EasyChair Preprint

№ 4868

Early Prediction Of Erythemato-Squamous Disease With Machine and Deep Learning Approaches

Harshvardhan Tiwari, Preeti V Patil, K R Sinchana,
Shiji K Shridhar and G Aishwarya

January 5, 2021

Early Prediction Of Erythemato-Squamous disease With Machine And Deep Learning Approaches

Harshvardhan Tiwari
Centre for Incubation, Innovation, Research and Consultancy, Jyothy Institute of Technology, Bengaluru, Karnataka, India
tiwari.harshvardhan@gmail.com

Preeti V Patil
Department of ISE,
Jyothy Institute of technology
preetivp2004@gmail.com

Sinchana K R
Department of ISE,
Jyothy Institute of technology
sinchanakr1207@gmail.com

Shiji K Shridhar
Department of ISE,
Jyothy Institute of technology
Shijiks20@gmail.com

Aishwarya G
Department of ISE,
Jyothy Institute of technology
gaishwarya03@gmail.com

Abstract:

Now a days skin diseases are a major health problems among the many common people.The classification and recognition systems have been improved in a great deal to help in the medical experts in diagnosing diseases.Here we develop the different machine leraning techniques,which can diagnose erythemato-squamous disease.The machine learning techniques applied to skin diseases prediction so far has better outcome over all the others.Here we apply five different machine learning techniques and then develop an ensamble approach that consist of all the five different machine leaning techniques as a single unit.We use informative

Dermatology data to analysis different data mining techniques to classify the skin disease and then, an ensemble machine learning method is applied.This study has focused on detection of erythemato-squamous on the dermatology dataset using different machine learning predictive techniques such as Logistic Regression [LR], Support Vector Machine [SVM], Random Forest [RF],Linear Regression,Decision Tree, AdaBoost Classifier.

Keywords:Erythemato-squamous disease,Dermatology dataset,Logistic Regression [LR], Support Vector Machine [SVM], Random Forest [RF],Linear Regression,Decision Tree,AdaBoost Classifier.

1.Introduction:

The skin is the outer most covering part of the body and it is the largest organ of the integumentary system.The skin has seven layers of ectodermal tissue and guards on the underlying muscles,bones,ligaments and internal organs.The skin protects the body from UV radiation infections, injuries, heat and harmful radiation, and also helps in the manufacture of vitamin D. The skin plays an important role in controlling body temperature, so it is important to maintain good health and protect the body from the skin diseases.Skin disease is one of the popular diseases among other diseases these days. Some skin conditions can be minor, temporary, and easily treated while others can be very serious, and even deadly. Dermatology diseases detection is really a difficult problem even for the experienced doctors.
Erythemato-Squamous is a class of dermatological diseases with redness of the skin due to the loss of the skin cells. Patient needs dermatologist that has the wide and substantial knowledge and experience in these of the diseases. The similarity of the clinical features of these diseases with erythema and scaling make the differential diagnosis of erythemato-squamous diseases is real difficult in dermatology .
A number of studies have shown that the diagnosis of one patient can differ with other patients significantly if the patient is examined by different physicians or even by the same physician at various times. By improving technology, the development of the systems which may help physicians is extremely important at the diagnosis. Computer-aided diagnostic [CAD] systems hold much promise for assisting in the differential diagnosis of diseases that bear minor differences in presentation with less time.Indeed, computer-aided diagnosis systems act as a second opinion for physicians.Several techniques such as data mining and pattern recognition can be used in these systems.
The diagnosis of erythemato-squamous disease is a complex problem and difficult to detect in dermatology. Besides that, it is a major cause of skin cancer. There are six groups in erythemato-squamous disease. They are psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, chronic dermatitis, and pityriasis rubra pilaris. They have little difference in their clinical and histopathological features. The disease is often found in an outpatient department of dermatology.Diagnosis of the disease is usually using the biopsy technique, but unfortunately histopathological features also greatly affect in the diagnosis. Another difficulty is the diagnosis may show the

features of other diseases at beginning stage and their characteristic features will appear at the following stage.

2.Related Work:

There have been several studies reported to focusing on the diagnosis of erythemato-squamous diseases using dermatology dataset.The first attempt was done by Guvenir, Demiroz and new classifier algorithm, the voting feature intervals-5 algorithm was developed. For performance of the evaluation, the accuracy was measured. And they achieved the accuracy of the classifier was 96.2%.
Menai and Altayash investigated the performance of boosting decision tree as an ensemble strategy for the diagnosis of erythemato-squamous disease. The result illustrated that the ensemble of unpruned decision tree had a better accuracy [96.72%] than other methods such as genetic algorithm [GA] and k-means clustering in other studies.
Xie,Xie has developed a Support Vector Machine [SVM] model with a novel hybrid feature of the selection method, called Improved F -score and Sequential Forward Floating Search [IFSFFS] which was a combination of Sequential Forward Floating Search [SFFS] and Improved F-score [IF] to carry out the optimal feature subset selection. The IF and SFFS based on SVM are evaluation criteria for filters and wrappers, respectively. They calculated the accuracy of training and testing data set for different splits. The average training and testing accuracy was 99.29% and 97.58%, respectively. Olatunji and Arif proposed and implemented a novel identification model for the diagnosis of ESD based on the Extreme Learning Machine [ELM]. The new model compared with classic Artificial Neural Networks [ANN]. ELM had some of the advantages, such as higher learning speed, the best performance and the ease of implementation. Results showed that the ELM had the greater degree of performance rather than classic ANN in both training and testing data set.

Ravichandran, Narayanamurthy used the Fuzzy Extreme Learning Machine (FELM) method to purpose the ESD classification. FELM consists of two methods: Fuzzy Logic and ELM. The achieved accuracy of the FELM was 92.84%, which was better than other mathods in terms of accuracy.

Accuracy of C4.5 classifier calculated as the 84.48% in the study of Polat and Guneş.They combined C4.5 method with one-against-all method and accuracy of 96.71% was achieved. In an expert system based on three classifier method including the decision tree, fuzzy weighted pre-processing and k-NN based weighted pre-processing has been designed in order to help the physicians for diagnosis of ESD. While, an ensemble of SVM based on the random subspace [RS] and the feature selection were employed for the diagnosis of ESD by Nanni.It was stated that classifiers using the different features offers complementary information about the classifiable patterns. Abdi and Giveki developed the diagnosis model for the automatic detection of ESD, using the particle swarm optimization [PSO] method and the SVM method based on Association Rules [AR]. The accuracy of 98.91% was achieved for AR-PSO-SVM model. Ubeyli used the SVM to classify the ESD.

3.Materials and Methods:

3.1 Description of dataset:

Here is the brief description of the patient dataset used here.This dataset is obtained from the University of California Irvine in Machine Learning.
This data set includes a total number of 34 discrete valued features and 366 samples are collected in the clinical setting.The 35^{th} nominal attribute is included which indicates the dermatological diagnosis of the patient who is suffering from erythemato-squamous disease.
Each sample contains the 12 clinical features and the 22 histopathological features obtained after the biposy of the patient skin disease.Some patients can be diagnosed without the biposy with some clinical features,but biposy is necessary for a correct diagnosis.The name of the six erythemato-squamous disesse are provided in the below given table 1.
In this dataset, the family history feature assumes value 1 or 0 depend on weather the disease has been observed or not observed in the family. Other clinical and histopathological features can assume,instead,a degree in the range is between 0-3: the value 0 indicates the absence of the particular feature, the largest possible amount of feature is represented by degree 3 while 1 and 2 denote intermediate values. The age feature is linearly valued and indicates the patient age in years. The data set contained eight missing values in the Age attribute in some instances. Hence, we consider 358 complete samples in our data system.

Table 1:THE UCI ERYTHEMATO-SQUAMOUS DISEASES DATASET

Erythemato-Squamous diseases (Number of patients)	**Features**	
	Clinical	Histopathological
Psoriasis (111)	Feature 1: Erythema	Feature 12 : Melanin incontinence
Seboreic dermatitis (60)	Feature 2 : Scaling	Feature 13 : Eosinophils in the infiltrate
Lichen planus (71)	Feature 3 : Definite borders	Feature 14 : PNL infiltrate
Pityriasis rosea (48)	Feature 4 : Itching	Feature 15 : Fibrosis of the papillary dermis
Cronic dermatitis (48)	Feature 5 : Koebner phenomenon	Feature 16 : Exocytosis
Pityriasis rubra pilaris (20)	Feature 6 : Polygonal papules	Feature 17 : Acanthosis
	Feature 7 : Follicular papules	Feature 18 : Hyperkeratosis
	Feature 8 : Oral mucosal involvement	Feature 19 : Parakeratosis
	Feature 9 : Knee and elbow involvement	Feature 20 : Clubbing of the rete ridges
	Feature 10 : Scalpinvolvement	Feature 21 : Elongation of the rete ridges
	Feature 11 : Family history, (0 or 1)	Feature 22 : Thinning of the suprapapillary epi dermis
	Feature 34 : Age (linear)	Feature 23 : Spongiform pustule
		Feature 24 : Munro microabcess
		Feature 25 : Focal hypergranulosis
		Feature 26 : Disappearance of the granular layer
		Feature 27 : Vacuolisation and damage of basal layer
		Feature 28 : Spongiosis
		Feature 29 : Saw-tooth appearance of retes
		Feature 30 : Follicular horn plug
		Feature 31 : Perifollicular parakeratosis
		Feature 32 : Inflammatory monoluclear infiltrate
		Feature 33 : Band-like infiltrate

3.2 Feature Importance Analysis

Feature importance analysis provides the feature score for each feature present in the dataset. High feature score for an attribute indicates that the attribute is more important for output variable. In the presented work their are 35 input features present in the dermetology dataset.

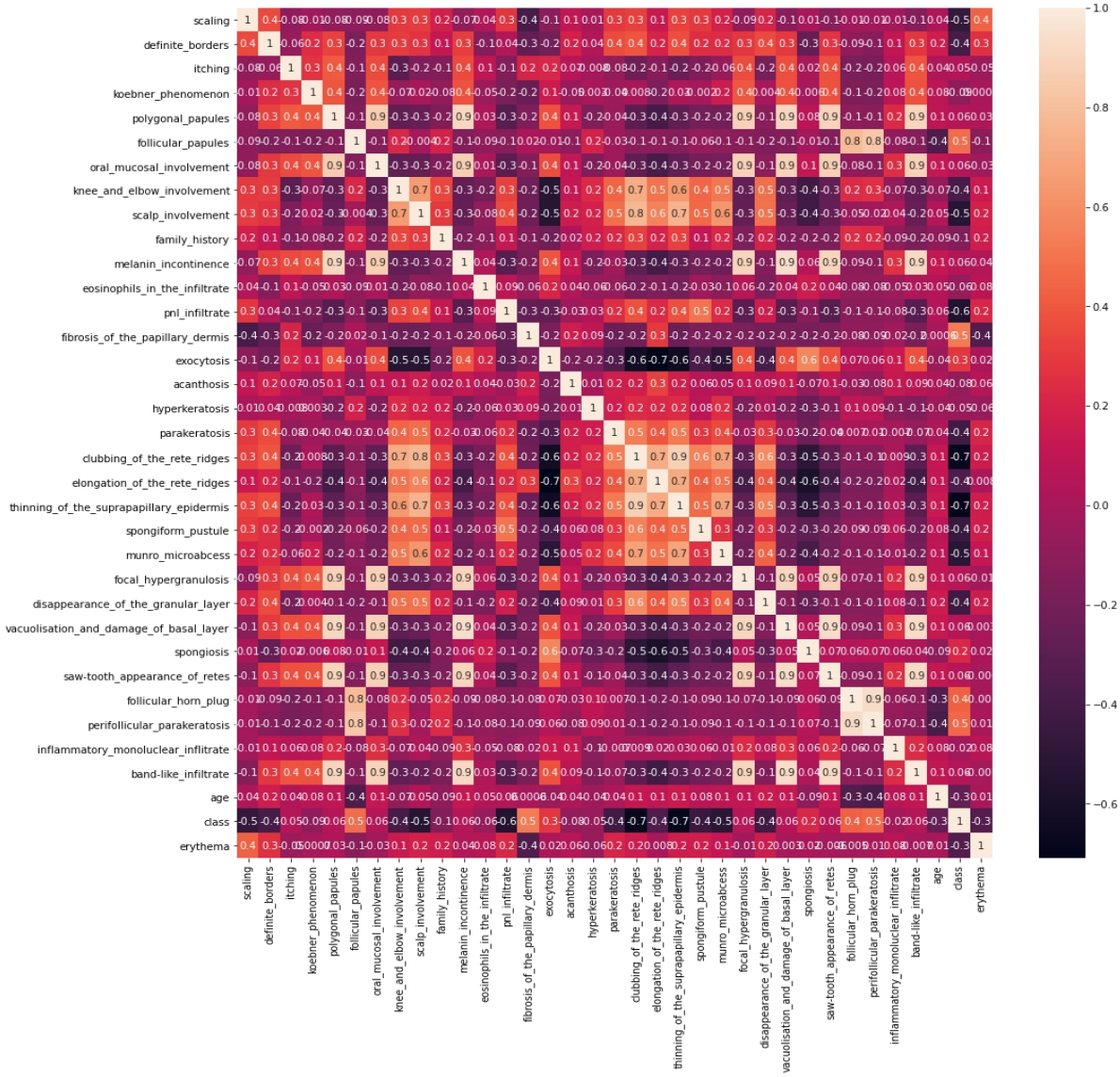


Fig 1:Correlation between features in Dermetology dataset.

3.3 Deep Learning application technique:

3.3.1.Logistic Regression:

It is a classification algorithm used to perdict the binary outcomes for a given dataset of independent variables.The dependent variables outcome is discrete.Here the response variable is categorical in nature.It also helps to calculate the possibility of the particular evnt taking place.It is a 'S' shaped curve[S-Sigmoid].It is used to slove the classification problem.In order to map the predicted values to the probabilities, we use the sigmoid function. The function maps any real value into another value between 0 and 1. In machine learning, we use sigmoid to map the predictions to the probabilities.

$$S(z)=1/1+e-z$$

Where,

S(z)=output between 0 and 1

z=input to the function

e=base of natural log

3.3.2.Linear Regression:

Linaer regression is a supervised learning.It is a statistical method thst helps to find the relationship between an independent variable and dependent variable,both are continuous.Here the response variables are continuous in nature.It also helps to estimate the dependent variable when there is a change in the independent variable.It is a straight line curve.It is used to slove the regression problems. If there appears to be no association between the proposed explanatory and dependent variables, then the fitting a linear regression model to the data probably will not provide a useful model. A valuable numerical measure of the association between two variables is the correlation coefficient , which has the value between -1 and 1 indicating the strength of the association of the observed data for the two variables.

A linear regression line has an equation of the form ***Y = a + bX***, where ***X*** is the explanatory variable and ***Y*** is the dependent variable. The slope of the line is ***b***, and ***a*** is the intercept [the value of ***y*** when ***x*** = 0].

3.3.3.Support Vector Machine[SVM]:

Support vector machine is a discriminatice classifier that is formally designed by a separative hyperplane.It is a representation example as the points in space that are mapped so that points of diiferent categories are seperated by a gap as wide possible.The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space[N - the number of features] that distinctly classifies the data points.It is mainly used to convert the low dimensional space to high dimensional space,it makes easier to seperate the data.It makes more secure.

3.3.4.Decision Tree:

Decision tree is a supervised learning.Here we take the decision using the tree structure.Here each branch node represents the choice and the leaf node represents the decision.Decision tree is one of the most predictive modelling approaches used in statistics,data mining and machine learning.Decision trees are constructed via an algorithmic approach that identifies the ways to split the data set based on the different conditions. It is one of the most widely used and practical methods for the supervised learning. Decision Trees are the non-parametric supervised learning method used for both classification and regression tasks.Tree models where the target variable can take a discrete set of values are called classification trees. Decision trees where the target variable that can take continuous values are called as regression trees. Classification And Regression Tree [CART] is general term for this.

3.3.5.Random Forest:

Random forest, like the name implies, it consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction.The fundamental concept behind random forest is a simple but powerful one -'the wisdom of crowds'.Random forest classifier is an ensemble method that trains several decision trees in parallel with bootstrapping followed by aggregation, jointly referred as bagging.

3.3.6.AdaBoost Classifier:

AdaBoost can be used to boost the performance of any machine learning algorithm. It is best used with weak learners. These are models that achieve accuracy just above random chance on a classification problem.The most suited and therefore the most common algorithm used with AdaBoost are decision trees with one level. Because these trees are so short and only contain one decision for classification, they are often called decision stumps.Each instance in the training dataset is weighted. The initial weight is set to:

$$weight(xi) = 1/n$$

Where ,

xi is the i'th training instance and n is the number of training instances.

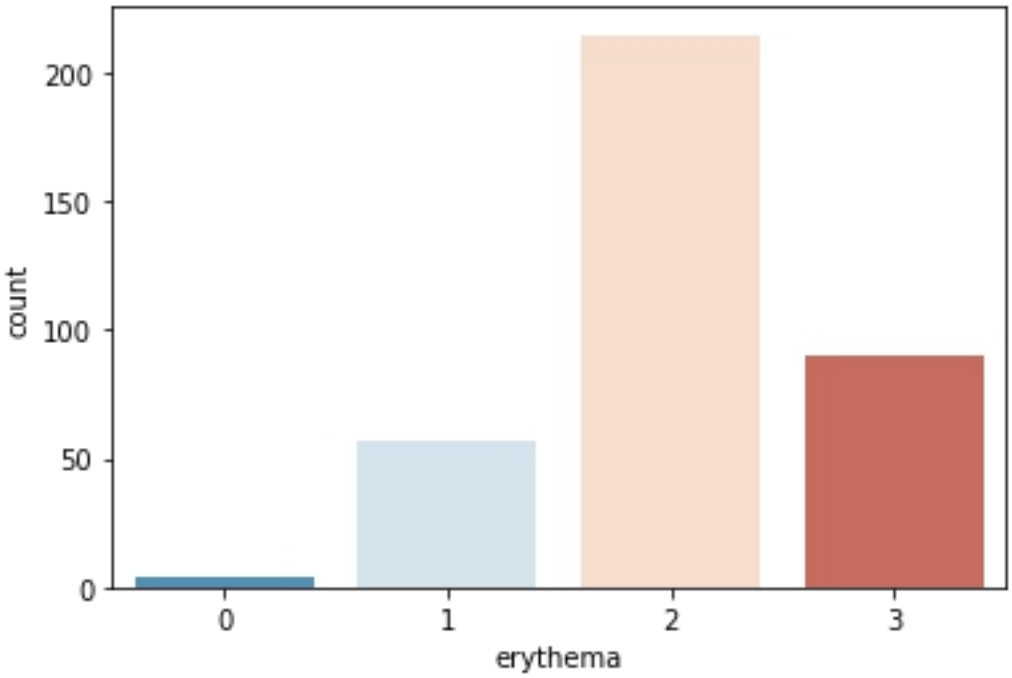
4.Results

For erythemato-squamous disease predictions, we considered 12 clinical features and the 22 histopathological features obtained after the biposy of the patient skin disease from 366 patients samples.We used five different type of techniques and three different types of models to learn and predict the findings. Later, predictions were performed and the performance of the deep learning applications models were evaluated.

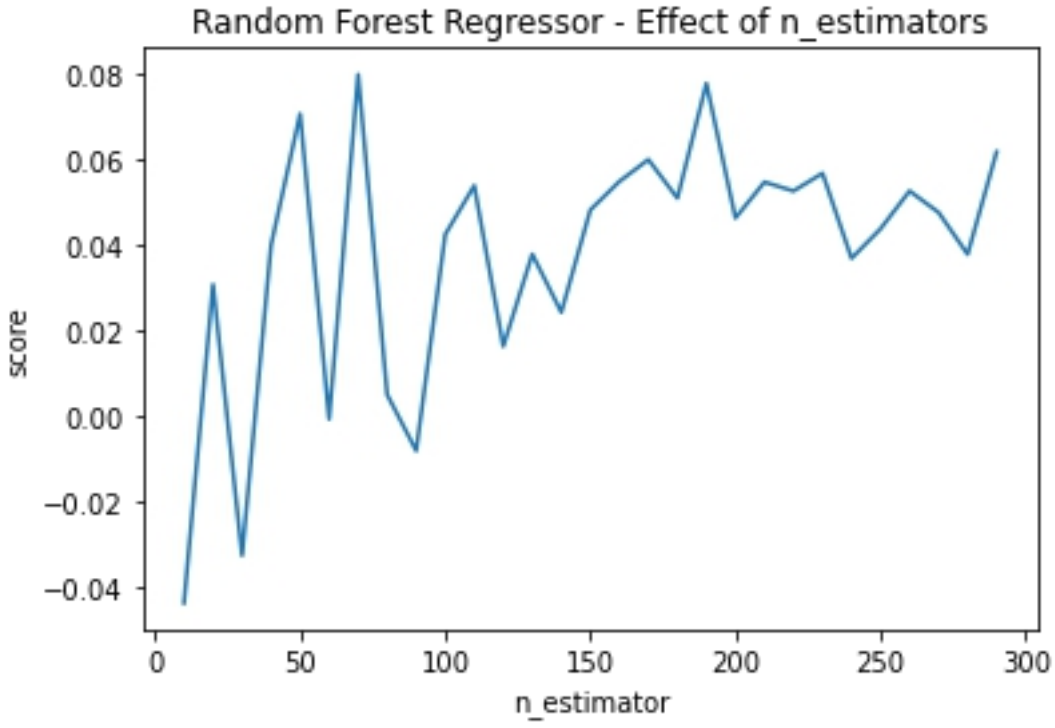
Techniques	**Training Set**	**Test Set**
Logistic Regression	0.67	0.46
Linear Regression	0.28	0.23
Decision Tree	1.00	0.43
Random Forest	0.88	0.09
Support Vector Machine	0.65	0.51
Ada Boost Classifier	0.69	0.65

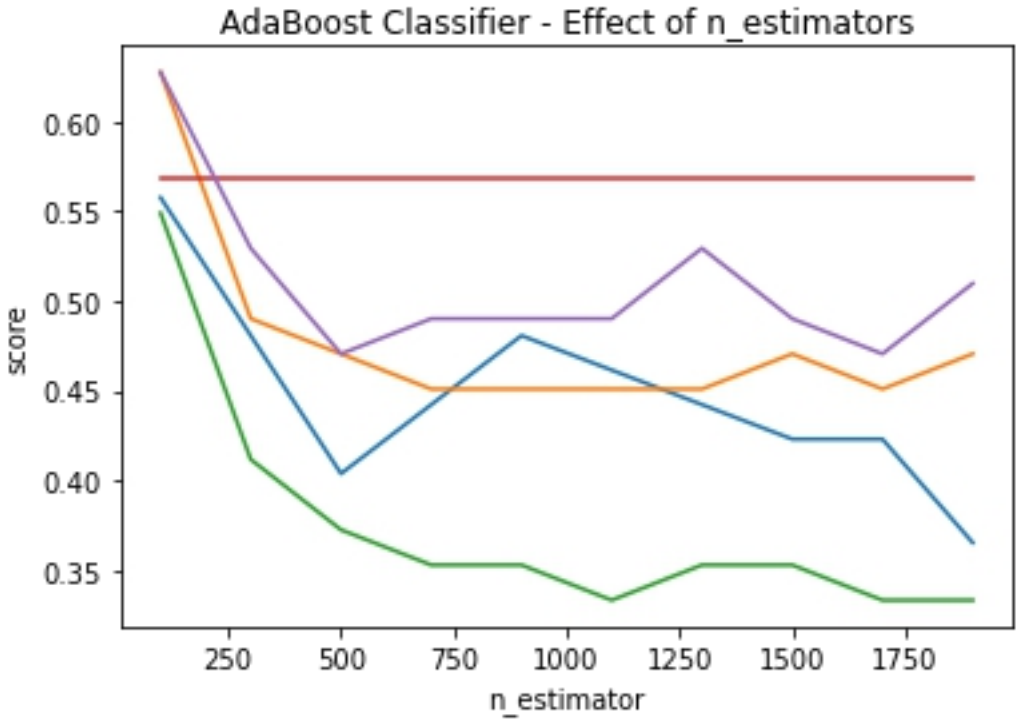
Table 2 :The results of all deep learning application techniques

Hence in the above table we can see the accuracy of the techniques used in this dataset in both training set and the test set.hence we conclude that the most accurate techniques fot the training set is the decision tree,its has the accuracy of 1.00.and for the tesr set is the the support vector machine with the accuracy of 0.51.



Overall patient count of the disease





5. Conclusion:

This paper measure the performance of 6 predictive technique built on Logistic regression,linear regression,SVM,RF,decission tree,ada boost classifier methods.These models are used to predict the erythemato-squamous disease using various parameters provided in the dermatology dataset.366 data samples are collected. In the first stage of the study, the data were standardized and then used as inputs for the deep learning models then classification was carried out and the performances of the technique were measured with precision, recall, accuracy.
In conclusion, we found evidence to suggest that deep learning application technique can be applied to predict erythemato-squamous disease infection with laboratory findings. Our experimental results indicate that may be useful to help prioritize scarce healthcare resources by assigning personalized risk scores using laboratory and blood analysis data. In addition to these, ourfindings on the importance of laboratory measurements towards predicting erythemato-squamous disease infection for patients increase our understanding of the outcomes of erythemato-squamous disease. Based on our study's results, we conclude that health- care systems should explore the use of predictive models that individual erythemato-squamous disease risk in order to improve healthcare are source prioritization and inform patient care.

6.Refrenences:

1. A new multiple classifier system for diagnosis of erythemato-squamous diseases based on rough set feature selection:By **3 authors:**
Behshad Lahijanian
University of Florida
Farzad V. Farahani
University of Central Florida
Mohammad Hossein Fazel Zarandi
Amirkabir University of Technology
2.S. Abdul-rahman, M. Yusoff, A. Mohamed, S. Mutalib, and A. K. Norhan, “Dermatology diagnosis with feature selection methods and artificial neural network,” 2012 IEEE-EMBS Conference on Biomedical Engineering and Sciences, no. December. pp. 371–376, 2012.
3. 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.
4. C. Science and S. Engineering, “Prediction of Different Dermatological Conditions Using Naïve Bayesian Classification,” International Journal of Advanced Research in Computer Science and Software Engineering, vol. 4, no. 1. pp. 864–868, 2014.
5. J. B. R. Gallagher, “Improving Differential Diagnosis of Pathologically Similar Dermatological Conditions to Minimize Invasive Procedures.”
6. F. V. Farahani, M. Fazel Zarandi, and A. Ahmadi, “Fuzzy rule based expert system for diagnosis of lung cancer,” in Fuzzy Information Processing Society (NAFIPS) held jointly with 2015 5th World Conference on Soft Computing (WConSC), 2015 Annual Conference of the North American, 2015, pp. 1–6.
7.I. N. Güvenir HA, Demiröz G, “Learning differential diagnosis of erythemato-squamous diseases using voting feature intervals,” Artif Intell Med, vol. 13, pp. 147–165, 1998.
8. N. Badrinath, G. Gopinath, K. Ravichandran, and R. G. Soundhar, “Estimation of automatic detection of erythemato-squamous diseases through adaboost and its hybrid classifiers,” Artif. Intell. Rev., pp. 1–18, 2013.
9. E. D. Ubeyli and I. Guler, “Automatic detection of erythematosquamous diseases using adaptive neuro-fuzzy inference systems,” Comput. Biol. Med., vol. 35, pp. 421–433, 2005.
10.L. Parthiban, “An Intelligent Agent for Detection of Erythemato Squamous Diseases using Co-Active Neuro-Fuzzy Inference System and Genetic Algorithm,” in Intelligent Agent & Multi-Agent Systems,

2009. IAMA 2009. International Conference on, 2009, pp. 1–6.
11. K. Revett, F. Gorunescu, A.-B. Salem, and E.-S. El-Dahshan, “Evaluation of the Feature Space of an Erythematosquamous Dataset Using Rough Sets,” Ann. Univ. Craiova, Math. Comp. Sci. Ser., vol. 36, no. 2, pp. 123–130, 2009.
12.S. Aruna, L. V. Nandakishore, and S. P. Rajagopalan, “A Hybrid Feature Selection Method based on IGSBFS and Naive Bayes for the Diagnosis of Erythemato - Squamous Diseases,” Int. J. Comput. Appl., vol. 41, no. 7, pp. 13–18, 2012.
13. S. Sarhan, E. Elharir, and M. Zakaria, “A Hybrid Rough-Neuro model For Diagnosing Erythemato-Squamous Diseases,” IJCSI Int. J. Comput. Sci. Issues, vol. 11, no. 1, 2014.
14. L. A. Lee, A. Z. Kouzani, and E. J. Hu, “Automated identification of lung nodules,” in 2008 IEEE 10th Work. Multimed. Signal Process, pp. 497–502.
15.T. G. Dietterich, “Machine-learning research,” AI Mag., vol. 18, no. 4, p. 97, 1997.
16.R. Polikar, “Ensemble Based Systems in Decision Making,” Circuits Syst. Mag. IEEE, vol. 6, no. 3, pp. 21–45, 2006.
17.S. L. a Lee, a Z. Kouzani, and E. J. Hu, “Random forest based lung nodule classification aided by clustering.,” Comput. Med. Imaging Graph., vol. 34, no. 7, pp. 535–42, Oct. 2010.
18. M. F. Akay, “Support vector machines combined with feature selection for breast cancer diagnosis,” Expert Syst. Appl., vol. 36, no. 2, pp. 3240–3247, 2009.
19. Z. Pawlak, “Rough sets,” Int. J. Parallel Program., vol. 11, no. 5, pp. 341–356, 1982.
20. H.-L. Chen, B. Yang, J. Liu, and D.-Y. Liu, “A support vector machine classifier with rough set-based feature selection for breast cancer diagnosisNo Title,” Expert Syst. Appl., vol. 38, no. 7, pp. 9014–9022, 2011.
21. Z. Pawlak, C. Science, A. Informatics, and P. Academy, “Why rough sets?,” 1996, pp. 738 – 743.
22 R. N. Khushaba, A. Al-Ani, and A. T. Al-Jumaily, “Feature subset selection using differential evolution and a statistical repair mechanism,” Expert Syst. Appl., vol. 38, no. 9, pp. 11515–11526, 2011.
23. Z. Pawlak, “Rough set approach to knowledge-based decision support,” Eur. J. Oper. Res., vol. 99, no. 1, pp. 48–57, 1997.
24. D. S. Johnson, “Approximation algorithms for combinatorial problems,” J. Comput. Syst. Sci., vol. 9, no. 3, pp. 256–278, 1974.
25. R. Jensen and Q. Shen, Computational intelligence and feature selection: rough and fuzzy approaches. John Wiley & Sons, 2008.
26.F. V. Farahani, A. Ahmadi, and M. H. F. Zarandi, “Lung Nodule Diagnosis from CT Images Based on Ensemble Learning,” in Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2015 IEEE Conference on, 2015, pp. 1–7.

27. B. E. Boser, I. M. Guyon, and V. N. Vapnik, “A Training Algorithm for Optimal Margin Classifiers,” Proc. 5th Annu. ACM Work. Comput. Learn. Theory, pp. 144–152, 1992.
28.V. Vapnik, The nature of statistical learning theory. New York: Springer, 1995.
29.V. Vapnik, Statistical learning theory. New York: Wiley, 1998.
30. A. Majid, S. Ali, M. Iqbal, and N. Kausar, “Prediction of human breast and colon cancers from imbalanced data using nearest neighbor and support vector machines.,” Comput. Methods Programs Biomed., vol. 113, no. 3, pp. 792–808, Mar. 2014.
31.T. M. Mitchell, Machine Learning. Elsevier, 1983.
32.a. H. El-Baz, “Hybrid intelligent system-based rough set and ensemble classifier for breast cancer diagnosis,” Neural Comput. Appl., vol. 26, no. 2, pp. 437–446, Oct. 2014.
33.K. Hornik, M. Stinchcombe, and H. White, “Multilayer feedforward networks are universal approximators,” Neural networks, 1989.
34. J. Kittler, I. C. Society, M. Hatef, R. P. W. Duin, and J. Matas, “On Combining Classifiers,” Pattern Anal. Mach. Intell. IEEE Trans., vol. 20, no. 3, pp. 226–239, 1998.
35. L. I. Kuncheva, J. C. Bezdek, and R. P. W. Duin, “Decision templates for multiple classifier fusion: an experimental comparison,” Pattern Recognit., vol. 34, no. 2, pp. 299–314, 2001.
36. L. I. Kuncheva, Combining Pattern Classifiers: Methods and Algorithms. John Wiley & Sons, 2004.