ICIC Express Letters
Volume **18**, Number **6**, June **2024**

 ISSN 1881-803X
pp. 575–585

IMBALANCED MEDICAL DATA ANALYSIS: DEVELOPING A MULTICLASS LOGISTIC REGRESSION CLASSIFICATION MODEL WITH PCA APPROACH

Adli Abdillah Nababan[1], Sutarman[2,*], Muhammad Zarlis[3] and Erna Budhiarti Nababan[3,4]

[1]Doctoral Program of Computer Science, Faculty of Computer Science and Information Technology
[2]Department of Mathematics, Faculty of Mathematics and Natural Sciences
[4]Department of Computer Science, Faculty of Computer Science and Information Technology
Universitas Sumatera Utara
Medan, North Sumatra 20155, Indonesia
adli.nababan@students.usu.ac.id; ernabrn@usu.ac.id
*Corresponding author: sutarman@usu.ac.id

[3]Department of Information Systems Management, BINUS Graduate Program – Master of Information Systems Management
Bina Nusantara University
West Jakarta, DKI Jakarta 11480, Indonesia
muhammad.zarlis@binus.edu

Received June 2023; accepted September 2023

Abstract. *Imbalanced datasets are prevalent in various machine learning applications, particularly in medical data, where a majority of instances represent healthy patients, while diseases comprise the minority class. This study addresses the challenge of class imbalance in multiclass medical datasets by proposing modified classifiers based on multiclass logistic regression. The gradient descent optimization method is utilized to minimize the cost function in the training model. Principal Component Analysis (PCA) is employed as a preprocessing step to eliminate noise and irrelevant information. Resampling techniques are avoided to preserve the original data distribution. Performance evaluations are conducted on four well-known multiclass medical datasets: thyroid, lymphography, dermatology, and ecoli. Comparative analysis with previous studies using the same datasets demonstrates the exceptional performance of the multiclass logistic regression model. The proposed approach achieves efficient medical diagnosis prediction by effectively discriminating between minority and majority classes in imbalanced datasets. The results highlight the superiority of the multiclass logistic regression model over other algorithms, emphasizing its potential for accurate diagnosis prediction in medical applications.*
Keywords: Classification, Medical data, Logistic regression, Multiclass, Optimization, PCA, Diagnosis, Prediction

1. **Introduction.** The era of digital data has witnessed a remarkable surge in database volumes, which poses a significant challenge in effectively harnessing this wealth of information. Data is commonly categorized into two main groups: binary data with two classes and multiclass data with more than two classes [1]. The issue of class imbalance is a noteworthy concern in both binary and multiclass datasets [2]. Typically, the majority class, with a larger number of instances, garners more attention, leaving the minority class relatively neglected due to its smaller size. Nonetheless, it is essential to recognize the importance of the minority class, as it may contain valuable insights, such as unique disease diagnoses [3]. Addressing the imbalanced nature of these datasets is pivotal for unlocking the concealed knowledge residing within the underrepresented minority classes.

DOI: 10.24507/icicel.18.06.575

The convergence of healthcare technology with Machine Learning (ML) has brought about a significant transformation in disease prediction, patient monitoring, and clinical decision-making. This advancement has resulted in improved patient outcomes and elevated healthcare standards [4]. ML algorithms have provided medical professionals with the ability to harness vast patient datasets, uncover underlying patterns, and make informed decisions. This transformative influence underscores the need for continuous research and development efforts to tackle the challenge of inaccurate disease prediction, as inaccuracies can pose serious risks to patient safety. Despite the wealth of medical data available, it is essential to focus on refining and improving disease prediction models to mitigate the potential harm associated with inaccuracies. Giving paramount importance to ongoing research and development endeavors is critical to achieving precise disease prediction, ultimately leading to optimized patient outcomes and reduced risks [5].

The issue of class imbalance has garnered considerable attention in Machine Learning (ML) research. Class imbalance is characterized by a substantial disparity between the majority class and the minority class, with the majority class significantly outweighing the minority class [6,7]. This imbalance is particularly pronounced in medical data, where the number of healthy patients (majority class) greatly surpasses the number of sick patients (minority class) [8]. Traditional ML algorithms, primarily designed for binary classification tasks with balanced class distributions in mind, face challenges when dealing with imbalanced data. When trained on such imbalanced datasets, these algorithms tend to exhibit bias towards the majority class, resulting in a decline in their performance. Addressing class imbalance is recognized as one of the ten significant challenges in machine learning research [9].

In the realm of machine learning classification tasks, the misclassification of minority classes is a common issue arising from the disproportionate focus on the majority classes, often leading to the neglect of minority classes [10]. Researchers have explored various techniques to address imbalanced data problems [11]. Among these challenges, classifying multiclass imbalanced data is notably more intricate compared to binary imbalanced data because it involves multiple minority classes [12].

Numerous approaches and techniques have been developed to tackle the challenges posed by imbalanced multiclass data [13], including data-level strategies, algorithm-level methods, and cost-sensitive approaches [14]. In the context of addressing imbalanced datasets, one data-level strategy involves utilizing the Synthetic Minority Oversampling Technique (SMOTE), initially introduced by [15]. The preference for oversampling over undersampling is driven by the concern that undersampling might lead to the loss of crucial information [16]. Nevertheless, it is essential to acknowledge that oversampling can introduce the risk of overfitting [17]. In light of the challenges posed by imbalanced datasets and the need to evaluate instance or classifier significance through complexity measurements [18], there is a pressing call to explore innovative techniques. These techniques should not only excel at classifying minority classes but also consider the majority classes to achieve optimal results and ensure accurate classification across all classes.

One such technique that can effectively address the issue of data imbalance is logistic regression. Logistic regression provides an efficient means to model the relationship between the dependent variable and multiple classes. It achieves this by utilizing a linear combination of independent variables, without the necessity of employing oversampling methods. Unlike traditional logistic regression that applies to binary cases [19-21], multiclass logistic regression can predict probabilities for different classes in each observation. This model employs the softmax function to generate expected class probabilities, transforming the model's output into a probability distribution indicating the likelihood of each class [22]. The aim of multiclass logistic regression is to estimate parameters that minimize prediction errors [23]. This is achieved by determining optimal weights for each independent variable, enabling accurate predictions for the correct class. To enhance this

process, we introduce the concept of restricting the number of features. Under this constraint, we explore all subsets of features with the specified number, applying ordinary logistic regression to each subset. The objective is to identify the subset among all possible combinations that exhibits the highest performance, as measured by accuracy in this study. This approach optimizes feature selection, refining parameter estimation in multiclass logistic regression, ultimately reducing prediction errors and bolstering the model's predictive capacity [24].

This research is dedicated to improving the performance of logistic regression models in the context of imbalanced multiclass medical data by reducing prediction errors. The primary goal is to devise strategies within the logistic regression framework that can adeptly handle the inherent class imbalance present in medical datasets.

To address this challenge, Principal Component Analysis (PCA) is employed as a critical tool. PCA is leveraged to transform the variables within these datasets into principal components, effectively capturing the majority of data variability while also identifying anomalies [25]. In recent times, PCA has gained prominence as a valuable component of classification models, with its primary objective being the elimination of irrelevant data to enhance predictive accuracy. The feature reduction framework plays a pivotal role in this research, serving as a means to streamline high-dimensional data into low-dimensional datasets containing uncorrelated attributes. This process aids in simplifying the data, making it more manageable and conducive to the accurate modeling of medical datasets with imbalanced multiclass distributions [26]. The optimization process employs gradient descent to optimize model parameters, aligning them with the training data. The study evaluates the performance of a logistic regression algorithm designed for imbalanced multiclass medical datasets and compares it to previous studies, providing insights into addressing this issue. In summary, our contributions include the recognition of the significance of class imbalance, a focus on enhancing disease prediction, the handling of multiclass imbalanced data, the utilization of logistic regression, the integration of Principal Component Analysis (PCA), and a rigorous performance evaluation. These elements collectively constitute the foundation of our research, showcasing its significance and potential impact on the field of machine learning in healthcare.

2. **Methodology.**

This methodology outlines the construction of a multiclass classification model using logistic regression and Principal Component Analysis (PCA) on imbalanced medical data. It begins by describing the medical datasets and proceeds to meticulously prepare the data, including handling missing values, outliers, and label encoding. PCA is then employed to reduce data dimensionality while preserving vital information, enhancing processing efficiency and minimizing overfitting risk. To ensure modeling consistency, the dataset is split into 70% for training and 30% for testing, with uniform data scaling applied to both subsets. Multiclass logistic regression is chosen for its suitability in multiclass scenarios, utilizing the softmax function for precise classification. The model parameters are iteratively optimized using gradient descent, aligning the model with ordinal encoded labels. Evaluation metrics such as accuracy, precision, recall, F1-score, and ROC AUC Score are employed to assess model effectiveness, especially considering imbalanced class distributions. This structured approach ensures precision and consistency throughout the research process, yielding robust and reproducible results.

2.1. **Datasets.**

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 [13]. 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
Lympography	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
Lympography	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

The level of class imbalance in a dataset can be evaluated using the Imbalance Ratio (IR) [27]. The IR is calculated by comparing the number of instances in the majority class to the number of instances in the minority class. A higher IR indicates a more pronounced imbalance within the dataset. On the other hand, in multiclass scenarios, the Imbalance Degree (ID) is used to measure the relative imbalance between the majority and minority classes, based on the percentage of occurrences in each class [9].

A higher ID signifies a greater imbalance within the dataset. These metrics provide insights into the degree of class imbalance and its impact on the learning of models. Particularly in classification tasks, datasets with a higher ID can present challenges due to the bias towards the majority class and the difficulties in detecting minority classes. It is important to note that datasets with a larger number of minority classes tend to exhibit a higher level of imbalance [28], especially in medical datasets.

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 [29]. 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 [30].

2.3. **Principal Component Analysis (PCA).** Data dimension reduction is performed using Principal Component Analysis (PCA) to reduce the number of features or variables in a data set without losing important information contained in the data. This aims to speed up processing time, reduce model complexity, and prevent overfitting. Feature reduction is employed to transform high-dimensional data into a lower-dimensional format with uncorrelated attributes. This study proposes a PCA framework to select a subset of relevant and uncorrelated features. The study considers Principal Component Analysis (PCA) for feature savings to select uncorrelated features [25]. The variance is calculated to find the spread of the medical dataset using Equation (1) to determine the deviation of data in the sample dataset.

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

where $\mathrm{Var}(x)$ = the variance of variable x; $\tilde{Z}_{ij}$ = the value of the ith data point of variable x or the jth feature; μ_j = the mean value of the jth 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 [26]. The covariance is calculated using Equation (2).

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

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

Finally, the Eigenvalues and Eigenvectors for the covariance matrices are calculated [31]. The Eigenvalues are then transformed (varimax orthogonal rotation) using Equation (3).

$$\mathrm{Det}(A-\lambda I) = 0 \tag{3}$$

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. **Building multiclass logistic regression.** To build a classification model, the initial step involves establishing the class boundary that will differentiate instances belonging to different classes [32]. 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 Equation (4) [24].

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

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, defined as Equation (5).

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

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.

The softmax function is used in multiclass logistic regression to calculate the predicted probability of the input x belonging to class j, denoted as $P(y=j|x)$. θ_j represents the weight vector specific to class j, and K represents the total number of classes. The softmax function computes the exponentiated score of each class's weight vector and normalizes it by the sum of all exponentiated scores across all classes. This normalization ensures that the predicted probabilities for all classes sum up to 1. Therefore, the softmax function provides a probability distribution over the possible classes for a given input x in multiclass logistic regression.

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 Equation (6).

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

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

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

The objective is to determine the optimal values for the weight parameters θ that maximize the log-likelihood function. To achieve this, different optimization algorithms, such as gradient descent, can be employed. These algorithms aim to estimate the parameters that provide the best fit to the data and maximize the likelihood of the observed ordinal encoded labels.

2.5. **Finding optimal values for model parameters.** After deriving the likelihood function for multiclass logistic regression with ordinal encoder vectors, the next step is to optimize the model parameters using gradient descent. Gradient descent is an iterative optimization algorithm that aims to find the optimal values for the weight parameters, maximizing the likelihood function. It updates the parameters iteratively in the direction of the steepest descent, gradually approaching the optimal solution. Here are the general steps for performing gradient descent in the context of multiclass logistic regression with ordinal encoder vectors.

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 Equation (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 Equation (8):

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

where y_{ij} represents an indicator variable, taking the value of 1 when the ordinal encoded label for sample i is j, and otherwise taking the value of 0.

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 Equation (9):

$$\theta = \theta - \alpha \nabla_\theta \text{Loss} \tag{9}$$

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 [1]. 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, are employed [33]. 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 [12]. 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 (i) collecting data, (ii) preprocessing, (iii) conducting the training and testing process, and (iv) validation. These methodological steps will be applied to addressing the problem of imbalanced multiclass medical data.

Step 1. Input the multiclass medical dataset for processing.

Step 2. Replace the missing values and remove outliers from the dataset and encode the labels.

Step 3. Apply PCA to reducing the dimensionality of the data. Find the principal components by considering the eigenvalues and select the top ones that collectively account for 98% of the variance in the data.

Step 4. The dataset will be divided into two, namely 70% for training and 30% for testing where the training data and test data will be scaled.

Step 5. Train the model using the multiclass logistic regression method.

Step 6. Use gradient descent to optimize the model parameters (weights and bias) by minimizing the loss function in the multiclass logistic regression model.

Step 7. Evaluate the performance of the trained model on both the training and testing datasets and compute metrics to assess the classification performance of the model.

This robust methodology ensures a systematic approach to addressing the intricacies of imbalanced multiclass medical data classification.

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 performance results of the proposed method for all medical datasets can be seen in Table 3.

TABLE 3. Performance results of the proposed method against all medical datasets

Datasets	Accuracy		Precision		Recall		F1-score		AUC	
	LG	LG+PCA	LG	LG+PCA	LG	LG+PCA	LG	LG+PCA	LG	LG+PCA
Thyroid	98.46%	98.46%	98.58%	98.58%	98.46%	98.46%	98.42%	98.42%	99.92%	100%
Lymphography	86.66%	84.44%	86.76%	84.63%	86.66%	84.44%	86.53%	84.49%	91.90%	92.75%
Dermatology	97.22%	99.07%	97.36%	99.13%	97.22%	99.07%	97.21%	99.07%	99.93%	99.94%
Ecoli	99.09%	99.09%	88.63%	88.37%	99.09%	99.09%	89.27%	89.11%	98.76%	98.84%

In our evaluation, we conducted a thorough examination of the classification performance of Logistic Regression (LG) and Logistic Regression with Principal Component Analysis (LG+PCA) across four diverse datasets: thyroid, lymphography, dermatology, and ecoli. Across the board, both LG and LG+PCA consistently delivered strong results, boasting high accuracy, precision, recall, F1-score, and AUC values. However, the key distinction emerged in the impact of PCA on these datasets. While LG alone performed exceptionally well in the thyroid and ecoli datasets, PCA did not significantly enhance the results, suggesting that the original features already contained sufficient discriminatory information. In contrast, the dermatology dataset saw a modest improvement with PCA, indicating its potential for enhancing classification in specific cases. Interestingly, the lymphography dataset displayed a slight dip in metrics with PCA, highlighting the importance of careful consideration when employing dimensionality reduction techniques, as their effects can vary depending on the dataset's characteristics.

In summary, the choice of whether to implement PCA should be driven by a comprehensive understanding of the dataset's unique properties and the specific classification goals. Table 4 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 presented table offers a comprehensive comparison of machine learning models applied to four distinct medical datasets, shedding light on their classification performance. In the case of the thyroid dataset, our proposed method, Multiclass Logistic Regression (M-LR), exhibited remarkable accuracy, reaching 98.46%, outperforming alternative approaches. Moreover, M-LR showcased strong precision, recall, and F1-score, affirming its capability to accurately categorize thyroid cases. Notably, the inclusion of Principal Component Analysis (PCA) as M-LR+PCA did not significantly alter the results, emphasizing the potency of logistic regression in this context. Shifting to the lymphography dataset, our M-LR model achieved an accuracy of 86.66%, surpassing certain prior methods. However, the introduction of PCA (M-LR+PCA) resulted in a minor drop in accuracy, precision, recall, and F1-score, suggesting that, for this dataset, logistic regression alone delivered superior performance. The dermatology dataset yielded interesting findings, with our M-LR model boasting an accuracy of 97.22%, making it competitive with prior approaches. However, what stands out is the substantial enhancement seen when PCA was incorporated (M-LR+PCA), which elevated precision to 99.13% and recall to 99.07%, culminating in an accuracy of 99.07%. These outcomes underscore the role of PCA in improving classification performance for dermatology data. Finally, in the ecoli dataset, our M-LR and M-LR+PCA models both delivered exceptional accuracy at 99.09%, surpassing prior work. Furthermore, precision and recall scores indicated robust classification performance, emphasizing the effectiveness of our approach for accurate classification in this context.

In summary, our findings demonstrate the consistency of our proposed M-LR method in delivering strong classification results across diverse medical datasets. Logistic regression, in particular, proves to be a potent tool for addressing imbalanced multiclass medical

Table 4. Comparison between other related classification models

Dataset: thyroid					
Authors	**Model**	**Accuracy**	**Precision**	**Recall**	**F1-score**
Febriantono et al. (2020) [13]	C5.0	94.42%	–	–	–
	C5.0+PSO	94.42%	–	–	–
	C5.0+PSO+META	95.81%	–	–	–
Islam et al. (2022) [34]	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**	**Accuracy**	**Precision**	**Recall**	**F1-score**
Febriantono et al. (2020) [13]	C4.5+PSO+META	83.33%	–	–	–
	C5.0+PSO+META	83.33%	–	–	–
Pathan et al. (2022) [35]	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**	**Accuracy**	**Precision**	**Recall**	**F1-score**
Prasetiyowati et al. (2022) [36]	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**	**Accuracy**	**Precision**	**Recall**	**F1-score**
Nababan et al. (2023) [37]	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%**

data. The impact of PCA, however, varies depending on the dataset, underlining the importance of tailored approaches to dimensionality reduction. Our work highlights the promise of M-LR for imbalanced multiclass medical data classification, both with and without PCA, offering valuable insights for future research in healthcare applications.

4. **Conclusions.** This study aims to develop a multiclass logistic regression classifier by utilizing PCA as part of feature reduction, then minimizing the cost function through the gradient descent optimization method. The dataset is divided into two parts, with 70% used for training and 30% for testing. Although the training and test data may have different scales, they have the same variance, allowing for accurate comparison during testing. The model's performance is evaluated using various metrics, including confusion metrics, accuracy, precision, recall, F1-score, and AUC. The experimental results demonstrate that the developed model performs well on four different datasets compared to other algorithms used in related research studies on the same dataset, without employing resampling techniques. These findings indicate the potential of logistic regression in predicting imbalanced medical data, particularly in multiclass scenarios. Future research efforts will focus on enhancing predictions for minority classes while ensuring the algorithm does not overlook the majority class. This is particularly important in medical

cases where accurate prediction of diseases necessitates considering the minority class. To improve the accuracy of predicting diseases, future research may explore cost-sensitive learning approaches and compare different resampling techniques individually. Additionally, combining feature selection, resampling, cost-sensitive learning methods, and other hybrid techniques could be explored as potential directions for handling imbalanced medical datasets in future research.

REFERENCES

[1] J. Tanha, Y. Abdi, N. Samadi, N. Razzaghi and M. Asadpour, Boosting methods for multi-class imbalanced data classification: An experimental review, *J. Big Data*, vol.7, no.1, DOI: 10.1186/s40537-020-00349-y, 2020.

[2] S. Narwane and S. Sawarkar, Machine learning and class imbalance: A literature survey, *Ind. Eng. J.*, vol.12, no.10, DOI: 10.26488/iej.12.10.1202, 2019.

[3] A. Nawaz, Y. Abbas, T. Ahmad, N. F. Mahmoud, A. Rizwan and N. M. A. Samee, A healthcare paradigm for deriving knowledge using online consumers' feedback, *Healthcare*, vol.10, https://api.semanticscholar.org/CorpusID:251776063, 2022.

[4] H. Habehh and S. Gohel, Machine learning in healthcare, *Curr. Genomics*, vol.22, no.4, pp.291-300, DOI: 10.2174/1389202922666210705124359, 2021.

[5] K. B. Johnson et al., Precision medicine, AI, and the future of personalized health care, *Clin. Transl. Sci.*, vol.14, no.1, pp.86-93, DOI: 10.1111/cts.12884, 2021.

[6] C. L. Liu and P. Y. Hsieh, Model-based synthetic sampling for imbalanced data, *IEEE Trans. Knowl. Data Eng.*, vol.32, no.8, pp.1543-1556, DOI: 10.1109/TKDE.2019.2905559, 2020.

[7] K. Gajowniczek and T. Ząbkowski, Imbtreeentropy and imbtreeauc: Novel R packages for decision tree learning on the imbalanced datasets, *Electron.*, vol.10, no.6, pp.1-23, DOI: 10.3390/electronics 10060657, 2021.

[8] N. Liu, X. Li, E. Qi, M. Xu, L. Li and B. Gao, A novel ensemble learning paradigm for medical diagnosis with imbalanced data, *IEEE Access*, vol.8, pp.171263-171280, DOI: 10.1109/ACCESS.2020. 3014362, 2020.

[9] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk and F. Herrera, *Learning from Imbalanced Data Sets*, Springer International Publishing, 2018.

[10] N. Santoso, W. Wibowo and H. Himawati, Integration of synthetic minority oversampling technique for imbalanced class, *Indones. J. Electr. Eng. Comput. Sci.*, vol.13, no.1, pp.102-108, DOI: 10.11591/ijeecs.v13.i1.pp102-108, 2019.

[11] Hartono, O. S. Sitompul, Tulus and E. B. Nababan, Biased support vector machine and weighted-SMOTE in handling class imbalance problem, *Int. J. Adv. Intell. Informatics*, vol.4, no.1, pp.21-27, DOI: 10.26555/ijain.v4i1.146, 2018.

[12] A. Mahadevan and M. Arock, A class imbalance-aware review rating prediction using hybrid sampling and ensemble learning, *Multimed. Tools Appl.*, vol.80, no.5, pp.6911-6938, DOI: 10.1007/s11042-020-10024-2, 2021.

[13] M. A. Febriantono, S. H. Pramono, Rahmadwati and G. Naghdy, Classification of multiclass imbalanced data using cost-sensitive decision tree C5.0, *IAES Int. J. Artif. Intell.*, vol.9, no.1, pp.65-72, DOI: 10.11591/ijai.v9.i1.pp65-72, 2020.

[14] A. Fernández, S. García, F. Herrera and N. V. Chawla, SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary, *J. Artif. Intell. Res.*, vol.61, pp.863-905, DOI: 10.1613/jair.1.11192, 2018.

[15] 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.6, pp.321-357, DOI: 10.1613/jair.953, 2002.

[16] F. Kamalov, Kernel density estimation based sampling for imbalanced class distribution, *Inf. Sci. (Ny).*, vol.512, pp.1192-1201, DOI: 10.1016/j.ins.2019.10.017, 2020.

[17] P. Vuttipittayamongkol and E. Elyan, Neighbourhood-based undersampling approach for handling imbalanced and overlapped data, *Inf. Sci. (Ny.)*, vol.509, pp.47-70, DOI: 10.1016/j.ins.2019.08.062, 2020.

[18] Hartono, S. Lestari, A. Rahmadsyah, R. M. F. Lubis and M. Gunawan, HAR-MI with COSTE in handling multi-class imbalance, *2020 8th Int. Conf. Cyber IT Serv. Manag. (CITSM 2020)*, pp.16-19, DOI: 10.1109/CITSM50537.2020.9268804, 2020.

[19] C. Salas-Eljatib, A. Fuentes-Ramirez, T. G. Gregoire, A. Altamirano and V. Yaitul, A study on the effects of unbalanced data when fitting logistic regression models in ecology, *Ecol. Indic.*, vol.85, no.9, pp.502-508, DOI: 10.1016/j.ecolind.2017.10.030, 2018.

[20] W. Ustyannie and S. Suprapto, Oversampling method to handling imbalanced datasets problem in binary logistic regression algorithm, *Indonesian J. Comput. Cybern. Syst. (IJCCS)*, vol.14, no.1, p.1, DOI: 10.22146/ijccs.37415, 2020.

[21] I. D. Mienye and Y. Sun, Performance analysis of cost-sensitive learning methods with application to imbalanced medical data, *Informatics Med. Unlocked*, vol.25, 100690, DOI: 10.1016/j.imu.2021.100690, 2021.

[22] W. H. Nugroho, S. Handoyo, Y. J. Akri and A. D. Sulistyono, Building multiclass classification model of logistic regression and decision tree using the chi-square test for variable selection method, *J. Hunan Univ. Nat. Sci.*, vol.49, no.4, pp.172-181, DOI: 10.55463/issn.1674-2974.49.4.17, 2022.

[23] P. Reverdy and N. E. Leonard, Parameter estimation in softmax decision-making models with linear objective functions, *IEEE Trans. Autom. Sci. Eng.*, vol.13, no.1, pp.54-67, DOI: 10.1109/TASE.2015.2499244, 2016.

[24] I. M. Chiu, W. H. Zeng, C. Y. Cheng, S. H. Chen and C. H. R. Lin, Using a multiclass machine learning model to predict the outcome of acute ischemic stroke requiring reperfusion therapy, *Diagnostics*, vol.11, no.1, DOI: 10.3390/diagnostics11010080, 2021.

[25] T. Huang, J. Li and W. Zhang, Application of principal component analysis and logistic regression model in lupus nephritis patients with clinical hypothyroidism, *BMC Med. Res. Methodol.*, vol.20, no.1, pp.1-7, DOI: 10.1186/s12874-020-00989-x, 2020.

[26] M. Z. F. Nasution, O. S. Sitompul and M. Ramli, PCA based feature reduction to improve the accuracy of decision tree C4.5 classification, *J. Phys. Conf. Ser.*, vol.978, no.1, DOI: 10.1088/1742-6596/978/1/012058, 2018.

[27] J. Ortigosa-Hernández, I. Inza and J. A. Lozano, Measuring the class-imbalance extent of multi-class problems, *Pattern Recognit. Lett.*, vol.98, pp.32-38, DOI: 10.1016/j.patrec.2017.08.002, 2017.

[28] R. Zhu, Z. Wang, Z. Ma, G. Wang and J. H. Xue, LRID: A new metric of multi-class imbalance degree based on likelihood-ratio test, *Pattern Recognit. Lett.*, vol.116, pp.36-42, DOI: 10.1016/j.patrec.2018.09.012, 2018.

[29] M. A. Jassim and S. N. Abdulwahid, Data mining preparation: Process, techniques and major issues in data analysis, *IOP Conf. Ser. Mater. Sci. Eng.*, vol.1090, no.1, 012053, DOI: 10.1088/1757-899x/1090/1/012053, 2021.

[30] L. B. V de Amorim, G. D. C. Cavalcanti and R. M. O. Cruz, The choice of scaling technique matters for classification performance, *Appl. Soft Comput.*, vol.133, 109924, DOI: 10.1016/j.asoc.2022.109924, 2023.

[31] W. K. Härdle and L. Simar, *Applied Multivariate Statistical Analysis*, Springer, 2019.

[32] A. Widodo and S. Handoyo, The classification performance using logistic regression and support vector machine (SVM), *J. Theor. Appl. Inf. Technol.*, vol.95, no.19, pp.5184-5193, 2017.

[33] M. Grandini, E. Bagli and G. Visani, Metrics for multi-class classification: An overview, *arXiv Preprint*, arXiv: 2008.05756, 2020.

[34] S. S. Islam, M. S. Haque, M. S. U. Miah, T. B. Sarwar and R. Nugraha, Application of machine learning algorithms to predict the thyroid disease risk: An experimental comparative study, *PeerJ Computer Science*, vol.8, pp.1-35, DOI: 10.7717/peerj-cs.898, 2022.

[35] S. Pathan, D. Rao and P. Kumar, Lymph node morbidity diagnosis using multiclass machine learning models, *2022 6th International Conference on Green Technology and Sustainable Development (GTSD)*, pp.1173-1176, DOI: 10.1109/GTSD54989.2022.9989185, 2022.

[36] M. I. Prasetiyowati, N. U. Maulidevi and K. Surendro, The accuracy of random forest performance can be improved by conducting a feature selection with a balancing strategy, *PeerJ Computer Science*, vol.8, pp.1-15, DOI: 10.7717/peerj-cs.1041, 2022.

[37] A. A. Nababan, Sutarman, M. Zarlis and E. B. Nababan, Improving the accuracy of k-nearest neighbor (k-NN) using Synthetic Minority Oversampling Technique (SMOTE) and Gain Ratio (GR) for imbalanced class data, *AIP Conf. Proc.*, vol.2714, no.1, DOI: 10.1063/5.0128413, 2023.