DOI Number: 10.37506/v11/i1/2020/ijphrd/193913

# Skin Diseases Prediction: Binary Classification Machine Learning & Multi Model Ensemble Techniques

**Vikas Chaurasia[1], Saurabh Pal[2]**

*[1]Research Scholor, MCA Dept., VBS Purvanchal University, Jaunpur,*
*[2]Dept. of MCA,VBS Purvanchal University, Jaunpur, UP, India*

## Abstract

Unlike many other diseases, the skin disease has more irritability. Dermatology sicknesses incorporates normal skin rashes to serious skin contaminations, which happens because of scope of things, like diseases, warm, allergens, framework issue and drugs. First regular skin issue are dermatitis. Atopic dermatitis is relating current (perpetual) condition that causes eager, aroused skin. Most much of the time it appears as patches on the face, neck, trunk or appendages. It will in general erupt sporadically so die down for a period. A large portion of the dermatological sicknesses are not reparable but rather most the treatments depend on the administration of the side effects related with it.

The focus of this research will be the Dermatology database. The problem is to determine the type of Eryhemato-Squamous disease like psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, cronic dermatitis and pityriasis rubra pilaris. The differential analysis of erythemato-squamous maladies is a genuine issue in dermatology. They all offer the clinical highlights of erythema and scaling, with next to no distinctions. Each pattern is a set of 33 numbers in the range linear values and one of them is nominal. The 80% of the dataset utilize for demonstrating and keep down 20% for approval. Objective is to accomplish best performer algorithm which will convey in dermatology informational collection so for this reason the gut feel recommends distance based calculations like k-Nearest Neighbors and Support Vector Machines may progress admirably. By using 10-fold cross validation and assess calculations utilizing the accuracy metric.

***Keywords:*** *Dermatology, Atopic dermatitis, Eryhemato-Squamous, k-Nearest Neighbors, Support Vector Machines.*

## Introduction

Around 1500 BC a medical document on skin ailments Ebers Papyrus was found in ancient Egypt. It portrays different skin maladies, including ulcers, rashes, and tumors, and recommends medical procedure and balms to treat the afflictions [1, 2]. From that point to now the skin sickness portion has indicated colossal development. The predominance of skin malady in India is 10 to 12 percent of the all out populace with Eczema and Psoriasis being the significant benefactors. Because of contamination, bright light, and an unnatural weather change, photosensitive skin issue like tanning, color obscuring, sunburn, skin malignant growths, and irresistible infections are expanding at a quicker pace. A one percent decrease in ozone prompts a two to four percent expansion in the occurrence of tumors. The seriousness of developing skin illnesses in India is additionally underlined by the way that the World Health Organization (WHO) has included skin infection under the most widely recognized non-transferable maladies in India. What's more, there is an absence of offices that give thorough skin related medicines under one rooftop." The circumstance is additionally compounded by the low accessibility of dermatologists in India. At present, there are around 6,000 dermatologists obliging a populace of more than 135 crore. This implies for each 100,000 individuals, just 0.49 dermatologists are

**Corresponding author:**
**Saurabh Pal**
Email: drsaurabhpal@yahoo.co.in

accessible in India when contrasted with 3.2 in numerous conditions of the US." Different tertiary consideration private setups come up short on the capacity to treat incessant, hereditary and pediatric skin diseases.

## Literature Survey

Investigating the ups and downs of the computerized skin disease conclusion system, several available arrangements are still under research and development. The difference between certain obstacles and shortcomings is that this arrangement subsequently attempts to overcome current problems in a variety of ways.

The different conclusions of erythema - squamous disease are a thorny problem in dermatology. They all provide clinical highlights of erythema and scales, with little difference. The disease at this gathering was psoriasis, sebaceous glands, lichen planus, pityriasis rosea, permanent dermatitis and pityriasis of the hair [3]. Some tests have been carried out in consideration of the discovery of erythematous squamous disease. These surveys link various technologies to specific issues and complete the unique correctness of the representation. In these investigations, the main work of the differential analysis of erythematous squamous disease is Table 1.

**Table 1: A few investigations which have dealt with skin disease mining**

| | | | |
|---|---|---|---|
| **Author** | **Year** | **Method** | **Classification accuracy** |
| Bojarczuk[4] | 2001 | A constrained-syntax genetic programmingC4.5 | 96.64%<br>89.12% |
| Chang et.al [5] | 2009 | decision tree<br>neural network | 80.33%<br>92.62% |
| Guvenir et al.[6] | 1998 | VFI5 | 96.2% |
| Ubeyli and Guler[7] | 2005 | ANFIS | 95.5% |
| Nani[8] | 2006 | LSVM<br>RS<br>B1_5<br>B1_10<br>B1_15<br>B2_5<br>B2_10<br>B2_15 | 97.22%<br>97.22%<br>97.5%<br>98.1%<br>97.22%<br>97.5%<br>97.8%<br>98.3% |
| Polat and Gunes[9] | 2009 | C4.5 and one-against-all | 96.71% |
| Ubeyli[10] | 2009 | CNN | 97.77% |
| Ubeyli and Dogdu[11] | 2010 | K-mean clustering | 94.22% |
| Lekka andMikhailov[12] | 2010 | Evolving fuzzy classification | 97.55% |
| Xie and Wang[13] | 2011 | IFSFS and SVM | 98.61% |
| A.A.L.C. Amarathunga et al[14] | 2015 | AdaBoost<br>BayesNet<br>J48,<br>MLP<br>NaiveBayes) | 85% for Eczema<br>95% for Impetigo<br>85% for Melanoma. |
| | | | |

## Method

Initially, differential expression analysis was used to select erythema-scaly, the most informative feature of significant differential expression, and then fed into the following classification process. Then, we use S-fold cross-validation technology to divide the initial data into k groups, training and test data sets. After that, multiple algorithms used for evaluation are learned from the training sets, each of which consist of k-1 of the k groups, and then applied to the corresponding test set, which is the remaining group of the S groups, to output the predicted class of the samples. Then, we will evaluate algorithms using the accuracy metric. This is a general indicator that quickly understands the correctness of a given model. We created a performance baseline on this issue and scrutinized many different issues of algorithms. Now we evaluate the same algorithms with a standardized copy of the dataset because we have the reason to suspect that the differing distributions of the raw data may be negatively impacting the skill of some of the algorithms. In this section we investigate tuning the parameters for most prominent algorithms that show promise from the spot-checking in the previous section. In next section, an ensemble model is another way to improve the accuracy by using four different machine algorithms; two boosting and two bagging methods. We will finalize the model by training it on the entire training dataset and make predictions for the hold-out validation dataset to confirm our findings. We can calculate from the entire training data set and apply the same transformation to the input properties of the validation data set. Finally, the algorithm-adjusted prediction is compared with the aggregate model to reduce the generalization error and obtain more accurate results [15-20].

### Evaluation of Algorithms

After the pre-processing of information collection, we evaluated the expected execution of six well-known classification techniques for finding skin diseases. Specifically, we apply Logistic Regression (LR), Linear Discriminant Analysis (LDA), k-Nearest Neighbor (KNN), Classification and Regression Tree (CART), Gaussian Bayesian (NB), Support Vector Machine (SVM). As the first time to arrange the classification model. These six classification strategies are highly accurate in practical applications.

## Ensemble Methods

In this research paper ensemble method is used as a method to find the accuracy of the skin disease dataset to improve the performance of algorithms. We will evaluate four different ensemble machine learning algorithms, two boosting Ada Boost (AB) and Gradient Boosting (GBM) and two bagging methods Random Forests (RF) and Extra Trees (ET).

### General Procedure of Boosting

The word "boost" refers to a set of algorithms that can transform weak learners into strong learners. Of course, weak learners are only slightly better than irregular guesses, while powerful learners are very close to perfect execution. If the appropriate response is positive, then any weak learner can be promoted to a strong learner, especially if it is difficult to obtain a strong learner rather than a weak learner. The promotion is done by continuously preparing a large number of learners and merging them into expectations, and later learners pay more attention to the mistakes of previous learners.

---

Input: Distribution of sample S;
Learning base algorithm A;
Total number of learning rounds N.
Process:
1. $S_1 = S$. # Initialization of distribution
2. for n = 1, . . . , N :
3. $h_n = A(S_n)$; # A weak learner trained from distribution $S_n$
4. $\mu_n = P_{x\sim Sn}(h_n(x) \neq f(x))$; # Evaluation of the error of $h_n$
5. $S_{n+1}$ = Adjustment of Distribution ($S_n$, $\mu_n$)
6. end
Output: H(x) = Combine Outputs ($\{h_1(x), \ldots, h_n(x)\}$)

---

Input: Distribution of sample S = $\{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$;
Learning base algorithm A;
Total number of learning rounds N.
Process:
1. for n = 1, . . . , N :
2. $h_n = A(S, S_{br})$ # $S_{br}$ is the bootstrap distribution
3. end
Output: $H(x) = \arg\max_{y \in Y} \sum_{n-1}^{N} (hn(x) = y)$

## Bagging Methods

The name Bagging originated from the contraction of Bootstrap AGGregatING. As the name implies, the two key elements of Bagging are bootstrap and aggregation. For training information collection, one probability is to examine various uncovered data subsets by all accounts, and after this preparation is the underlying learner from each subset. In any case, because we do not have endless training data information, such a program will provide few and non-representative examples, resulting in poor implementation of the underlying learners. The bag receives a guided loop to create a diverse base learner. It applies a bootstrap check to get a subset of the data used to prepare the underlying learner,

## Results

Here we utilized distance based algorithms like k-Nearest Neighbors and Support Vector Machines We have utilized 10-fold cross validation. The data set isn't excessively little and this is a decent standard test saddle setup. We will assess algorithms utilizing the exactness metric. This is a gross metric that will give a fast thought of how right a given model is. Progressively valuable on binary arrangement issues like this one. Making a standard of act on this issue and spot-check various distinctive algorithms. We will choose a suite of various algorithms fit for dealing with this classification issue. The algorithms all utilization defaults tuning parameters. On comparing the algorithms mean accuracy values are given in following table 2.

**Table 2: Output of Evaluating Algorithms**

| Algorithms | Mean Accuracy Values |
|---|---|
| LR | 0.979425 (0.022806) |
| LDA | 0.962299(0.024175) |
| KNN | 0.855747 (0.051314) |
| CART | 0.935057 (0.028180) |
| NB | 0.890230 (0.072177) |
| SVM | 0.921034 (0.027220) |

### Evaluation of Algorithms with Standardize Data

We speculate that the differing distributions of the raw data might be adversely affecting the ability of a portion of the algorithms. How about we assess similar algorithms with an standardized copy of the data set. This is the place the data is changed with the end goal that each attribute has a mean estimation of zero and a standard deviation of one. We likewise need to maintain a strategic distance from data leakage when we change the data. A decent method to keep away from data leakage is to utilize pipelines that standardize the data and construct the model for each fold in the cross validation test bridle.

**Table 3: Output of Evaluating Algorithms on the Scaled Dataset**

| Algorithms | Mean Accuracy Values |
|---|---|
| ScaledLR | 0.972529(0.025776) |
| ScaledLDA | 0.962299(0.024175) |
| ScaledKNN | 0.969195(0.018533) |
| ScaledCART | 0.938391(0.029970) |
| ScaledNB | 0.869655(0.087102) |
| ScaledSVM | 0.969310(0.023769) |

We can see that LR is as yet progressing admirably. We can likewise observe that the standardization of the data has lifted the aptitude of SVM to be the most precise algorithm tried up until this point. See Table 3.

The outcomes propose delving further into the LR and SVM algorithms. Almost certainly, setup past the default may yield significantly increasingly precise models.

**Ensemble Methods**

Another way that we can improve the execution of algorithms on this issue is by utilizing ensemble strategies. We will utilize a similar test tackle as previously, 10-fold cross validation. No data standardization is utilized for this situation since each of the ensemble algorithms depend on decision trees that are less sensitive to information distributions. See table 4.

**Table 4: Output of Evaluating Algorithms**

| Algorithms | Mean Accuracy Values |
|---|---|
| AB | 0.588391 (0.062920) |
| GBM | 0.959080 (0.036807) |
| RF | 0.959195 (0.036403) |
| ET | 0.969310 (0.041881) |

**Finalization of Model**

In view of the results, The LR demonstrated the most guarantee as a low intricacy and stable model for dermatology dataset. In this section we will conclude the model via preparing it on the whole training dataset and make predictions for the hold-out validation dataset to affirm our findings. A part of the findings was that LR performs better when the dataset is standardized so all characteristics have a mean estimation of zero and a standard deviation of one. We can ascertain this from the whole training dataset and apply the equivalent change to the input properties from the validation dataset.

We can see that we accomplish an exactness of about 99% on the held-out validation dataset. A score that more and improved to our desires evaluated above amid the tuning of LR. See table 5.

**Table 5: Output of Evaluating SVM on the Validation Dataset.**

| **accuracy_score** | **0.9864864864864865** |
|---|---|
| confusion_matrix | [[11 0 0 0 0 0]<br>[ 0 11 0 0 0 0]<br>[ 0 0 13 0 0 1]<br>[ 0 0 0 4 0 0]<br>[ 0 0 0 0 24 0]<br>[ 0 0 0 0 0 10]] |
| classification_report | precision recall f1-score support<br>cronic dermatitis 1.00 1.00 1.00 11<br>lichen planus 1.00 1.00 1.00 11<br>pityriasis rosea 1.00 0.93 0.96 14<br>pityriasis rubra pilaris 1.00 1.00 1.00 4<br>psoriasis 1.00 1.00 1.00 24<br>seboreic dermatitis 0.91 1.00 0.95 10<br>avg / total 0.99 0.99 0.99 74 |

## Conclusion

Skin disease is a disturbing and real medical issue around the world. In spite of the fact that the machine learning strategies have been increasingly more generally utilized in disease expectation, nobody technique beats all the others. In this paper, we exhibited six diverse order models and multi-model ensemble way to deal with the prediction of skin disease. The outcomes demonstrate that differential expression analysis is important to diminish the dimensionality of data and to choose effective data, along these lines expanding the accuracy of the prediction and decreasing the computational time to a substantial degree. The multi-model ensemble method at that point uses the expectations of numerous diverse classification models as input. The classification technique decreases the generation error and acquires more data by utilizing the principal organize predictions as highlights than it is trained in isolation. Also, by utilizing classification techniques, the mind boggling connections among the classifiers are found out consequently, in this way empowering the order strategy to accomplish better prediction.

**Ethical Clearance**- No ethical clearance is needed for this research paper.

**Funding**: This study was not funded by any funding

agency

**Competing Interests** None declared

## References

1. Mukhopadhyay AK. Dermatology in India and Indian dermatology: A Medico-historical perspective. Indian dermatology online journal. 2016 Jul;7(4):235.
2. Hartmann A. Back to the roots–dermatology in ancient Egyptian medicine. JDDG: Journal der Deutschen Dermatologischen Gesellschaft. 2016 Apr;14(4):389-96.
3. Güvenir HA, Demiröz G, Ilter N. Learning differential diagnosis of erythemato-squamous diseases using voting feature intervals. Artificial intelligence in medicine. 1998 Jul 1;13(3):147-65.
4. Bojarczuka CC, Lopesb HS, Freitasc AA. Data mining with constrained-syntax genetic programming: applications in medical data set. algorithms. 2001;6:7.
5. Chang CL, Chen CH. Applying decision tree and neural network to increase quality of dermatologic diagnosis. Expert Systems with Applications. 2009 Mar 1;36(2):4035-41.
6. Güvenir HA, Emeksiz N. An expert system for the differential diagnosis of erythemato-squamous diseases. Expert Systems with Applications. 2000 Jan 1;18(1):43-9.
7. Übeylı ED, Güler I. Automatic detection of erythemato-squamous diseases using adaptive neuro-fuzzy inference systems. Computers in biology and medicine. 2005 Jun 1;35(5):421-33.
8. Nanni L. An ensemble of classifiers for the diagnosis of erythemato-squamous diseases. Neurocomputing. 2006 Mar 1;69(7-9):842-5.
9. Polat K, Güneş S. A novel hybrid intelligent method based on C4. 5 decision tree classifier and one-against-all approach for multi-class classification problems. Expert Systems with Applications. 2009 Mar 1;36(2):1587-92.
10. Übeyli ED. Combined neural networks for diagnosis of erythemato-squamous diseases. Expert Systems with Applications. 2009 Apr 1;36(3):5107-12.
11. Übeyli ED, Doğdu E. Automatic detection of erythemato-squamous diseases using k-means clustering. Journal of medical systems. 2010 Apr 1;34(2):179-84.
12. Lekkas S, Mikhailov L. Evolving fuzzy medical diagnosis of Pima Indians diabetes and of dermatological diseases. Artificial Intelligence in Medicine. 2010 Oct 1;50(2):117-26.
13. Xie J, Wang C. Using support vector machines with a novel hybrid feature selection method for diagnosis of erythemato-squamous diseases. Expert Systems with Applications. 2011 May 1;38(5):5809-15.
14. Amarathunga AA, Ellawala EP, Abeysekara GN, Amalraj CR. Expert system for diagnosis of skin diseases. International Journal of Scientific & Technology Research. 2015 Jan;4(01):174-8.
15. https://scikit-learn.org/stable/modules/feature_selection.html
16. Perriere G, Thioulouse J. Use of correspondence discriminant analysis to predict the subcellular location of bacterial proteins. Computer Methods and Programs in Biomedicine. 2003 Feb 1;70(2):99-105.
17. Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician. 1992 Aug 1;46(3):175-85.
18. Chaurasia V, Pal S, Tiwari BB. Prediction of benign and malignant breast cancer using data mining techniques. Journal of Algorithms & Computational Technology. 2018 Jun;12(2):119-26.
19. Chaurasia V, Pal S, Tiwari BB. Prediction of benign and malignant breast cancer using data mining techniques. Journal of Algorithms & Computational Technology. 2018 Jun;12(2):119-26.
20. Yadav, D., Pal, S. To Generate an Ensemble Model for Women Thyroid Prediction Using Data Mining Techniques. *Asian Pacific Journal of Cancer Prevention.* 2019; 20 (4).