







Multiclass Logistic Regression Classification with PCA for Imbalanced Medical Datasets

Adli A. Nababan^{1*}, Sutarman², Muhammad Zarlis³, Erna B. Nababan⁴

¹ Department of Informatics Engineering, Faculty of Science and Technology, Universitas Prima Indonesia, Medan 20118, Indonesia

² Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Sumatera Utara, Medan 20155, Indonesia

³ Department of Information Systems Management, BINUS Graduate Program-Master of Information Systems Management, Bina Nusantara University, Jakarta 11480, Indonesia

⁴ Department of Computer Science, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan 20155, Indonesia

Corresponding Author Email: adliabdillahnababan@unprimdn.ac.id

Copyright: ©2024 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/mmep.110911>

ABSTRACT

Received: 2 June 2024

Revised: 25 August 2024

Accepted: 5 September 2024

Available online: 29 September 2024

Keywords:

imbalanced learning, medical informatics, logistic regression, imbalanced medical datasets, preprocessing techniques, gradient descent optimization, class imbalance, medical applications

The challenge of class imbalance in multiclass medical datasets is addressed in this study, through the proposal of modified classifiers premised on multiclass logistic regression. The principal aim is to augment the accuracy of medical diagnosis predictions by decisively managing imbalanced datasets with innovative methodologies. Performance evaluations are conducted on renowned multiclass medical datasets including thyroid, lymphography, dermatology, and ecoli. Prior to model development, Principal Component Analysis (PCA) is employed as a preprocessing measure to bolster data quality. The bespoke classifiers are trained via gradient descent optimization and evaluated through various metrics such as accuracy, precision, recall, and f1-score. A comparative analysis with preceding studies underscores the superior performance of the proposed model, accentuating its advantageous position over other algorithms. This research underscores the potential of the proposed model to furnish precise medical diagnosis predictions amidst class imbalance, capably distinguishing between minority and majority classes. In conclusion, this study delineates the promising potential of multiclass logistic regression for precise medical diagnoses in the realm of imbalanced datasets.

1. INTRODUCTION

The burgeoning volume of data in the digital age poses challenges in the effective utilization of critical information, particularly in binary and multiclass datasets. Class imbalance has surfaced as a significant challenge [1], predominantly in medical data, where the minority class often encapsulates crucial insights, such as rare disease diagnoses. Addressing this imbalance is vital for harnessing the valuable knowledge concealed within the underrepresented minority classes.

The amalgamation of healthcare technology and machine learning (ML) has instigated a revolution in disease prediction, patient monitoring, and clinical decision-making, thereby amplifying patient outcomes and healthcare quality [2]. ML algorithms have equipped medical practitioners with the ability to leverage vast patient data for informed decision-making. Nonetheless, the persistent issue of inaccurate disease prediction necessitates continual research and development to circumvent risks to patient safety [3].

Class imbalance poses a formidable issue in ML research, with this imbalance being distinctly noticeable in medical data, where healthy patients considerably outnumber their sick

counterparts [4]. Such imbalance can induce bias towards the majority class in conventional ML algorithms, adversely affecting their performance. Consequently, addressing class imbalance has been identified as one of the top ten challenges in ML research [5].

In ML classification tasks, the misclassification of minority classes is a prevalent issue, attributed to the disproportionate emphasis on majority classes [6, 7]. Various techniques have been explored to mitigate the issues posed by imbalanced data, with the classification of multiclass imbalanced data introducing additional complexities [8, 9]. However, the exploration of novel techniques that effectively classify minority classes while delivering optimal results across all classes remains a necessity [10].

Logistic regression emerges as an effective approach for addressing data imbalance, facilitating the modeling of relationships between the dependent variable and multiple classes [11]. In contrast to traditional logistic regression, multiclass logistic regression can predict probabilities across diverse classes [12-14], with the objective being to estimate parameters that minimize prediction errors and furnish accurate predictions for each class [15, 16].

This research is geared towards tackling the challenges presented by imbalanced multiclass medical data by harnessing logistic regression. The emphasis is placed on devising techniques and methodologies that adeptly manage the imbalanced distribution of classes in medical datasets, integrating Principal Component Analysis (PCA) as a preprocessing step to capture data variability, and utilizing gradient descent optimization for model parameter optimization during training. The study evaluates the performance of a logistic regression algorithm specifically engineered to handle imbalanced data in multiclass medical datasets, drawing comparisons with previous studies to gauge advancements in addressing this issue.

2. METHODS

2.1 Datasets

This study utilizes four imbalanced medical datasets obtained from the UCI Machine Learning Repository. The selected datasets were chosen for their common use in multiclass imbalanced data classification research [9]. The level of class imbalance in the datasets is evaluated using the imbalance ratio (IR) and the imbalance degree (ID). The IR compares the number of instances in the majority class to that in the minority class, while the ID measures the relative imbalance between majority and minority classes based on their occurrence percentages. Higher IR and ID values indicate a more pronounced imbalance within the datasets. In classification tasks, higher ID values can present challenges due to bias towards the majority class and difficulties in detecting minority classes, particularly in medical datasets with a larger number of minority classes [17].

In this study, four imbalanced medical datasets were utilized, which were obtained from the UCI Machine Learning Repository. The selected datasets include thyroid, lymphography, and ecoli. These specific datasets were chosen due to their common usage in classification research involving multiclass imbalanced data. A detailed description of these datasets can be found in Table 1 and the data distribution for each class can be seen in Table 2.

Table 1. Description of the medical dataset

Datasets	Number of Instances	Number of Features	Number of Classes
Thyroid	215	5	3
Lymphography	148	18	4
Dermatology	358	34	6
Ecoli	336	7	8

Table 2. Data distribution for each class

Datasets	Class Distribution	IR	ID
Thyroid	150/35/30	5.0	1.55
Lymphography	2/81/61/4	40.5	1.76
Dermatology	111/60/71/48/48/20	5.55	2.33
Ecoli	143/77/52/35/20/5/2/2	71.5	2.75

2.2 Data preprocessing

Data Preprocessing involves analyzing and improving datasets to create new datasets that are appropriate for further procedures. It includes various steps like modifying or

cleaning data, reducing data, and transforming data [18]. The dataset will be split into two portions: 70% for training and 30% for testing. Z-score normalization will be used to standardize the data during the process. Both the training and test data will undergo data scaling, which aims to standardize the input features in the dataset. This ensures that features with different scales but the same variance can be accurately compared [19].

2.3 Principal Component Analysis (PCA)

In the field of machine learning, data dimension reduction plays a crucial role in processing high-dimensional data efficiently. Principal Component Analysis (PCA) is a widely-used technique that aims to reduce the number of features or variables in a dataset while preserving the essential information contained within the data.

To improve the performance of classification models and enhance the accuracy of medical diagnosis predictions, this study proposes a PCA framework to select a subset of relevant and uncorrelated features from the multiclass medical datasets used in the research [20]. To quantify the spread of data within the medical datasets, we calculate the variance using Eq. (1):

$$Var(x) = \sigma^2 = \frac{1}{n} \sum_{i=1}^n (\tilde{Z}_{ij} - \mu_j)^2 \quad (1)$$

where, $Var(x)$ =the variance of variable x ; \tilde{Z}_{ij} =the value of the i -th data point of variable x ; or the j -th feature; μ_j =the mean value of the j -th feature.

After that, the covariance is calculated to find the relationship between classes, where a value of zero indicates that there is no relationship between the two dimensions [21]. The covariance is calculated using Eq. (2):

$$Cov(x, y) = \frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \mu_{xj})(y_{ij} - \mu_{yj}) \quad (2)$$

where, $Cov(x, y)$ =the covariance between variables x and y ; x_{ij} =the value of the i -th data point of variable x ; μ_{xj} =the mean value of variable x ; y_{ij} =the value of the i -th data point of variable y ; μ_{yj} =the mean value of variable y .

Finally, the Eigenvalues and Eigenvectors for the covariance matrices are calculated [22]. The Eigenvalues are then transformed using Eq. (3):

$$Det(A - \lambda I) = 0 \quad (3)$$

where, Det =the determinant of the matrix; A =the value square matrix; λI =the scalar, and the identity matrix.

In this study, PCA was applied to both training and testing attributes from medical data sets that are expected to yield good results when applied to correlated attributes.

2.4 Build multiclass logistic regression

To build a classification model, the initial step involves establishing the class boundary that will differentiate instances belonging to different classes [23]. The number of boundaries required is determined by the number of classes to be distinguished. In binary classification, a single decision boundary is adequate. However, in multiclass classification

scenarios with more than two classes, the number of decision boundaries needed is equal to $k-1$, where k represents the total number of class instances being separated. In the machine learning approach, the logistic regression classification model is employed for binary classes and utilizes the sigmoid function, defined as Eq. (4) [24]:

$$h_{\theta}(x) = \frac{1}{1 + e^{-\theta^T x}} \quad (4)$$

where, $h_{\theta}(x)$ =the predicted probability that the input data; x belongs to the positive class (class 1); θ =the parameter vector of the logistic regression model; θ^T =denotes the transpose of θ ; x =the feature vector of the input data; e =Euler's number.

In the case of multiclass classification, the logistic regression model utilizes the Softmax function. The Softmax function calculates the probabilities of the input data x belonging to each class. It takes the linear combination of the input features and the corresponding model parameters θ_j for each class and transforms these values into probabilities between 0 and 1. The probabilities are then normalized by the sum of probabilities over all classes, ensuring that the predicted probabilities for all classes sum up to 1. The class with the highest probability is chosen as the predicted class for the input data x in multiclass classification. Softmax defined as Eq. (5):

$$P(y = j|x) = \frac{e^{(\theta_j^T x)}}{\sum_{k=1}^K e^{(\theta_k^T x)}} \quad (5)$$

where, $P(y=j|x)$ =the predicted probability that the input data; x belongs to class j out of K classes; θ_j =the parameter vector corresponding to class; j in the Softmax function; x =the feature vector of the input data; e =Euler's number; $\sum_{k=1}^K$ =the sum over all classes from 1 to K .

By incorporating the Softmax function with the ordinal encoder, the logistic regression model for multiclass cases can effectively predict the probability of an input belonging to each class based on its features.

The likelihood function for multiclass logistic regression with ordinal encoder vectors can be derived by extending the binary logistic regression likelihood to the multiclass scenario. Suppose we have a dataset with N samples and K classes. Each sample i is represented by a feature vector x_i and its associated class label y_i . To facilitate computation, the class labels are encoded using an ordinal encoder, which assigns unique numerical labels to each class. This results in a vector of ordinal encoded labels y_i for each sample. In multiclass logistic regression, the likelihood function can be defined as follows Eq. (6):

$$L(\theta) = \prod_{i=1}^N P(y_i|x_i; \theta) \quad (6)$$

where, θ represents the model's weight parameters. The probability $P(y_i|x_i; \theta)$ can be calculated using the Softmax function as Eq. (5). To maximize the likelihood function, we can take the logarithm of $L(\theta)$ and convert the product into a sum as follows Eq. (7):

$$\log L(\theta) = \sum_{i=1}^N \log P(y_i|x_i; \theta) \quad (7)$$

where, $\log L(\theta)$ represents the logarithm of the likelihood function, which is the natural logarithm of the probability of the observed data given the model parameters θ . N is the total number of data samples in the dataset.

By taking the logarithm of the likelihood function, we convert the product of probabilities into a sum of logarithms. This is a common practice in statistics and machine learning as it simplifies computations and avoids numerical underflow that can occur when dealing with small probabilities. Maximizing the log-likelihood function is equivalent to maximizing the likelihood function itself since the logarithm is a monotonic function. The goal of maximizing the log-likelihood is to find the optimal values of the model parameters θ that best explain the observed data and result in the highest probability of the true classes given the input data.

2.5 Find optimal values for model parameters

Gradient descent is an essential optimization algorithm in machine learning, used to minimize the loss function iteratively by adjusting model parameters. In our work, it plays a vital role in training the multiclass logistic regression model to achieve better classification on imbalanced medical datasets. By applying gradient descent, we seek to find the optimal weights and bias parameters. In the following subsection, we explore how gradient descent is used in the training process and its significance in enhancing model accuracy.

The next step after deriving the likelihood function for multiclass logistic regression with ordinal encoder vectors is to optimize the model parameters using gradient descent. It iteratively updates the weight parameters to maximize the likelihood function, moving towards the optimal solution. The general steps for performing gradient descent in the context of multiclass logistic regression with ordinal encoder vectors are explained below.

Step 1: Initialize the weight parameters θ with small random values.

Step 2: Compute the predicted probabilities for each sample using the Softmax function as follows Eq. (5).

Step 3: Compute the loss function, which quantifies the difference between the predicted probabilities and the true ordinal encoded labels. In multiclass logistic regression, as follows Eq. (8):

$$\text{Loss} = - \sum_{i=1}^N \sum_{j=1}^K y_{ij} \log P(y = j|x_i; \theta) \quad (8)$$

where, y_{ij} is an indicator variable that is equal to 1 if the ordinal encoded label for sample i is j , and 0 otherwise.

Step 4: Computing the gradient of the loss function concerning the weight parameters indicates the direction and magnitude of the steepest ascent in the likelihood space. It represents how the loss function changes as the weight parameters are adjusted, providing crucial information for optimizing the model through gradient-based optimization algorithms. By following the gradient, the algorithm can iteratively update the weight parameters in the direction that maximizes the likelihood and reduces the loss.

Step 5: Update the weight parameters using the gradient descent update, as follows Eq. (9):

$$\theta = \theta - \alpha \nabla_{\theta} \text{Loss} \quad (9)$$

where, α is the learning rate, controlling the step size in each

iteration of the optimization process.

Step 6: Continue iterating steps 2 to 5 until either convergence is achieved or a predefined number of iterations is reached.

By iteratively updating the weight parameters using the gradient descent algorithm, we can gradually improve the model's performance and find the optimal parameter values that maximize the likelihood function for the given dataset and ordinal encoded labels.

2.6 Multiclass model evaluation

To evaluate the performance of a classification model, it is essential to use specific measures. These measures can be obtained by utilizing a confusion matrix [25]. By analyzing the elements of the confusion matrix, one can assess the performance of the classification model. In this study, various metrics, including accuracy, precision, recall, f1-score, and ROC AUC [26]. The matrix used to evaluate classification results in multiclass classification problems can be seen in Figure 1.

		Predicted Classification			
		C ₁	C ₂	...	C _n
Actual Classification	C ₁	N ₁₁	N ₁₂	...	N _{1n}
	C ₂	N ₂₁	N ₂₂	...	N _{2n}
	⋮	⋮	⋮	⋮	⋮
	C _n	N _{n1}	N _{n2}	...	N _{nn}

Figure 1. Confusion matrix for multiclass classification

Accuracy is a widely used metric in multiclass classification and is calculated directly from the confusion matrix. A higher accuracy score indicates better performance of the classification model. It is calculated based on the number of correctly predicted classifications. The accuracy value can be estimated using the following Eq. (10):

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (10)$$

When dealing with medical datasets, relying solely on overall accuracy as a measure of classification performance may not be reliable, particularly for positive (minority) classes. Therefore, in this study, precision metrics are utilized, which are commonly employed in data mining and statistical testing. In the following Eq. (11), precision is calculated by dividing the number of positive samples that are correctly classified by the total number of predicted positive instances. This approach provides a more robust evaluation of classification performance in medical datasets, considering the specific challenges posed by minority classes:

$$Prec = \frac{TP}{TP + FP} \quad (11)$$

The True Positive Rate also referred to as recall in information retrieval, represents the proportion of relevant objects correctly identified among the objects that were retrieved. It quantifies the ability of a classification model to accurately identify and retrieve relevant instances from a given dataset. The recall calculation can be seen in the formula in Eq. (12):

$$Rec = TP_{rate} = \frac{TP}{TP + FN} \quad (12)$$

Essentially, the f1-score can be understood as the harmonic mean of recall and precision. It provides a balanced measure that considers both the ability to retrieve relevant instances (recall) and the accuracy of the retrieved instances (precision). By incorporating both recall and precision into a single metric, the f1-score offers a comprehensive evaluation of the classification model's performance as follows Eq. (13):

$$f1 - score = \frac{2}{precision^{-1} + recall^{-1}} = 2 \times \frac{precision \times recall}{precision + recall} \quad (13)$$

Another effective metric for evaluating the performance of a multiclass logistic model in a scenario with imbalanced class distribution is the ROC AUC Score. This metric provides a reliable assessment of the model's performance in multiclass classification tasks [11]. An illustration of the ROC and AUC curves can be seen in Figure 2.

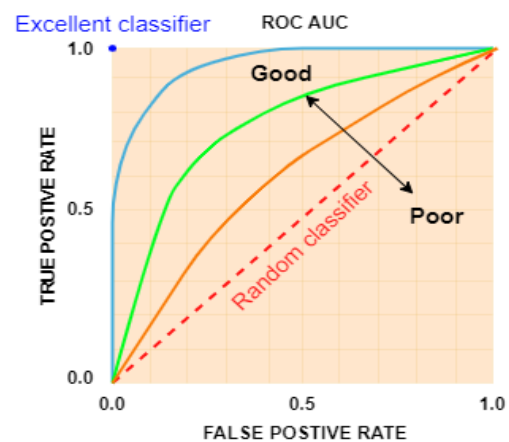


Figure 2. ROC curve and AUC illustration

These metrics provide insight into various aspects of a model's performance, enabling a comprehensive assessment of its ability to handle imbalanced medical data.

2.7 Training process

This study involves several research steps or stages, including (1) collecting data, (2) preprocessing, (3) applying PCA, (4) splitting the dataset, (5) training model (6) Utilizing gradient descent and (7) Evaluating the model's performance. These methodological steps will be applied to address the problem of imbalanced multiclass medical data.

1. Input the multiclass medical dataset for processing:

At the beginning of the training process, the dataset containing multiclass medical data is collected and prepared

for further analysis. This dataset will be used to train and evaluate the multiclass logistic regression model.

2. Preprocess the dataset:

In this step, the dataset is preprocessed to handle any missing values and outliers. Missing values can be imputed or removed, depending on the nature and significance of the missing data. Outliers, which are extreme values that deviate significantly from the rest of the data, may be identified and either removed or treated appropriately to avoid their negative impact on the model.

3. Apply Principal Component Analysis (PCA):

PCA is utilized to reduce the dimensionality of the dataset. It transforms the original features into a new set of uncorrelated features called principal components. By selecting the top principal components that capture most of the variance in the data, we can effectively reduce the number of features, making the dataset more manageable and preventing overfitting.

4. Split the dataset:

The dataset is divided into two subsets: A training set and a testing set. The training set, which comprises 70% of the data, is used to train the multiclass logistic regression model. The testing set, representing 30% of the data, is used to evaluate the model's performance and assess its ability to generalize to new, unseen data.

5. Train the multiclass logistic regression model:

The multiclass logistic regression model is trained using the training set. The model learns to make predictions for each class in the dataset based on the input features. It aims to find the optimal weights and bias parameters that minimize prediction errors and improve the accuracy of class predictions.

6. Utilize gradient descent for optimization:

Gradient descent is an optimization algorithm used to adjust the model's parameters (weights and bias) during training. It iteratively updates these parameters in the direction that minimizes the loss function, which measures the difference between the predicted and actual class labels. By using gradient descent, the model can fine-tune its parameters to achieve better classification performance.

7. Evaluate the model's performance:

Finally, the trained model is evaluated on both the training and testing datasets to assess its performance. Various metrics such as accuracy, precision, recall, f1-score, and ROC AUC are computed to measure the model's ability to classify instances correctly across all classes. This evaluation step helps determine the model's effectiveness in handling the imbalanced multiclass medical data and provides insights into its overall classification capabilities.

3. RESULTS AND DISCUSSION

In this section, we conducted experiments to test the multiclass logistic regression model on a personal computer. The computer used was equipped with an Intel Core i5 processor, 4 GB RAM, and running on the Windows 10 operating system. The implementation of the model was carried out using the Python programming language within the Jupyter Notebook® application. We present the results and discuss the findings of our research on the application of the developed classification model. The methodology involved conducting simulations on four multiclass imbalance datasets: thyroid, lymphography, dermatology, and ecoli.

The training process begins by collecting a dataset containing multiclass medical data, which will be used to train

and evaluate the multiclass logistic regression model. Next, the dataset undergoes preprocessing to handle any missing values and outliers. Missing values are either imputed or removed, while outliers are identified and appropriately dealt with to avoid negative impacts on the model.

3.1 Dataset dimension reduction results

Principal Component Analysis (PCA) was used to reduce the dimensionality of the data while ensuring that at least 98.00% of the total variability in the original data is captured. For each dataset, a specific number of principal components was retained to achieve this level of variability explanation: 5 for the thyroid dataset, 18 for the lymphography dataset, 26 for the dermatology dataset, and 7 for the e-coli dataset. This dimensionality reduction process aids in preserving essential information while making the dataset more manageable and conducive to analysis.

3.2 Dataset scaling results for training and test data

After preprocessing and dimensionality reduction, the dataset is split into two subsets: a training set and a testing set. The training set, consisting of 70% of the data, is used to train the multiclass logistic regression model, while the testing set, representing 30% of the data, is used to evaluate the model's performance on unseen data.

The next step involves training the multiclass logistic regression model using the training set. The model learns to make predictions for each class in the dataset based on the input features. It aims to find the optimal weights and bias parameters that minimize prediction errors and improve the accuracy of class predictions.

3.3 Adjust the model parameters

To fine-tune the model's parameters during training, gradient descent optimization is utilized.

The provided results (Figures 3-6) are from training a Multiclass Logistic Regression model with PCA (Principal Component Analysis) on four different datasets: thyroid, lymphography, dermatology, and ecoli. Each dataset has its own set of features and corresponding target labels, and the goal is to classify the data into multiple classes using the logistic regression model.

For the thyroid dataset, the model's performance improves significantly over the training epochs. Both the train and test losses decrease steadily, indicating that the model is learning and generalizing well. Similarly, the train and test accuracies increase over time, reaching close to 100%, which suggests that the model successfully learned to classify the thyroid data accurately.

In the lymphography dataset, the logistic regression model with PCA also shows improvements in train and test losses, which decrease as the model trains. The train and test accuracies show a positive trend, but the final accuracies are comparatively lower than in the thyroid dataset, reaching around 90%. Although the model performs reasonably well, it might benefit from further optimization or a more complex model for better accuracy.

For the dermatology dataset, the logistic regression model with PCA demonstrates excellent performance. The train and test losses decrease significantly over the epochs, while the train and test accuracies increase steadily and reach approximately 98%. This indicates that the model effectively

learned to classify the dermatology data with high accuracy.

Lastly, in the ecoli dataset, the logistic regression model with PCA performs reasonably well. The train and test losses decrease gradually, and the train and test accuracies improve throughout training, reaching around 85%. Although the model shows promising results, it may have some difficulty capturing complex patterns in the data, which could be addressed with more sophisticated models.

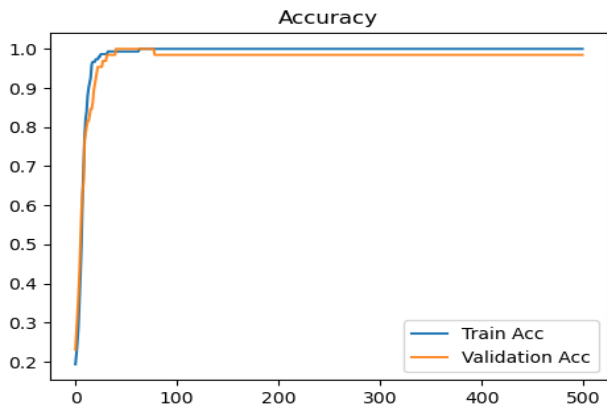


Figure 3. Performance results for the training and test datasets on the accuracy curve (per epoch) for the thyroid data

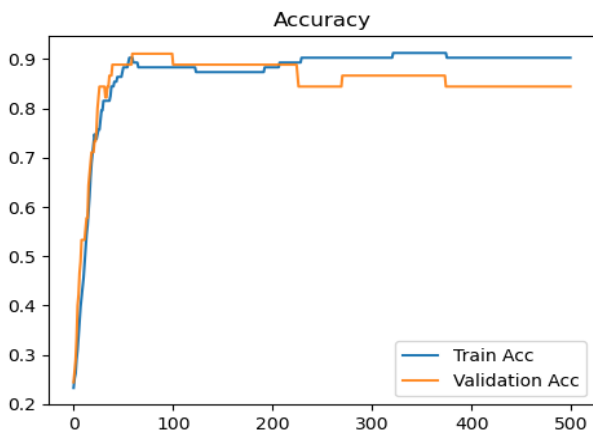


Figure 4. Performance results for the training and test datasets on the accuracy curve (per epoch) for the lymphography data

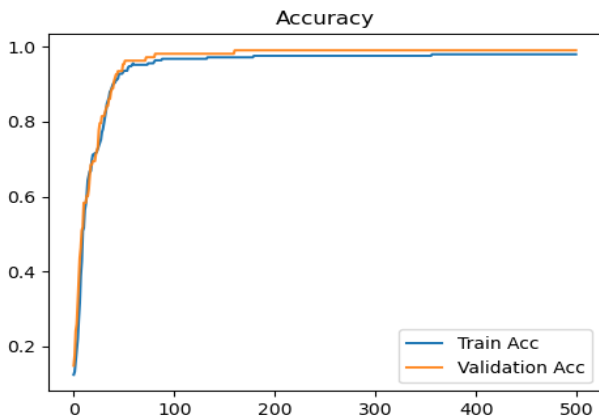


Figure 5. Performance results for the training and test datasets on the accuracy curve (per epoch) for the dermatology data

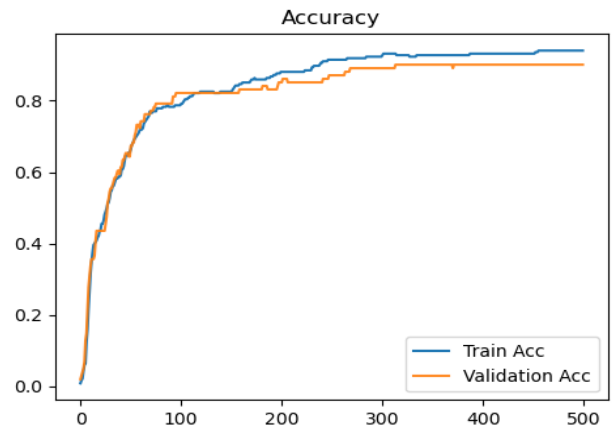


Figure 6. Performance results for the training and test datasets on the accuracy curve (per epoch) for the ecoli data

3.4 Evaluating the model performance

Finally, the trained model is evaluated on both the training and testing datasets to assess its performance. Various metrics, such as accuracy, precision, recall, f1-score, and ROC AUC, are computed to measure the model's ability to classify instances correctly across all classes. This evaluation step provides insights into the model's effectiveness in handling the imbalanced multiclass medical data and its overall classification capabilities.

3.4.1 Confusion matrix and individual class performance

The following is a visualization of the performance model confusion matrix in classifying thyroid data as shown in Figure 7 and Table 3 presents the individual class performance of the multiclass logistic regression model on the thyroid dataset:

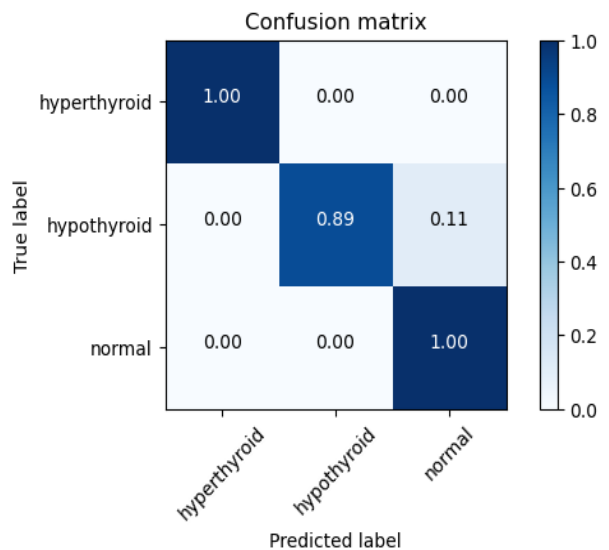


Figure 7. Confusion matrix for thyroid dataset

Table 3. The model performance results for each class individually on the thyroid dataset

Class	Acc	Rec	Prec	F1-S	AUC
Normal	0.985	1.000	0.978	0.989	1.000
Hyperthyroid	1.000	1.000	1.000	1.000	1.000
Hypothyroid	0.985	0.889	1.000	0.941	1.000

The following is a visualization of the performance model confusion matrix in classifying data lymphography as shown in Figure 8 and Table 4 presents the individual class performance of the multiclass logistic regression model on the lymphography dataset:

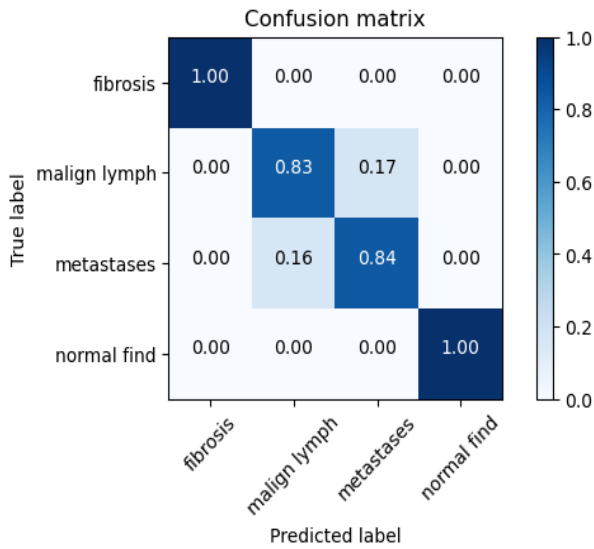


Figure 8. Confusion matrix for lymphography dataset

Table 4. The model performance results for each class individually on the lymphography dataset

Class	Acc	Rec	Prec	F1-S	AUC
Normal	1.000	1.000	1.000	1.000	1.000
Metastases	0.844	0.840	0.875	0.857	0.922
Malign lymph	0.844	0.833	0.789	0.811	0.936
Fibrosis	1.000	1.000	1.000	1.000	1.000

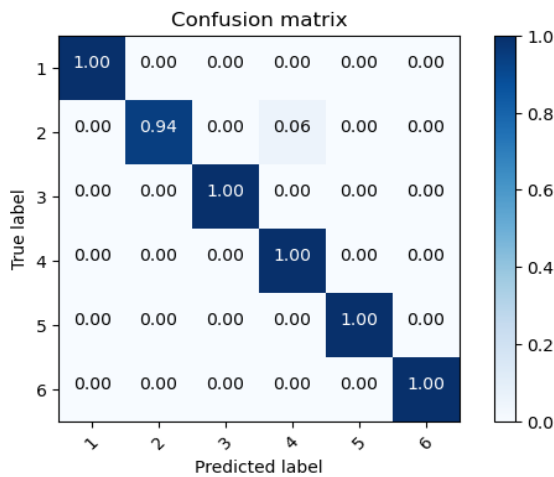


Figure 9. Confusion matrix for dermatology dataset

Table 5. The model performance results for each class individually on the dermatology dataset

Class	Acc	Rec	Prec	F1-S	AUC
Psoriasis	1.000	1.000	1.000	1.000	1.000
Seboric	0.991	0.994	1.000	0.971	0.998
Lichen	1.000	1.000	1.000	1.000	1.000
Pityriasis	0.991	1.000	0.938	0.968	0.999
Cronic	1.000	1.000	1.000	1.000	1.000
Pirubra	1.000	1.000	1.000	1.000	1.000

The above is a visualization of the performance model confusion matrix in classifying dermatology data as shown in Figure 9 and Table 5 presents the individual class performance of the multiclass logistic regression model on the lymphography dataset.

The following is a visualization of the performance model confusion matrix in classifying dermatology data as shown in Figure 10 and Table 6 presents the individual class performance of the multiclass logistic regression model on the ecoli dataset:

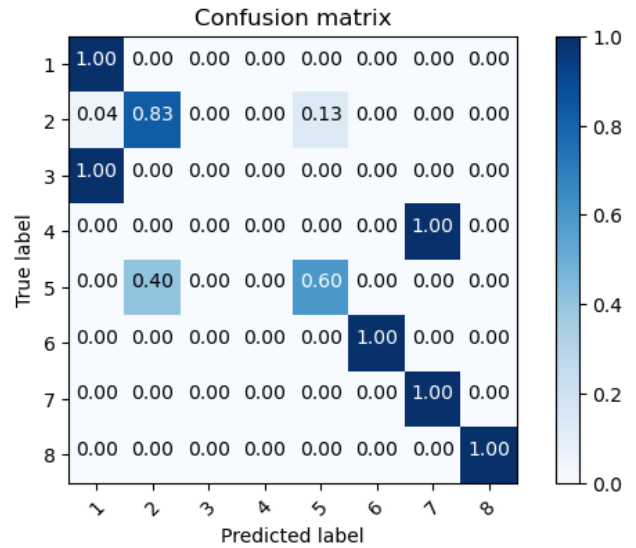


Figure 10. Confusion matrix for ecoli dataset

Table 6. The model performance results for each class individually on the ecoli dataset

Class	Acc	Rec	Prec	F1-S	AUC
cp	0.980	1.000	0.956	0.977	1.000
im	0.921	0.826	0.826	0.826	0.974
imL	0.990	0.000	0.000	0.000	0.920
imS	0.990	0.000	0.000	0.000	0.800
imU	0.931	0.600	0.667	0.632	0.970
om	1.000	1.000	1.000	1.000	1.000
omL	0.990	1.000	0.500	0.667	1.000
pp	1.000	1.000	1.000	1.000	1.000

3.4.2 The performance results of the multiclass logistic regression model without PCA

Overall, the results indicate that the Logistic Regression model performs well on these medical datasets, achieving high accuracy and demonstrating strong discriminative power. The model shows good performance in correctly classifying instances across various datasets without utilizing dimensionality reduction through PCA. Table 7 presents the performance results of the multiclass logistic regression model without PCA on all medical datasets.

Table 7. The performance results of the multiclass logistic regression model without PCA against all medical datasets

Datasets	Acc	Prec	Rec	F1-S	AUC
Thyroid	98.46%	98.58%	98.46%	98.42%	99.92%
Lymphography	86.66%	86.76%	86.66%	86.53%	91.90%
Dermatology	97.22%	97.36%	97.22%	97.21%	99.93%
Ecoli	99.09%	88.63%	99.09%	89.27%	98.76%

3.4.3 The performance results of the multiclass logistic regression model with PCA

Looking at the results, we can observe that the model achieved high performance on most datasets. The thyroid dataset shows excellent accuracy, precision, recall, f1-score, and AUC, indicating the model's ability to classify instances accurately and effectively. Table 8 presents the performance results of the multiclass logistic regression model with PCA against all medical datasets.

Table 8. The performance results of the multiclass logistic regression model with PCA against all medical datasets

Datasets	Acc	Prec	Rec	F1-S	AUC
Thyroid	98.46%	98.58%	98.46%	98.42%	100%
Lymphography	84.44%	84.63%	84.44%	84.49%	92.75%
Dermatology	99.07%	99.13%	99.07%	99.07%	99.94%
Ecoli	99.09%	88.37%	99.09%	89.11%	98.84%

The lymphography dataset has slightly lower performance compared to the others, especially in terms of accuracy and f1-score. This may be due to the dataset's higher degree of class imbalance, making it more challenging for the model to correctly classify minority classes. On the other hand, the dermatology dataset demonstrates exceptional performance across all metrics, showcasing the model's strong ability to handle imbalanced multiclass data effectively. The ecoli dataset also exhibits high performance, with particularly high accuracy, recall, and AUC values. However, it shows a relatively lower precision and f1-score, indicating some challenges in avoiding false positives.

Overall, the results indicate that the multiclass logistic regression model with PCA performs well on these medical datasets, achieving high accuracy and demonstrating strong discriminative power. However, further optimizations may be necessary to address the challenges posed by class imbalance, especially on datasets with more pronounced imbalances (Figures 11-15).

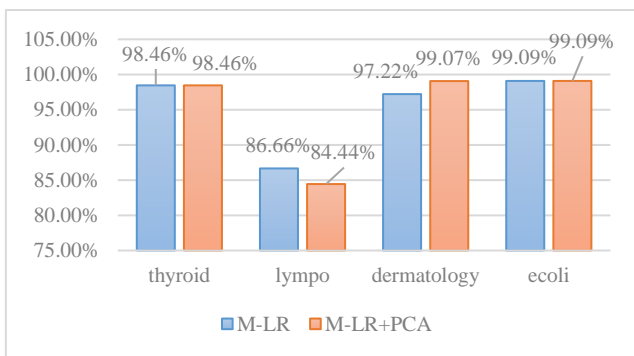


Figure 11. Result of accuracy measurement

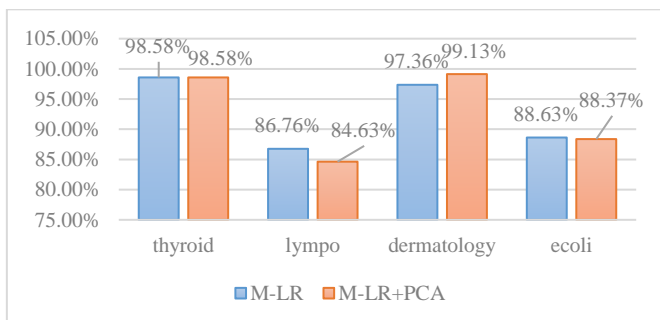


Figure 12. Result of precision measurement

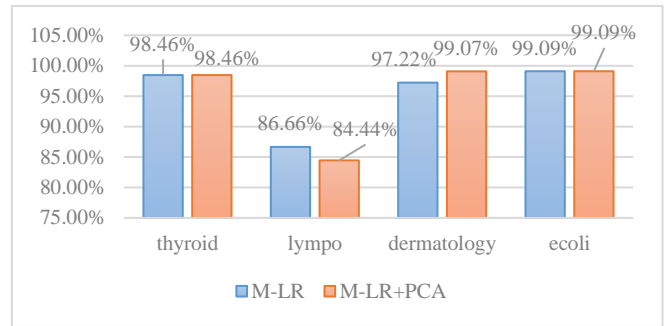


Figure 13. Result of recall measurement

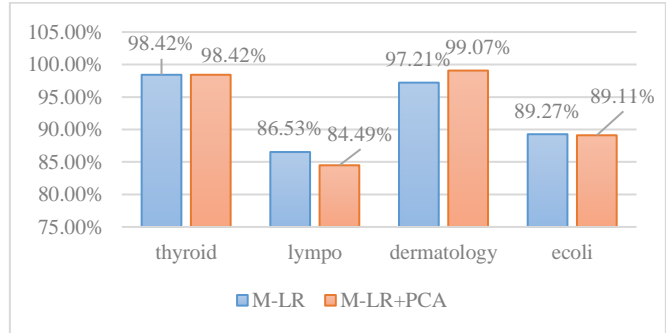


Figure 14. Result of f1-score measurement

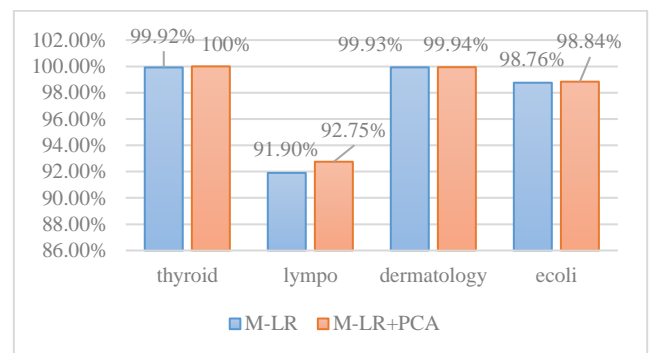


Figure 15. Result of AUC measurement

3.4.4 The results of the analysis compare the performance of Logistic Regression with and without PCA

The results show that both models generally perform well in classifying instances for all datasets, as indicated by high accuracy values. The Logistic Regression model with PCA achieved a perfect AUC score of 100% for the thyroid dataset, indicating excellent discriminative power in distinguishing different classes. For the thyroid, dermatology, and ecoli datasets, the Logistic Regression model with PCA achieved similar or slightly better performance in terms of accuracy, precision, recall, and f1-score compared to the model without PCA.

However, for the lymphography dataset, the Logistic Regression model without PCA outperformed the model with PCA in all performance metrics, with notably higher accuracy and AUC values. The results suggest that PCA can be beneficial for some datasets by reducing dimensionality and capturing most of the variance in the data, leading to improved performance. However, for datasets with less pronounced imbalances or complexities, the model without PCA can still achieve competitive results. Overall, both models demonstrate robust performance, and the choice of whether to use PCA depends on the specific dataset and its characteristics.

3.4.5 Comparison with related classification models

Table 9 shows the results of the multiclass logistic regression comparison that was developed with several related classification model developments from previous studies using the same dataset. It should be emphasized that some studies carried out a balanced strategy on the proposed method, but in this study, we compared it to their method without using a balanced strategy.

The performance of our logistic regression models in different datasets showed promising results, surpassing some of the best-performing models reported in related studies. In the Thyroid dataset, our M-LR and M-LR+PCA models achieved an impressive accuracy of 98.46%, outperforming the models by Febriantono et al. using C5.0 and C5.0+PSO. Similarly, our models outperformed the ANN, CatBoost, and XGBoost models proposed by Islam et al. Possible reasons for our models' superior performance lie in the effectiveness of logistic regression for handling imbalanced multiclass data and the feature enhancement provided by PCA.

In the Lymphography dataset, our M-LR model achieved an accuracy of 86.66%, slightly outperforming the best model by Pathan et al. using Robust Classifiers. However, our M-LR+PCA model's accuracy was slightly lower at 84.44%. The dataset's characteristics and class distribution may have influenced the results. Nonetheless, our logistic regression models demonstrated good performance, highlighting their

robustness for multiclass medical datasets.

For the Dermatology dataset, our M-LR model achieved an accuracy of 97.22%, performing better than the CBF and Random Forest models proposed by Prasetyowati et al. [29]. Moreover, our M-LR+PCA model achieved the highest accuracy of 99.07%. This demonstrates the effectiveness of logistic regression for classifying dermatology data, and PCA further improved the model's ability to capture relevant information, as evident from the higher accuracy achieved with M-LR+PCA.

In the Ecoli dataset, both our logistic regression models (M-LR and M-LR+PCA) achieved an outstanding accuracy of 99.09%, significantly outperforming the k-NN model. This further emphasizes the suitability of logistic regression for handling multiclass imbalanced datasets. The incorporation of PCA helped in reducing dimensionality without significant information loss, leading to enhanced performance.

In conclusion, our logistic regression models, especially when combined with PCA, demonstrated competitive and robust performance compared to other models in related studies. The effectiveness of logistic regression for handling imbalanced multiclass data and the dimensionality reduction provided by PCA contributed to the superior results. However, it is important to consider dataset characteristics and class distributions as potential factors influencing model performance.

Table 9. Comparison between other related classification models

Dataset: thyroid					
Authors	Model	Acc	Prec	Rec	F1-S
Febriantono et al. [9]	C5.0	94.42%	-	-	-
	C5.0+PSO	94.42%	-	-	-
	C5.0+PSO+META	95.81%	-	-	-
Islam et al. [27]	ANN	95.87%	95.70%	95.90%	95.70%
	CatBoost	95.38%	95.50%	95.38%	95.38%
	XGBoost	95.33%	95.39%	95.33%	95.32%
Our Works	M-LR	98.46%	98.58%	98.46%	98.42%
	M-LR+PCA	98.46%	98.58%	98.46%	98.42%
Dataset: lymphography					
Authors	Model	Acc	Prec	Rec	F1-S
Febriantono et al. [9]	C4.5+PSO+META	83.33%	-	-	-
	C5.0+PSO+META	83.33%	-	-	-
Pathan et al. [28]	Robust Classifiers	85.00%	-	-	-
Our Works	M-LR	86.66%	86.76%	86.66%	86.53%
	M-LR+PCA	84.44%	84.63%	84.44%	84.49%
Dataset: dermatology					
Authors	Model	Acc	Prec	Rec	F1-S
Prasetyowati et al. [29]	CBF	94.92%	94.93%	94.92%	94.92%
	Random Forest	97.01%	96.90%	96.91%	96.90%
Our Works	M-LR	97.22%	97.36%	97.22%	97.21%
	M-LR+PCA	99.07%	99.13%	99.07%	99.07%
Dataset: ecoli					
Authors	Model	Acc	Prec	Rec	F1-S
Nababan et al. [30]	k-NN	75.94%	48.04%	42.83%	44.45%
Our Works	M-LR	99.09%	88.63%	99.09%	89.27%
	M-LR+PCA	99.09%	88.37%	99.09%	89.11%

4. CONCLUSIONS

The amalgamation of logistic regression and PCA demonstrated robust performance, standing competitive in

comparison to other models examined in related studies. The efficacy of logistic regression in handling imbalanced multiclass data and the role of PCA in dimensionality reduction served as pivotal factors propelling the models'

superior performance. However, the characteristics of the datasets and class distributions were noted to influence the models' outcomes. Collectively, the results accentuate the potential of logistic regression and PCA to perform accurate classification and prediction of diseases in healthcare applications.

In this investigation, the devised models consistently surpassed some of the best-performing models documented in related studies across various datasets. This was evidenced by achieving high scores in accuracy, precision, recall, and f1-score, reinforcing their robustness in precise classification of medical data. This positions them as invaluable tools for disease prediction and decision-making within the healthcare domain.

The objectives of addressing the issue of imbalanced multiclass medical data and enhancing logistic regression via PCA were successfully fulfilled. The integration of PCA with logistic regression effectively managed imbalanced data while preserving crucial information through dimensionality reduction.

Through our experiments and analysis, two salient observations were discerned. Firstly, logistic regression emerged as a formidable candidate for imbalanced multiclass medical datasets, delivering reliable and accurate classification. Secondly, PCA was instrumental in elevating the model's performance by selecting pertinent and uncorrelated features, thereby improving accuracy and interpretability. These findings bear considerable implications for the medical sector, as accurate disease prediction and classification can contribute to enhanced patient outcomes and more astute clinical decision-making.

Certain assumptions were made during the study to facilitate the analysis. Firstly, the selected medical datasets were presupposed to mirror real-world scenarios apt for evaluating the logistic regression model integrated with PCA. Secondly, effective preprocessing steps, including the handling of missing values and outliers, were undertaken under the assumption of not significantly impacting the model's performance. Despite the encouraging results, certain limitations persist. The performance of the logistic regression model combined with PCA heavily relies on the selected number of principal components, which might necessitate adjustments for optimal performance with different datasets. Furthermore, the scalability to larger datasets and the model's behavior under varying practical scenarios call for additional exploration.

This study paves the way for future research. The exploration of different feature selection techniques in conjunction with logistic regression could further enhance model performance. Investigating alternative machine learning algorithms, such as ensemble methods and deep learning models, may potentially improve classification accuracy and address class imbalance. Conducting experiments on larger and more diverse medical datasets can validate the generalizability of the proposed approach. The incorporation of advanced medical features and domain-specific knowledge can augment disease prediction models. An extension of the proposed model to handle time-series medical data would enable the prediction and monitoring of medical conditions over time.

REFERENCES

[1] Ali, H., Salleh, M.N.M., Saedudin, R., Hussain, K.,

Mushtaq, M.F. (2019). Imbalance class problems in data mining: A review. *Indonesian Journal of Electrical Engineering and Computer Science*, 14(3): 1560-1571. <https://doi.org/10.11591/ijeecs.v14.i3.pp1552-1563>

[2] Casalino, G., Castellano, G., Zaza, G. (2020). A mHealth solution for contact-less self-monitoring of blood oxygen saturation. In *2020 IEEE Symposium on Computers and Communications (ISCC)*, pp. 1-7. <https://doi.org/10.1109/ISCC50000.2020.9219718>

[3] Ghorbani, R., Ghousi, R., Makui, A., Atashi, A. (2020). A new hybrid predictive model to predict the early mortality risk in intensive care units on a highly imbalanced dataset. *IEEE Access*, 8: 141066-141079. <https://doi.org/10.1109/ACCESS.2020.3013320>

[4] Liu, N., Li, X., Qi, E., Xu, M., Li, L., Gao, B. (2020). A novel ensemble learning paradigm for medical diagnosis with imbalanced data. *IEEE Access*, 8: 171263-171280. <https://doi.org/10.1109/ACCESS.2020.3014362>

[5] Fernández, A., García, S., Galar, M., Prati, R.C. (2018). *Learning from Imbalanced Data Sets*, Springer.

[6] Liu, C.L., Hsieh, P.Y. (2019). Model-based synthetic sampling for imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 32(8): 1543-1556. <https://doi.org/10.1109/TKDE.2019.2905559>

[7] Gajowniczek, K., Ząbkowski, T. (2021). ImbTreeEntropy and ImbTreeAUC: Novel R packages for decision tree learning on the imbalanced datasets. *Electronics*, 10(6): 657. <https://doi.org/10.3390/electronics10060657>

[8] Sitompul, O.S., Nababan, E.B. (2018). Biased support vector machine and weighted-smote in handling class imbalance problem. *International Journal of Advances in Intelligent Informatics*, 4(1): 21-27. <https://doi.org/10.26555/ijain.v4i1.146>

[9] Febriantono, M.A., Pramono, S.H., Rahmadwati, R., Naghdy, G. (2020). Classification of multiclass imbalanced data using cost-sensitive decision tree C5.0. *IAES International Journal of Artificial Intelligence*, 9(1): 65-72. <https://doi.org/10.11591/ijai.v9.i1.pp65-72>

[10] Santoso, N., Wibowo, W., Himawati, H. (2019). Integration of synthetic minority oversampling technique for imbalanced class. *Indonesian Journal of Electrical Engineering and Computer Science*, 13(1): 102-108. <https://doi.org/10.11591/ijeecs.v13.i1.pp102-108>

[11] Mahadevan, A., Arock, M. (2021). A class imbalance-aware review rating prediction using hybrid sampling and ensemble learning. *Multimedia Tools and Applications*, 80: 6911-6938. <https://doi.org/10.1007/s11042-020-10024-2>

[12] Salas-Eljatib, C., Fuentes-Ramirez, A., Gregoire, T.G., Altamirano, A., Yaitul, V. (2018). A study on the effects of unbalanced data when fitting logistic regression models in ecology. *Ecological Indicators*, 85: 502-508. <https://doi.org/10.1016/j.ecolind.2017.10.030>

[13] Ustyannie, W., Suprpto, S. (2020). Oversampling method to handling imbalanced datasets problem in binary logistic regression algorithm. *Indonesian Journal of Computing and Cybernetics Systems*, 14(1): 1-10. <https://doi.org/10.22146/ijccs.37415>

[14] Mienye, I.D., Sun, Y. (2021). Performance analysis of cost-sensitive learning methods with application to imbalanced medical data. *Informatics in Medicine Unlocked*, 25: 100690. <https://doi.org/10.1016/j.imu.2021.100690>

- [15] Nugroho, W.H., Handoyo, S., Akri, Y.J., Sulistyono, A.D. (2022). Building multiclass classification model of logistic regression and decision tree using the chi-square test for variable selection method. *Journal of Hunan University Natural Sciences*, 49(4): 172-181. <https://doi.org/10.55463/issn.1674-2974.49.4.17>
- [16] Reverdy, P., Leonard, N.E. (2015). Parameter estimation in Softmax decision-making models with linear objective functions. *IEEE Transactions on Automation Science and Engineering*, 13(1): 54-67. <https://doi.org/10.1109/TASE.2015.2499244>
- [17] Zhu, R., Wang, Z., Ma, Z., Wang, G., Xue, J.H. (2018). LRID: A new metric of multi-class imbalance degree based on likelihood-ratio test. *Pattern Recognition Letters*, 116: 36-42. <https://doi.org/10.1016/j.patrec.2018.09.012>
- [18] Jassim, M.A., Abdulwahid, S.N. (2021). Data mining preparation: Process, techniques and major issues in data analysis. *IOP Conference Series: Materials Science and Engineering*, 1090(1): 012053. <https://doi.org/10.1088/1757-899x/1090/1/012053>
- [19] de Amorim, L.B., Cavalcanti, G.D., Cruz, R.M. (2023). The choice of scaling technique matters for classification performance. *Applied Soft Computing*, 133: 109924. <https://doi.org/10.1016/j.asoc.2022.109924>
- [20] Huang, T., Li, J., Zhang, W. (2020). Application of principal component analysis and logistic regression model in lupus nephritis patients with clinical hypothyroidism. *BMC Medical Research Methodology*, 20(1): 1-7. <https://doi.org/10.1186/s12874-020-00989-x>
- [21] Nasution, M.Z.F., Sitompul, O.S., Ramli, M. (2018). PCA based feature reduction to improve the accuracy of decision tree c4. 5 classification. *Journal of Physics: Conference Series*, 978(1): 012058. <https://doi.org/10.1088/1742-6596/978/1/012058>
- [22] Härdle, W.K., Simar, L. (2019). *Applied multivariate statistical analysis*.
- [23] Firdausanti, N.A., Ningrum, R.A., Qomariyah, S. (2022). Comparisons of logistic regression and support vector machines in classification of echocardiogram dataset. *Inferensi*, 5(2): 85-90. <https://doi.org/10.12962/j27213862.v5i2.14121>
- [24] Chiu, I.M., Zeng, W.H., Cheng, C.Y., Chen, S.H., Lin, C.H.R. (2021). Using a multiclass machine learning model to predict the outcome of acute ischemic stroke requiring reperfusion therapy. *Diagnostics*, 11(1): 80. <https://doi.org/10.3390/diagnostics11010080>
- [25] Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N., Asadpour, M. (2020). Boosting methods for multi-class imbalanced data classification: An experimental review. *Journal of Big Data*, 7: 1-47. <https://doi.org/10.1186/s40537-020-00349-y>
- [26] Grandini, M., Bagli, E., Visani, G. (2020). Metrics for multi-class classification: An overview. *arXiv preprint arXiv:2008.05756*. <http://arxiv.org/abs/2008.05756>
- [27] Islam, S.S., Haque, M.S., Miah, M.S.U., Sarwar, T.B., Nugraha, R. (2022). Application of machine learning algorithms to predict the thyroid disease risk: An experimental comparative study. *PeerJ Computer Science*, 8: e898. <https://doi.org/10.7717/PEERJ-CS.898>
- [28] Pathan, S., Rao, D., Kumar, P. (2022). Lymph node morbidity diagnosis using multiclass machine learning models. In *2022 6th International Conference on Green Technology and Sustainable Development (GTSD)*, pp. 1173-1176. <https://doi.org/10.1109/GTSD54989.2022.9989185>
- [29] Prasetyowati, M.I., Maulidevi, N.U., Surendro, K. (2022). The accuracy of random forest performance can be improved by conducting a feature selection with a balancing strategy. *PeerJ Computer Science*, 8: e1041. <https://doi.org/10.7717/peerj-cs.1041>
- [30] Nababan, A.A., Sutarman, S., Zarlis, M., Nababan, E.B. (2023). Improving the accuracy of k-nearest neighbor (k-NN) using synthetic minority oversampling technique (SMOTE) and gain ratio (GR) for imbalanced class data. *AIP Conference Proceedings*, 2714(1). <https://doi.org/10.1063/5.0128413>