SSRG International Journal of Computer Science and Engineering
ISSN: 2348–8387 / https://doi.org/10.14445/23488387/IJCSE-V12I12P101

Volume 12 Issue 12, 1-7, December 2025

Original Article

Enhancing Diagnostics with Logistic Regression

Ruby Pilakhwal[1], Raj Trivedi[2], Ankita Patel[3], Dhruvi Pandya[4]

[1,2,4]*Computer Engineering, Gandhinagar University, Gujarat, India.*
[3] *Information Technology, Gandhinagar University, Gujarat, India.*

[1]*Corresponding Author : rubypilkhwal04@gmail.com*

Received: 10 October 2025 Revised: 13 November 2025 Accepted: 02 December 2025 Published: 13 December 2025

Abstract - *Machine learning is significant in clinical decision support, especially regarding dermatological disorders, whose symptoms significantly overlap, and whose features are mostly discrete. The Logistic Regression (LR) can be employed in such a case, as it gives clear probabilistic results with which a clinician can work, but the literature in this field rarely presents a systematic comparison between LR and other classical algorithms in consistent assessment environments. This paper analyses the LR with the Dermatology dataset on the WEKA system and compares the performance with Naive Bayes, Support Vector (SMO) machine, and Decision Tree (J48). This work is novel due to its interpretability-based evaluation and the systematic analysis of classical models with the help of a discrete and symptom-based dataset of dermatology, a field where journal publications and literature tend to prioritize the precision of data over transparency or rigorous methods in data analysis. The study makes use of 10-fold cross-validation to examine the predictive accuracy and error behavior of each model in order to gain insight into the diagnostic potential of each model. Results indicate that LR has competitive accuracy and stable performance, but is interpretable, which is crucial in clinical applications. Even though SMO has a minor difference in accuracy, LR offers reliable and interpretable predictions, which apply to structured clinical data. The results provide an essential methodological foundation and emphasize the applicability of interpretable models to dermatology diagnostics, where there is a clear, understandable line of decision-making that can aid credible and reliable automated services.*

Keywords - *Classification, Data mining, Logistic regression, Machine learning, Weka.*

1. Introduction

Machine learning has been a key part of medical data analysis, especially in tasks that require precise and transparent decision-making. Clinical data sets can include a heterogeneous set of attributes, overlapping symptom patterns, and slight differences between classes of disease-i.e., algorithmic support is crucial to enhance the reliability of the diagnosis. Classification models are popularly used for this purpose, but it matters a lot how well they deal with medical data constraints and how interpretable their predictions are for healthcare professionals.

Logistic Regression (LR) remains one of the mainstays in applied medical research owing to both its computational simplicity and clinically meaningful and probabilistic results. Unlike complicated, black-box systems, LR gives practitioners the ability to understand the contribution of individual features to the prediction process, which is so important in areas like dermatology where interpretability plays a role in treatment decisions. The Dermatology dataset, commonly cited with studies on the diagnosis of various skin diseases, contains a combination of clinical and histopathological features used for the description of six skin diseases with overlapping symptoms. Although used extensively, existing studies usually focus on broad accuracy comparisons and have limited discourse about how classical machine learning models do in controlled evaluation cases using standard evaluation tools.

Despite the popularity of the dataset, there are a number of research gaps. Prior work often uses a variety of algorithms without giving a clear examination as to why certain models are more appropriate than others for dermatological classification. Many studies also focus exclusively on accuracy and do not focus on in-depth evaluation metrics, which may better reflect the clinical relevance (e.g., class-wise performance, error distribution). Furthermore, the relative performance of interpretable methods such as Logistic Regression against other traditional algorithms has not been systematically analyzed in the WEKA environment, where results can be strongly influenced by parameter settings, validation strategies, and characteristics of the data, which in turn, can dramatically affect the results.

The current study aims to fill these gaps, as it performs a focused evaluation of Logistic Regression on the Dermatology dataset with 10-fold cross-validation on WEKA. The work compares LR against Naive Bayes, SMO, and J48 in an attempt to provide a structured comparison of their behavior on clinical data. The contribution of the study is the focus on interpretability-based model evaluation, the thorough methodological description, and the comparative analysis that was conducted based on standardized evaluation procedures. This way, better insights could be gained on the suitability of Logistic Regression for dermatological diagnosis, and a transparent baseline could be established for other possible research regarding medical classification tasks.

2. Related Work

Machine learning has been widely applied in medical diagnosis, and there are several research papers that highlight the value of Logistic Regression (LR) in the analysis of structured clinical datasets. Previous research has shown that LR makes stable probabilistic predictions and is easily interpreted to fit the needs of industries where clinicians are seeking clarity when making decisions. Studies such as Maalouf (2011) and Sperandei (2014) are focused on the importance of LR in the model of categorical medical features; however, these works focus mainly on methodological explanations, but not comparative evaluations on dermatology-specific data.

Recent studies focus on the diverse classification methods for clinical diagnosis, such as Naive Bayes, Support Vector Machine, and Decision Tree. These algorithms have been tested and evaluated in tools like WEKA to see their performance on structured datasets, and it can be determined that the preprocessing, attribute selection, and cross-validation play a big role in the predictive accuracy. While these studies discuss the behavior of general classifiers, in most cases, the key interpretability trade-offs are not examined, which are essential for dermatological applications where clinical justification is critical.

In parallel, advanced models, such as Convolutional Neural Networks (CNNs), attention-based models, and ensemble methods, have demonstrated strong performance in dermatology, especially for image-based diagnosis. Works based on datasets like ISIC and HAM10000 report a high accuracy; however, these models often represent black box systems and need large-scale image data, and are therefore less accessible for structured tabular datasets, such as the Dermatology dataset in this study.

Despite the increasing amount of research, there is a lack of studies that conduct a focused comparison of LR with other classical interpretable algorithms using standardised evaluation settings in WEKA. Existing literature does not include a detailed evaluation of the behavior of LR on discrete dermatological features, and does not fully address the methodological considerations (e.g., preprocessing, validation strategies), which could affect the model performance. This work is a response to these lacunae and a structured comparative gathering with an emphasis on interpretability, as well as a solid premise for future study on the dermatological classification.

3. Materials and Methods

3.1. Dataset Description

This Dermatology dataset contains clinical and histopathological points of reference that are applied to diagnose six dermatological diseases, which include psoriasis, seborrheic dermatitis, lichen planus, pityriasis rosea, chronic dermatitis, and pityriasis rubra pilaris. The data contains 366 instances, 34 attributes, the majority of which are discrete values and indicate the levels of severity or the presence of certain symptoms on a scale between 0 and 3. These characteristics include not only combinations of observable changes (itchiness, scaling, polygonal papules, etc) but also a mixture of microscopic changes (spongiosis, papillary dermal fibrosis, perivascular inflammation, etc).

The overall analysis of the distribution of classes shows that there is a moderate imbalance of the six types of diseases, which represents the realistic variation in the prevalence of dermatology cases. Given that the dataset comprises tabular structured data as opposed to image data, it fits well into classical machine learning models where the relationships among features are interpretable. None of the cases are discarded, and every possible record is stored to maintain diversity in diagnosis.

The dataset is also best suited to comparative analysis due to its discrete attribute format, which enables the use of algorithms, including Logistic Regression, Naive Bayes, SMO, and J48, to be tested under the same conditions. This research works with the dataset using its full form in order to accomplish an extensive evaluation of a classifier's performance on the clinically meaningful features.

3.2. Attribute Characteristics

The characteristics applied during the analysis are typical dermatological ones, like itching, scaling, spongiosis, fibrosis of the papillary dermis, polygonal papules, and the Koebner phenomenon. Table 3.1 is a brief description of the chosen attributes along with their potential values. None of the attributes are lost in the analysis, and every feature has a clinically relevant aspect, and no feature selection methods are used. Such an attribute structure reduces the number of complex transformation steps required and allows algorithms like Logistic Regression, Naive Bayes, SMO, and J48 to identify relevant patterns of diagnosis using the data directly.

The variety of clinical and microscopic presentations also ensures that the dataset covers a variety of dimensions of dermatological manifestations, which allows conducting a thorough comparative analysis of the performance of the models.

Table 3.1. Description of Dataset Attributes

Attribute Name	Description/Type	Values
Itching	Sensation causing desire to scratch	{0,1,2,3}
Koebner phenomenon	Skin reaction to injury	{0,1,2,3}
Polygonal papules	Flat-topped bumps	{0,1,2,3}
PNL infiltrate	Presence of polymorphonuclear leukocytes	{0,1,2,3}
Fibrosis of the papillary dermis	Scarring of dermis	{0,1,2,3}
Clubbing of the rete ridge	Thickening of the epidermal ridge	{0,1,2,3}
Spongiosis	Edema in epidermis	{0,1,2,3}
Perifollicular parakeratosis	Parakeratosis around follicles	{0,1,2,3}
Class	Disease classification	{1,2,3,4,5,6}

3.3. Data Preprocessing

Before conducting any form of modeling, the Dermatology data is subjected to a systematic preprocessing process within WEKA so that all the attributes are reliably added to the performance of the classifier. Because the dataset contains primarily discrete ordinal attributes, an initial check is done in the Preprocess tab so as to confirm the existence of missing values as well as to check attribute-wise statistics. The dataset contains a few missing entries that are treated with the ReplaceMissingValues filter of WEKA, which is employed to ensure the uniformity of all classifiers without losing out on potentially useful clinical data.

Normalization is used to transform the values of the attributes to the same range. Even though the original values are picked to a 0-3 scale, normalization provides an opportunity to avoid the uneven impact on the parameter estimation stage, especially when using algorithms like the Logistic Regression and the SMO. Normalization is better than standardization since the dataset does not have real continuous variables, and normalizing discrete variables is an effort to maintain the ordinal nature of the severity of the symptoms.

An analysis of the distribution of classes is also marked by a slight imbalance in the six categories of the diseases, but SMOTE-based oversampling techniques are not applied as the data set is not large, and the study's aim is to make a controlled comparison of the classical algorithms under the same conditions. The methodological fairness of preprocessing is ensured by applying all of the preprocessing to models in a consistent way. The resulting dataset is then exported in ARFF format and is ready to be classified in the WEKA Explorer environment.

3.4. Logistic Regression Model

Logistic Regression is used to model the probability of each class based on a linear combination of input attributes. The sigmoid (logistic) function is expressed as:

$$P(Y = 1 \mid x) = \frac{1}{1+e-(\beta 0+\beta 1x1+\beta 2x2+\cdots+\beta nxn)} \quad (1)$$

Model parameters are estimated by maximizing the log-likelihood function:

$$LL(\beta) = \sum_{\{i=1\}}^{\{n\}} \, [yiln(pi) + (1 - yi)ln(1 - pi)] \quad (2)$$

These equations are used to measure the agreement of the model with the real class labels and guarantee the optimal choice of the parameters during training.

This paper uses the WEKA inbuilt Logistic classifier to apply the Logistic Regression in the proposed research, where the ridge regularization of the coefficient estimation is done to stabilize the coefficient estimation and normalize the coefficient estimation, especially in cases where there is mild multicollinearity among the attributes.

No changes have been made to the ridge parameter, which is kept at the default value of (1.0×10^−8), which offers the right regularization and does not affect interpretability. WEKA implements multinomial logistic regression, which allows the six diagnostic types of the Dermatology data to be accommodated with automatic coefficient sets to be estimated in each of the diagnostic types.

The choice of Logistic Regression in analysis is because it is clearer and can be used to analyze data that has discrete and ordinal clinical characteristics. LR, unlike complex black-box models, provides interpretable coefficients and can thus assist in elucidating the effect that individual symptoms and other histopathological indicators have on diagnostic outcomes. The property proves useful in medical practice where clinicians need to know the boundaries of decisions in an understandable format. Probabilistic output of the model also supports confidence-based decision-making, and as such, LR represents a significant basing point to compare the performance of the classifier on the structured dermatological data.

3.5. WEKA Implementation

WEKA 3.8 is used to implement the Logistic Regression model. The dataset is first loaded into the Preprocess panel, where filters such as normalization are applied to scale the attribute values. The interface displays attribute statistics, missing value counts, and class-wise histograms that help in understanding the distribution of dermatological features, as shown in Figure 1. Classification is then performed using the Logistic algorithm under the Functions category.

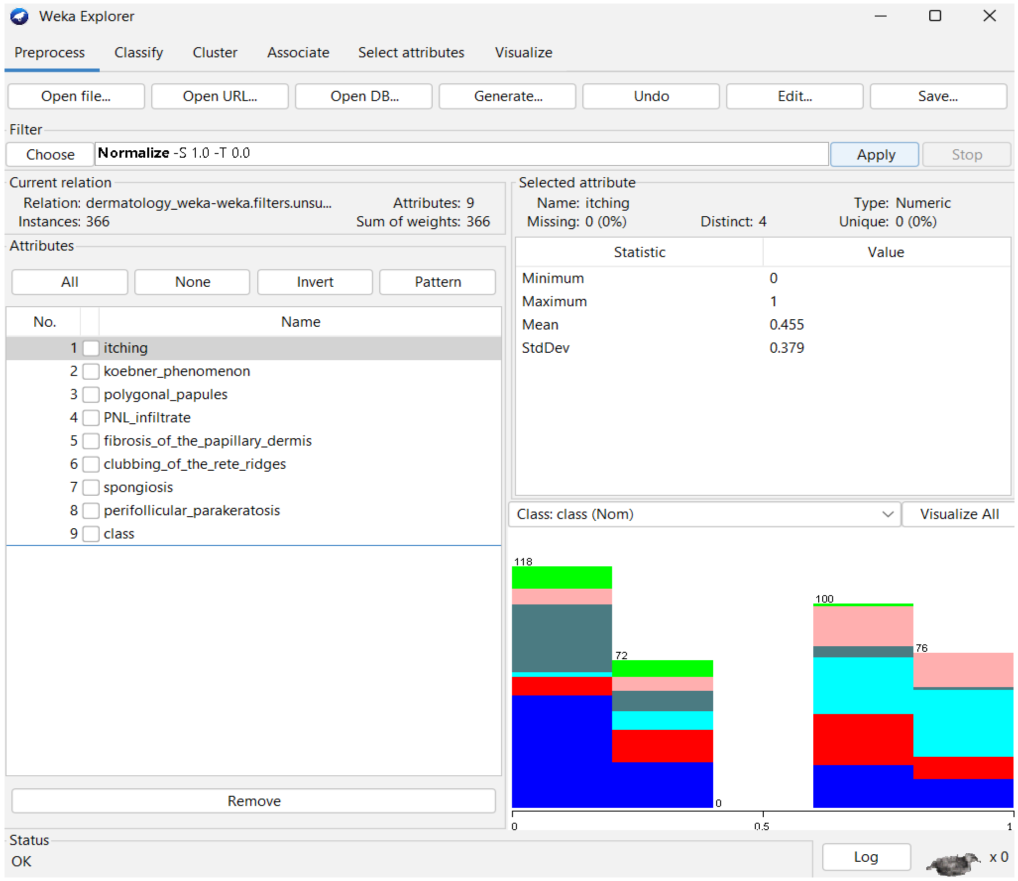

Fig. 1 WEKA Preprocess panel showing selected attribute statistics, filters, and class distribution histogram

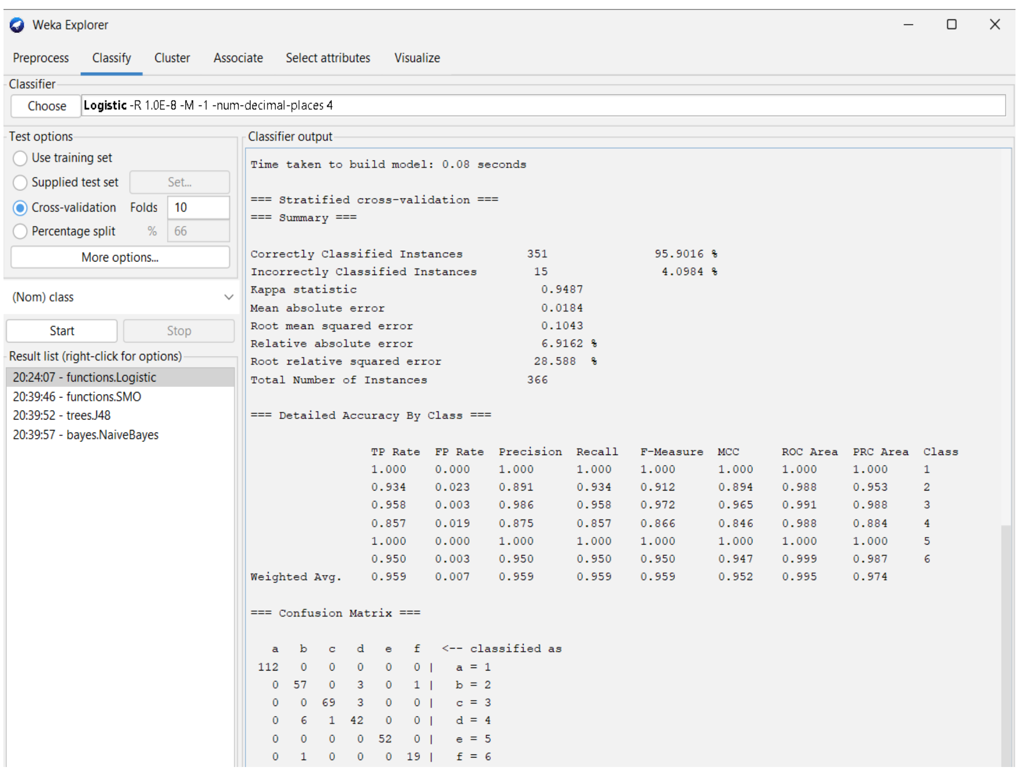

Fig. 2 WEKA classifier output for Logistic Regression

3.6. Evaluation Procedure

All the classifiers are tested in WEKA through a 10-fold cross-validation approach; in this method, the dataset is split into 10 smaller parts, and each part serves as the test set, and the rest of the folds serve as the training data. This method has a good approximation of model performance and minimizes the chances of overfitting data sets composed of small numbers of observations. WEKA automatically produces the accuracy score, error rate, and predictions by classes for each classifier, as shown in Figure 2. Other products of evaluation, such as the confusion matrix, Kappa value, and precision-recall values, are utilized to indicate the diagnostic reliability. Validation procedure and the same random seed are used in all the models in order to provide consistent and comparable results.

4. Results and Discussion

4.1. Classification Performance

The Logistic Regression model was tested on the Dermatology data with 10-fold cross-validation in WEKA. The findings are presented in Table 1. The model demonstrates high predictive accuracy and low error values, which shows that the model remains stable in performance across all folds. Agreeing with the results of the Kappa, the substantial consensus of predicted and actual classes is also an indication that the classifier learns useful diagnostic information using the discrete clinical features. The small MAE and RMSE indicate that the probabilistic model outputs of the Logistic Regression are calibrated, which is significant in addressing real-world clinical exposure.

Table 1. Classification Metrics of Logistic Regression

Metric	Value
Correctly Classified Instances	94.375%
Incorrectly Classified Instances	5.625%
Kappa Statistic	0.9323
Mean Absolute Error	0.0179
Root Mean Squared Error	0.1349
Relative Absolute Error	7.95%
Root Relative Squared Error	31.73%

4.2. Comparative Analysis with Other Classifiers

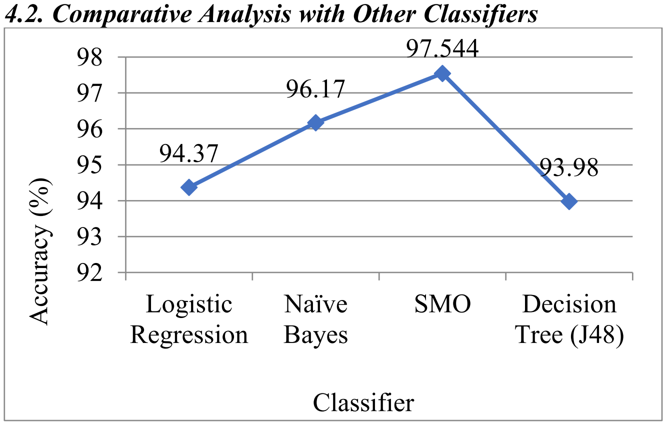

Fig. 3 Accuracy comparison of classifiers

Other classifiers evaluated for comparison include Naive Bayes, SMO, and J48, and they are tested on the same training and validation conditions. Figure 3 and Table 4.2 contain the values of accuracy of each model. SMO is the most precise, and the second in order of accuracy is Naive Bayes. The model of Logistic Regression is a competitive one, which demonstrates that it is beneficial to operate with categorical independent variables of the data. Even though it does not reach optimal accuracy as the SMO, LR has the advantage that none other can; it is interpretable and thus is more acceptable in clinical practices, whereby it is important to make transparent decisions.

Table 2. Comparison of Classifier Accuracies

Classifier	Accuracy (%)
Logistic Regression	94.37%
Naïve Bayes	96.17%
SMO(SVM)	97.54%
Decision Tree (J48)	93.98%

Among the tested algorithms, SMO achieved the highest classification accuracy (97.54%), followed by Naïve Bayes (96.17%). Logistic Regression also performed competitively, indicating that it is a robust and interpretable model for medical classification. Although SMO slightly outperformed other models, Logistic Regression remains advantageous for clinical settings because its coefficients support interpretability and feature relevance analysis.

4.3. Discussion

The comparison outcomes indicate the following important differences between the responses of each model to the aspects of the Dermatology dataset. SMO is advantageous in that it allows modelling non-linear relationships, and this aspect likely improves its accuracy by a small margin. Naive Bayes works best because of the structured and discrete nature of the data. Although with somewhat worse numeric results, Logistic Regression has much more conveniently interpretable coefficients that indicate the contribution of each symptom to diagnostic results, which is a significant need in the medical decision support domain of computerized medicine. The consistency of the model in cross-fold performance also testifies to the reliability of the model in structured clinical data. In general, the results indicate that any interpretable model, like the Logistic Regression, is still applicable in dermatological diagnosis and can be practically used in situations where transparency is more important than the marginal gains in terms of accuracy.

4.4. Class-Wise Metrics, Confusion Matrix, and Error Analysis

The Logistic Regression model provides a detailed class-wise analysis as shown in Table 3. The results show that the performance of all six types of diseases in dermatology is persistently at high performance, with the accuracy, recall, and the F1-scores of 0.85 or higher in all the classes mentioned above. The accuracy and recall of Class 1 and Class 5 are perfect (1.000), that is, the model is capable of correctly identifying these classes. The recall of Class 4 is somewhat lower (0.857), and this means weak overlap in the pattern of symptoms between it and the nearest classes. This is expected in dermatological data sets where there are cases in which certain clinical features are shared across a number of conditions.

Table 3. Class-wise Performance Metrics for Logistic Regression

Class	Precision	Recall	F-Measure	ROC Area	PRC Area
1	1.000	1.000	1.000	1.000	1.000
2	0.891	0.934	0.912	0.988	0.953
3	0.986	0.958	0.972	0.991	0.988
4	0.875	0.857	0.866	0.988	0.884
5	1.000	1.000	1.000	1.000	1.000
6	0.950	0.950	0.950	0.999	0.987

The confusion matrix is also indicated in Table 4 to reveal the rate of correct and incorrect predictions. The majority of the misclassifications are between Class 2 and Class 4, where there are a few cases of Class 2 that represent Class 4. It coincides with the clinical expectations, as these conditions overlap with each other, causing the same symptoms of scaling and erythema. Class 6 has a single false classification to Class 2, and this may be explained by the fact that the histopathological features are similar.

Table 4. Confusion Matrix for Logistic Regression

Actual \ Predicted	1	2	3	4	5	6
Class 1	112	0	0	0	0	0
Class 2	0	57	0	3	0	1
Class 3	0	0	69	3	0	0
Class 4	0	1	3	42	0	0
Class 5	0	0	0	0	52	0
Class 6	0	1	0	0	0	19

The overall pattern of the model has been a balanced predictive model, with the most confusion experienced in the classes that had an inherent overlap in the severity of the symptoms. The overlaps imply that even with such overlap, the ROC and the PRC values in all classes are strong, thereby making the model have a high discriminative capacity. These findings support the hypothesis that structurally defined dermatological data could be reliably diagnosed with the assistance of the Logistic Regression tool and that the estimates of its probability could be adequately utilized in medical practice.

5. Conclusion

This paper is an assessment of the Logistic Regression on the diagnosis of dermatology using the Dermatology dataset in the WEKA environment. The model has a high predictive power with high accuracy, probabilities well-calibrated, and balanced class-wise performance regardless of the six disease types. Comparative analysis reveals that SMO has the best accuracy of all; however, Logistic Regression can offer the strength that the individual contributions of each symptom can be interpreted, which is very important in medical decision-making. The confounding analysis of errors and the confusion matrix prove that the majority of the misclassifications are made between clinically related disorders, which naturally overlap when it comes to dermatological patterns of symptoms.

The novelty of the work is in its interpretability-oriented assessment and comparative evaluation of classical machine learning models in a uniform preprocessing and validation environment, which fills an empty methodology in the literature on dermatology-associated topics. Though the result suggests high diagnostic potential, the study is constrained by the size of the dataset, and external validation sets are absent. Future studies may use larger or multi-institutional datasets, examine the use of ensemble learning, or incorporate an explainability method, like SHAP or LIME, to further improve clinical trust and model transparency. Altogether, the paper confirms Logistic Regression as an adequate and interpretable baseline of organized dermatological data and a good methodology for further studies.

Acknowledgments

The authors would like to express their sincere gratitude to Ankita Patel and Dhruvi Pandya, faculty members at Gandhinagar University, for their continuous guidance, constructive feedback, and valuable academic support throughout the course of this research work. Their expertise and encouragement played a significant role in shaping the methodology and analysis presented in this study.

References

[1] Ganesh Khekare et al., "Logistic and Linear Regression Classifier Based Increasing Accuracy of Non-Numerical Data for Prediction of Enhanced Employee Attrition," *International Conference on Advance Computing and Innovative Technologies in Engineering*, Greater Noida, India, pp. 758-761, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[2] Thomas Johnson III, "Machine Learning Evaluations Using WEKA," *Honors Thesis*, Elizabeth City State University, North Carolina, USA, 2020. [Google Scholar] [Publisher Link]

[3] Priyanka Rajendra, and Shahram Latifi, "Prediction of Diabetes Using Logistic Regression and Ensemble Techniques," *Computer Methods and Programs in Biomedicine Update*, vol. 1, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[4] Amina Aboulmira, Hamid Hrimech, and Mohamed Lachgar, "Skin Diseases Classification with Machine Learning and Deep Learning Techniques: A Systematic Review," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 10, pp. 921-931, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[5] Taye Girma Debelee, "Skin Lesion Classification and Detection Using Machine Learning Techniques: A Systematic Review," *Diagnostics*, vol. 13, no. 19, pp. 1-40, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[6] Masao Iwagami et al., "Comparison of Machine-Learning and Logistic Regression Models for Prediction of 30-Day Unplanned Readmission Using Electronic Health Records," *Journal of Medical Systems*, vol. 3, no. 8, pp. 1-16, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[7] Marta Bistroń, and Zbigniew Piotrowski, "Comparison of Machine Learning Algorithms Used for Skin Cancer Diagnosis," *Applied Sciences*, vol. 12, no. 19, pp. 1-13, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[8] G. Thippanna, and G. Ravi Kumar, "Logistic Regression for Multiclass Classification of Dermatology Disorders: A Comparison of One-vs-One and One-vs-Rest Approaches," *International Journal of Data and Information Engineering*, vol. 14, no. 2, pp. 60-63, 2025. [Google Scholar] [Publisher Link]