JOURNAL OF SCIENTIFIC RESEARCH & TECHNOLOGY (JSRT) VOLUME-1 ISSUE-4 JULY 2023
Registered under MSME Government of India

Comparison of Machine Learning Techniques in the diagnosis of Erythematous Squamous Disease

Shubam Sharma[1], Prof. Vinod Sharma[2]

[1]*M-Tech Student, University of Jammu, Jammu, India, shubamsharma3108@gmail.com*

[2]*Professor, University of Jammu, Jammu, India*

ABSTRACT

Diagnosing a medical condition and its root cause is an involved procedure that calls for much investigation and knsowledge. Before making any health-related choices, it's crucial to have accurate information. Making a doctor's appointment is the first step in getting a diagnosis if you suspect you're unwell. The primary goal of this study is to develop a system for illness classification using machine learning techniques. The first step is to gather and prepare the data for analysis. Uncertainty in the dataset caused by noise, outliers, or missing values will be removed during pre-processing. Separate sets of Training data and Test data will be created from the preprocessed dataset. Machine learning algorithms will be used to train the model, and the resulting model will be put to the test on a separate dataset. The effectiveness of the algorithms being used in production will then be evaluated.

Keywords – Machine learning, Erythematous Squamous, logistic regression.

I. INTRODUCTION

1.1 Erythematous Squamous

Most people with eczema have erythematous squamous, a persistent skin condition characterized by an itchy, scaly rash. Although it may manifest anywhere on the body, the face and chest are common initiation sites. It may take some time for this form of eczema to cure completely, even without therapy. The emotional, social, and quality of life of a person with an erythemato-squamous n illness is adversely impacted, as is the mental health of the person's family and community. Because of the high expense of treating these illnesses using pharmaceuticals imported from other nations, they have a negative impact on the economy and result in a loss of workers.

Figure: 1 Image of erythematous squamous disease

1.2 Medical diagnosis with machine learning

The field of medicine has found great success in using Machine Learning (ML), a subfield of AI, to the task of illness diagnosis. In addition to being effective in diagnosing more prevalent illnesses, Ml methods have shown to be at least as good at spotting more obscure ones. Medical diagnosis and decision-making might undergo a dramatic transformation thanks to machine learning. Diagnosis is the process through which a doctor attempts to identify the underlying medical conditions that are producing a patient's symptoms. Existing machine learning diagnostic methods, on the other hand, are entirely correlational, meaning that they only look for illnesses that show a high degree of association with the patient's symptoms. We demonstrate that incorrect or even harmful diagnoses may emerge from a failure to differentiate between correlation and causation. Consequently, we build counterfactual diagnostic algorithms and reframe diagnosis as a counterfactual inference job. Through the use of a test set of clinical scenarios, we evaluate our counterfactual algorithms against both the gold-standard associative algorithm and 44 medical professionals. In contrast to the associative algorithm, which achieves clinical accuracy in the top 48% of physicians, our counterfactual approach achieves clinical accuracy in the top 25%. For machine learning to be useful in medical diagnosis, our findings indicate that causal reasoning must be included in the gradient. Machine learning is a significant area of computer science concerned with developing predictive models and algorithms that can learn from existing data and information. Spam detection is one example of a machine learning challenge. Genetic factors contribute to the genesis of chronic illnesses, which are characterized by their gradual progression, inability to be cured by current medical treatments, and a myriad of other contributing variables. The field of machine learning examines algorithms that may be programmed to learn and improve on their own. It's generally considered to fall within the umbrella of AI.

Machine learning is often broken down into four distinct subfields, each with its own unique approach to learning:

- Supervised Machine Learning
- Unsupervised Machine Learning
- Semi-Supervised Machine Learning
- Reinforcement Learning

II. LITERATURE REVIEW

Han J et al. (2022) It examines the potential of data mining methods for massive datasets, including their applicability, value, efficiency, and scalability. Data mining is introduced before the authors get into detail on how to go about preprocessing, characterizing, and storing data. Data categorization and model development; cluster analysis; outlier identification; and frequent pattern, association, and correlation mining for huge data sets are then introduced as well as other data mining methodologies. This chapter provides a comprehensive introduction to deep learning, including its core concepts and methodology. Finally, the book discusses the future of data mining, including its potential uses and areas of future study. Includes a brand new chapter on deep learning, covering topics such as how to make better use of convolutional, recurrent, and graph neural networks in training deep learning models. data mining trends and research frontiers are discussed in detail, as are data mining applications (including sentiment analysis, truth discovery, and information propagation), data mining methodology and systems, and data mining and society. Offers a thorough, useful introduction to the ideas and methods behind optimizing your data.

Raghupathi W et al.(2010) Data mining techniques for classification, clustering, and association are briefly summarized, along with their individual benefits and downsides. There are three instances of data mining's use in healthcare, and some suggested rules for using data mining techniques in the areas of classification, clustering, and association are provided. We introduce how data mining technologies (in each area of classification, clustering, and association) have been used for a variety of purposes, including research in the biomedical and healthcare fields, and highlight the successful application of data mining by health-related organizations in predicting health insurance fraud and under-diagnosed patients and identifying and classifying at-risk individuals in terms of health to reduce healthcare costs. Data mining can be used to find connections between health conditions and a disease, between diseases, and between drugs; this article discusses the technology that allows for the prediction of healthcare costs (including length of hospital stay), disease diagnosis and prognosis, and the discovery of hidden biomedical and healthcare patterns from related databases. In the article's last section, the challenges that prevent data mining from being used in therapeutic settings are discussed.

Hemalatha et al. (2011) The healthcare systems provide easy access to a wealth of information. However, there is a dearth of efficient analytic tools that can reveal latent connections and tendencies in data. The fields of business and science have found several uses for knowledge discovery and data mining. Data mining methods, when applied to the healthcare system, may provide very useful information. Children have also been protected from illnesses and infections including polio, DPG, BCG, and measles by preventative measures like immunization and vaccination. In this analysis, we take a look at how data mining may be used in the medical

field. In this research, we take a quick look at the feasibility of using classification-based data mining methods like decision tree and Artificial Neural Network to a large dataset such as immunization records.

Xie J et al. (2010) In order to better identify erythemato-squamous disorders, this article created a diagnostic model based on Support Vector Machines (SVM) using a unique hybrid feature selection technique. To pick the best feature subset from the whole feature set, we developed a hybrid feature selection approach we call IFSFFS (Improved F-score and Sequential Forward Floating Search). The first thing we did in our IFSFFS was to extend the F-score so that it could be used to measure the discriminating between more than two sets of real data. Then, to get to the best feature subset selection, we suggested combining SFFS with our enhanced F-score. For filters, we provide an enhanced F-score, while for wrappers, SFFS and SVM form an assessment framework. Grid search with ten-fold cross-validation is used to get the optimal values for the kernel function of the support vector machine (SVM). Five different random training-test splits of the UCI machine learning database's erythemato-squamous illnesses dataset have been used in the experiments. Our SVM-based model using IFSFFS also obtained the best classification accuracy in experiments using just 14 features.

III. JUSTIFICATION

- Clinical experts may benefit from the use of artificial intelligence (Machine learning) to create intelligent systems that can interpret the diagnosis instantly and accurately.
- The diagnostic profession relies heavily on machine learning methods. Clinical research into every illness begins with a diagnosis. In the area of diagnostics, artificial intelligence (AI) methods (machine learning) may be used to construct an architecture that can be used to extract the relevant facts from the collected data.
- The goal of most forms of automation is to lessen the workload of humans. The diagnosis of systemic lupus erythematous will proceed more quickly and easily now that it is automated. The employment of machines to do mundane tasks like washing vehicles and tidying the house is on the rise throughout the globe. It would be very helpful if machine learning could be used to automatically diagnose erythematous squamous.

IV. OBJECTIVES

- To investigate erythematous squamous disease in depth.
- To investigate the use of machine learning methods in the medical diagnosing process.
- Performance criteria including accuracy, recall, precision, and f-score will be used to evaluate several machine learning approaches to erythematous squamous disease categorization.

V. RESEARCH METHODLOGY

The primary goal of this study is to develop a system for illness classification using machine learning techniques. The first step is to gather and prepare the data for analysis. Uncertainty in the dataset, such as outliers, missing values, or unusual patterns, will be removed by pre-processing. Separate sets of Training data and Test data will be created from the preprocessed dataset. We'll use an ML technique to teach the model some stuff, and then we'll put it to the test on the Test dataset. The effectiveness of the algorithms being used in production will then be evaluated.

ALGORITHM USED FOR PREDICTION AND RESULTS

SUPPORT VECTOR MACHINE

Separating hyperplanes provide the formal definition for a Support Vector Machine (SVM), a discriminative classifier. The method generates a hyperplane that best classifies fresh samples based on the labeled training data provided by the supervisor (supervisor learning). This hyperplane is a line in two space that separates the plane into two equal sections, one for each category. For the purposes of classification and regression analysis, support vector machines (SVM, also known as support vector networks) are supervised learning models that use specific learning techniques. It is a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting) because, given a set of training examples, the SVM training algorithm constructs a model that assigns new examples to one of the two categories. In a support

vector machine (SVM) model, each sample is represented as a point in space, and each category is mapped onto its own independent space. New cases are mapped into the same space, and their classification is predicted according to which side of the divide they lie. By implicitly translating their input into high-dimensional feature spaces, SVMs are able to execute non-linear classification as effectively as linear classification.
In the absence of labels, it is not feasible to use supervised learning techniques; instead, unsupervised learning must be used to discover natural grouping of the data into groups and then map new data to these created clusters. The support vector clustering method, developed by Hava Siegel Ma'am and Vladimir Vapnik, is one of the most popular clustering algorithms used in industry. It use the statistics of support vectors, first introduced in the support vector machine technique, to classify un-labelled data.
For the purposes of classification, regression, and other tasks like outliers identification, a Support vector machine creates a hyperplane or group of hyperplanes in high or infinite dimensional space. The hyperplane with the biggest distance to the closest training data point of each class (so called functional margins) achieves a good separation intuitively, as in general the higher the margin, the smaller the generalization error of the classifier.

RANDOM FOREST CLASSIFIER

The ensemble algorithm that makes up Random Forest Classifier. When it comes to object classification, ensemble algorithms are those that use many algorithms of the same or different types. For instance, we might perform prediction using our Naive Bayes, SVM, and Decision Tree models, and then poll the whole group to decide which class the test item should be placed in. The principle of "divide and conquer" lies at the heart of the performance-enhancing method of "ensemble learning." The theory behind this method of education is that by working together, a group of weak students may become a strong one. This technique, as its name implies, generates a forest comprised of many decision trees. In general, the greater the number of trees, the more reliable and precise the algorithm will be. Similar methods to building decision trees, such as information gain and the Gini index, are used to create this forest. As a kind of classification based on supervised learning, it is among the most effective and robust algorithms available. The number of trees formed is proportional to the accuracy achieved. For the purpose of model representation, random forests produce many trees to represent a single tree.
When deciding how to categorize a new item, each tree "votes" for a predetermined category based on the values of a single attribute. If many trees have the same categorization, the forest will choose the one with the most votes. Random forest has several benefits, including its ability to do both classification and regression tasks, its resistance to model overfitting, and its scalability to huge, high-dimensional data sets. The results of this method may be used in conjunction with any other classification or prediction models to aid in feature selection depending on the relevance of each variable.
As a user-friendly way for adjusting those two variables, random forests are a great choice. The first is the number of trees, denoted by n-tree, with a default value of 500; the second is the number of trees to randomly choose as candidates at each split, denoted by m-tree. Random forests have these key drawbacks: While it excels at classifying data, it struggles with regression because of its inability to provide exact continuous nature predictions, its inability to extrapolate outside the range of the training data, and its potential to over fit datasets that are very noisy.

K-NEAREST NEIGHBOUR:

The k-nearest neighbors algorithm (k-NN) is a regression and classification technique that does not rely on parameters. In both circumstances, we feed in the k training examples that are closest to the target in the feature space as input. Whether k&-NN is used for classification or regression determines the results:
Class membership is the result of k-NN categorization. There are several ways to categorize a given thing based on a majority vote among the object's closest k neighbors (k is a positive integer, usually rather small). If k is equal to 1, then the item is classified according to the category of its closest neighbor. The object's property value is the result of k-NN regression. The average of the values of the k closest neighbors is used to get this value.
Lazy learning, of which k-NN is an example, involves approximating the function locally and leaving all computation for the classification stage. When compared to other machine learning algorithms, the k-NN is quite straightforward.
One helpful strategy that may be used in both classification and regression is to give greater weight to the contributions of closer neighbors than those of farther away. One typical method is assigning a weight of 1/d, where d is the distance to the neighbor, to each neighbor.
For k-NN class 2, the object property value (for k-NN regression), is known, hence these objects serve as

neighbors. Although no training step is actually performed, this may be considered of as the algorithm's training set.
The k-NN method has the distinction of being very dependent on the structure of data.
Uses for the K-Nearest Neighbors Algorithm

- Collection of financial characteristics for credit ratings vs. database comparison of individuals with comparable financial profiles. Credit ratings are designed to provide similar scores to individuals with comparable financial profiles. Since this information already exists, they want to utilize it to anticipate a new customer's credit rating without having to manually calculate it.

- Should the bank provide the person a loan? Would someone be likely to stop making loan payments? Is that individual more similar to others who have loan defaulted or died owing money?

- In the field of political science, categorizing a possible voter as "will vote," "will not vote," "will vote Democratic," "will vote Republican," etc.

- Examples of such sophisticated applications may include optical character recognition (OCR) applied to handwriting, picture recognition, and even video recognition.

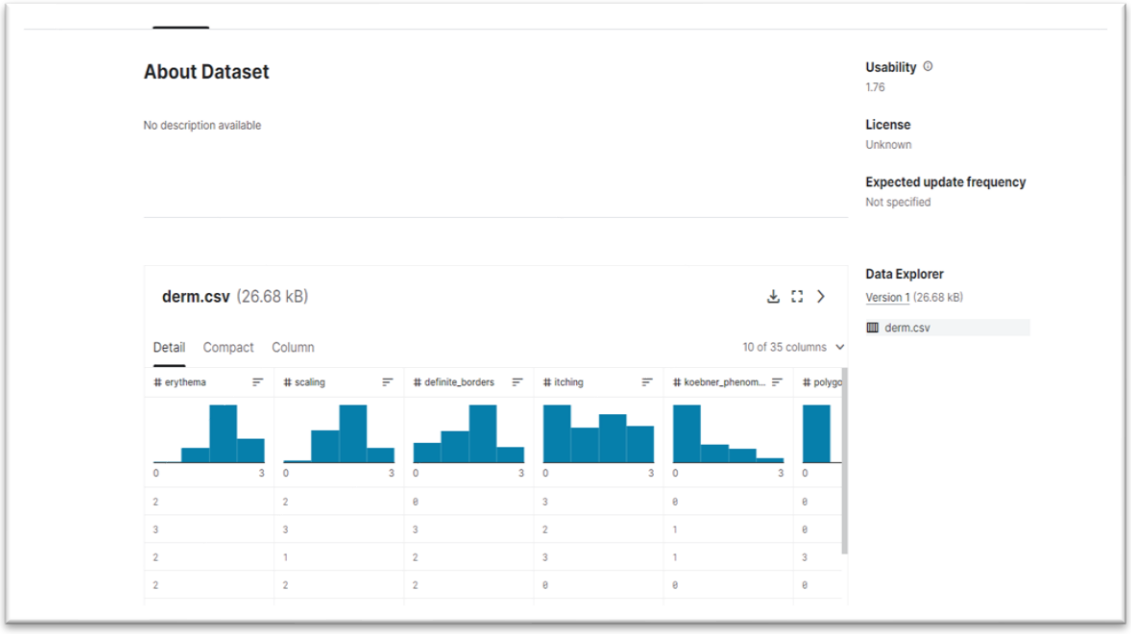
LOGISTIC REGRESSION:

A logistic function is the foundation of the simplest version of the statistical model known as "logistic regression," which is used to represent a binary dependent variable. Logistic regression, often known as logit regression, is a method for estimating the parameters of a logistic model (a kind of binary regression) in the context of regression analysis. In mathematics, a binary logistic model uses an indicator variable whose values may only be 0 or 1, like in the case of a pass/fail dependent variable. Logistic regression models use a linear combination of one or more independent variables ("predictors"), each of which may be either a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value), with the log- odds (the logarithm of the odds) for the value labelled "1" being the linear result.
Log-odds to probability are converted using a function called the logistic function, thus the name. The probability of the value labeled "1" may range from 0 (surely the value "0") to 1 (definitely the value "1"), therefore the labeling. The various names come from the fact that the unit of measurement for the log-odds scale is called a logit, which is short for "logistic unit." The logistic model is distinguished from similar models with different sigmoid functions by the fact that an increase in one independent variable multiplicatively scales the odds of the given outcome at a constant rate, with each dependent variable having its own parameter; this generalizes the odds ratio for a binary independent variable.

Multinomial logistic regression is used to model categorical outputs with more than two values, and ordinal logistic regression is used if the multiple categories are ordered, such as in the proportional odds ordinal logistic model. Although the model itself does not perform statistical classification (it is not a classifier), it can be used to create one by doing things like picking a cutoff value and labeling inputs with a probability greater than the cutoff as one class and inputs with a probability lower than the cutoff as the other. This is a common way to create a binary classifier. In contrast to linear least squares, the coefficients are not often calculated using a closed-form equation. Joseph Berkson is mostly credited with pioneering and popularizing the logistics regression as a broad statistical model.

DATA COLLECTION

Source of Data : using the secondary data from Kaggle



Data Sample as a CSV File:

erythema	scaling	definite_b	itching	koebner_p	polygonal_	follicular_	oral_mucc	knee_and_	scalp_invo	family_his	melanin_ir	eosinophil	pnl_infiltra	fibrosis_of	exocytosis	acanthosis	hyperkerat	parakerat	clubbing_c	elongation	thinning_o	spongiform	mu
2	2	0	3	0	0	0	0	1	0	0	0	0	0	0	3	2	0	0	0	0	0	0	
3	3	3	2	1	0	0	0	1	1	1	0	0	1	0	1	2	0	2	2	2	2	2	
2	1	2	3	1	3	0	3	0	0	0	1	0	0	0	1	2	0	2	0	0	0	0	
2	2	2	0	0	0	0	0	3	2	0	0	0	3	0	0	2	0	3	2	2	2	2	
2	3	2	2	2	2	0	2	0	0	0	1	0	0	0	1	2	0	0	0	0	0	0	
2	3	2	0	0	0	0	0	0	0	0	0	2	1	0	2	2	0	2	0	0	0	1	
2	1	0	2	0	0	0	0	0	0	0	0	0	0	3	1	3	0	0	0	2	0	0	
2	2	3	3	3	3	0	2	0	0	0	2	0	0	0	2	3	0	0	0	0	0	0	
2	2	1	0	2	0	0	0	0	0	0	0	0	0	0	2	1	0	1	0	0	0	0	
2	2	1	0	1	0	0	0	0	0	0	0	0	0	0	3	2	0	2	0	0	0	0	
3	3	2	1	1	0	0	0	2	2	1	0	0	0	0	0	3	2	3	2	2	2	1	
2	2	0	3	0	0	0	0	0	0	0	0	0	2	0	2	2	0	0	0	0	0	1	
3	3	1	2	0	0	0	0	0	1	0	0	0	2	0	3	1	0	1	0	0	0	0	
2	3	3	0	0	0	0	0	1	1	1	0	0	1	0	0	2	1	2	1	2	3	0	
2	2	3	3	0	3	0	2	0	0	0	2	0	0	0	1	1	1	1	0	0	0	0	
1	1	0	1	3	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0	
2	2	1	3	0	0	0	0	0	0	0	0	0	2	0	2	1	0	1	0	0	0	0	
3	3	3	0	0	0	0	0	3	3	1	0	0	2	0	0	2	0	2	3	3	3	2	
2	1	3	3	3	3	0	0	2	0	0	3	0	0	0	3	2	0	1	0	0	0	0	
1	1	0	3	0	0	0	0	0	0	0	0	0	0	3	0	3	2	2	0	3	0	0	
2	1	1	2	0	0	3	0	1	2	0	0	0	1	0	0	1	2	2	0	1	0	1	
3	2	2	0	0	0	0	0	0	0	0	0	0	2	0	2	2	1	2	0	2	1	2	
2	2	0	2	0	0	0	0	0	0	0	0	0	0	1	1	3	1	2	0	2	1	0	
2	2	2	3	2	2	0	2	0	0	0	3	2	0	0	0	2	1	1	0	0	0	0	
2	0	0	3	0	0	0	0	0	0	0	0	0	0	2	2	2	0	0	0	3	0	0	
2	1	1	0	1	0	0	0	2	0	0	0	0	0	0	0	2	2	2	2	2	2	1	
1	1	0	1	0	0	3	0	1	0	0	0	0	1	0	0	1	1	1	0	0	0	0	
1	2	2	3	0	0	0	0	0	0	0	0	0	0	1	1	2	1	1	0	3	0	0	

Performance Analysis of SVM:

```
In [21]: Y_predict

Out[21]: array([4, 1, 3, 1, 2, 1, 4, 2, 2, 1, 1, 4, 2, 4, 3, 1, 3, 1, 4, 4, 1, 3,
                3, 1, 2, 2, 1, 6, 1, 3, 5, 5, 1, 2, 5, 4, 1, 5, 5, 1, 5, 6, 4, 2,
                4, 4, 1, 6, 3, 1, 1, 1, 5, 1, 2, 1, 4, 5, 3, 3, 5, 6, 1, 1, 4, 3,
                2, 3, 2, 2, 2, 5, 2, 1, 2, 3, 4, 1, 1, 2, 3, 4, 4, 5, 2, 3, 3, 3,
                3, 3, 3, 2], dtype=int64)
```

```
In [26]: print('Accuracy Score with linear kernel')
         print(metrics.accuracy_score(Y_test,Y_predict))

         Accuracy Score with linear kernel
         0.9456521739130435
```

RandomForest Classifier

Performance Analysis of KNN:

```
In [275]: `y_pred = regressor.predict(X_test_scaled)
```

```
In [276]: from sklearn.metrics import r2_score
          r2 = r2_score(y_test, y_pred)
          print("R-squared score:", r2)

          R-squared score: 0.9277787861758922
```

JOURNAL OF SCIENTIFIC RESEARCH & TECHNOLOGY (JSRT) VOLUME-1 ISSUE-4 JULY 2023
Registered under MSME Government of India

LOGISTIC REGRESSION

VI. RESULT AND DISCUSSION

Experimental Setup:

```
2]: from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
    classifier.fit(X_train, Y_train)
    Y_pred = classifier.predict(X_test)
    accuracy = accuracy_score(Y_test, Y_pred)
    print("Accuracy:", accuracy)
    print("Classification Report:")
    print(classification_report(Y_test, Y_pred))
    print("Confusion Matrix:")
    print(confusion_matrix(Y_test, Y_pred))
```

```
Accuracy: 0.9130434782608695
Classification Report:
              precision    recall  f1-score   support

           1       0.91      1.00      0.95        21
           2       0.82      0.88      0.85        16
           3       0.95      0.95      0.95        19
           4       0.88      0.94      0.91        16
           5       1.00      0.92      0.96        12
           6       1.00      0.62      0.77         8

    accuracy                           0.91        92
   macro avg       0.93      0.88      0.90        92
weighted avg       0.92      0.91      0.91        92

Confusion Matrix:
[[21  0  0  0  0  0]
 [ 1 14  0  1  0  0]
 [ 0  0 18  1  0  0]
 [ 0  1  0 15  0  0]
 [ 1  0  0  0 11  0]
 [ 0  2  1  0  0  5]]
```

The Intel Core i3 2.0 GHz CPU with 4 GB of RAM was used for all the tests.
Microsoft Windows 10 is utilized with the Python 3.7.0 release.

Experimental Result:

Python 3.7.2 was used to implement the suggested methodology's individual steps, and the code was tested on a PC running Windows 10 with an Intel i3 CPU. Thirty-three of the database's properties have linear values, while one is nominal. There were 366 cases with a total of 8 blanks.

We eliminated the cases whose values were missing since their number was so little in comparison to the overall number of cases. Seventy-five percent of the remaining 358 instances were utilized for teaching, while the remaining 25 percent were used for testing.

The accuracy of the used classifiers and the corresponding confusion matrices are shown below.

The success rate of a classification system is measured by how many cases out of all those evaluated were properly labeled.

Classification Accuracy No. of correctly classified instances/ Total no. of instances

Accuracy Score of Classifiers:

Classifier logistic regression received the lowest accuracy score (91%), while classifier KNN achieved the best accuracy score (96%). The same 94% accuracy may be expected by SVM.

Additionally, 92% accuracy was shown using Random forest.

CLASSIFIER	ACCURACY SCORE
SVM	94%
RANDOM FOREST	92%
KNN	96%
LOGISTIC REGRESSION	91%

VII. FUTURE SCOPE:

- Since Erythemato-Squamous illness differential diagnosis is a real-world issue, our model may be used to real-world data, potentially assisting physicians in making accurate prognoses when only clinical parameters are available.
- The model's efficacy may be further improved by using Deep Learning methods.
- Logistic regression classifier for erythemato-squamous diseases: an enhancement.

REFERENCES

1. Güvenir HA, Demiröz G, Ilter N. Learning differential diagnosis of erythemato-squamous diseases using voting feature intervals. Artificial Intelligence in Medicine.1998-13-03.
2. Aruna S, Nandakishore L, Rajagopalan S. A Hybrid Feature Selection Method based onIGSBFS and Naïve Bayes for the Diagnosis of Erythemato-Squamous Diseases. International Journal of Computer Applications (0975–8887)2012.
3. Nanni L. An ensemble of classifiers for the diagnosis of erythemato-squamous diseases. Neuro computing. 2006;69(7):842–5.
4. Han J, Kamber M. Data mining : concepts and techniques. 3rd ed. Burlington, MA:Elsevier;2011.
5. Raghupathi W. Data mining in health care. Healthcare Informatics: Improving Efficiency and Productivity. 2010:211–23.
6. Hemalatha M, Megala S. Mining Techniques in Health Care: A Survey of Immunization. Journal of Theoretical and Applied Information Technology (JATIT)2011;25(2).