ISSN: 2754-494X

Journal of Dermatology Research Reviews & Reports

SCIENTIFIC Research and Community

Research Article

Open Access

Classification of Erythematosquamous Dermatosis by the Method of Random Forest

Dwaipayan Mukhopadhyay[1], Dieudonne D Phanord[2], Rohan J Dalpatadu[2], Laxmi P Gewali[3] and Ashok K Singh[4*]

[1]Ph.D. Candidate, Department of Mathematical Sciences, University of Nevada Las Vegas, USA

[2]Professor, Department of Mathematical Sciences, University of Nevada Las Vegas, USA

[3]Laxmi P Gewali, Professor, Department of Computer Science, University of Nevada Las Vegas, USA

[4]Chair & Professor, Department of Resorts Gaming & Golf Management, University of Nevada Las Vegas, USA

***Corresponding author**

Ashok K Singh, Chair & Professor, Department of Resorts Gaming & Golf Management, University of Nevada Las Vegas, USA. Tel: 702-845-0391.

Received: December 28, 2022; **Accepted:** January 18, 2023; **Published:** January 20, 2023

Introduction

Machine Learning (ML) methods have found wide applications in dermatology (Chan et al., 2020) [1]. Thomsen, Iversen, Titlestad & Winther (2020) reviewed 2175 publications and found that the most common usage of ML methods was in the binary classification of malignant melanoma from images [2]. Adamson and Smith have a word of advice about usage of ML methods in diagnosis of skin diseases that inclusivity must be kept in mind for classification results to be accurate [3]. Steele et al. searched PubMed, Embase, and CENTRAL, and found that the performance of ML methods was variable, and overall accuracy measure was not a good measure for sub-group accuracy [4].

Erythematosquamous Dermatosis has symptoms of itchy skin or pruritus; possible causes for pruritus include an underlying medical condition, contact with an irritant or a reaction to a medication [5]. An important gene associated with this skin disease (MalaCards) is IL22RA1 (Interleukin 22 Receptor Subunit Alpha 1). Back in 1998, Guvenir, Demiroz, and Ilter introduced a new classifier called Voting Feature Intervals (VFI) in which each feature voted on a class, with the class getting the most votes declared as the predicted class value [6,7]. The overall accuracy of VFI was reported to be 99.2%. Data for this article is given in Dua and Graff [8]. Singh, Sinha and Yadav used logistic regression, support vector machine and K-Nearest neighbor classifiers on this dataset and computed accuracy measures [9]. Rathore et el. (2022) used the XGBoost model on this dataset [10]. We will use this data set and apply the method of random forest which uses a bagging algorithm: a Random Forest (RF) model randomly selects features to use from the set of all features, grows a large number of trees, then uses majority vote to classify the response for each observation. We will also compute the accuracy-based importance measure for each feature, and then fit a reduced RF model using most important features.

Data

The dataset has 366 subjects with measurements on 34 features and a categorical response variable [8]. One of these features is age, a continuous variable; family. history is binary, eosinophils. Infiltrate is (0,1,2) ordinal, and the remaining 31 features are all (0,1,2,3) ordinal. The response variable is categorical with 6 levels:

C1: psoriasis
C2: seboreic dermatitis
C3: lichen planus
C4: pityriasis rosea
C5: cronic dermatitis
C6: pityriasis rubra pilaris

There are 8 missing values in the Age column, which are all removed yielding the final dataset of 358 observations on 35 variables. This dataset is split into a 75% training set of 269 rows and a 25% test set of 89 rows; RF model is then fit to the training set, and accuracies are computed for both training and test sets separately.

Method: Random Forest Classifier

The method of random forest is a decision-trees based supervised learning method for categorical or continuous response variable Y. It randomly selects a subset of observations and a subset of features at a time to fit a large number of decision trees to predict Y and then averages (mode for classification, mean for regression) these predicted Y values for the predicted Y. Random forest is one of the most accurate predictive methods and it reduces overfitting [11, 12]. All computations and graphs are done in the statistical software environment R [13]. Even though the R package random Forest yields out of bag (OOB) accuracy, we report all accuracy measures for multi-class classification for both training and test datasets [14].

Accuracy Measures For Multi-Class Classification

Commonly used measures for multi-level classifiers (accuracy, precision, recall and F1 [15] are briefly described. These measures are calculated from the confusion matrix shown in Table 1.

Table 1: Confusion matrix for the 6-class classification problem

($C_{i,j}$ = number of times true response of j get predicted as i; i, j = 1, 2, …, 6).

	Observed					
Predicted	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**
C1	C1,1	C1,2	C1,3	C1,4	C1,5	C1,6
C2	C2,1	C2,2	C2,3	C2,4	C2,5	C2,6
C3	C3,1	C3,2	C3,3	C3,4	C3,3	C3,4
C4	C4,1	C4,2	C4,3	C4,4	C4,3	C4,4
C5	C5,1	C5,2	C5,3	C5,,4	C5,5	C5,6
C6	C6,1	C6,2	C6,3	C6,,4	C6,5	C6 ,6

The one vs all binary performance measures accuracy, precision, recall, F1 and the overall prediction accuracy [15, 16. are calculated from the following formulas:

$$\text{Overall prediction accuracy} = \frac{\sum_{j=1}^{6} C_{j,j}}{\sum_{i=1}^{6}\sum_{j=1}^{6} C_{j,j}} = \frac{\text{Sum of diagonal elements of the confusion matrix}}{\text{Sum of all elements of the confusion matrix}}$$

$$\text{Precision}_j = \frac{C_{j,j}}{\sum_{k=1}^{6} C_{j,k}} = \frac{j-th \text{ diagonal element of the confusion matrix}}{\text{Sum of j-th row of the confusion matrix}}$$

$$\text{Recall}_j = \frac{C_{j,j}}{\sum_{k=1}^{6} C_{k,j}} = \frac{j-th \text{ diagonal element of the confusion matrix}}{\text{Sum of j-th column of the confusion matrix}}$$

$$F1_j = \frac{2 \times \text{Precision}_j \times \text{Recall}_j}{(\text{Precision}_j + \text{Recall}_j)} = \text{the harmonic mean of Precision and Recall for class } j$$

The Area Under the Curve (AUC) for each class is given by [15. (Molin et al 2021):

$$AUC_j = \frac{1}{2}\left(\frac{TP_j}{TP_j + FN_j} + \frac{TN_j}{TN_j + FP_j}\right), j=1,2,...6.$$

where TPj = true positive, TNj = true negative, FPj = false positive and FNj = false negative for the j-th class, shown below as elements of the confusion matrix CM_j for the j-th class (j = 1, …,6):

$$CM_j = \begin{bmatrix} & \textit{Observed} & \\ & C_j = 1 & C_j = 0 \\ C_j = 1 & TP_j & FP_j \\ C_j = 0 & FN_j & TN_j \end{bmatrix}$$

For multi-class classification problems, macro- and micro-averages of the above measures [15-17. are also included.

$$recall_{macro} = \frac{\sum_{j=1}^{6} recall_j}{6}$$

$$precision_{macro} = \frac{\sum_{j=1}^{6} precision_j}{6}$$

$$recall_{micro} = \frac{\sum_{j=1}^{6} TP_j}{\sum_{j=1}^{6} TP_j + \sum_{j=1}^{6} FN_j}$$

$$precision_{micro} = \frac{\sum_{j=1}^{6} TP_j}{\sum_{j=1}^{6} TP_j + \sum_{j=1}^{6} FN_j}$$

$$AUC_{macro} = \frac{\sum_{j=1}^{6} AUC_j}{6}$$

There is no well-accepted multi-class Receiver Operating Characteristic Analysis and hence micro-averaged AUC's are not computed [15, 17].

ML literature recommends splitting the original dataset into a training set and a test set and reporting all performance metrics for both training and test sets [16].

Results

The package randomForest was used with 250 trees (parameter ntree = 250) to fit a random forest classifier to the training data set using all 34 features. Variable importance for each feature was computed using decrease in accuracy as the measure of feature importance. The RF model fitted to the training set was then used to predict the response for the test set, Tables 2 and 3 show the accuracy measures of the RF classifier for the training and test datasets, respectively.

Table 2: The confusion matrix and the prediction accuracy of the RF classifier using all 34 features for the training set

Observed										
Predicted	**1**	**2**	**3**	**4**	**5**	**6**	**Recall**	**Precision**	**F1**	**AUC**
1	84	0	0	0	0	0	1	1	1	1
2	0	38	0	0	0	0	1	1	1	1
3	0	0	57	0	0	0	1	1	1	1
4	0	0	0	35	0	0	1	1	1	1
5	0	0	0	0	38	0	1	1	1	1
6	0	0	0	0	0	17	1	1	1	1
Macro average							1	1	1	1
Micro average							1	1	1	

Table 3: The confusion matrix and the prediction accuracy of the RF classifier using all 34 features for the test set

Observed										
Predicted	**1**	**2**	**3**	**4**	**5**	**6**	**Recall**	**Precision**	**F1**	**AUC**
1	27	0	0	0	0	0	1.00	1.00	1.00	1.00
2	0	20	0	1	0	0	0.91	0.95	0.93	0.95
3	0	0	14	0	0	0	1.00	1.00	1.00	1.00
4	0	2	0	12	0	0	0.92	0.86	0.89	0.95
5	0	0	0	0	10	0	1.00	1.00	1.00	1.00
6	0	0	0	0	0	3	1.00	1.00	1.00	1.00
Macro average							1.00	0.95	1.00	0.98
Micro average							0.95	1.00	1.00	

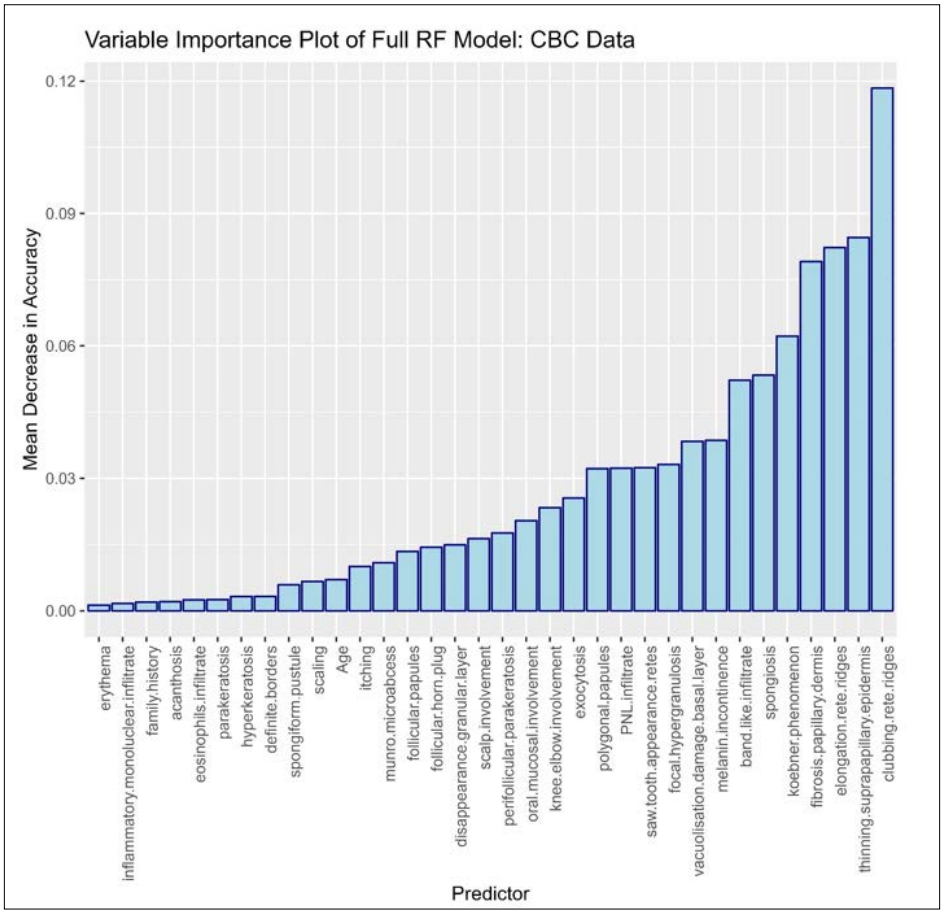

Figure 1: Variable Importance Plot for the Full RF Model

Figure 1, the variable importance plot of the Full RF model (i.e., the RF model with all features in the model), shows the mean decrease in prediction accuracy for each feature if the feature is removed from the model.

We next drop the bottom 17 features and fit the Reduced RF Model. Tables 4 and 5 show the confusion matrix and the prediction accuracy of the reduced RF classifier.

Table 4: The confusion matrix and the prediction accuracy of the RF classifier using best 17 features for the training set

Observed										
Predicted	**1**	**2**	**3**	**4**	**5**	**6**	**Recall**	**Precision**	**F1**	**AUC**
1	84	0	0	0	0	0	1	1	1	1
2	0	38	0	5	0	0	1	0.88	0.94	1
3	0	0	57	0	0	0	1	1	1	1
4	0	0	0	30	0	0	0.86	1	0.92	1
5	0	0	0	0	38	0	1	1	1	1
6	0	0	0	0	0	17	1	1	1	1
Macro average							0.97	0.96	0.97	1.00
Micro average							0.98	0.98	0.98	

Table 5: The confusion matrix and the prediction accuracy of the RF classifier using best 17 features for the test set

Observed										
Predicted	**1**	**2**	**3**	**4**	**5**	**6**	**Recall**	**Precision**	**F1**	**AUC**
1	27	0	0	0	0	0	1.00	1.00	1.00	1.00
2	0	20	0	1	0	0	0.91	0.95	0.93	0.95
3	0	0	14	0	0	0	1.00	1.00	1.00	1.00
4	0	2	0	12	0	0	0.92	0.86	0.89	0.95
5	0	0	0	0	10	0	1.00	1.00	1.00	1.00
6	0	0	0	0	0	3	1.00	1.00	1.00	1.00
Macro average							0.95	0.93	0.94	0.98
Micro average							0.95	0.95	0.95	

It can be seen from Tables 4 and 5 that for both training and test datasets, the RF classifier based on top 17 features is quite accurate.

Figure 2: Shows the Variable Importance Plot of the reduced RF model

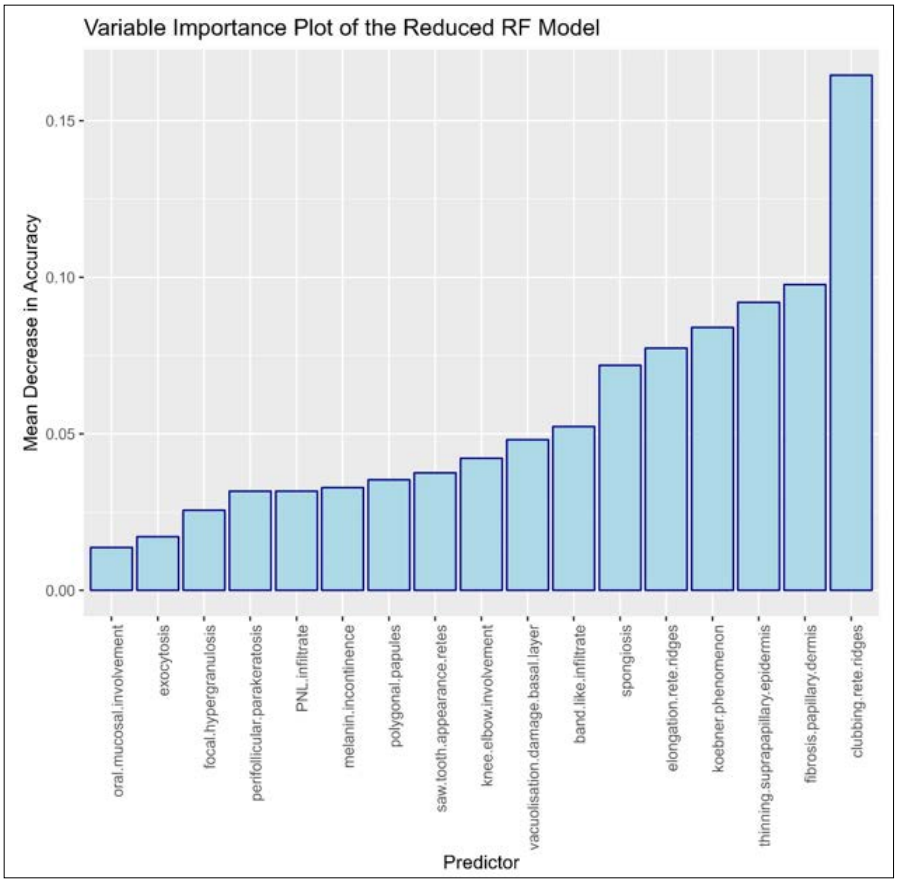

Figure 2: Variable Importance Plot for the Reduced RF Model

Conclusions

We have successfully demonstrated that the method of RF classifier is able to classify Erythematosquamous Dermatosis with high accuracy.

References

1. Chan S, Reddy V, Myers B, Thibodeaux Q, Brownstone N, et al. (2020) Machine Learning in Dermatology: Current Applications, Opportunities, and Limitations. Dermatol Ther (Heidelb) 10: 365-386.
2. Kenneth Thomsen, Lars Iversen, Therese Louise Titlestad, Ole Winther (2020) Systematic review of machine learning for diagnosis and prognosis in dermatology. Journal of Dermatological Treatment 31: 496-510.
3. Adamson AS, Smith A (2018) Machine Learning and Health Care Disparities in Dermatology. JAMA Dermatol 154:1247-1248.
4. Steele L, Tan X, Olabi B, Gao J, Tanaka R et al. (2022) Determining the clinical applicability of machine learning models through assessment of reporting across skin phototypes and rarer skin cancer types: a systematic review. J Eur Acad Dermatol Venereol. Accepted Author Manuscript. https://doi.org/10.1111/jdv.18814
5. Cleveland Clinic (2022) https://my.clevelandclinic.org/health/diseases/11879-pruritus
6. MalaCards. https://www.malacards.org/card/erythematosquamous_dermatosis
7. Guvenir HA, Demiroz G, Ilter N (1998) Learning differential diagnosis of erythemato-squamous diseases using voting feature intervals, Artificial Intelligence in Medicine 13: 147-165.
8. Dua D, Graff C (2019) UCI Machine Learning Repository http://archive.ics.uci.edu/ml Irvine, CA: University of California, School of Information and Computer Science.
9. S K Singh, A Sinha, S Yadav (2022) "Performance Analysis of Machine Learning Algorithms for Erythemato-Squamous Diseases Classification," 2022 IEEE International Conference on https://ieeexplore.ieee.org/document/9793000
10. Abhishek Singh Rathore, Siddhartha Kumar Arjaria, Manish Gupta, Gyanendra Chaubey, Amit Kumar Mishra, et al. (2022) Erythemato-Squamous Diseases Prediction and Interpretation Using Explainable AI, IETE Journal of Research, DOI: 10.1080/03772063.2022.21149537
11. Hastie T, Tibshirani R, Friedman J (2009) Springer series in statistics, The elements of statistical learning: Data mining, inference, and prediction (2nd ed.) 587-590.
12. Schonlau M, Zou R Y (2020) The random forest algorithm for statistical learning. The Stata Journal 20: 3-29.
13. R Core Team (2021) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
14. A Liaw and M Wiener (2002) Classification and Regression by randomForest. R News 2: 18-22.
15. Molin, Nicole, Molin, Clifford, Dalpatadu, et al. (2021) Prediction of Obstructive Sleep Apnea Using FFT of Overnight Breath Recordings Machine Learning with Applications, Volume 4: 100022.
16. Tutz G (2011) Regression for categorical data. Cambridge University Press 210-214.
17. Sokolova M, Lapalme G (2009) A systematic analysis of performance measures for classification tasks. Information Processing and Management 45: 427-437.