Evaluation of Machine Learning Classifiers in Identifying Erythemato-Squamous Diseases

1st Praveen Kumar Nalli
Department of AI
Shri Vishnu Engineering College for Women
Bhimavaram, India
praveenkumar.nalli@svecw.edu.in

2nd Katta Eekshitha
Department of AI
Shri Vishnu Engineering College for Women
Bhimavaram, India
eekshitha5904@gmail.com

3rd Daka Kavya
Department of AI
Shri Vishnu Engineering College for Women
Bhimavaram, India
dakakavya2501@gmail.com

4th Gayatri Rishitha
Department of AI
Shri Vishnu Engineering College for Women
Bhimavaram, India
gayatririshitha44@gmail.com

5th Gona Veena Sudha Rani
Department of AI
Shri Vishnu Engineering College for Women
Bhimavaram, India
veenagona36@gmail.com

6th Kotapuri Lavanya
Department of AI
Shri Vishnu Engineering College for Women
Bhimavaram, India
kotapuri21b01a5457@svecw.edu.in

Abstract—This research pioneers a machine learning-enhanced diagnostic system for erythemato-squamous diseases (ESD), utilizing Naive Bayes, Random Forest, and Support Vector Machine (SVM) algorithms to categorize various ESD types. The system, leveraging a dataset inclusive of clinical, historical, and demographic data, attained a remarkable 98.64% accuracy rate with SVM, underscoring its potential to significantly refine ESD diagnosis and treatment modalities. Subsequent validation against an additional dataset corroborated its reliability, heralding substantial progress in dermatological healthcare and pioneering new avenues for the application of machine learning in medical diagnostics.

Keywords—Erythemato Squamous Diseases (ESD), Naïve Bayes, Random Forest, Support Vector Machine.

I. INTRODUCTION

Erythemato-Squamous Diseases (ESDs) encompass a spectrum of skin conditions characterized by erythema and scaling, among others, comprising psoriasis, seborrheic dermatitis, lichen planus, pityriasis rosea, and chronic dermatitis. The overlap in clinical presentations of ESDs complicates their diagnosis, necessitating comprehensive examinations and, in some instances, biopsy for accurate differentiation. Precise diagnosis is vital for effective treatment, as each disease demands a specific therapeutic strategy. Recent advancements in dermatological research, particularly the incorporation of machine learning for clinical image pattern recognition, are enhancing diagnostic accuracy for these diseases.

The diagnosis of this condition typically entails both clinical and histopathological examinations [1]. Clinical evaluations, being non-invasive, concentrate on symptom observation such as lesion location and size, pustules, color alterations, and similar features. On the other hand, histopathological assessments necessitate skin biopsies to ascertain potential viral etiologies. The complexity in differentiating among ESDs, given their shared clinical manifestations, has propelled the exploration of data mining techniques to enhance the precision of identification.

The research on machine learning classifiers for diagnosing Erythemato-Squamous Diseases (ESD) is driven by the significant diagnostic challenge these diseases pose in dermatology, owing to their overlapping initial symptoms, distinct disease-specific characteristics, and common clinical and histopathological features. Diagnosing ESDs typically involves clinical and histopathological examinations. Clinical assessments are non-invasive, focusing on symptoms observation, while histopathological evaluations require skin biopsies to explore potential viral causes.

The challenge of distinguishing between ESDs has led to the exploration of data mining techniques for more accurate identification. The data mining model utilizes a combination of statistics, artificial intelligence, machine learning, and database management to discern and interpret key information. Within this domain, various models and algorithms offer a balance between accuracy and interpretability. Although models like Artificial Neural Networks (ANN) deliver high predictive accuracy, their complex structures render them less interpretable. This study aims to utilize machine learning to enhance the diagnostic accuracy and efficiency for ESDs, applying advanced classification techniques to navigate the complexities of these diseases.

The data mining model operates by discovering and interpreting important information, utilizing a blend of various fields like statistics, artificial intelligence, machine learning, and database management [2]

Within this domain, various models and algorithms exist, with a trade-off between accuracy and transparency. While models like Artificial Neural Networks (ANN) often yield high predictive accuracy, they lack interpretability due to their intricate structures.

The Dataset consists of 2 different types of Features

a) Clinical Features

b) Histopathological Features

The clinical features of erythemato-squamous diseases (ESD) include a range of dermatological symptoms. These typically encompass various skin lesions, scaling patterns, and inflammation levels. The manifestation of these symptoms can vary significantly between different ESD types, such as psoriasis and seborrhoeic dermatitis. Additionally, factors like patient age, disease duration, and specific affected areas are considered critical in distinguishing between these conditions. This comprehensive approach to symptom analysis is essential for accurate diagnosis and effective treatment planning in dermatology.

Histopathological examination of erythemato-squamous disorders (ESD) in great detail identifies diagnostically important features of affected tissues. Distinct patterns, such as hyperkeratosis or parakeratosis, and alterations in the epidermal layer's thickness characterize these illnesses. Disruptions in the normal keratinization of the stratum corneum are common. A typical aspect of the dermis that contributes to the overall pathophysiology of the disease is the infiltration of inflammatory cells, especially lymphocytes. Psoriasis and other skin conditions can manifest with distinctive symptoms, such as neutrophil buildup indicated by Kogoj's spongiform pustules or Munro's microabscesses. The dermoepidermal junction changes and rete ridge elongation are also important indicators. Accurate diagnosis and efficient treatment methods in dermatology rely on a thorough understanding of the many forms of ESD, which can be achieved by combining these microscopic results with clinical data. Research and practice in dermatology benefit greatly from this degree of granular analysis. These are the symptoms which can be analysed under the microscope.

In Section II, the review of prior literature was presented, followed by Section III discussing the diagnostic methodology, Section IV outlining the results, and Section V summarizing the conclusions.

II. LITERATURE REVIEW

This portion is dedicated to gaining a comprehensive understanding of the literature concerning the application of machine learning in diagnosing skin diseases. Abhishek Singh Rathore et. al [3] has worked on Random Forest & XGBoost Models in which Random Forest achieved a classification accuracy of 98.21%. some of the recent works done on ESD classification is presented in Table I.

TABLE I.   REFERENCES OF THE PREVIOUS WORK

Reference No	Author	No. of Features used	Classifiers	Accuracy (%)
[4]	Juanying Xie, Chunxia Wang et.al	21	Combined SVM with a brand-new hybrid feature selection technique	98.61
[5]	Mohammad Javad Abdi, Davar Giveki et.al	24	Used PSO–SVM based association rules	98.91
[6]	Loris Nanni	50% features selected at random	Used ensemble of SVM based random subspace method (RS)	99.2
[7]	Tong Liu, Liang Hu et.al	20	ELM classifier	98.89
[8]	Efosa Charles Igodan et.al	two-third of the 34-original features	Three embedded feature selection techniques and four distinct filtering algorithms	92.9

The biggest issue dermatologists deal with when treating skin conditions is that symptoms from one type of disease might overlap with those from other classifications, making diagnosis and therapy very challenging [9].

III. METHODOLOGY

The ES-Diseases dataset was employed for analysis, preceded by essential pre-processing steps to ensure data integrity and reliability. This phase involved null value handling within critical features like clinical and histopathological sections, utilizing methods such as imputation or removal.
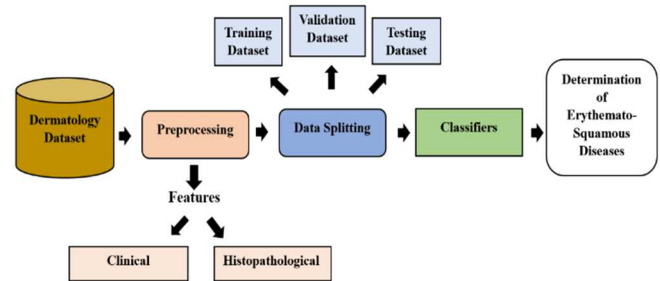

Fig. 1. Proposed Frame work for ESD

Fig. 1 depicts the steps followed while dealing with the dataset to diagnosis of ESD.

The ES-Diseases dataset was employed for analysis, preceded by essential pre-processing steps to ensure data integrity and reliability. This phase involved null value handling within critical features like clinical and histopathological sections, utilizing methods such as imputation or removal.

Prior to model fitting, the dataset underwent pre-processing, specifically involving the removal of null values within pivotal features like clinical and histopathological sections. Null value handling techniques, including imputation and contextual removal, were applied to ensure the dataset's reliability for subsequent analysis.

Following pre-processing, the dataset underwent crucial partitioning for model training, testing, and validation. The division involved creating distinct segments for training, testing, and validation subsets. This approach ensured robust evaluation, preventing over fitting.

The pre-processed dataset was used to test the prediction capabilities of various classifiers. These classifiers were used, specifically:

A. Naive Bayes(NB)

The Naive Bayes classifier is a fundamental and widely-used machine learning algorithm based on Bayes' Theorem, with the 'naive' assumption of independence between every pair of features. It is particularly effective for classification tasks, where it calculates the probability of a class based on the presence of features in a given sample. Despite its simplicity, Naive Bayes can yield surprisingly accurate results, especially in cases of text classification, spam detection, and even medical diagnosis.

The strength of Naive Bayes lies in its efficiency and scalability, handling large datasets with ease. It performs well with categorical input variables compared to numerical variables. For numerical data, different types of Naive Bayes models like Gaussian Naive Bayes can be used, which assume a normal distribution for the numerical variables. Its simplicity in understanding and implementation, coupled with its

effectiveness in handling complex real-world datasets, makes it a popular choice in the machine learning community.

Mathematically expressed as

$$P\left(\frac{y}{x}\right) = \frac{P\left(\frac{x}{y}\right).P(y)}{P(y)} \qquad (1)$$

Where

$P(y/x)$ is posterior probability of class y given features x

$P(x/y)$ is the Likelihood

$P(y)$ is the prior probability of class y, and

$P(x)$ is the probability of feature x.

B. Random Forest

The Random Forest algorithm is a powerful and flexible machine learning model, well-known for its precision and effectiveness in performing classification and regression tasks. It functions as an ensemble learning method, combining the forecasts of various machine learning algorithms to provide more accurate predictions than any single model could. In essence, Random Forest builds numerous decision trees during training and delivers the most common outcome of the classes (for classification) or the average prediction (for regression) from these individual trees.

Random Forest is particularly praised for its ability to handle large datasets with higher dimensionality [11]. It can manage thousands of input variables without variable deletion and is highly effective against overfitting, thanks to its ensemble approach. This model is also adept at handling missing values and maintains accuracy for a large proportion of data missing. Furthermore, its feature importance score, which is a helpful tool for feature selection, makes it a valuable tool for predictive modelling in various domains including but not limited to finance, healthcare, and e-commerce.

$$\text{Final Prediction: } \sum_{i=1}^{n}\left(\frac{1}{N}\right).fi(x) \qquad (2)$$

where

N is the number of trees,

$fi(x)$ denotes the prediction of the ith tree.

C. Support Vector Machine (SVM)

The Support Vector Machine (SVM) is an effective and adaptable supervised machine learning algorithm, primarily utilized for classification purposes, though it can also be used for regression [10]. Its fundamental objective is to identify the most suitable hyper plane that effectively segregates data points from distinct classes in a high-dimensional space. This hyper plane is selected to enlarge the margin between the data points of different classes, thereby not just separating the classes, but also ensuring there is maximum space surrounding each class [14].

A key feature of SVM is its use of kernel functions, which enable it to handle non-linear relationships by transforming the input space into a higher-dimensional space where a linear separation is possible. This makes SVM exceptionally effective in complex datasets where the relationship between features is not linear [15]. SVMs are known for their accuracy

and robustness, particularly in scenarios where the distinction between different classes is clear. They are widely used in applications ranging from image recognition to bioinformatics.

Mathematically, the linear SVM separates classes using the equation:

$$w^T.x + b = 0 \qquad (3)$$

Where

w represents the weights,

x represents input vector, and

b represents bias term.

IV. RESULTS

The dataset used for performance evaluation focuses on Erythemato-squamous Disease and comprises six classes and 34 features. These classes are seborrheic dermatitis, psoriasis, lichen planus, pityriasis rosea, chronic dermatitis, and pityriasis rubra pilaris.

To assess the effectiveness of the models, the dataset is split into two segments: validation and testing. Both the validation and testing sets are included in each dataset, with the validation set making up 30% of the total dataset and the testing set constituting 20%.

In comparison to existing studies, the implementation of machine learning enhancements in this research demonstrates superior accuracy by leveraging all 34 features as

TABLE II.   VALIDATION AND TESTING ACCURACIES OF DIFFERENT CLASSIFIERS

Model	Validation Accuracy (%)	Testing Accuracy(%)
NB	82.95	83.78
Random forest	96.59	97.29
SVM	95.45	98.64

Table II gives the Validation and Testing Accuracies of different classifiers.
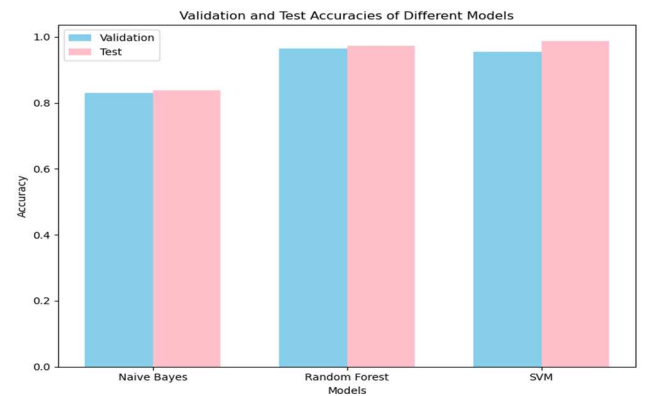

Fig. 2.   Visualization of Accuracies of different models

Fig. 2 visualizes the validating and testing accuracies of specific classification models.

The performance evaluation encompassed various metrics:

A. Accuracy

This calculation assesses the proportion of correctly identified cases from the total, providing a comprehensive performance score.

B. Precision

The capability of the model to identify relevant cases is illustrated through the proportion of accurately forecasted positive instances to the aggregate predicted positives.

C. Recall (Sensitivity, True Positive Rate)

It represents the model's ability to find pertinent cases and is expressed as the ratio of correctly predicted positive observations to all actual positives.

D. Specificity (TNR)

Specificity is an important metric in fields where the cost of false positives is high. For example, in medical diagnostics, a test with high specificity means that it is good at identifying people who do not have a disease, thereby minimizing the risk of unnecessary worry or treatment for those individuals.

It is often used in conjunction with sensitivity (also known as recall or true positive rate), which measures the proportion of actual positives that are correctly identified. Together, sensitivity and specificity give a comprehensive picture of a model's performance in distinguishing between positive and negative cases.

E. F1 Score

Harmonic mean of precision and recall, providing balanced performance assessment.

These metrics facilitated a comprehensive evaluation of each classifier's strengths and weaknesses when handling the ES-Diseases dataset.

TABLE III.   EVALUATION METRICS OF DIFFERENT MODELS

Models	Precision (%)	Recall (%)	Specificity (%)	F1 Score (%)
NB	83.78	96.76	83.78	83.78
Random Forest Classifier	97.29	99.44	97.29	97.29
SVM	98.64	99.73	98.64	98.64

Table III presents evaluation metrics for various classification models, where the SVM model demonstrates superior performance with a precision rate of 98.65%, indicating exceptionally high accuracy in identifying true positive cases. Its recall rate stands at an almost perfect 99.73%, complemented by a specificity of 98.65%, showcasing its ability to correctly identify both positive and negative cases. Furthermore, the F1 score, also at 98.65%, reflects the model's excellent overall performance by effectively balancing precision and recall.

TABLE IV. PERFORMANCE COMPARISION

Reference No	Classifiers	Accuracy(%)
[10]	KNN, DT	94.7, 94
[11]	SVM	92.79
[12]	SVM	98.5
[13]	CNN-ensemble with random forest	90.7
Proposed Models	Naïve Byes	83.78
	Random Forest	97.29
	Support Vector Machine	98.64

Table IV gives the comparison of various models with the proposed model. Delving into the accuracy metrics of various classifiers, we observe a wide range of performance levels. The Support Vector Machine (SVM) model emerges as the top performer, achieving an accuracy rate of 98.64%, which underscores its exceptional ability to manage the task at hand efficiently. Conversely, the Naïve Bayes classifier is at the lower end of the spectrum, with an accuracy rate of 83.78%, indicating its challenges in capturing complex patterns as effectively as its more advanced counterparts. Remarkably, the integration of KNN with Decision Trees, alongside the standalone SVM and Random Forest classifiers, demonstrate high accuracy rates exceeding 94%. This highlights their strong performance in classification tasks. However, the CNN-ensemble combined with Random Forest falls somewhat short of expectations when compared to simpler models.

V.   CONCLUSION

The investigation underscores that machine learning markedly improves the diagnosis of ESD by leveraging high-accuracy classifiers like Naive Bayes, Random Forest, and SVM. These models adeptly categorize ESD subtypes utilizing 34 attributes, highlighting the efficacy of clinical and histopathological data in enhancing diagnostic precision. The study champions the integration of data mining with dermatological practices, advocating for interdisciplinary methods. It posits that future endeavours should concentrate on refining these models and broadening datasets to boost diagnostic accuracy further. Integrating these advancements into clinical settings could signify a transformative leap in medical diagnostics.

REFERENCES

[1] Mohamed El Bachir Menai, "Random forests for automatic differential diagnosis of erythemato squamous diseases," International Journal of Medical Engineering and Informatics (IJMEI), Vol. 7, No. 2, 2015.

[2] L. Wang and T. Z. Sui, "Application of Data Mining Technology Based on Neural Network in the Engineering," in 2007 International Conference on Wireless communications, Networking and Mobile Computing, 2007, pp. 5544-5547.

[3] Abhishek Singh Rathore, Siddhartha Kumar Arjaria, Manish Gupta, Gyanendra Chaubey, Amit Kumar Mishra & Vikram Rajpoot (2022) Erythemato-Squamous Diseases Prediction and Interpretation Using Explainable AI, IETE Journal of Research, DOI: 10.1080/03772063.2022.2114953.

[4] Juanying Xie, Chunxia Wang, "Using support vector machines with a novel hybrid feature selection method for diagnosis of erythemato-squamous diseases" in 2011 Expert Systems with Applications Volume 38 pp 5809–5815

[5] M. J. Abdi and D. Giveki, "Automatic detection of erythemato-squamous diseases using PSO-SVM based on association rules," Eng. Appl. Artif. Intell., vol. 26, no. 1, pp. 603-608, 2013.

[6] Loris Nanni, An ensemble of classifiers for the diagnosis of erythemato-squamous diseases, Neurocomputing, Volume 69, 2006, Pages 842-845.

[7] Tong Liu, Liang Hu, Chao Ma, Zhi-Yan Wang & Hui-Ling Chen (2015) A fast approach for detection of erythemato-squamous diseases based on extreme learning machine with maximum relevance minimum redundancy feature selection, International Journal of Systems Science, 46:5, 919-931.

[8] Igodan, E.C.; Thompson, A.F.-B.; Obe, O.; Owolafe, O. Erythemato Squamous Disease Prediction using Ensemble Multi-Feature Selection Approach. Int. J. Comput. Sci. Inf. Secur. IJCSIS 2022, 20, 95–106

[9] Verma, A.K., Pal, S. Prediction of Skin Disease with Three Different Feature Selection Techniques Using Stacking Ensemble Method. Appl Biochem Biotechnol 191, 637–656 (2020). https://doi.org/10.1007/s12010-019-03222-8

[10] Igodan, Efosa & Obe, Olumide & Thompson, Bethy & Owolafe, Otasowie. (2022). Erythemato-Squamous Disease Prediction using Ensemble Multi-Feature Selection Approach. International Journal of Computer Science and Information Security,. 20. 95-106.8

[11] Verma AK, Pal S, Kumar S. Classification of Skin Disease using Ensemble Data Mining Techniques. Asian Pac J Cancer Prev. 2019 Jun 1;20(6):1887-1894. doi: 10.31557/APJCP.2019.20.6.1887. PMID: 31244314; PMCID: PMC7021628.

[12] Igodan, Efosa & Obe, Olumide & Thompson, Aderonke & Owolafe, Otasowie. (2022). Prediction of erythemato squamous-disease using ensemble learning framework. 10.1049/PBHE050E_ch7.

[13] H.A Güvenir, Jiang P. 'CNN-based diagnosis system on skin cancer using ensemble method weighted by cubic precision'. TechRxiv.Preprint,2021.https://doi.org/10.36227/techrxiv.16964893.v1. Assessed 27 March 2022.

[14] M. Venkata Subbarao, S. Lakshmi Praveena, T. Naga Sharmila, Praveen Kumar Nalli, G. Challa Ram, D. Ramesh Varma, Srinivasarao Alluri, Tabassum Nazeer Syed, "Machine Learning Algorithms for Erythemato-Squamous Disease Classification: Feature Rankings and Performance Analysis", Proceedings On Engineering Sciences. DOI: 10.24874/PES.SI.24.03.009

[15] M. V. Subbarao, U. L. S. Rani, V. H. Mythreya, G. P. Kumar, R. Priyakanth and U. S. Abudhagir, "Investigations on Feature Ranking Algorithms for Date Fruit Genotype Detection," 2023 8th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2023, pp. 1506-1513.