Comparative Analysis of Machine Learning Approaches for Classifying Erythemato-Squamous Skin Diseases

Bhavana Kaushik[1](✉), Ankur Vijayvargiya[2], Jayant Uppal[3], and Ankit Gupta[4]

[1] School of Computer Science, University of Petroleum and Energy Studies, Dehradun, India
kau.bhavana@gmail.com
[2] Accenture, Noida, India
[3] People Strong, Noida, India
[4] Samsung R&D, Noida, India

Abstract. In the recent era, Machine Learning and Artificial Intelligence have come to a very great development point as we can use ML algorithms to predict the type of Erythemato-Squamous (Skin) diseases of the skin. In Dermatology, differential diagnosis of skin diseases is quite challenging in real life because most skin diseases share many histopathological features. And in this work, Psoriasis, Lichen Planus, Seborrheic Dermatitis, Chronic Dermatitis, Pityriasis Rosea, and Pityriasis Rubra Pilaris are among the skin illnesses for which eight different algorithm analytical comparison is done. Moreover, each classifier algorithm is discussed in detail with its pros and cons. The machine learning algorithms like Support Vector Machine, Decision tree, Random Forest, KNN, Naïve Bayes, Gradient Boosting, XGBoost, and Multilayer Perception have been proven to be successful in preserving state information through exact segmentation/classification. Random forest, Gradient Boosting, and XGBoost outperform all other methods and give an accuracy of 100% on the given ESD dataset. While Support Vector Machine gives the least accuracy of 72.97%. The paper also discusses the difficulties connected with skin disease segmentation or categorization. Furthermore, the study proposes future potential directions that include real-time analysis.

Keywords: Erythemato-Squamous Diseases · Machine Learning · Classification · Skin Diseases · Comparative Analysis

1 Introduction

In recent times we observed that many people are suffering from skin diseases or skin cancer which can be curable at an early stage but now they are not curable thus after watching the continues advancement and development in technology and especially in AI & ML we decided to combine them and take the help from various machine learning algorithms to predict the skin diseases at the earliest stage so that patient can be cured within time and it can also help all medical field and especially in dermatology to predict the diseases at the earliest stage. Many times doctor is not able to get the type of skin

S. Kadry and R. Prasath (Eds.): MIKE 2023, LNAI 13924, pp. 67–77, 2023.
https://doi.org/10.1007/978-3-031-44084-7_7

disease in the earlier stage because at the beginning stage all types of skin diseases show the same symptoms and also share the same histopathological features so identifying in the earliest stage is also a very challenging task for the doctors thus this time is wasted and can cause the disease to grow very quickly and can cause to death and cancer. This wasted time is very crucial for the patient and in this time if our algorithm can make the accurate and right decision then we can save the life of the patient. One of the hardest challenges facing today's health care organisations (hospitals, medical facilities) is the provision of high-quality services at fair pricing. Providing patients with accurate diagnoses and efficient treatments are examples of quality care [1]. The majority of hospitals presently use a hospital information system to manage their patient or healthcare data. Typically, these systems output enormous amounts of data in the form of statistics, text, charts, and images [2]. Regrettably, these data are rarely used to guide clinical decisions. Utilising pertinent computer-based information and/or decision assistance technologies can yield the desired results. A critical question is raised in this context: "How can we transform data into useful knowledge that enables clinicians to make informed therapeutic decisions?" This is the main motivation behind studying. Erythemato-squamous diseases (ESDs) are very prevalent skin conditions. The six different varieties are psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, chronic dermatitis, and pityriasis rubra pilaris. By a little margin, all of them exhibit the medical symptoms of erythema and scaling [3]. **Psoriasis -** It is believed that psoriasis is an immune system issue. As skin cells accumulate to form scales, itchy, dry areas start to appear. Triggers include illnesses, stress, and the common cold. The most typical sign is a rash on the skin, but it can also affect the joints or nails. The treatment's goals are to get rid of scales and slow down skin cell growth. Medication, light treatment, and topical ointments can all help [4]. **Seborrheic dermatitis -** This skin ailment is characterised by flaking patches and red skin, mainly on the scalp. It can also appear on oily body parts like the face, upper chest, and back. Seborrheic dermatitis can create obstinate dandruff in addition to scaly patches and red skin. Self-care and medicinal shampoos, creams, and lotions are used in the treatment. Treatments may need to be repeated [5]. **Lichen planus -**An inflammatory skin and mucous membrane disease. When the immune system mistakenly targets skin or mucous membrane cells, lichen planus develops. Lichen planus shows on the skin as reddish, itchy pimples with a flat top. It creates lacy, white patches on mucous membranes, such as the mouth, and sometimes severe ulcers. Lichen planus is frequently self-resolving. Topical treatments and antihistamines may help if symptoms are bothersome [6]. **Pityriasis Rosea -** A hasty that twitches as a huge spot on the abdomen, chest, or back and then spreads out into a pattern of smaller lesions. The cause of pityriasis rosea is unknown, but it is thought to be caused by a viral infection. The illness creates a rash on the torso, upper legs, and upper arms that is mildly irritating. Pityriasis rosea is frequently self-resolving. Antihistamines, steroid cream, and, in rare situations, antiviral medicines can all assist [7]. **Chronic dermatitis -** Dermatitis is a word used to define a group of itchy, inflammatory skin disorders marked by epidermal abnormalities. Dermatitis affects one out of every five people at some time in their livesIt can have many different patterns and many different causes. Eczema and dermatitis are commonly used interchangeably. The term "eczematous dermatitis" is occasionally used. Acute, chronic, or a combination of the two types of dermatitis are possible. Dermatitis, another name for acute eczema, is a rapidly growing red rash that can blister and enlarge. Eczema (also known as dermatitis) is a chronic, uncomfortable skin condition. Typically, it is thicker, darker than the surrounding skin, and heavily scraped [8]. **Pityriasis Rubra Pilaris (PRP) –** It is a term used to describe a collection of rare skin illnesses characterized by scaling patches that are reddish orange in colour and have well-defined edges. They can cover the entire body or convinced parts like the prods and laps, palms, and soles. Islands of sparing are patches of uncomplicated skin, especially on the stem and limbs, which are frequently seen. The palms and soles are commonly affected, becoming swollen and yellowish in appearance (palmoplantar keratoderma). PRP is frequently misdiagnosed as psoriasis or another skin disorder [9].

Skin cells (squamous) deteriorate, and erythema (redness of the skin) is caused by infection with one of these skin illnesses. Dermatologists frequently examine patients both clinically and based on histological features [10]. Clinical examinations involve looking at the colour, presence of zits, their size, location, and other symptoms. For each person/patient, the aforementioned examinations result in 12 clinical and 22 histological variables. Investigating these variables may have ambiguous and illogical results since they may cross paths, particularly in the early stages of ESD. So, there is a need to identify the appropriate classification technique to solve this problem and give a better result/prediction. This paper reviews all conventional techniques present to perform the differential diagnosis of ESD and also to identify the challenges existing with approaches. The contributions of this study are as follows:

(i) Analytical comparison of all relevant conventional machine learning techniques for differential diagnosis of ESD is done.
(ii) Also, the challenges and issues associated with each technique is identified.
(iii) The study discusses the future potential directions that include real-time analysis.
(iv) Discuss the difficulties connected with skin disease segmentation or categorization.

The structure of this paper is as follows: The description of material and methods used in this comparative analysis is given in Sect. 2. Section 3 contains the result and discussions of the analysis. Prominent challenges and future scope are described in the Sect. 4 and lastly Sect. 5 concludes the study.

2 Material and Methods

2.1 Materials

We used a standardized dermatology data set from the “University of California, School of Information and Computer Science’s machine learning repository, or UCI. It has 34 properties, 12 of which are clinical and 22 of which are histological”. Age and family history are continuous characteristics in the data set, with values ranging from 0–1. Every additional clinical and histological feature was given a degree from 0 to 3, where 0 meant the feature wasn’t present, 3 meant it was present to its fullest extent, and 1, 2 meant it was present to a relatively moderate level. Naive Bayes, Random Forest, Support Vector Machines, XGBoost, Multi-layered perceptron, K-nearest neighbors, Decision tree, Gradient boosting DT are among the ML Classification Algorithms investigated in this paper [11] Table 1 shows the six classes of ESD.

Table 1. Six classes of ESD

Keys	Values (Class Labels)
1	Psoriasis
2	Seborrheic Dermatitis
3	Lichen Planus
4	Pityriasis Rosea
5	Chronic Dermatitis
6	Pityriasis Rubra Pilaris

2.2 Methods

2.2.1 Support Vector Machine (SVM)

The objective of the SVM, where n is the number of variables, is to identify the hyper plane in n-dimensional space. The hyper-plane is selected so that there is as little space as feasible between the closest data points and support vectors of the two distinct modules. The hyper plane can be pictured as both a plane and a line in three dimensions. Hyper plane separates the data points of two different classes [12]. Numerous of the prevailing (non)convex soft-margin losses can be observed as one of the substitutes of the L0/1 soft-margin loss. SVM have gained huge consideration for the last two decades due to its wide-ranging usage, so many researchers have established optimization procedures to solve SVM with various soft-margin losses [13].

For the prediction corresponding to a new input can be obtained using Eq. 1:

$$f(x) = B0 + sum(ai * (x, xi)) \tag{1}$$

where *f(x)* is used to calculate the inner dot product which is the sum of the multiplication of each pair of the input values i.e., *x* as the new input and *xi* as each of the support vectors present in the training set. *B0* and *ai* are the coefficients evaluated from the training data [14]. Figure 2 shows the visuals of the Support Vector Machines (SVMs) Model with all necessary features like absolute hyperplane with maximum margin, hyperplane with positive and negative trends and the nearest data points as the support vectors.

2.2.2 Random Forest

It is, as the name implies, a group of various decision trees, each of which predicts some class, and the class having the most votes is accepted as the predicted class. These decision trees produce distinct outcomes relatively. This concept is highly effective in the reduction of prediction errors if predicted through a single decision tree. In this approach, an individual tree may be in the wrong direction, but the common direction could be in the right direction [15].

2.2.3 Naive Bayes

This classifier is based on the Bayes theorem and approaches the probabilistic strategy in classification through prediction. Equation 2 states the approach of Bayes theorem:

$$P(A|B) = P(B|A)P(A)/P(B) \tag{2}$$

where we discover the chance of happening of A assuming that B had already occurred. In this concept A is considered as the hypothesis and B is considered as the evidence. This approach is best when the features are not affected by each other [16].

2.2.4 Decision Tree

As the name suggests, we can find an analogy between a tree and a result tree. A result tree is similar to a tree by having split conditions as a node, directing edges as branches, and decisions as leaves. The formation of a tree involves feature decision and branching conditions and holds the decision by preventing further branching. This approach follows the greedy concept by splitting the branches with lower prediction cost i.e., the class with the maximum data points should be classified initially at 0 level/root node [17].

2.2.5 K-nearest Neighbours

As the name suggests, this algorithm finds the separate clusters of data points present in proximity i.e., near to each other based on the distance between the two data points. In this classification approach, the K refers to the number of neighbors as the class labels and the mode of k labels is considered as the predicted outcome. The efficiency of this algorithm decreases with an increase in the number of predictor variables [18]. For a new input having real values, the distance is most likely to be measured through Euclidean distance given by Eq. 3:

$$EuclideanDistance(x, xi) = sqrt(sum((xj - xij)^2)) \tag{3}$$

where x is the new input and xi is the existing point covering all the j input attributes [19].

2.2.6 Gradient Boosting DT

As the name suggests, in this approach small steps are initiated from a point in a direction by enhancing the weak learners to make them strong. It consists of a cost function, feeble leaner, and preservative sequential approach to improving the presentation of the predictive model [20]. This classifier algorithm is highly used to optimize the user-defined cost functions by using the gradients in the loss function to make them controlled and realistic [21].

2.2.7 Multi-Layered Perceptron

As the name suggests, it refers to the neural networks or system of input, output layers, and various hidden layers between them with multiple neurons connected. A perceptron is referred to as a neuron with a random activation function. This algorithm uses the technique of backpropagation, a repetitive approach of combining the weights and the inputs which are achieved through the threshold function to minimize the cost function [22].

2.2.8 XGBoost

This is known as an extreme gradient boosting algorithm. In this approach the framework of gradient boosting is conserved. It is a highly optimized algorithm in terms of software as well as hardware resources usage for supercilious prediction outcomes in a quick time with minimal computing cost. This approach involves Gradient descent methodology as gradient boosting for strengthening the weak learners like CARTs [23].

3 Results and Discussions

This section will analyse all the classifier accuracy and performance using the confusion matrix and will give the proper comprehension regarding the best classifier algorithm. To do this analysis few pre-processing steps is to be applied on used dataset. First the count of null values of attributes will be checked it is seen that there are 8 rows of age attribute with null values. Now, describe the dataset to check the composition of the dataset. After this, the descriptive statistic of the dataset is collected, which is shown in Table 2.

Table 2. Description of dataset without replacing null values.

Parameters	Values
count	358.000000
mean	36.296089
std	15.324557
min	0.000000
25%	25.000000
50%	35.000000
75%	49.750000
max	75.000000

Now, in lieu of the null values with the median value of the attribute and describe the dataset again. Table 3 describes the altered dataset, after replacing the null values present in the age attribute.

Now, it can be clearly seen that there is only a slight difference in the mean frequency of the dataset. Mean frequency difference percentage = 0.18%. Now, import all the 8 classifier algorithms i.e., Naive Bayes, Support Vector Machines, Random Forest, XGBoost, Multi-layered perceptron, K-nearest neighbors, Decision tree, Gradient

Table 3. Description of dataset after replacing null values.

Parameters	Values
count	366.000000
mean	36.363388
std	15.037366
min	7.000000
25%	25.000000
50%	35.000000
75%	48.000000
max	75.000000

boosting DT. By meeting the internal classes ration in both sets, divide the dataset into a train set and a test set with a test size of 0.1. Now, the next step is to check the distribution of classes in training and test set. Figure 1 shows that the distribution of both the training and test set are in proportion.

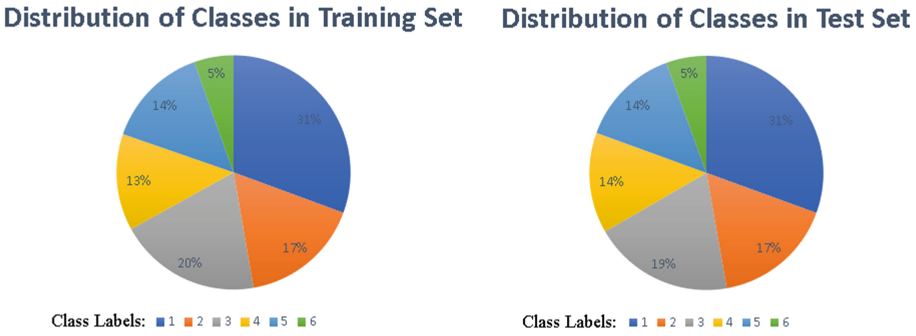


Fig. 1. Class distribution in training and test set

The internal ratio of classes is same in both the sets as shown in Fig. 10. After training these models with various classifier algorithms mentioned above and test for the accuracy scores. Figure 2 describes the accuracy scores of the classifier algorithms. It is calculated as the number of accurate forecasts produced divided by the total number of predictions, then multiplied by 100 [24]. Plot the confusion matrices for each of the classification algorithms. The confusion matrix serves as a performance metric for categorisation using machine learning. The output of the machine learning classification problem can be two or more classes, hence it is a performance evaluation for that problem [24]. The Figs. 3 shown below are the confusion matrices of the classifiers algorithms which compares the predicted class of ESD by classifiers algorithm to the actual class of ESD for a particular set of attribute values from test set.

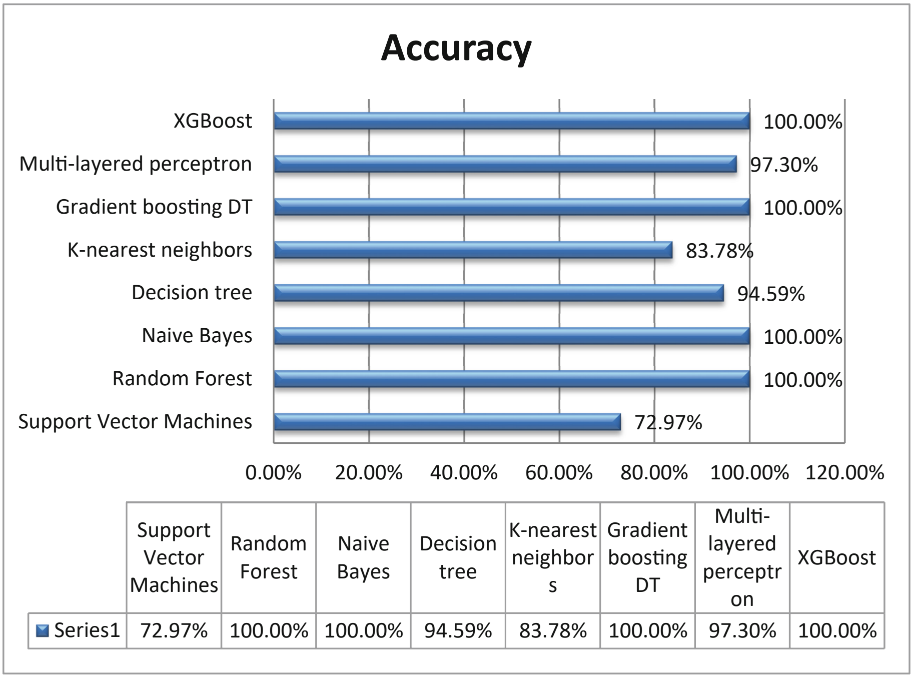
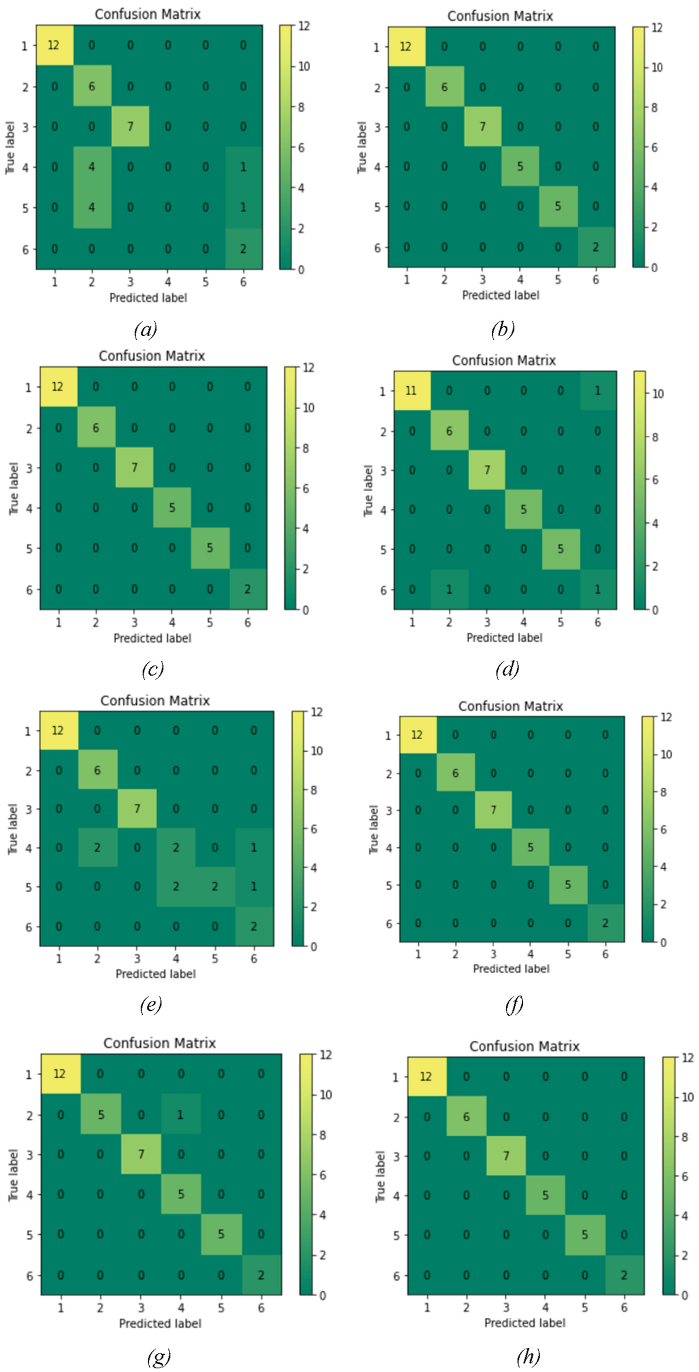


Fig. 2. Accuracies of classifier algorithms

From the above confusion matrices, we can clearly see predictions as per the accuracy scores of all the classes of ESD. The algorithms with 100 percent accuracy or the Ensemble of classifier algorithms with the best accuracies were employed for better classification about the prediction of the differential analysis of erythemato-squamous diseases constructed on their accuracies (ESD). By satisfying the internal classes ratio in both sets and checking the classification distribution in the trained and trial sets, the dataset was divided into a train set and a trial set with a test size of 0.1. After that, we can see that the internal class ratio is the same in both sets, and we also trained these models using the various classifier techniques discussed above, and we tested their accuracy: “Support Vector Machines (72.97%), Random Forest (100.0%), Naive Bayes (100.0%), Decision Tree (94.59%), K-nearest neighbors (83.78%), Gradient Boosting DT (100.0%), Multi-layered perceptron (97.3%), and XGBoost (100.0%) were the most popular”.

4 Challenges and Future Scope

Challenges in the current study is as follows: (i) In the data set, there is 2.2 percent missing data for the age attribute. The mean frequency was used to replace missing data with true values. As a result, training M.L. model would have been more effective if we had used real data. Our dataset size is small so it can lead to many problems like overfitting, Measurement errors, Missing values, Sampling Bias, etc. Due to this model



(a) *(b)* *(c)* *(d)* *(e)* *(f)* *(g)* *(h)*

Fig. 3. (a): SVM Confusion matrix, Figure (b): Random Forest Confusion matrix, Figure (c): Naive Bayes Confusion matrix, Figure (d): Decision tree Confusion matrix, Figure (e): KNN Confusion matrix, Figure (f): Gradient boosting DT Confusion matrix, Figure (g): MLP Confusion matrix, Figure (h): XGBoost Confusion matrix accuracy will be low and can produce very bad results also at sometimes. Like in if have the biased data then it can lead to the worst prediction [25]. (ii) The system will take time even if we use the best method with massive data. In some circumstances, this may result in the use of more CPU power. Furthermore, the data may take more storage space than is available. (iii) Vast quantity of data for training and testing is acquired. As a result of this technique, data inconsistencies may emerge. This is due to the fact that some data is updated on a frequent basis. As a result, we'll have to wait for more information. If this is not the case, the old and new data may produce contradictory results.

Future scope of the current study is as follows: (i) Automatic diagnosis of these illness groupings could aid physicians in making decisions. (ii) Medical testing in hospitals must be kept to a minimum. They can achieve these results by utilising appropriate computer-based info and/or decision-making technology. (iii) We can improve our app by using a larger dataset and creating an app that can predict a huge number of diseases. (iv) We can establish the link between clinical and histological features using our feedback methodologies.

5 Conclusion

The classification of psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, chronic dermatitis, and pityriasis rubra pilaris as erythemato-squamous disorders (ESD) is made possible by machine learning models that employ various classification algorithms, such as Support Vector Machines, Random Forest, and Naive Baye. We can see from the preceding calculations that utilizing the attribute's median to replace missing values results in a very small percentage change in the dataset's mean frequency. Furthermore, while comparing eight other classification methods, we can find that XGBoost fared exceptionally well with a 100% accuracy rate for three of them, namely Random Forest, Naive Bayes, and Gradient Boosting DT. XGBoost could be useful since it uses a mixture of software and hardware enhancement methods to provide supercilious prediction results with minimal computer assets in a small quantity of period.

References

1. Greiner, K.E.A.C.: Health Professions Education: A Bridge to Quality. National Academies Press (US), Washington (2003)
2. Brook, C.: What is a Health Information System? (2020)
3. Elsayad, M., Al-Dhaifallah, M., Nassef, A.M.: Analysis and diagnosis of erythemato-squamous diseases using CHAID decision trees. Yasmine Hammamet, Tunisia (2018)
4. Staff, M.C.: Psoriasis (2020)
5. Gary, C.W., Sara, P.M., Jaboori, K.A.: Diagnosis and treatment of seborrheic dermatitis, no. Feb 1, 2015 Issue (2015)
6. Usatine, R.P.: Diagnosis and Treatment of Lichen Planus. no. Jul 1, 2011 Issue (2011)
7. Staff, M.C.: Pityriasis rosea (2020)
8. Oakley, D.A.: Dermatitis (1997)
9. Vanessa Ngan, D.A.O.: Pityriasis rubra pilaris (2015)
10. Verma, S.S.P., Kumar, S.: Comparison of skin disease prediction by feature selection using ensemble data mining techniques. Inf. Med. Unlocked, 1–16 (2019)

11. Ilter, N., Guvenir, H.A.: Dermatology Data Set. (01 January 1998). [Online]. Available: https://archive.ics.uci.edu/ml/datasets/dermatology
12. Gandhi, R.: Support Vector Machine — Introduction to Machine Learning Algorithms. (7 June 2018). [Online]. Available: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47
13. Wang, H., Shao, Y., Zhou, S., Zhang, C., Xiu, N.: Support Vector Machine Classifier via Soft-Margin Loss. IEEE Trans. Pattern Analy. Mach. Intell. 1–11 (2021)
14. Brownlee, J.: Support Vector Machines for Machine Learning. (20 April 2016). [Online]. Available: https://machinelearningmastery.com/support-vector-machines-for-machine-learning/
15. Yiu, T.: Understanding Random Forest. (12 June 2019). [Online]. Available: https://towardsdatascience.com/understanding-random-forest-58381e0602d2
16. Gandhi, R.: Naive Bayes Classifier. (5 May 2018). [Online]. Available: https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c
17. Gupta, P.: Decision trees in machine learning (18 May 2017). [Online]. Available: https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052
18. Harrison, O.: Machine Learning Basics with the K-Nearest Neighbors Algorithm (11 September 2018). [Online]. Available: https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761
19. Brownlee, J.: K-Nearest Neighbors for Machine Learning (15 April 2016). [Online]. Available: https://machinelearningmastery.com/k-nearest-neighbors-for-machine-learning/
20. Brownlee, J.: A Gentle Introduction to the Gradient Boosting Algorithm for Machine Learning (9 September 2016). [Online]. Available: https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/
21. Singh, H.: Understanding Gradient Boosting Machines (November 4 2018). [Online]. Available: https://towardsdatascience.com/understanding-gradient-boosting-machines-9be756fe76ab
22. Bento, C.: Multilayer Perceptron Explained with a Real-Life Example and Python Code: Sentiment Analysis (21 September 2021). [Online]. Available: https://towardsdatascience.com/multilayer-perceptron-explained-with-a-real-life-example-and-python-code-sentiment-analysis-cb408ee93141
23. Morde, V.: XGBoost Algorithm: Long May She Reign! (8 April 2019). [Online]. Available: https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d
24. Agrawal, R.: The 5 Classification Evaluation metrics every Data Scientist must know (17 September 2019). [Online]. Available: https://towardsdatascience.com/the-5-classification-evaluation-metrics-you-must-know-aa97784ff226. Accessed 9 Decemeber 2021
25. EduPristine: Problems of Small Data and How to Handle Them (2016)