Journal of Future Artificial Intelligence and Technologies

E-ISSN: 3048-3719

Research Article

Analyzing Preprocessing Impact on Machine Learning Classifiers for Cryotherapy and Immunotherapy Dataset

De Rosal Ignatius Moses Setiadi [1],*, Hussain Md Mehedul Islam [2], Gustina Alfa Trisnapradika [1], and Wise Herowati [1]

[1] Faculty of Computer Science, Dian Nuswantoro University, Semarang, Indonesia;
e-mail : moses@dsn.dinus.ac.id; gustina.alfa@dsn.dinus.ac.id; wise@dsn.dinus.ac.id
[2] Software Engineer, The Mathworks, Inc., United States; e-mail: mehadi.cuet@gmail.com
* Corresponding Author : De Rosal Ignatius Moses Setiadi

Abstract: In the clinical treatment of skin diseases and cancer, cryotherapy and immunotherapy offer effective and minimally invasive alternatives. However, the complexity of patient response demands more sophisticated analytical strategies for accurate outcome prediction. This research focuses on analyzing the effect of preprocessing in various machine learning models on the prediction performance of cryotherapy and immunotherapy. The preprocessing techniques analyzed include advanced feature engineering, Synthetic Minority Over-sampling Technique (SMOTE), and Tomek links as resampling techniques and their combination. Various classifiers, including support vector machine (SVM), Naive Bayes (NB), Decision Tree (DT), Random Forest (RF), XGBoost, and Bidirectional Gated Recurrent Unit (BiGRU), were tested. The findings of this study show that preprocessing methods can significantly improve model performance, especially in the XGBoost model. Random Forest also gets the same results as XGBoost, but it can work better without significant preprocessing. The best results were 0.8889, 0.8889, 0.6000, 0.9037, and 0.8790, respectively, for accuracy, recall, specificity, precision, and f1 on the Immunotherapy dataset, while on the Cryotherapy dataset, respectively, they were 0.8889, 0.8889, 0.6000, 0.9037, and 0.8790. This study confirms the potential of customized preprocessing and machine learning models to provide deep insights into treatment dynamics, ultimately improving the quality of diagnosis.

Keywords: BiGRU classification; Feature engineering; Small dataset; SMOTE; Tomek links.

Received: April, 20th 2024
Revised: May, 28th 2024
Accepted: May, 29th 2024
Published: June, 1st 2024
Curr. Ver.: July, 3rd 2024

1. Introduction

In recent years, machine learning (ML) has revolutionized various fields, including healthcare, by providing powerful data analysis and prediction tools. Its ability to process and learn from large amounts of data has made ML an invaluable asset in clinical data analysis and prediction, thereby improving the accuracy of diagnostic and treatment results. To increase the effectiveness of clinical treatment, especially in the context of treating cancer and skin diseases, immunotherapy and cryotherapy have emerged as promising methods because their effectiveness can be tailored to the needs of each patient.

Immunotherapy, which harnesses the patient's immune system to fight disease, and cryotherapy, which relies on applying extremely low temperatures to destroy abnormal tissue, offer less invasive approaches than conventional methods such as chemotherapy and radiotherapy[1]–[4]. However, the complexity of patient response to these two therapies indicates the need for a more granular approach to predicting treatment outcomes, considering individual characteristics and associated clinical variables. In this context, ML offers significant potential to improve the classification or prediction of treatment success, facilitating more accurate personalization of treatment[5]. This classification can not only improve our understanding of treatment dynamics but also optimize clinical decision-making and treatment strategies, ultimately improving patient health outcomes.

DOI : 10.62411/faith.2024-2

Each ML method for this task has its strengths and. The Support Vector Machine (SVM) is known for its effectiveness in handling high dimensional data and generating highly accurate models. However, it is often less efficient on huge datasets and requires careful parameter tuning[6]. On the other hand, Naive Bayes (NB) is valued for its simplicity and rapid implementation[7], but its assumption of feature independence may not hold in complex clinical scenarios, potentially compromising prediction accuracy. Decision trees (DT) allow easy interpretation and are capable of handling non-linear data, but they are susceptible to overfitting, especially in very deep trees. To overcome this weakness, Random Forest (RF) aggregates many decision trees to improve stability and accuracy[8], [9], but in turn, these models can become very complex and require more computing resources. XGBoost further refines the ensemble learning approach with more efficient optimization and better overfitting management[10], but still requires careful parameter tuning to achieve optimal performance.

Deep learning methods based on a recurrent neural network (RNN), such as the bidirectional gated recurrent unit (BiGRU), can also be used in this case. BiGRU stands out for its ability to manage long-term dependencies in data[11], [12]. BiGRU is more efficient than other recurrence models, such as long-sort-term memory (LSTM), because it requires less computation but still requires a large dataset for training to produce a robust model. Thus, selecting an appropriate ML algorithm should be based on specific data characteristics, desired model complexity, and available computational resources to produce accurate and efficient predictions in the context of clinical settings.

The characteristics of medical datasets often reflect unique challenges in developing classification models. Medical datasets typically feature a limited number of samples and exhibit imbalance[13]. Additionally, certain datasets may be constrained by a limited number of features, which can introduce challenges such as outliers or extreme values diverging from typical patterns. These anomalies can hinder the effective training of models[14]. The Cryotherapy and Immunotherapy dataset[15]–[18] is no exception. Dataset preprocessing is crucial to determining model performance. Feature engineering is a method for creating new features from existing features. This helps reveal important aspects of the data that may not be immediately obvious but are highly relevant for predictions[19], [20].

Resampling techniques such as SMOTE-Tomek can be used to overcome the imbalance problem in the dataset. The Synthetic Minority Over-sampling Technique (SMOTE) functions by creating synthetic samples from minority classes to balance the class distribution[21]–[23]. Meanwhile, Tomek links identify and delete pairs of samples from different classes close to each other. Combining SMOTE and Tomek links creates an oversampling and undersampling process to form a balanced dataset that does not overlap between classes. Based on the literature above, the objectives of this research are:

1. Analyze various machine learning and BiGRU methods to classify Cryotherapy and Immunotherapy Datasets.
2. Analyze how various preprocessing techniques, such as feature engineering and SMOTE-Tomek, can affect the performance of different classification methods in small and imbalanced medical datasets.
3. Determine the most effective combination of preprocessing strategies and machine learning models to increase the accuracy and robustness of predictions and provide more accurate and useful insights in clinical decision-making.

Furthermore, this research aspires to guide the development of more adaptive clinical systems that are responsive to individual treatment needs. The remainder of this paper will delve deeper into the details of proposed method and the outcomes obtained, as well as a discussion of the implications of the research findings for current clinical practice.

2. Related Works

Recent advancements in machine learning have shown promising results in medical datasets classification, particularly with Cryotherapy and Immunotherapy datasets. One notable study [3],uses the J48 decision tree by introducing attributes constructed through genetic programming to classify immunotherapy and cryotherapy datasets. The construction of new attributes through genetic programming to expand the information space shows a significant improvement in J48 classification accuracy, from about 14% for the immunotherapy dataset to 5% for the cryotherapy dataset. The conclusions of this study underscore the importance of new attributes in improving classification accuracy while recommending further use of this method to improve the efficiency of classification models in medical applications.

Furthermore, a comprehensive analysis conducted by another research [2] evaluated various machine learning methods, such as NB, RF, SVM, K-Nearest Neighbors (KNN), and Artificial Neural Network (ANN) on cryotherapy and immunotherapy datasets. This research also tested and analyzed several optimization techniques, such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and oversampling. Random Forest showed the best performance with an accuracy of 0.95, a sensitivity of 0.88, and a specificity of 0.98 on the cryotherapy dataset. Meanwhile, the immunotherapy dataset resulted in accuracy, sensitivity, and specificity of 0.86, 0.87, and 0.84, respectively.

Research [24] compares KNN, NB, and DT methods for classifying immunotherapy datasets. The KNN method is implemented by calculating the Euclidean distance between data points, NB based on conditional probability, and DT using entropy and gain measurements to build a decision tree. The results of this study show that NB provides the highest classification accuracy, namely 0.81, followed by DT 0.8, and KNN 0.74.

Research [4] discusses an optimized machine learning approach by adding synthetic data for various health datasets such as cancer, heart disease, diabetes, cryotherapy, and immunotherapy. Synthetic data is generated using the Generative Adversarial Network (GAN) method and RF as a classifier. The results show that the improved data enables the use of visual learning as a new approach in data analysis, offering a beneficial synergy between good quality data and optimal classification performance. Focusing more on observing immunotherapy and cryotherapy datasets, the use of synthetic data can increase classifier performance by around 5% to 16% for accuracy, recall, precision, specificity, and f1.

Despite these advancements, the literature reveals a gap in systematically exploring the combined impact of preprocessing techniques, such as advanced feature engineering and resampling methods like SMOTE-Tomek, on classification outcomes. This research aims to bridge that gap, providing a nuanced understanding of preprocessing's role in enhancing classifier performance in the context of small and imbalanced medical datasets. 3. Proposed Method

3. Proposed Method

This study proposes a comprehensive methodology designed to investigate the impact of various preprocessing techniques on the performance of machine learning classifiers in the context of Cryotherapy and Immunotherapy datasets. Given the challenges associated with small and imbalanced medical datasets, such as limited samples and the presence of outliers, our approach emphasizes advanced feature engineering, the SMOTE, and Tomek links for data preprocessing. The ultimate goal is to identify the most effective combination of preprocessing strategies and machine learning models to enhance prediction accuracy and robustness, thereby offering more precise insights for clinical decision-making. Transitioning from the theoretical underpinnings to practical application, this research proposes to perform performance analysis of methods such as XGBoost, SVM, NB, DT, RF, and BiGRU to classify Cryotherapy and Immunotherapy datasets. This shift from a broad methodological framework to the specific analysis of machine learning models underscores our commitment to not only addressing the inherent challenges of medical dataset analysis but also to exploring the potential of these models to yield actionable insights. Figure 1 illustrates the stages of the proposed method, visually guiding the reader through the sequential steps of our research process, from data preprocessing to model evaluation.

3.1. Dataset Collection

This study utilizes cryotherapy [18]and immunotherapy [17] datasets, both of which are prominent in the context of wart treatment. Each dataset classifies treatment outcomes into two categories: failed or successful. Additionally, both datasets share six common features, with the immunotherapy dataset containing one unique feature. Each dataset comprises 90 records. A detailed description of the features shared by and unique to each dataset is provided in Table 1.

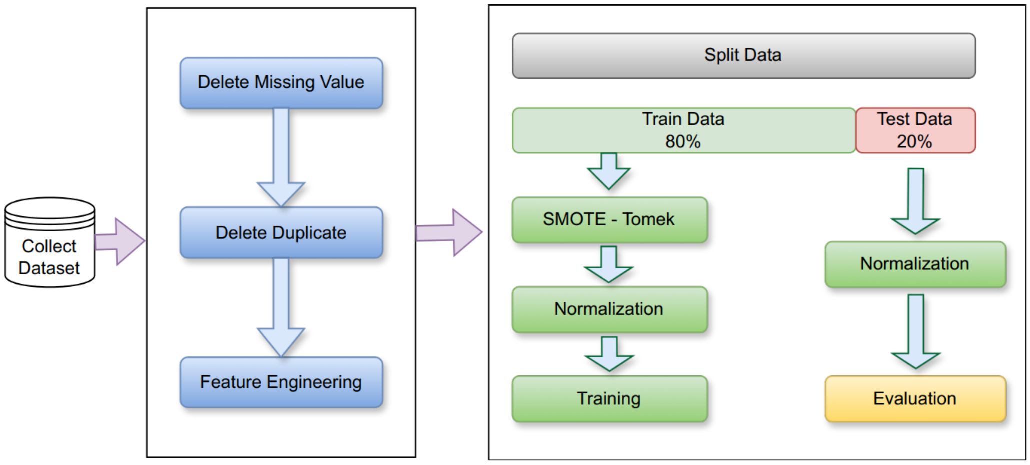

Figure 1. Step-by-step proposed method illustration.

Table 1. Details of Datasets Features.

Feature Name	Cryotherapy	Immunotherapy	Description
Sex	☑	☑	Gender of the patient (1 = Male, 2 = Female)
Age	☑	☑	Age of the patient in years
Time	☑	☑	Duration of treatment in months
Number_of_Warts	☑	☑	Total number of warts present
Type	☑	☑	Type of wart (numerical category)
Area	☑	☑	Area of warts in mm²
Induration_diameter	X	☑	The diameter of the induration around the wart in mm
Result_of_Treatment	☑	☑	Outcome of the treatment (1 = Successful, 0 = Unsuccessful)

Table 1 elaborates on the characteristics of each dataset, including patient gender, age, treatment duration, total number of warts, wart type, affected area, and, exclusively for the immunotherapy dataset, the diameter of the induration surrounding the wart. The outcome of the treatment, indicating success or failure, is also documented for both datasets.

3.2. Delete Missing Value and Duplicate Records

In data mining, especially for classification tasks, it is essential to ensure data cleanliness by removing missing values and duplicate records. In preparing the cryotherapy and immunotherapy datasets for analysis, we meticulously addressed missing values and duplicate records to ensure the integrity and reliability of our findings. Our approach was guided by the principle of minimizing data loss while preserving the quality of the dataset. Missing values can reduce the accuracy and effectiveness of classification models by causing errors in calculations and analysis. Removing or imputing these values helps improve the reliability of predictions[13], [25]. Meanwhile, duplicate records can cause bias in the data and inefficient processing. Eliminating duplicates ensures that each data is considered unique, reduces redundancy, and increases data processing speed[26], [27]. These steps are essential to achieve accurate and representative classification results from the dataset.

3.3. Feature Engineering

In addressing the unique challenges the Cryotherapy and Immunotherapy datasets presented, including their small size and imbalance, feature engineering emerged as a pivotal preprocessing step. This process involved creating new features from the existing data to uncover patterns not immediately apparent yet crucial for enhancing our classification models' predictive accuracy and robustness.

This process is instrumental in allowing models to assimilate more complex and relevant information from the dataset, significantly improving prediction accuracy[28]. Several engineering feature techniques include logarithmic transformation, interaction, and derivative/ratio. Logarithmic transformation can reduce the effects of outliers and bring the data distribution closer to a normal distribution, making it easier for some algorithms to model data more efficiently. Interaction features combine two or more variables to form a new feature that reflects their combined effect. This is especially important in cases where the effect of one variable on the target variable may depend on the value of another variable. Meanwhile, derivative features or ratios involve creating new features through arithmetic operations such as dividing or subtracting two variables, providing insight into relative proportions or differences.

3.4. Data Splitting, SMOTE-Tomeks, and Normalization

This section delves into the strategic implementation of data splitting, the SMOTE-Tomek resampling approach, and data normalization as foundational preprocessing steps to optimize model performance.

Data splitting is a critical initial step, partitioning the dataset into training (80%) and testing (20%) subsets. This segregation is essential for evaluating the model's performance on unseen data. Subsequently, the training data undergoes preprocessing using the SMOTE-Tomek technique to address class imbalance—a common challenge in medical datasets.

SMOTE is designed to mitigate imbalance by generating synthetic instances of the minority class, rather than merely duplicating existing ones. This method involves selecting two or more similar instances within the minority class and calculating synthetic points between them to create new samples. [29]. The process can be summarized as follows[21]:

1. Identify the k nearest neighbors for each minority class sample, typically using Euclidean distance.
2. Randomly select one of these neighbors and create a new sample along the vector that connects the original sample to this neighbor.
3. Generate the new sample by interpolating between the original sample's features and its neighbor's features, scaled by a random factor between 0 and 1.

This approach not only augments the minority class but also ensures a richer, more diverse dataset for training the model. Let's say x is a minority sample and x_{nn} is one of its nearest neighbors, the synthetic sample (x_{new}) can be calculated by Equation (1), where λ is a random number between 0 and 1.

$$x_{new} = x + \lambda \cdot (x_{nn} - x) \tag{1}$$

Meanwhile, Tomek links is an undersampling technique which enhance data quality by eliminating overlaps between classes [30]. So, if two samples are closest neighbors and come from different classes, then the two samples form a Tomek link. If samples x_i from class C_i and x_j from class C_j are the only nearest neighbors of each other, they form a Tomek link. In other words, pairs of samples from different classes are identified as the closest neighbors to be deleted.

In machine learning practice, it is important not to apply SMOTE-Tomek to testing data to preserve the purity and original distribution of the dataset, thereby allowing objective evaluation of model performance on previously unseen data. SMOTE-Tomek is only applied to training data to overcome class imbalances and clarify boundaries between classes. Furthermore, the use of the Standard Scaler for normalization of both training and testing data helps reduce bias in models sensitive to feature scales, such as SVM and KNN. It speeds up the convergence process in algorithms that use gradient descent. Standard Scaler converts features into a distribution with a mean of zero and a standard deviation of one, ensuring that all features contribute equally to the model's predictions. The standard scaler transformation can be calculated with Equation (2).

$$z = \frac{(x - \mu)}{\sigma} \tag{2}$$

Where x is the original value of the feature, μ and σ are the mean and standard deviation of the feature, respectively.

3.5. Training and Testing Evaluation

After preprocessing and balancing the training data, several classification models such as SVM, Naive Bayes, Random Forest, Decision Tree, and XGBoost were tested in this research. The training process entails adjusting the model's internal parameters based on given data and targets, looking for patterns or relationships that can accurately predict outcomes. Additionally, the BiGRU DL model was utilized, necessitating converting the training data into a suitable format, specifically 3D data, to accommodate the GRU's input structure requirements. This model was then trained with specified parameters, such as the number of epochs and optimizer, to minimize the loss function and enhance accuracy.

Upon completion of the training phase, the subsequent step involves testing the model with data it has not previously encountered, to assess its predictive capabilities on unseen data. This evaluation employs several metrics: a confusion matrix, accuracy, precision, recall, specificity, and F1 score metrics[31]–[35] Where each of these metrics provides different insights into aspects of model performance, namely:

- Confusion Matrix provides a visual representation of classification performance.
- Accuracy measures the total proportion of correct predictions.
- Precision is the ratio of true positive predictions to all positive predictions.
- Recall (or sensitivity) measures the model's ability to find all actual positive cases.
- Specificity measures the ability of the model to determine true negatives from tested negatives.
- The F1 Score is the harmonic average of precision and recall, balancing the two.

In the context of medical datasets, which are often characterized by their small size and imbalance, it is important to select the most appropriate measurement tools to depict model performance accurately. Relying solely on accuracy as a metric may not always yield the most precise representation, particularly in the case of imbalanced datasets. This is due to the potential bias towards predicting the majority class more frequently, thereby neglecting the minority classes that hold significant importance[6]. Recall, or sensitivity, is crucial as it quantifies the model's proficiency in identifying all true positive instances. High recall indicates that the model successfully detects nearly all positive cancer cases, substantially reducing the risk of failing to provide necessary treatment to diagnosed individuals. On the other hand, specificity measures the model's ability to truly identify negatives, which is important to avoid false diagnoses that can cause anxiety or unnecessary treatment in healthy patients. High specificity ensures that individuals who do not have the tested condition are completely excluded[36]. These two metrics are of utmost importance in helping to balance identifying as many cases as possible and avoiding false alarms, which is especially important in the clinical setting.

4. Results and Discussion

This section discusses results and processes, starting with initial data processing, including removing missing values, duplicate records, and data normalization. The process of removing missing values and duplicate records must be carried out to minimize noise. Even though both datasets did not have missing values, duplicate data was found in the cryotherapy dataset. So, the number of records for each dataset is 89 and 90. Next, feature engineering is carried out to obtain four new features, namely log_Area, age_time_interaction, Intensity, and Time_per_wart. In more detail, the engineering features carried out are as follows:

1. Logarithmic transformation of the Area feature: the natural logarithm of $(1 + x)$ is carried out for each element x in the Area to form the log_Area feature. This transformation aims to reduce the influence of extreme values or outliers in the 'Area' feature. Equation (3)

$$\text{log\_Area} = \log(1 + \text{Area}) \tag{3}$$

2. Create an interaction feature between Age and Time: This feature is the product of two variables, 'Age' and 'Time', and is used to capture the interaction between these two variables. To generate the age_time_interaction feature, Equation (4) is used.

$$\text{age\_time\_interaction} = \text{Age} \times \text{Time} \tag{4}$$

3. The third feature is Intensity which is obtained with Equation (5). This feature attempts to assess the density of warts per unit area. This can be very informative, especially in a

medical context, as the density of warts can relate to the severity or type of treatment required. This feature shows how many warts there are relative to the affected area.

$$\text{Time\_per\_Wart} = \frac{\text{Time}}{\text{Number\_of\_Warts}} \quad (5)$$

4. Lastly is the Time_per_Wart feature, which can be calculated from Equation (6). This feature measures the average treatment time required per wart. This can provide a perspective on treatment efficiency, namely whether the overall treatment time is adequate for the number of warts to be treated.

$$\text{Time\_per\_Wart} = \frac{\text{Time}}{\text{Number\_of\_Warts}} \quad (6)$$

The four aforementioned features are combined into the dataset, which is subsequently partitioned into 80% for training and 20% for testing, resulting in 71 training records for the cryotherapy dataset and 72 for the immunotherapy dataset. The SMOTE-Tomek technique is applied to the training data, as illustrated in Figures 2 and 3.

Next, the training and testing process for the above data was carried out for several ML and BiGRU models, where the configuration of each model used in this research is presented in Table 2 for the ML model and Table 3 for BiGRU. Next, evaluation of the testing data is carried out using various measuring tools that have been described previously. The measurement results of each method are presented in Tables 4 to 11.

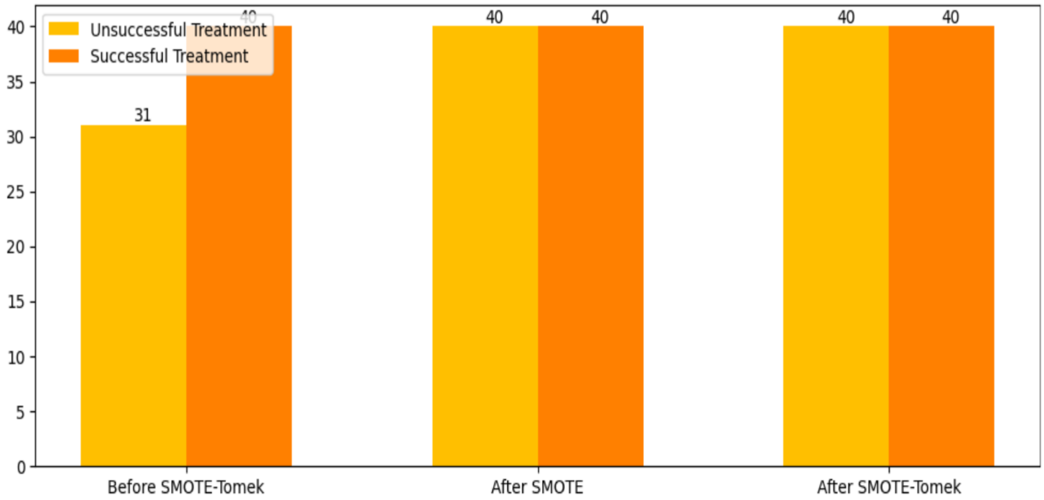

Figure 2. Before and after SMOTE-Tomeks on Cryotherapy dataset.

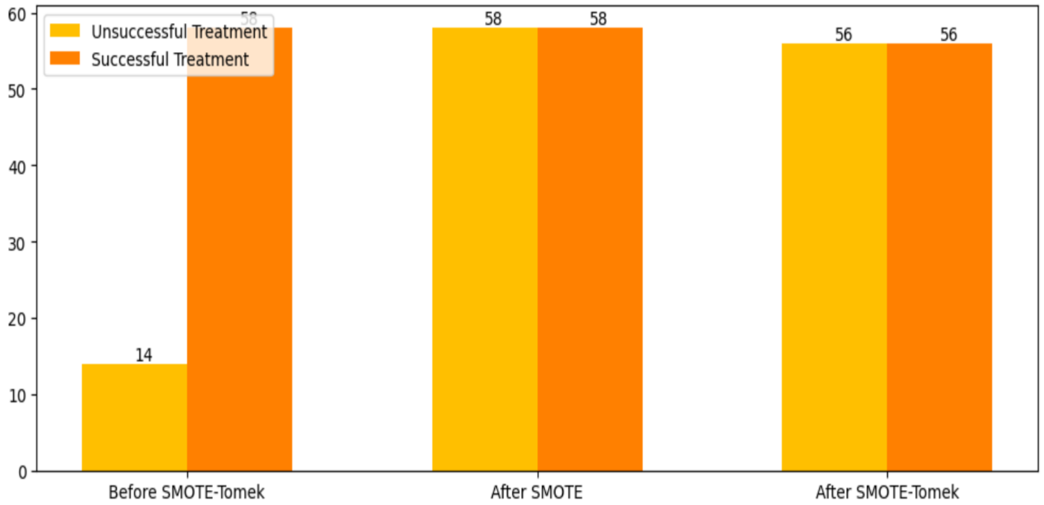

Figure 3. Before and after SMOTE-Tomeks on Immunotherapy dataset.

Table 2. ML Models Configuration.

Method	Configuration
XGBoost	The selected random seed was set to 42 to ensure consistency of results.
SVM	Kernel = RBF and random seed = 42.
Naïve Bayes	Using Gaussian Naive Bayes, features are fitted, assuming a normal distribution.
Random Forest	Using random seed = 42, n_estimators=100
Decision Tree	Using random seed = 42

Table 3. BiGRU Model Configuration

Component	Configuration
Input Layer	Bidirectional Layer that packs GRU with 32 units. 3D input data is required.
Hidden Layer	Second Bidirectional Layer with 16 GRU units. Returns no sequences (return_sequences=False).
Output Layer	Dense layer using softmax activation function for binary classification.
Compiler	The model is compiled with the optimizer 'adam' and the loss function 'categorical_crossentropy'.
Training	Model dilatih dengan 50 epochs

Table 4. Performance Evaluation after Feature Engineering and SMOTE-Tomek in Immunotherapy Dataset.

Method	Accuracy	Recall	Specificity	Precision	F1 score
XGBoost	**0.8889**	**0.8889**	**0.6000**	**0.9037**	**0.8790**
SVM	0.7778	0.7778	0.4000	0.7630	0.7579
Naïve Bayes	0.7222	0.7222	**0.6000**	0.7407	0.7293
Random Forest	**0.8889**	**0.8889**	**0.6000**	**0.9037**	**0.8790**
Decision Tree	0.7778	0.7778	**0.6000**	0.7778	0.7778
BiGRU	0.7222	0.7222	0.4000	0.7063	0.7119

Table 5. Performance Evaluation after Feature Engineering is only in the Immunotherapy dataset.

Method	Accuracy	Recall	Specificity	Precision	F1 score
XGBoost	0.7778	0.7778	0.2000	0.8301	0.7185
SVM	0.7778	0.7778	0.2000	0.8301	0.7185
Naïve Bayes	0.6667	0.6667	0.4000	0.6667	0.6667
Random Forest	**0.8333**	**0.8333**	**0.4000**	**0.8646**	**0.8062**
Decision Tree	0.7222	0.7222	0.4000	0.7063	0.7119
BiGRU	0.7222	0.7222	0.2000	0.6806	0.6771

Table 6. Performance Evaluation after SMOTE-Tomek is only in the Immunotherapy dataset.

Method	Accuracy	Recall	Specificity	Precision	F1 score
XGBoost	0.7778	0.7778	0.2000	0.8301	0.7185
SVM	**0.8333**	**0.8646**	0.4000	**0.8333**	**0.8062**
Naïve Bayes	0.7778	0.7778	**0.6000**	0.7778	0.7778
Random Forest	**0.8333**	0.8333	0.4000	**0.8333**	**0.8062**
Decision Tree	**0.8333**	0.8333	0.4000	**0.8333**	**0.8062**
BiGRU	0.6111	0.6343	0.4000	0.6111	0.6210

Table 7. Performance Evaluation without preprocessing enhancement in Immunotherapy dataset.

Method	Accuracy	Recall	Specificity	Precision	F1 score
XGBoost	0.7778	0.7778	0.2000	0.8301	0.7185
SVM	0.7778	0.7778	0.2000	0.8301	0.7185
Naïve Bayes	**0.8333**	**0.8333**	**0.4000**	**0.8646**	**0.8062**
Random Forest	**0.8333**	**0.8333**	**0.4000**	**0.8646**	**0.8062**
Decision Tree	**0.8333**	**0.8333**	**0.4000**	**0.8646**	**0.8062**
BiGRU	0.7222	0.7222	0.2000	0.6806	0.6771

Table 8. Performance Evaluation after Feature Engineering and SMOTE-Tomek in Cryotherapy dataset.

Method	Accuracy	Recall	Specificity	Precision	F1 score
XGBoost	**0.9444**	**0.9444**	**1.0000**	**0.9495**	**0.9439**
SVM	0.8889	0.8889	0.9000	0.8889	0.8889
Naïve Bayes	0.8333	0.8333	0.7000	0.8788	0.8318
Random Forest	**0.9444**	**0.9444**	**1.0000**	**0.9495**	**0.9439**
Decision Tree	0.8333	0.8333	0.8000	0.8395	**0.9439**
BiGRU	0.8889	0.8889	0.9000	0.8889	0.8889

Table 9. Performance Evaluation after Feature Engineering only in Cryotherapy dataset.

Method	Accuracy	Recall	Specificity	Precision	F1 score
XGBoost	**0.9444**	**0.9444**	**1.0000**	0.9495	0.9439
SVM	**0.9444**	**0.9444**	0.9000	**0.9506**	**0.9446**
Naïve Bayes	0.8333	0.8333	0.7000	0.8788	0.8318
Random Forest	**0.9444**	**0.9444**	**1.0000**	0.9495	0.9439
Decision Tree	0.8333	0.8333	0.8000	0.8395	0.8338
BiGRU	0.8333	0.8333	0.8000	0.8395	0.8338

Table 10. Performance Evaluation after SMOTE-Tomek only in the Cryotherapy dataset.

Method	Accuracy	Recall	Specificity	Precision	F1 score
XGBoost	0.8889	0.8889	0.9000	0.8889	0.8889
SVM	0.8889	0.8889	0.9000	0.8889	0.8889
Naïve Bayes	0.7778	0.7778	0.6000	0.8519	0.7722
Random Forest	**0.9444**	**0.9444**	**1.0000**	**0.9495**	**0.9439**
Decision Tree	0.8889	0.8889	1.0000	0.9074	0.8860
BiGRU	0.8333	0.8333	0.8000	0.8395	0.8338

Table 11. Performance Evaluation without preprocessing enhancement in Cryotherapy dataset.

Method	Accuracy	Recall	Specificity	Precision	F1 score
XGBoost	**0.9444**	**0.9444**	**1.0000**	**0.9495**	**0.9439**
SVM	0.8889	0.8889	0.9000	0.8889	0.8889
Naïve Bayes	0.7778	0.7778	0.6000	0.8519	0.7722
Random Forest	**0.9444**	**0.9444**	**1.0000**	**0.9495**	**0.9439**
Decision Tree	**0.9444**	**0.9444**	**1.0000**	**0.9495**	**0.9439**
BiGRU	0.8889	0.8889	0.9000	0.8889	0.8889

Based on the test data analysis, it is evident that XGBoost outperforms other models,, especially with the implementation of comprehensive preprocessing such as Feature Engineering (FE) combined with SMOTE-Tomek. This observation is notably apparent in the Cryotherapy dataset, where FE increases the specificity to 100%, indicating perfect identification of the negative class. FE significantly boosts the performance of the XGBoost model, corroborating the initial analysis.

Random Forest demonstrates consistent performance across various preprocessing modes and datasets. Random Forest is particularly robust, showing lower performance variations than other models. Decision Tree consistently performs moderately and does not significantly benefit from advanced preprocessing techniques. This may be due to the high variance nature of Decision Trees where additional preprocessing does not necessarily mean better generalization.

There is marked variability in the performance of SVM and NB across different preprocessing techniques. In most cases, SVM tends to perform slightly better than NB, likely due to its ability to manage non-linear data limits more effectively through the use of kernel tricks. The suboptimal performance of BiGRU may stem from its requirement for a larger dataset to learn dependencies in sequential data effectively. Moreover, a modest improvement in performance with simpler preprocessing suggests potential overfitting when applying more complex preprocessing strategies.

Overall, this analysis highlights that preprocessing impacts machine learning models differently, with XGBoost and Random Forest standing out for their strong performance across the board. This insight emphasizes the importance of selecting appropriate preprocessing techniques tailored to the specific characteristics of the model and the dataset at hand.

5. Comparison

In this section, a comparative evaluation is carried out between the results obtained from the best-performing models (XGBoost and Random Forest) on this study against several others from previous studies. Further comparison results are presented in Tables 10 and 11.

Table 10. Comparison in Immunotherapy dataset.

Method	Accuracy	Recall	Specificity	Precision	F1 score
Ref [2]	0.8300	**0.9300**	0.5000	-	**0.9000**
Ref [4]	0.8800	0.8200	**0.6000**	0.8100	0.8100
Our (best model)	**0.8889**	0.8889	**0.6000**	**0.9037**	0.8790

Table 11. Comparison in Cryotherapy dataset.

Method	Accuracy	Recall	Specificity	Precision	F1 score
Ref [2]	0.8900	0.8600	0.9300	-	0.8900
Ref [4]	**0.9700**	**0.9500**	0.9800	0.9400	0.9100
Our (best model)	0.9444	0.9444	**1.0000**	**0.9495**	**0.9439**

The analysis of these tables indicates that the best model tested does not uniformly outperform across all metrics. Still, considering its relevance for clinical applications, the balance of recall and specificity is more important. The Immunotherapy dataset shows high recall, while specificity is the best. Meanwhile, the cryotherapy dataset produced perfect specificity, showing the absence of false positives, which is very important in preventing unnecessary treatment. The recall value is also close to the previous best value. Despite references to higher accuracy in other works, the equilibrium between recall and specificity offered by our model presents significant advantages in clinical settings, facilitating more accurate and safer treatment decisions.

6. Conclusions

This research has succeeded in analyzing the impact of preprocessing techniques on various ML and DL methods. Through comprehensive experiments, we managed to identify the most effective combination of preprocessing strategies and machine learning models. The results indicate that the XGBoost method can significantly enhance prediction performance when combined with appropriate preprocessing techniques such as feature engineering and SMOTE-Tomek links. Meanwhile, Random Forest demonstrates robust performance even without preprocessing, whereas BiGRU is less effective in this context. Although the proposed model does not outperform all metrics compared to previous methods, the balance achieved between recall and specificity makes this approach highly valuable in clinical practice.

For future research, we suggest further investigation of model parameter optimization techniques to improve the effectiveness of these already promising models. Hyperparameter tuning through grid or random search approaches can help find ideal configurations that may not be achieved through default settings. Additionally, a broader exploration of ensemble algorithms and model combination techniques could provide improved stabilization and more consistent performance.

Author Contributions: Conceptualization: D.R.I.M.S; Methodology: D.R.I.M.S; Software: D.R.I.M.S; Validation: G.A.T, W.H. and H.M.M.I.; Formal analysis: G.A.T and W.H.; Investigation: G.A.T, W.H. and H.M.M.I.; Resources: D.R.I.M.S; Data curation: G.A.T, W.H. and H.M.M.I.; Writing—original draft preparation: D.R.I.M.S; Writing—review and editing: G.A.T, W.H. and H.M.M.I.; Visualization: H.M.M.I.; Supervision: D.R.I.M.S; Project administration: G.A.T and W.H.; Funding acquisition: All.

Data Availability Statement: The dataset used is a public dataset that can be accessed on the UCI Machine Learning Repository with the URL:https://archive.ics.uci.edu/dataset/428/immunotherapy+dataset and https://archive.ics.uci.edu/dataset/429/cryotherapy+dataset.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] R. Jain, R. Sawhney, and P. Mathur, "Feature Selection for Cryotherapy and Immunotherapy Treatment Methods Based on Gravitational Search Algorithm," in *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, Mar. 2018, pp. 1–7. doi: 10.1109/ICCTCT.2018.8550983.

[2] A. Cüvitoğlu and Z. Işik, "Evaluation Machine-Learning Approaches for Classification of Cryotherapy and Immunotherapy Datasets," *Int. J. Mach. Learn. Comput.*, vol. 8, no. 4, pp. 331–335, Aug. 2018, doi: 10.18178/ijmlc.2018.8.4.707.

[3] S. Khatri, D. Arora, and A. Kumar, "Enhancing Decision Tree Classification Accuracy through Genetically Programmed Attributes for Wart Treatment Method Identification," *Procedia Comput. Sci.*, vol. 132, pp. 1685–1694, 2018, doi: 10.1016/j.procs.2018.05.141.

[4] A. Y. Mahmoud, D. Neagu, D. Scrimieri, and A. R. A. Abdullatif, "Early diagnosis and personalised treatment focusing on synthetic data modelling: Novel visual learning approach in healthcare," *Comput. Biol. Med.*, vol. 164, no. August, p. 107295, Sep. 2023, doi: 10.1016/j.compbiomed.2023.107295.

[5] O. Jaiyeoba, E. Ogbuju, O. T. Yomi, and F. Oladipo, "Development of a Model to Classify Skin Diseases using Stacking Ensemble Machine Learning Techniques," *J. Comput. Theor. Appl.*, vol. 2, no. 1, pp. 22–38, May 2024, doi: 10.62411/jcta.10488.

[6] F. S. Gomiasti, W. Warto, E. Kartikadarma, J. Gondohanindijo, and D. R. I. M. Setiadi, "Enhancing Lung Cancer Classification Effectiveness Through Hyperparameter-Tuned Support Vector Machine," *J. Comput. Theor. Appl.*, vol. 1, no. 4, pp. 396–406, Mar. 2024, doi: 10.62411/jcta.10106.

[7] M. A. Araaf, K. Nugroho, and D. R. I. M. Setiadi, "Comprehensive Analysis and Classification of Skin Diseases based on Image Texture Features using K-Nearest Neighbors Algorithm," *J. Comput. Theor. Appl.*, vol. 1, no. 1, pp. 31–40, Sep. 2023, doi: 10.33633/jcta.v1i1.9185.

[8] F. Mustofa, A. N. Safriandono, A. R. Muslikh, and D. R. I. M. Setiadi, "Dataset and Feature Analysis for Diabetes Mellitus Classification using Random Forest," *J. Comput. Theor. Appl.*, vol. 1, no. 1, pp. 41–48, Jan. 2023, doi: 10.33633/jcta.v1i1.9190.

[9] M. I. Akazue, I. A. Debekeme, A. E. Edje, C. Asuai, and U. J. Osame, "UNMASKING FRAUDSTERS: Ensemble Features Selection to Enhance Random Forest Fraud Detection," *J. Comput. Theor. Appl.*, vol. 1, no. 2, pp. 201–211, Dec. 2023, doi: 10.33633/jcta.v1i2.9462.

[10] F. Omoruwou, A. A. Ojugo, and S. E. Ilodigwe, "Strategic Feature Selection for Enhanced Scorch Prediction in Flexible Polyurethane Form Manufacturing," *J. Comput. Theor. Appl.*, vol. 1, no. 3, pp. 346–357, Feb. 2024, doi: 10.62411/jcta.9539.

[11] S. Ali, A. Hashmi, A. Hamza, U. Hayat, and H. Younis, "Dynamic and Static Handwriting Assessment in Parkinson's Disease: A Synergistic Approach with C-Bi-GRU and VGG19," *J. Comput. Theor. Appl.*, vol. 1, no. 2, pp. 151–162, Dec. 2023, doi: 10.33633/jcta.v1i2.9469.

[12] M. Diaz, M. Moetesum, I. Siddiqi, and G. Vessio, "Sequence-based dynamic handwriting analysis for Parkinson's disease detection with one-dimensional convolutions and BiGRUs," *Expert Syst. Appl.*, vol. 168, no. August 2020, p. 114405, Apr. 2021, doi: 10.1016/j.eswa.2020.114405.

[13] M. Misdram *et al.*, "Analysis of Imputation Methods of Small and Unbalanced Datasets in Classifications using Naïve Bayes and Particle Swarm Optimization," in *2020 International Seminar on Application for Technology of Information and Communication (iSemantic)*, Sep. 2020, pp. 115–119. doi: 10.1109/iSemantic50169.2020.9234225.

[14] G. Charizanos, H. Demirhan, and D. İçen, "A Monte Carlo fuzzy logistic regression framework against imbalance and separation," *Inf. Sci. (Ny).*, vol. 655, no. 1, p. 119893, Jan. 2024, doi: 10.1016/j.ins.2023.119893.

[15] F. Khozeimeh, R. Alizadehsani, M. Roshanzamir, A. Khosravi, P. Layegh, and S. Nahavandi, "An expert system for selecting wart treatment method," *Comput. Biol. Med.*, vol. 81, no. January, pp. 167–175, Feb. 2017, doi: 10.1016/j.compbiomed.2017.01.001.

[16] F. Khozeimeh *et al.*, "Intralesional immunotherapy compared to cryotherapy in the treatment of warts," *Int. J. Dermatol.*, vol. 56, no. 4, pp. 474–478, Apr. 2017, doi: 10.1111/ijd.13535.

[17] A. R. R. M. Khozeimeh Fahime and P. Layegh, "Immunotherapy Dataset." 2018. doi: 10.24432/C5DC72.

[18] A. R. R. M. Khozeimeh Fahime and P. Layegh, "Cryotherapy Dataset." 2018. doi: 10.24432/C5FC7C.

[19] D. Gibert, J. Planes, C. Mateu, and Q. Le, "Fusing feature engineering and deep learning: A case study for malware classification," *Expert Syst. Appl.*, vol. 207, no. June, p. 117957, Nov. 2022, doi: 10.1016/j.eswa.2022.117957.

[20] A. M. Q. Farhan, S. Yang, A. Q. S. Al-Malahi, and M. A. Al-antari, "MCLSG:Multi-modal classification of lung disease and severity grading framework using consolidated feature engineering mechanisms," *Biomed. Signal Process. Control*, vol. 85, no. January, p. 104916, Aug. 2023, doi: 10.1016/j.bspc.2023.104916.

[21] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, no. February 2017, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[22] F. O. Aghware *et al.*, "Enhancing the Random Forest Model via Synthetic Minority Oversampling Technique for Credit-Card Fraud Detection," *J. Comput. Theor. Appl.*, vol. 1, no. 4, pp. 407–420, Mar. 2024, doi: 10.62411/jcta.10323.

[23] N. Cahyana, S. Khomsah, and A. S. Aribowo, "Improving Imbalanced Dataset Classification Using Oversampling and Gradient Boosting," in *2019 5th International Conference on Science in Information Technology (ICSITech)*, Oct. 2019, pp. 217–222. doi: 10.1109/ICSITech46713.2019.8987499.

[24] N. Reska and K. Tsabita, "Comparison of KNN, naive bayes, and decision tree methods in predicting the accuracy of classification of immunotherapy dataset," *J. Student Res. Explor.*, vol. 1, no. 2, pp. 104–121, Jul. 2023, doi: 10.52465/josre.v1i2.170.

[25] A. Ali, M. Abu-Elkheir, A. Atwan, and M. Elmogy, "Missing values imputation using Fuzzy K-Top Matching Value," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 35, no. 1, pp. 426–437, Jan. 2023, doi: 10.1016/j.jksuci.2022.12.011.

[26] M. S. Sunarjo, H. Gan, and D. R. I. M. Setiadi, "High-Performance Convolutional Neural Network Model to Identify COVID-19 in Medical Images," *J. Comput. Theor. Appl.*, vol. 1, no. 1, pp. 19–30, Aug. 2023, doi: 10.33633/jcta.v1i1.8936.

[27] E. B. Wijayanti, D. R. I. M. Setiadi, and B. H. Setyoko, "Dataset Analysis and Feature Characteristics to Predict Rice Production based on eXtreme Gradient Boosting," *J. Comput. Theor. Appl.*, vol. 1, no. 3, pp. 299–310, Feb. 2024, doi: 10.62411/jcta.10057.

[28] A. Zheng, Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists. Sebastopol, CA: O'Reilly Media, 2018.

[29] Y. Chachoui, N. Azizi, R. Hotte, and T. Bensebaa, "Enhancing algorithmic assessment in education: Equi-fused-data-based SMOTE for balanced learning," *Comput. Educ. Artif. Intell.*, vol. 6, p. 100222, Jun. 2024, doi: 10.1016/j.caeai.2024.100222.

[30] R. M. Pereira, Y. M. G. G. Costa, C. N. Silla Jr., and C. N. Silla, "MLTL: A multi-label approach for the Tomek Link undersampling algorithm," *Neurocomputing*, vol. 383, pp. 95–105, Mar. 2020, doi: 10.1016/j.neucom.2019.11.076.

[31] F. M. Firnando, D. R. I. M. Setiadi, A. R. Muslikh, and Syahroni Wahyu Iriananda, "Analyzing InceptionV3 and InceptionResNetV2 with Data Augmentation for Rice Leaf Disease Classification," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 1, pp. 1–11, 2024.

[32] R. K. Rachman, D. R. I. M. Setiadi, A. Susanto, K. Nugroho, and H. M. M. Islam, "Enhanced Vision Transformer and Transfer Learning Approach to Improve Rice Disease Recognition," *J. Comput. Theor. Appl.*, vol. 1, no. 4, pp. 446–460, Apr. 2024, doi: 10.62411/jcta.10459.

[33] T. R. Noviandy, K. Nisa, G. M. Idroes, I. Hardi, and N. R. Sasmita, "Classifying Beta-Secretase 1 Inhibitor Activity for Alzheimer's Drug Discovery with LightGBM," *J. Comput. Theor. Appl.*, vol. 1, no. 4, pp. 358–367, Mar. 2024, doi: 10.62411/jcta.10129.

[34] M. T. H. Khan Tusar, M. T. Islam, A. H. Sakil, M. N. H. N. Khandaker, and M. M. Hossain, "An Intelligent Telediagnosis of Acute Lymphoblastic Leukemia using Histopathological Deep Learning," *J. Comput. Theor. Appl.*, vol. 2, no. 1, pp. 1–12, May 2024, doi: 10.62411/jcta.10358.

[35] M. Çiftçi, M. U. Türkdamar, and C. Öztürk, "Leveraging YOLO Models for Safety Equipment Detection on Construction Sites," *J. Comput. Theor. Appl.*, vol. 1, no. 4, pp. 492–506, May 2024, doi: 10.62411/jcta.10453.

[36] D. R. I. M. Setiadi, K. Nugroho, A. R. Muslikh, S. Wahyu, and A. A. Ojugo, "Integrating SMOTE-Tomek and Fusion Learning with XGBoost Meta-Learner for Robust Diabetes Recognition," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 1, pp. 23–38, 2024.