarized in Table 1.

Table 1. Three Main Feature Groups of the Dataset

Category	Features		
Environmental & Demographic Features	Age	Temperature	Wind Speed
	Gender	Humidity	
Medical History Feature	Asthma History	High Cholesterol	Diabetes
	Obesity	HIV/AIDS	Nasal Polyps
	Asthma	High Blood Pressure	
Symptom Features	Nausea	Joint Pain	Abdominal Pain
	High Fever	Chills	Fatigue
	Runny Nose	Knee Ache	Dizziness
	Severe Headache	Chest Pain	Vomiting
	Cough	Shivering	Headache
	Weakness	Trouble Seeing	Fever
	Body Aches	Sore Throat	Sneezing
	Rapid Heart Rate	Rapid Breathing	Diarrhea
	Pain Behind Eyes	Swollen Glands	Rashes
	Shortness of Breath	Facial Pain	Sinus Headache
	Reduced Smell and Taste	Skin Irritation	Itchiness
	Throbbing Headache	Confusion	Back Pain
	Pain Behind the Eyes		

Each sample in the dataset is labeled with a target variable representing the associated disease class. For model evaluation, the dataset is randomly divided into 80% training data and 20% testing data. This data splitting strategy is consistent with previous research on machine learning-based medical classification and ensures reliable performance assessment [19].

3.2. Data Preprocessing

Preprocessing is executed to guarantee the uniformity, quality, and equilibrium of data prior to its utilization in model training. The primary phases consist of:

3.2.1. Data Cleaning

Duplicate data was removed to prevent it from affecting class distribution. A total of 209 duplicate data points were found out of a total of 428 duplicate rows. The initial dataset had 5,200 rows and 51 columns. After the duplicates were removed, the total remaining data was 4,981 rows, as shown in Table 2.

Table 2. Example of Duplicate Dataset After Cleaning

Age	Gender	Temperature	Humidity	Wind Speed	nausea	...	prognosis	count
6	0	37.056	0.83	6.369	0	...	Influenza	2
17	1	20.665	0.97	0.798	0	...	Migraine	2
32	0	9.680787	0.8775	14.917992	0	...	Migraine	2
47	0	18.723843	0.8375	9.760625	0	...	Migraine	2
100	1	13.615	0.93	0.145	0	...	Stroke	3

3.2.2. Normalization

Every numerical feature is scaled using the Min–Max normalization method to stay within the interval (0, 1). Examples of the dataset before and after normalization are presented in Table 3 and Table 4, respectively. This process is very important for distance-based algorithms such as KNN [20]. The normalization formula used is as follows:

$$\boldsymbol{X'} = \frac{\boldsymbol{X - X_{min}}}{\boldsymbol{X_{max} - X_{min}}} \tag{1}$$

where X′ represents the normalized value, X denotes the original feature value, and X_{min} and X_{max} indicate the minimum and maximum feature values, respectively.

Table 3. Example Dataset Before Normalization

Age	Gender	Temperature	Humidity	Wind Speed	nausea	joint_pain	...	knee_ache
18	1	13.356019	0.77	13.100704	0	0	...	0
57	1	40.241	0.73	3.702	0	0	...	0
72	0	39.37	0.91	15.619	0	0	...	0
76	0	28.936	0.97	19.81	0	0	...	0
12	0	24.713	0.87	12.667	0	0	...	0

Table 4. Example Dataset After Normalization

Age	Gender	Temperature	Humidity	Wind Speed	nausea	joint_pain	...	knee_ache
0.171717	1.0	0.507493	0.634437	0.418326	0.0	0.0	...	0.0
0.565657	1.0	0.986547	0.570861	0.117981	0.0	0.0	...	0.0
0.717172	0.0	0.971027	0.856954	0.4988	0.0	0.0	...	0.0
0.757576	0.0	0.785107	0.952318	0.632727	0.0	0.0	...	0.0
0.111111	0.0	0.709859	0.793377	0.404466	0.0	0.0	...	0.0

3.2.3. Data Balancing Using SMOTE

To tackle class imbalance, SMOTE is employed, generating synthetic data for minority classes based on the proximity between samples [21]. The number of data samples for each class before and after applying SMOTE is presented in Table 5.

The SMOTE technique involves generating new samples by calculating the distance between minority data and its k-nearest neighbors through the following formula:

$$\mathbf{X_{new}} = \mathbf{X_i} + \boldsymbol{\lambda}(\mathbf{X_{nn}} - \mathbf{X_i}) \tag{2}$$

where X_i represents minority data, X_{nn} denotes one of the k-nearest neighbors, and λ is a random value ranging from 0 to 1.

Table 5. Number of Datasets after Balancing using SMOTE

Label Encoding Prognosis	Initial Data Amount	Amount of Training Data (80%)	Amount of Data After SMOTE	Amount of Data After Under sampling Labels 4 and 8
Label 0 (*Arthritis*)	301	241	600	600
Label 1 (*Common Cold*)	309	247	600	600
Label 2 (Dengue)	308	246	600	600
Label 3 (*Eczema*)	311	249	600	600
Label 4 (*Heart Attack*)	968	774	774	600
Label 5 (*Heat Stroke*)	323	258	600	600
Label 6 (Influenza)	632	506	600	600
Label 7 (Malaria)	319	255	600	600
Label 8 (*Migraine*)	897	717	717	600
Label 9 (Sinusitis)	301	241	600	600
Label 10 (Stroke)	312	250	600	600

To ensure that this study not only yields comparative results but also actionable methodological guidance, an experimental design was developed to systematically evaluate preprocessing decisions under controlled conditions. Specifically, four preprocessing pipelines were constructed, varying the presence and order of normalization and SMOTE. Each pipeline was evaluated using multiple algorithms with different learning characteristics (tree-based and distance-based models) and assessed using accuracy-oriented and fairness-oriented metrics. This design allows the results to be interpreted as a decision-making guide for selecting a preprocessing strategy based on dataset characteristics, model type, and research objectives.

3.3. Classification Algorithms

3.3.1. Decision Tree (DT)

The Decision Tree algorithm constructs a classification model by partitioning the dataset according to attribute values that yield the greatest information gain [22]. Each node represents an attribute condition, while the leaf shows the class prediction result. The process of selecting the best attributes follows the Entropy and Information Gain criteria as follows:

$$\boldsymbol{Entropy(S) = -\sum_{i=1}^{n} p_i log^2(p_i)} \tag{3}$$

$$\boldsymbol{Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S|}{|Sv|} Entropy(Sv)} \tag{4}$$

3.3.2. Random Forest (RF)

Random Forest is an ensemble technique that merges multiple Decision Trees to enhance accuracy and stability [23]. Every tree is developed with a random selection of data and features (bagging), and the outcomes are subsequently merged through a majority voting process. Random Forest is known to be resistant to overfitting and capable of handling high-dimensional data.

3.3.3. K-Nearest Neighbors (KNN)

KNN is a non-parametric method that categorizes data by measuring the closeness between points. The distance utilized is Euclidean distance, expressed by the formula:

$$\mathbf{d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}} \tag{5}$$

The value of k is established through experimental observations. KNN is sensitive to data scaling, so normalization is a crucial step [24][25].

3.4. Evaluation Metrics

The evaluation of model performance is performed utilizing four primary metrics frequently employed in classification [26]:

- Accuracy

$$\mathbf{Accuracy = \frac{TP+TN}{TP+TN+FP+FN}} \tag{6}$$

- Precision

$$\mathbf{Precision = \frac{TP}{TP+FP}} \tag{7}$$

- Recall

$$\mathbf{Precision = \frac{TP}{TP+FN}} \tag{8}$$

- F1-Score

$$\mathbf{Precision = 2 \times \frac{Precision \times Recall}{Precision+Recall}} \tag{9}$$

where TP, TN, FP, and FN denote True Positive, True Negative, False Positive, and False Negative, respectively.

Additionally, an Information Gain analysis was conducted to determine the features that were most significant for classifying the disease. A higher gain value indicates that the attribute has a greater influence on class separation [27].

3.5. Experimental Setup

All experiments were performed using the Python 3.13 programming language, employing the scikit-learn library for implementing classification models, imbalanced-learn for SMOTE, and pandas–numpy for data handling. The data was divided 80:20 for training and testing. Each model was run 10 times (cross validation) to ensure the stability of the results.

The experiment was conducted in two main stages:

- Stage 1: The model undergoes testing exclusively on normalized data.
- Stage 2: The model undergoes testing on data that has been normalized and balanced employing SMOTE.

The evaluation of outcomes from these two phases establishes the foundation for determining how much data preprocessing influences the effectiveness of machine learning algorithms in disease classification.

4. Result

This section outlines the outcomes of a comparative experiment involving three machine learning algorithms: RF, DT, and KNN across four data preprocessing scenarios: (1) no preprocessing (v5(0)), (2) Min–Max normalization (v5(1)), (3) normalization plus SMOTE (v6), and (4) SMOTE plus normalization (v7).

The primary aim of this section is to assess how preprocessing methods and their order impact the effectiveness of models in categorizing diseases affected by atmospheric conditions, health background, and clinical indicators.

4.1. Performance on Raw Data

In the original situation without preprocessing (v5(0)), the outcomes reveal a significant disparity in performance among the algorithms, as shown in Table 6. The Random Forest model achieved the best accuracy at 0.987, with Decision Tree close behind at 0.976, whereas KNN managed only 0.324.

The poor performance of KNN is due to a lack of feature normalization. Because this algorithm relies on Euclidean distance to measure the proximity between data points, differences in scale between features, for example, between temperature (range -15–40) and symptom frequency (0–1), cause larger-scale features to dominate the distance calculation. This results in KNN failing to find relevant neighbors.

In contrast, RF and DT are relatively unaffected by feature scale because they rely on threshold splitting rather than distance calculations. These results are in line with the theory proposed by Supangat et al. [1] and W Sena & Emanuel [4], who argue that tree-based models are capable of handling raw data without complex preprocessing.

Table 6. Model Performance Results on Raw Data

	Accuracy	**Balanced Accuracy**	**Precision**	**Recall**	**F1-Score**
Random Forest	0.987	0.986	0.987	0.986	0.987
Decision Tree	0.976	0.971	0.974	0.971	0.972
KNN	0.324	0.217	0.210	0.217	0.208

4.2. Performance after Normalization

The application of Min–Max normalization has a significant impact, especially on the KNN algorithm, as presented in Table 7. After data normalization, KNN accuracy increased dramatically from 0.324 to 0.968, an increase of approximately +199% compared to the raw data. Enhancements were also observed in the macro recall and F1-score metrics, which attained 0.967 and 0.970, respectively.

The Random Forest and Decision Tree models experienced only minor changes (≤0.01) because both algorithms are independent of feature scaling. This confirms the statement by W Sena & Emanuel [4] that feature scaling is a prerequisite for distance-based models, but not necessarily essential for tree-based models.

Table 7. Model Performance Results on Normalized Data

	Accuracy	Balanced Accuracy	Precision	Recall	F1-Score
Random Forest	0.985	0.985	0.985	0.985	0.985
Decision Tree	0.974	0.970	0.971	0.970	0.970
KNN	0.968	0.967	0.973	0.967	0.970

Normalization has been proven to optimize KNN performance, approaching the performance of RF and DT. This shows that after feature scaling, all three algorithms perform relatively well, with RF remaining the most stable and efficient model.

4.3. Effect of SMOTE after Normalization

Scenario v6 (Normalization → SMOTE) was applied to balance class distribution without changing feature scale. The results showed that applying SMOTE improved model fairness (as indicated by Balanced Accuracy) without a significant decrease in accuracy, as presented in Table 8.

In the Random Forest model, accuracy remained high (0.985), but Balanced Accuracy increased from 0.985 to 0.987, indicating an improvement in the ability to recognize minor classes. The KNN model experienced a slight decrease in accuracy (0.968 → 0.964), but Balanced Accuracy increased from 0.967 to 0.971, indicating an improvement in the detection of rare cases.

In comparison, the Decision Tree model showed a drop in accuracy across every metric (from 0.974 to 0.964). This phenomenon is consistent with the research of Sediatmoko et al. [7] and Nanda et al. [10], which explains that synthetic data generated from SMOTE can expand decision boundaries unreasonably, thereby reducing the generalization ability of DT.

Table 8. Model Performance Results in Normalization Scenario → SMOTE

	Accuracy	Balanced Accuracy	Precision	Recall	F1-Score
Random Forest	**0.985**	**0.987**	0.983	0.987	0.985
Decision Tree	0.964	0.962	0.958	0.962	0.960
KNN	0.964	0.971	0.962	0.971	0.965

These results show that the combination of normalization before SMOTE (pipeline v6) is the most ideal configuration for achieving a balance between accuracy and performance equality across all classes. Therefore, this pipeline was selected as the main model for further feature analysis.

4.4. Effect of SMOTE before Normalization

Scenario v7 (SMOTE → Normalization) was conducted to evaluate the effect of changes in the preprocessing order. Unlike v6, the results of v7 show a decrease in performance across all models and metrics, as presented in Table 9.

The accuracy of Random Forest dropped from 0.985 to 0.977, whereas the accuracies for Decision Tree and KNN fell to 0.951 and 0.946, respectively. This decline occurred because SMOTE calculated the distance between data points before the features were on a uniform scale, resulting in synthetic points that did not represent the actual distribution. This caused the model to learn from biased distance patterns. The confusion matrices for Random Forest, Decision Tree, and KNN are presented in Figures 3, 4, and 5, respectively.

This decline in performance was also confirmed by research by Nanda et al. [10], which emphasized the importance of applying SMOTE in standard feature space.

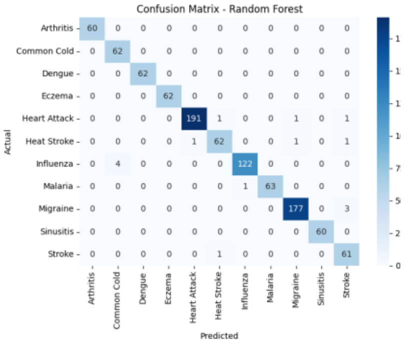

Figure 3. Confusion Matrix Random Forest

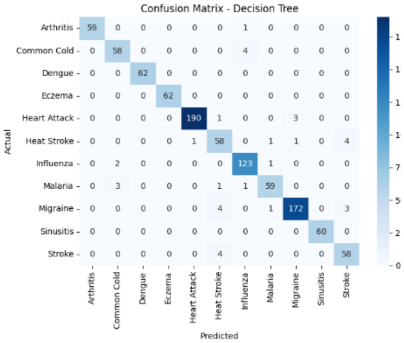

Figure 4. Confusion Matrix Decision Tree

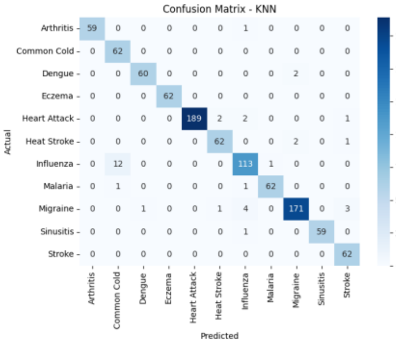

Figure 5. Confusion Matrix KNN

Table 9. Model Performance Results in the SMOTE Scenario.

	Accuracy	Balanced Accuracy	Precision	Recall	F1-Score
Random Forest	0.977	0.982	0.970	0.982	0.975
Decision Tree	0.951	0.957	0.939	0.957	0.947
KNN	0.946	0.959	0.934	0.959	0.944

Thus, the sequence of Min–Max Normalization → SMOTE proved to be much more effective than the SMOTE method applied before normalization. Sorting errors cause imbalances in distance representation and reduce the quality of oversampling data.

Although SMOTE is widely used to address class imbalance, the experimental results in this study indicate that its application should be selective, not universal. The effectiveness of SMOTE depends on several conditions related to data characteristics, preprocessing sequence, and learning algorithm.

First, SMOTE is only beneficial when class imbalance significantly impacts minority class recognition, as reflected by low recall or balanced accuracy. In this study, the normalized dataset without SMOTE (v5(1)) achieved high overall accuracy across all models. However, applying SMOTE after normalization (v6) resulted in consistent improvements in Balanced Accuracy, particularly for Random Forest (0.985 → 0.987) and KNN (0.967 → 0.971). This suggests that SMOTE is most appropriate when the research objective emphasizes fair classification across all classes, rather than simply maximizing overall accuracy.

Second, SMOTE should only be applied after feature scaling when distance-based algorithms or distance-dependent oversampling methods are used. As shown in scenario v7 (SMOTE → Normalization), applying SMOTE to an unnormalized feature space degrades performance across all models. Because SMOTE generates synthetic samples based on Euclidean distance, the unequal feature scales lead to distorted neighborhood structure and unrealistic synthetic examples. This confirms that SMOTE is not suitable when feature distributions are not standardized.

Third, the results show that SMOTE is model-dependent. Tree-based models respond differently to synthetic data. While Random Forest benefits from SMOTE due to its ensemble nature and robustness to noise, Decision Tree performance deteriorates across all metrics in scenario v6. This supports previous findings that SMOTE can overextend the decision boundary for single-tree models, increasing overfitting and reducing generalization. Therefore, SMOTE should be avoided or tuned carefully when using unstable learners such as standalone Decision Trees.

Finally, SMOTE should not be applied when baseline performance already demonstrates high accuracy and balanced class recognition, as unnecessary oversampling can introduce noise without significant performance gains. In such cases, normalization alone (v5(1)) is sufficient to achieve near-optimal results, especially for KNN.

4.5. Feature Information Gain Analysis

In the optimal model (Random Forest in v6), the Mutual Information method was utilized for feature importance analysis. The findings indicate that the Clinical Symptoms feature group has the greatest impact on information contribution, with Environment & Demographics coming next, and Medical History last, as summarized in Table 10.

Table 10. Information Gain Feature Analysis

Feature Group	Number of Features	Total Information Gain	Relative Contribution
Clinical Symptoms	37	5.32	67.34%
Environment & Demographics	5	1.94	24.56%
Medical History	8	0.64	8.1%

These findings indicate that patients' current clinical symptoms are the most informative indicator in determining disease classification, while medical history has relatively little influence. These results support the general pattern of symptom-based medical diagnostic systems [16].

4.6. Summary of Performance Comparison

Overall, the results of the four scenarios show the following consistent patterns, as summarized in Table 11:

Table 11. Performance Results Comparison

Analysis Aspects	Key Findings
1. Effect of Normalization	Normalization significantly improves KNN performance; RF and DT remain relatively stable.
2. Effect of SMOTE	SMOTE improves fairness (Balanced Accuracy), especially for RF and KNN.
3. Preprocessing Order	Normalization → SMOTE (v6) is superior to SMOTE → Normalization (v7).
4. Best Model (Overall)	Random Forest (v6): Accuracy 0.985, Balanced Accuracy 0.987, F1-score 0.985.
5. Information Gain	The Clinical Symptoms feature group provides the highest information contribution (5.32 or 67.34%), followed by Environment & Demographics (1.94 or 24.56%), and Medical History (0.64 or 8.1%).

Pipeline v6 has been proven to deliver the most optimal and balanced results. Applying Min–Max normalization prior to using SMOTE enhances the model's capacity to identify minor classes while maintaining overall accuracy.Based on the comparative evaluation across preprocessing pipelines and learning algorithms, this study provides practical guidance for applying normalization and SMOTE in medical classification tasks.

Normalization should be treated as a mandatory preprocessing step for distance-based algorithms such as KNN, as the absence of feature scaling leads to severely distorted distance calculations and substantial performance degradation, while normalization alone is sufficient to raise KNN performance to a level comparable with tree-based models. In contrast, SMOTE should be applied selectively and only after feature scaling, as its effectiveness depends on the presence of class imbalance that negatively affects minority class recall and on a normalized feature space that preserves meaningful distance relationships; applying SMOTE prior to normalization consistently results in performance deterioration and should therefore be avoided. Furthermore, the suitability of SMOTE is strongly model-dependent: ensemble-based models such as Random Forest benefit from oversampling due to their robustness to synthetic noise, whereas single Decision Tree models are more sensitive to oversampling and may suffer from reduced generalization. Finally, when baseline model performance already demonstrates high accuracy and balanced recall, additional oversampling is often unnecessary and may introduce noise without yielding meaningful gains, making normalization alone the most efficient and reliable preprocessing strategy.

Although the observed performance differences between the preprocessing scenarios and algorithms only ranged from approximately 0.4% to 1%, this variation remains significant in the context of medical classification and imbalanced datasets. In applied machine learning, even small improvements in accuracy can result in a significant number of correctly classified instances; for example, a 0.5% accuracy improvement in a dataset of 5,200 samples represents approximately 26 cases, which is nontrivial in healthcare applications where misclassification can have clinical consequences. Furthermore, overall accuracy alone is insufficient to evaluate imbalanced data, which is why this study emphasizes balanced accuracy, macro recall, and macro F1 score as complementary metrics. The small changes in accuracy observed were consistently accompanied by improvements in balanced accuracy and recall, suggesting improved minority class recognition rather than random fluctuations. Previous studies have shown that consistent directional improvements across multiple evaluation metrics, even if numerically modest, can be analytically significant and have practical implications, particularly in high-stakes areas such as medical decision support systems [29]. Therefore, these small performance differences are sufficient to provide information for comparative analysis and justify conclusions regarding the relative effectiveness of pre-processing strategies.

5. Discussion

The findings of this research present several key insights into the efficacy of algorithms and data preprocessing for disease classification. First, Min–Max normalization proved crucial for distance-based algorithms such as KNN. Without normalization, large-scale feature values dominate the Euclidean distance calculation, causing a bias that drastically reduces accuracy (from 0.324 to 0.968). This is in line with the findings of W Sena & Emanuel [4], which emphasize the importance of feature scaling in distance-based models.

Second, Random Forest showed the most stable and consistent performance across all scenarios. This stability is due to its ensemble nature and bagging mechanism, which reduces the variance of individual models [16]. RF is less affected by distribution changes due to SMOTE, making it the best model for complex and heterogeneous health data.

Third, the application of SMOTE after normalization (v6) provides the best balance between accuracy and balanced accuracy, indicating that the preprocessing sequence has a significant methodological impact. When SMOTE is applied before normalization (v7), the performance of all models decreases because the distance between features is calculated on an uneven scale. This phenomenon is in line with the theory of Chawla et al. [28] and the results of the study by Sediatmoko et al. [7], which show that SMOTE is sensitive to feature scale.

Fourth, the information gain results reinforce the clinical relevance of this study. The Symptoms feature had the highest information contribution (5.32), followed by Environment & Demographics (1.94), and finally Medical History (0.64). This pattern indicates that weather-based disease prediction is more influenced by current clinical symptoms than past medical conditions, consistent with the symptom-driven classification approach in the study by Suresh et al. [16].

Overall, Random Forest in the v6 scenario (Normalization → SMOTE) can be considered the most balanced model for weather-related disease classification, maintaining high accuracy (0.985) while improving fairness (Balanced Accuracy 0.987). However, if the primary goal is efficiency and pure accuracy without a focus on class balance, then the v5(1) pipeline (normalization only) remains a practical choice with nearly equivalent results.

6. Conclusion

This research indicates that the choice of algorithms and the order of data preprocessing have a substantial impact on the classification outcomes of diseases affected by weather, medical history, and clinical symptoms. Of the four testing scenarios, the combination of Min–Max normalization followed by SMOTE (v6) proved to provide the most optimal and balanced performance, with Random Forest emerging as the best model due to its stability and ability to maintain high accuracy (0.985) and balanced accuracy (0.987) in various data conditions. Conversely, Decision Tree tended to experience a decline in performance on oversampled data, while KNN showed a dramatic improvement after normalization, emphasizing the importance of feature scaling for distance-based models. Information gain analysis showed that the Symptoms feature group collectively provided the highest information contribution (5.32), far surpassing the Environment & Demographics (1.94) and Medical History (0.64) groups. This indicates that clinical symptoms are the strongest predictors in this classification model. This study offers empirical and practical insights for developing effective disease classification systems and highlights the significance of designing appropriate preprocessing pipelines in applying machine learning within the healthcare sector.

7. References

[1] P. Khan et al., “Machine Learning and Deep Learning Approaches for Brain Disease Diagnosis: Principles and Recent Advances,” *IEEE Access*, vol. 9, pp. 37622–37655, 2021, doi: 10.1109/ACCESS.2021.3062484.

[2] M. M. Ahsan, S. A. Luna, and Z. Siddique, “Machine-Learning-Based Disease Diagnosis: A Comprehensive Review,” *Healthcare*, vol. 10, no. 3, p. 541, Mar. 2022, doi: 10.3390/healthcare10030541.

[3] S. Asif et al., “Advancements and Prospects of Machine Learning in Medical Diagnostics: Unveiling the Future of Diagnostic Precision,” *Arch. Comput. Methods Eng.*, vol. 32, no. 2, pp. 853–883, Mar. 2025, doi: 10.1007/s11831-024-10148-w.

[4] I. G. W. Sena and A. W. R. Emanuel, “MOBILE LEGEND GAME PREDICTION USING MACHINE LEARNING REGRESSION METHOD,” *JURTEKSI (Jurnal Teknologi dan Sistem Informasi)*, vol. 9, no. 2, pp. 221–230, Mar. 2023, doi: 10.33330/jurteksi.v9i2.1866.

[5] Z. Azam, M. M. Islam, and M. N. Huda, “Comparative Analysis of Intrusion Detection Systems and Machine Learning-Based Model Analysis Through Decision Tree,” *IEEE Access*, vol. 11, pp. 80348–80391, 2023, doi: 10.1109/ACCESS.2023.3296444.

[6] Z. Azam, Md. M. Islam, and M. N. Huda, “Comparative Analysis of Intrusion Detection Systems and Machine Learning-Based Model Analysis Through Decision Tree,” *IEEE Access*, vol. 11, pp. 80348–80391, 2023, doi: 10.1109/ACCESS.2023.3296444.

[7] N. S. Sediatmoko, Y. Nataliani, and I. Suryady, “Sentiment Analysis of Customer Review Using Classification Algorithms and SMOTE for Handling Imbalanced Class,” *Indonesian Journal of Information Systems*, vol. 7, no. 1, pp. 38–52, Aug. 2024, doi: 10.24002/ijis.v7i1.8879.

[8] A. Saad Hussein, T. Li, C. W. Yohannese, and K. Bashir, “A-SMOTE: A New Preprocessing Approach for Highly Imbalanced Datasets by Improving SMOTE,” *Int. J. Comput. Intell. Syst.*, vol. 12, no. 2, p. 1412, 2019, doi: 10.2991/ijcis.d.191114.002.

[9] T. Wongvorachan, S. He, and O. Bulut, “A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining,” *Information*, vol. 14, no. 1, p. 54, Jan. 2023, doi: 10.3390/info14010054.

[10] N. N. A. Nanda, Y. Farida, and W. D. Utami, “Implementation of SMOTE to Improve the Performance of Random Forest Classification in Credit Risk Assessment in Banking,” *INTENSIF: Jurnal Ilmiah Penelitian dan Penerapan Teknologi Sistem Informasi*, vol. 9, no. 2, pp. 158–177, Jul. 2025, doi: 10.29407/intensif.v9i2.23930.

[11] A. I. Marqués, V. García, and J. S. Sánchez, “A literature review on the application of evolutionary computing to credit scoring,” *J. Oper. Res. Soc.*, vol. 64, no. 9, pp. 1384–1399, Sep. 2013, doi: 10.1057/jors.2012.145.

[12] S. Bhatore, L. Mohan, and Y. R. Reddy, “Machine learning techniques for credit risk evaluation: a systematic literature review,” *Journal of Banking and Financial Technology*, vol. 4, no. 1, pp. 111–138, Apr. 2020, doi: 10.1007/s42786-020-00020-3.

[13] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O’Sullivan, “A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction,” *Frontiers in Bioinformatics*, vol. 2, Jun. 2022, doi: 10.3389/fbinf.2022.927312.

[14] M. M. Ahsan, S. A. Luna, and Z. Siddique, “Machine-Learning-Based Disease Diagnosis: A Comprehensive Review,” *Healthcare*, vol. 10, no. 3, p. 541, Mar. 2022, doi: 10.3390/healthcare10030541.

[15] T. Rahmawati, Alexander Wirapraja, and E. C. Soesilo, “Sistem Pendukung Keputusan Penentuan Dosen Pembimbing Tugas Akhir Menggunakan Fuzzy Dan Simple Additive Weighting Berbasis Android: Studi Kasus IKADO Surabaya,” *KONSTELASI: Konvergensi Teknologi dan Sistem Informasi*, vol. 2, no. 1, Apr. 2022, doi: 10.24002/konstelasi.v2i1.5632.

[16] T. Suresh, T. A. Assegie, S. Rajkumar, and N. Komal Kumar, “A hybrid approach to medical decision-making: diagnosis of heart disease with machine-learning model,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 2, p. 1831, Apr. 2022, doi: 10.11591/ijece.v12i2.pp1831-1838.

[17] A. S. Barkah, S. R. Selamat, Z. Z. Abidin, and R. Wahyudi, “Impact of Data Balancing and Feature Selection on Machine Learning-based Network Intrusion Detection,” *JOIV : International Journal on Informatics Visualization*, vol. 7, no. 1, p. 241, Feb. 2023, doi: 10.30630/joiv.7.1.1041.

[18] A. Shan, I. Amir, and M. Kamal, “Weather-related Disease Prediction Dataset,” May 2024, *Zenodo*. doi: 10.5281/zenodo.11366485.

[19] Q. H. Nguyen et al., “Influence of Data Splitting on Performance of Machine Learning Models in Prediction of Shear Strength of Soil,” *Math. Probl. Eng.*, vol. 2021, pp. 1–15, Feb. 2021, doi: 10.1155/2021/4832864.

[20] Y. Dimas Pratama and A. Salam, “Comparison of Data Normalization Techniques on KNN Classification Performance for Pima Indians Diabetes Dataset,” *Journal of Applied Informatics and Computing*, vol. 9, no. 3, pp. 693–706, Jun. 2025, doi: 10.30871/jaic.v9i3.9353.

[21] P. Soltanzadeh and M. Hashemzadeh, “RCSMOTE: Range-Controlled synthetic minority over-sampling technique for handling the class imbalance problem,” *Inf. Sci. (Ny)*., vol. 542, pp. 92–111, Jan. 2021, doi: 10.1016/j.ins.2020.07.014.

[22] M. B. Al Snousy, H. M. El-Deeb, K. Badran, and I. A. Al Khlil, “Suite of decision tree-based classification algorithms on cancer gene expression data,” *Egypt. Informatics J.*, vol. 12, no. 2, pp. 73–82, Jul. 2011, doi: 10.1016/j.eij.2011.04.003.

[23] H. A. Salman, A. Kalakech, and A. Steiti, “Random Forest Algorithm Overview,” *Babylonian J. Mach. Learn.*, vol. 2024, pp. 69–79, Jun. 2024, doi: 10.58496/BJML/2024/007.

[24] A. Pandey and A. Jain, “Comparative Analysis of KNN Algorithm using Various Normalization Techniques,” *Int. J. Comput. Netw. Inf. Secur.*, vol. 9, no. 11, pp. 36–42, Nov. 2017, doi: 10.5815/ijcnis.2017.11.04.

[25] I. Handayani, “Application of K-Nearest Neighbor Algorithm on Classification of Disk Hernia and Spondylolisthesis in Vertebral Column,” *Indonesian Journal of Information Systems*, vol. 2, no. 1, pp. 57–66, Aug. 2019, doi: 10.24002/ijis.v2i1.2352.

[26] H. M and S. M.N, “A Review on Evaluation Metrics for Data Classification Evaluations,” *Int. J. Data Min. Knowl. Manag. Process*, vol. 5, no. 2, pp. 01–11, Mar. 2015, doi: 10.5121/ijdkp.2015.5201.

[27] F. Gong, L. Jiang, H. Zhang, D. Wang, and X. Guo, “Gain ratio weighted inverted specific-class distance measure for nominal attributes,” *Int. J. Mach. Learn. Cybern.*, vol. 11, no. 10, pp. 2237–2246, Oct. 2020, doi: 10.1007/s13042-020-01112-8.

[28] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[29] O. Graham and P. Henderson, “Advancing Explainable Artificial Intelligence for Clinical Decision Support: Techniques, Challenges, and Evaluation Frameworks in High-Stakes Medical Environments,” May 28, 2025. doi: 10.20944/preprints202505.2281.v1.