Predictive Tool for Dermatology Disease Diagnosis using Machine Learning Techniques

M. Sudha, B. Poorva

***Abstract:** Prediction of skin diseases is more complex as many diseases have the same symptoms at the early stage but may vary at the later stages while the disease becomes incurable. So we can use data mining algorithms to classify the diseases based on the input symptoms. In this paper, the best algorithm suitable for classification of data into six dermatological diseases is determined by comparison with few other algorithms. Naive Bayes tends to show higher accuracy of 99.31% , Random forest exhibits 97.80% and SVM reveals 94.35% when test size is 40% in jupyter notebook. Linear regression and K Nearest Neighbors when trained with 80% of the data displays 82.14% and 94.44% accuracy respectively. Naive Bayes can be used for the prediction of several other diseases and is best for classification of data and thus helps doctors predict the disease more accurately and with comparatively lesser time.*

***Index Terms:** KNearest Neighbors, ,Linear Regression, Naive Bayes, Predictive Model, Random forest, Support Vector machine,*

I. INTRODUCTION

Skin is the largest organ of human body which looks very simple but is actually complicated. Skin diseases may be benign but without any proper medication turn out to be fatal. Many diseases have early symptoms but most of the diseases have similar symptoms that make it complicated to detect the disease at an early stage. The accurate prediction of different erythemato-squamous diseases is an existing problem in dermatology. The diseases of this sort are psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, cronic dermatitis, and pityriasis rubra pilaris. Usually the diagnosis of diseases mentioned above involves biopsy and other examinations that take a lot of time but unfortunately these diseases share many histopathological features as well. Another hitch of differential diagnosis is that a disease may show the features of another disease at the beginning stage and may have the characteristic features at the following stages. So in this paper, all the features are taken into account, helping the doctors to identify and distinguish the six different diseases mentioned above and give them confidence to provide better treatment.

A Psoriasis: Psoriasis is a genetic and long lasting autoimmune disease. It is triggered by environmental factors and form skin patches which are red, dry, itchy and scaly. The five main types of this disease are plague, guttate, inverse, pustular and erythrodemic. Infections and psychological stress might be the cause for the disease. It is not contagious and diagnosis is based on symptoms that often worsen during winter. Men and women are affected with equal frequency. Various treatments help in controlling the symptoms as there is no cure for this disease. 2 to 4 percent of the population is affected.

B Seborrhoea Dermatitis: A genetic disease called Seborrhoea dermatitis is a chronic skin disorder in which the skin becomes red, scaly, greasy, itchy and inflammatory. The oil-producing areas like scalp, face and chest are mostly affected. The environmental factors are the major cause for this disease. The symptom based diagnosis is made. People around the age of 50 are most commonly affected in which males are more than females. It may worsen during the stress and winter. In adults about 2% of people and up to 40% of babies are affected.

C. Lichen Planus : Lichen planus is a persistent inflammatory and immune transmitted disease which affects skin, nails, hair and mucous membranes. The cause for this disease is thought to be result of autoimmune process with an unknown initial trigger. There is no remedy for this disease yet. Different medications and procedures are used to control the symptoms.

D Pityriasis Rosea : Pityriasis rosea is a kind of skin rash that initially develops with single red and slightly scaly area. It is not contagious and most commonly occurs in those between the age of 10 and 35. The major cause for this disease is human herpesvirus 6 (HHV6) or human herpesvirus 7 (HHV7) and the diagnosis is done based on the symptoms. About 1.3% of people are affected at some point in time.

E Chronic Dermatitis : Chronic dermatitis is a disease that results in inflammation of the skin which involves the coalition of irritation, allergy and poor venous return. There are four types of this disease which are apotic dermatitis, allergic contact dermatitis, irritant contact dermatitis and stasis dermatitis. The type of the disease is generally resolved by the history of person and the location of the rash. The exact cause for this disease is unclear and if there are any signs of skin infection, antibiotics are required. About 7% of people are affected around the world.

F Pityriasis Rubra : Pilaris Pityriasis rubra pilaris cites a group of long-term disorders that triggers reddish orange, scaling plaques and keratotic follicular papules. The symptoms include reddish orange patches, twinge flaking, uncomfortable itching, stiffening of the skin on the feet and hands and thickened bumps around hair follicles. Currently, there is no known cause and cure for this disease.

Revised Manuscript Received on July 05, 2019

M. Sudha, Associate Professor, School of Information Technology & Engineering, VIT Vellore, India.
B. Poorva, PG Scholar, School of Information Technology & Engineering, VIT Vellore, India.

DOI: 10.35940/ijitee.G5376.078919

*Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication*

II. BACKGROUND

The data collected from southern part of Kerala, India depicts the classification of diseases based on Naïve Bayes and J48 which is the implementation of basic decision tree ID3(Iterative Dichotomizer3) algorithms after preprocessing by WEKA tool. Several attributes are considered for classifying eight types of diseases. The results show that all the eight diseases depend on all the attributes and that the correct rate of classification is 99.13% by decision tree and naive bayes shows 87.4%. Naive Bayes classification requires only 0.01s and J48 requires 0.03s to obtain the results. On comparing the performance of various algorithms, the result showed that Naive bayes generated less precision and true Positive rate as compared with J48 algorithm. J48 is more coherent in all aspects like TP-rate, FP-rate, Precision, Recall and ROC area.[1]

Naïve Bayes classification has an indispensable role in medical data mining [15]. It has manifested great performance in terms of accuracy thus if attributes are independent of each other we can deploy it in medical fields. The analysis performed within this research are based on data surveyed from various tertiary health care centres in Kottayam and Alappuzha districts of Kerala. In the statistics collected, there may be chances of missing values and the Naïve Bayes handles it at random by replacing sparse numerical data with zeros and categorical data with zero vectors. Naive Bayes depends on counting methodologies to calculate probabilities.[13] Columns should be binned to minimize the cardinality as appropriate. It is mentioned that the numerical data is to be binned into ranges of values like low, medium and high while categorical data is binned into meta classes which significantly reduce the discriminating power of the algorithms. It is also considered to have a minimum error rate compared to other classifiers.[2]

This paper is based on using an approach of mining and image processing of the data to predict skin diseases using certain features identified in the digital image of the skin. The collected data is difficult to classify and is done using the data mining principles, the decision tree which will classify the skin disease images and to determine a suitable medicine for it. The system can distinguish between the normal and infected skin using data mining principles to recognize the skin diseases based on the skin diseases images available in the database[14]. In this paper of filters, operators, adapters and then transformation and extraction is done. The system mainly focuses on collecting on answers from the user for certain questions in the system and also the images and combine both the answers and images after processing to predict and display the skin disease so that the doctors can provide the treatment.[3]

The effect of combining two models such as Support vector machine and Artificial neural networks by applying a confidential weighted voting scheme to categorize the six different sets of Erythemato-Squamous diseases shows the highest accuracy of 99.25% and 98.99% at training and testing stages respectively.[4] Thus by developing a decision support tree will diagnose the problem uniformly and intelligently can help even doctors will relatively less or mere experience to diagnose and treat these six diseases effectively. This paper summarizes the prediction of dermatology diseases by the utilization of data mining approaches for classifying the data into different categories. Classification involves the necessity to extract rules that can partition the data into different categories. Association mining is suitable for extracting rules and consists of two predominant steps, namely frequent patterns discovery and rule construction. Feature extraction and feature selection are two different approaches of dimensionality reduction which are used in the preprocessing and variable selection step. Classification approach of data mining is better than association and clustering. Gini -based decision tree shows better performance than ID3 and C4.5 . Artificial neural networks is better for rule extraction. Apriori algorithm is the more suitable technique for association mining compared with SEMT, Predictive Apriori, Tertius algorithms.[5]

Data mining depicts a very crucial role in the prediction of psoriasis disease by taking into consideration the major clinical and histopathological attributes so that it’s stages are identified and prevented from spreading to other parts. This prediction is used to analyse the relationship by using the regression equation. This work has initiated the relationship between input and response attributes for improving disease prediction in diagnostic domains. The Response Surface Methodology (RSM) is used in this paper to predict the disease. The evaluation of RSM model depicts the developed empirical relationship and it has the greatest conformity with test results. It provides the relationship among input parameters are considered as control factors and output class such as psoriasis status of the patient. The Analysis of Variance(ANOVA) is performed for understanding the mathematical analysis of the outcome. The developed empirical model best fits disease prediction requirements[6]. This model has presented the other dimensional way of predicting the psoriasis disease and RSM proves to be an effective mathematical model that is helpful for medical practitioners and researches. Implementing integrated business intelligence implication for healthcare provides rapid and accurate information on time, based on enterprise resource planning by applying data mining techniques. Database is created that contains details about all the diseases and its treatment to be followed for recovery. The specific descriptions of several dermatology disease, its symptoms, treatment procedures to be followed, precautions taken, and patient records will be retained in ERP centralized database after processing.The intelligence system enables the user to retrieve the resolution based on all the details of the concerned patient. In this phase, knowledge discovery techniques are executed on the patient records maintained in the hospital centralized ERP database. Classification mining should be primarily induced on data to transform it into different categories of dermatology cases. Then the relationship between symptoms and disease is extracted from the database. The dermatologist can then easily predict the disease and provide appropriate treatment that is already illustrated in the database.[7]Classification and Regression Tree (CART) and Cross Industry Standard Process for Data Mining (CRISP-DM) methodology is used to predict differential diagnosis of ErythmatoSquamous Diseases(ESD) using classification and regression techniques. Total accuracy of the model was 93.69% (Standard Deviation 24.42). Hence, the CART algorithm can predict differential diagnosis of ESD with 93.69% accuracy. Although, the CART classifier achieved a better accuracy than some methods such as ANN and C4.5, but it is strongly recommended that ensemble methods should be used to classify differential diagnosis of ESD.[8]

Decision Support Systems(DSS) are used in the healthcare diagnosis systems to improve the efficiency of analysis time buy implying classification and association data mining techniques. Apriori is a seminal algorithm that deploys candidate generation for determining the frequent patterns and finally different pattern finding algorithms such as decision tree, classification and clustering are followed.12 classifier algorithms and 3 decision tree algorithms were tested. Apriori algorithm was chosen for extracting association rules .Among the Bayesian classifiers, with a minimum difference in accuracy of 0.1 BayesNet proved to be the best . BayesNet is differentiated from FT decision tree by an accuracy of 0.5 points. Data Mining, being an experimental approach, resulted in describing that the best algorithm to be used cannot be extended to other databases without some experimentation over them. This is implemented because of the distinctive instances and attributes of datasets, that can lead other algorithms to perform better than the best ones related in this research. [9]
Support vector machines(SVM) with an advanced hybrid feature selection pattern to diagnose erythemato-squamous diseases. The optimal features are determined from the original features by combining the advantage of filter and wrapper methods of the recommended hybrid feature selection method, IFSFS(Improved F-score and sequential forward search) where, the improved F-score
is an evaluation metric of the filter method, and SFS of the wrapper method. The grid search technique helps to identify the best parameters of kernel functions of the support vector machine. The experimental results show that the proposed SVM-based model containing 21 features with IFSFS exhibits 98.61% classification accuracy.[10]

The primary aim in this work is building a dermatology classification model by using the clinical and histopathological features which play a key role in dermatology classification problem. A well-known feature selection method, maximum relevance minimum redundancy(m RMR), is used to select relevant features. Then, support vector machine(SVM), a popular classification methodology is applied on the model. The after effects demonstrate that the accuracy of classification model is improved obviously through feature selection. Some optimal features that are selected are regarded as the most significant features to classify the skin diseases for patients at different stages of age. [11]

The best method for achieving maximum accuracy is established based on the analysis and results of the survey on existing techniques of classification and regression. K- NN, decision tree, Neural network, Support Vector machine ,Bayesian belief network are the various classification techniques models discussed in this model that could be used in health care systems. An analysis report depicted the weaknesses of various classification and regression models in detail, which after establishing data mining models can predict the diseases with greater accuracy. [12] , [16] and [17] machine learning models are widely adopted in various disease diagnosis process. [18] and [19]data mining approach has been applied in predicting diabetes. [20] recently computational diagnostic models are widely adopted in medical disease diagnosis and in preventive analysis.

III. MATERIALS AND METHODS

A .Dataset and Attribute Information:

The data is collected from the UCI repository. This database contains 366 instances, 34 attributes. The family history feature has the value 1 if any of these diseases has been observed in the family, and 0 otherwise. The age feature is a linear value representing age of the patient. Every other feature (clinical and histopathological) takes a value in the range of 0 to 3. Here, 0 indicates that the feature was not present, 3 indicates the largest amount possible, and 1, 2 indicate the relative intermediate values. The names and id numbers of the patients have been removed from the database.

B .Data Set Characteristics:

Multivariate Number of Instances: 366

Area: Life

Attribute Characteristics: Categorical, Integer
Number of Attributes: 33

Date Donated: 1998-01-01
Associated Tasks: Classification

Attribute number	Clinical Attribute	Value
1	Erythema	Nominal(takes values 0,1,2,3)
2	Scaling	Nominal(takes values 0,1,2,3)
3	Definite borders	Nominal(takes values 0,1,2,3)
4	Itching	Nominal(takes values 0,1,2,3)
5	Koebner Phenomenon	Nominal(takes values 0,1,2,3)
6	Polygonal papules	Nominal(takes values 0,1,2,3)
7	Follicular papules	Nominal(takes values 0,1,2,3)
8	Oral mucosal involvement	Nominal(takes values 0,1,2,3)
9	Knee and elbow involvement	Nominal(takes values 0,1,2,3)
10	Scalp involvement	Nominal(takes values 0,1,2,3)
11	Family history	0 or 1
34	Age	Linear
Attribute number	Histopathological attributes	Value
12	Melanin incontinence	Nominal(takes values 0,1,2,3)
13	Eosinophils in the infiltrate	Nominal(takes values 0,1,2,3)
14	PNL infiltrate	Nominal(takes values 0,1,2,3)
15	Fibrosis of the papillary dermis	Nominal(takes values 0,1,2,3)

16	Exocytosis	Nominal(takes values 0,1,2,3)
17	Acanthosis	Nominal(takes values 0,1,2,3)
18	Hyperkeratosis	Nominal(takes values 0,1,2,3)
19	Parakeratosis	Nominal(takes values 0,1,2,3)
20	Clubbing of the rite ridges	Nominal(takes values 0,1,2,3)
21	Elongation of the rite ridges	Nominal(takes values 0,1,2,3)
22	Thinning of the suprapillary epidermis	Nominal(takes values 0,1,2,3)
23	Spongiform pustule	Nominal(takes values 0,1,2,3)
24	Munro microabcess	Nominal(takes values 0,1,2,3)
25	Focal hypergranulosis	Nominal(takes values 0,1,2,3)
26	Disappearance of the granular layer	Nominal(takes values 0,1,2,3)
27	Vacuolisation and damage of basal layer	Nominal(takes values 0,1,2,3)
28	Spongiosis	Nominal(takes values 0,1,2,3)
29	Saw-tooth appearance of retes	Nominal(takes values 0,1,2,3)
30	Follicular horn plug	Nominal(takes values 0,1,2,3)
31	Perifollicular parakeratosis	Nominal(takes values 0,1,2,3)
32	Inflammatory mononuclear infiltrate	Nominal(takes values 0,1,2,3)
33	Band-like infiltrate	Nominal(takes values 0,1,2,3)

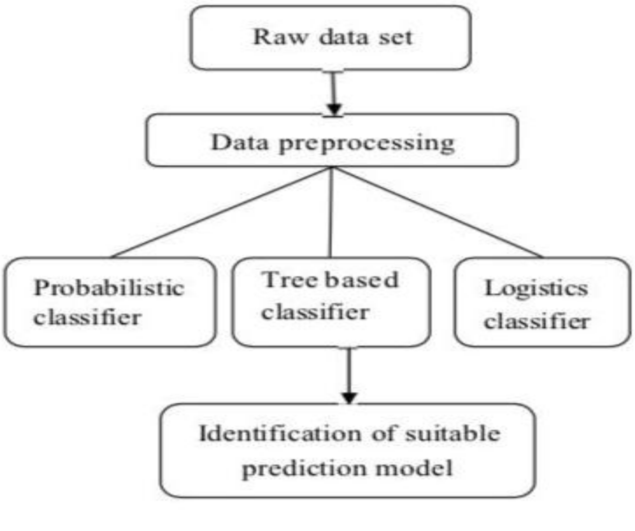


Figure 1. Predictive Modelling for Dermatology Disease Diagnosis

A. Naive Bayes Algorithm :

This classification technique is based on Bayes' theorem which requires non dependence among the parameters. Weka tool is used for pre-processing and then Naive applied on the simple discrete numerical dataset for classification. It performs the probabilistic prediction of the data. It is based on the Bayes theorem

P(H/X)=[P(X/H). P(H)] / P(X) where X is the sample set, H is the hypothesis that X belongs to class C,P(X/H) is the posterior probability of H. Naive Bayes prediction requires each conditional probability be non -zero because it is severely affected b zero probability error and is overcome by using the Laplacian correction. It is the best approach for classification of categorical data.

B. Random Forest:

Random forest is the extension of Decision tree algorithm. Constructing a decision tree classifier does not require any prior sector understanding, hence it is applicable for exploratory knowledge discovery. It can handle multi dimensional data. Decision trees also have their own complexities like overfitting and extremely low accuracy in certain cases. So to overcome this limitation of decision tree, random forest is used. It involves building several decision trees and then following the bagging technique which involves bootstrapping for resampling the data to find the output of each tree and aggregating to add up the output of all trees.

Step1: Randomly select any number of dependent attributes from the total attributes.

Step 2: Calculate the node using best split point, based on the attributes.

Step3: Repeat the same till all the attributes are split into nodes.

Step 4: Build several such trees and then aggregate it.

C. Support Vector machine:

Support Vector machine technique is used to classify the data based on prior knowledge and retrieval. The demographic variables are chosen and then the dependent variables are mapped to the respective demographic variables. Then many number of decision boundaries are obtained, out of which one optimal boundary is selected for classification. The higher the boundary separates the variables, better will be the classification. Support vector machine is to find the maximum margin hyper plane that will distinguish the data into different classes. SVM has its own limitations like, when the variables are more dependent on each other, they are not linearly separable, so to overcome it we have to change the geometry of the plane.

D. Linear Regression:

Linear regression is the prediction of scores of one variable with respect to the other variables, thus dependent variables are easily classified. There can be only one independent variable or the class label while several dependent variables can be considered. The higher the prediction of one variable, higher will be the prediction of other variables. The best-fitting line is determined by this methodology to make the decision which is called the regression line. The distance of the variables from the regression line is the error prediction. Smaller the distance, error of prediction is smaller.

E. K NearestNeighbors:

KNN deals with clustering of the data into various clusters with more similarities. It is a supervised learning algorithm and is non-parametric, meaning it does not make any assumptions about the data.

Step 1: A centroid is selected by using various methods for different type of dataset.

Step 2: Then that centroid becomes the random seed based on which, the points in the neighbourhood are clustered.

Step 3: K is the count of clusters to be created. It is defined before initiation of clustering.

Step 4: The algorithm is iterated till every value is clustered without missing any value and stops when there is no change in the clustering of data.

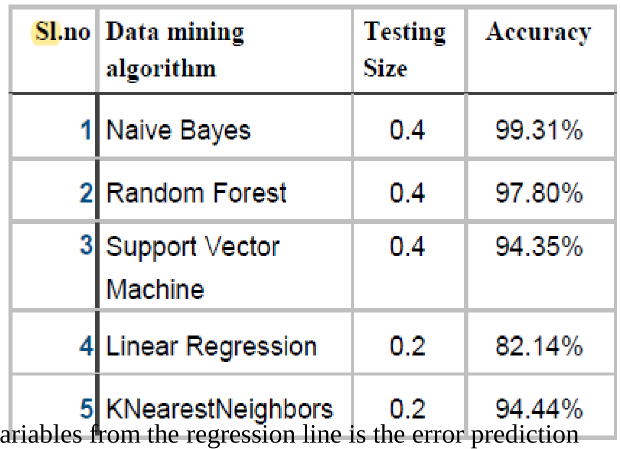
F. Performance Evaluation Metric

Accuracy Rate (AcR) = (Tp+ Tn) / (Tp+ Tn+ Fp+ Fn) where, Tp - True positive , Tn - True negative , Fp - False positive and Fn - False negative.

The accuracy (AcR) is the percent of the sum of predictions that were true. It is estimated using the above equation . The false positive rate (Fp) is the percent of negative cases that were imprecisely classified as positive. The true negative rate (Tn) is defined as the amount of negative cases that are classified correctly as negative. The false negative rate (Fn) is the percent of positive cases that were misclassified as negative.

IV. RESULTS

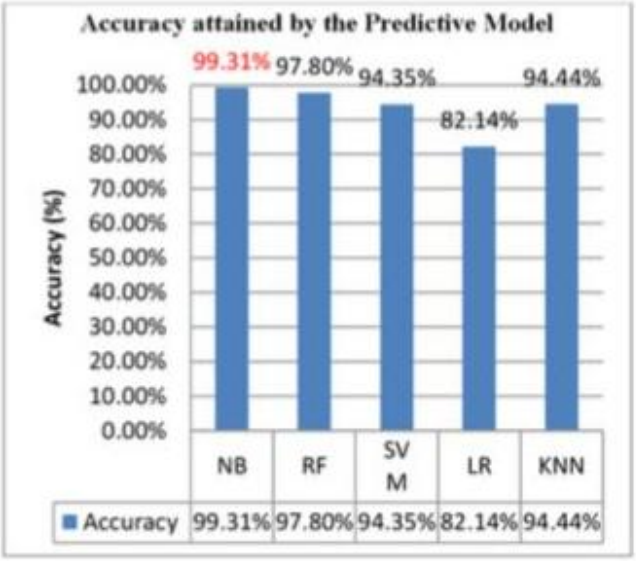
The results of this study revealed that different data mining classification techniques show different accuracies with various training and testing sizes. The test sizes and accuracy of the respective algorithms is shown in Table 1.



Sl.no	Data mining algorithm	Testing Size	Accuracy
1	Naive Bayes	0.4	99.31%
2	Random Forest	0.4	97.80%
3	Support Vector Machine	0.4	94.35%
4	Linear Regression	0.2	82.14%
5	KNearestNeighbors	0.2	94.44%

variables from the regression line is the error prediction

The jupyter Notebook environment is used to execute the interactive python codes for the classification of data. Weka tool is used for preprocessing, missing values were removed, and checked for any outliers. Certain algorithms were executed in the Weka tool itself. Random forest and Linear regression algorithms executed in the Weka tool showed accuracies 97.12% and 93.7% respectively while the results of these algorithms are better in ipython codes. Accuracy of the various data mining classification algorithms is determined for different test sizes, to classify the dermatology diseases into six different diseases which have more similar characteristics and symptoms thus leading to misperception of the disease by the doctors. These data mining algorithms train the data and then test them to predict the accuracy of the system. The more accurate the algorithm is, it would be effective to implement it in medical fields to predict the disease and thus obtain better results.



V.CONCLUSIONS

The Naive Bayes algorithm proves to be more accurate than the other algorithms in classification of a dermatology dataset into six different diseases with 99.31% accuracy. This proves that despite the zero probability limitation of Naive Bayes, it is the best for classification of data .It is uncomplicated to execute and righteous outcome is obtained in most cases. Thus it can be used for classification of data in medical fields for prediction of several other diseases also. Followed by Naive Bayes , accuracy. Linear Regression can be used to detect whether the disease is present or not but for classifying it into different diseases, it shows less accuracy .From the experimental results, probabilistic classifier is an appropriate predictive tool for dermatological issues.

VI.REFERENCES

1. Manjusha KK, Sankaranarayanan K, Seena P. Data Mining in Dermatological Diagnosis: A Method for Severity Prediction. International Journal of Computer
2. Applications. 2015 Jan 1;117(11).
3. Manjusha KK, Sankaranarayanan K, Seena P. Prediction of different dermatological conditions using naive bayesian classification. International Journal of Advanced Research in Computer Science and Software Engineering. 2014 Jan;4(1).
4. Kadhim QK. Classification of human skin diseases using data mining. International Journal of Advanced Engineering Research and Science. 2017;4(1).
5. Sharma DK, Hota HS. Data mining techniques for prediction of different categories of dermatology diseases. Journal of Management Information and Decision Sciences. 2013 Jul 1;16(2):103.
6. Barati E, Saraee MH, Mohammadi A, Adibi N, Ahmadzadeh MR. A survey on utilization of data mining approaches for dermatological (skin) diseases prediction. Journal of Selected Areas in Health Informatics (JSHI). 2011;2(3):1-1.
7. Sudha J, Aramudhan M, Kannan S. Development of a mathematical model for skin disease prediction using response surface methodology. BIOMEDICAL RESEARCH-INDIA. 2017 Jan 1;28:S355-9.
8. Al-Ghamdi FA, Al-Ghamdi AS. Integrated Business Intelligent System for E-Health: A Case for Dermatology Diseases. Journal of Cosmetics, Dermatological Sciences and Applications. 2014 Jan 23;4(01):53.
9. Maghooli K, Langarizadeh M, Shahmoradi L, Habibi- koolaee M, Jebraeily M, Bouraghi H. Differential diagnosis of Erythmato-Squamous Diseases using classification and regression tree. Acta Informatica Medica. 2016 Oct;24(5):338.

10. Chimieski BF, Fagundes RD. Association and classification data mining algorithms comparison over medical datasets. Journal of health informatics. 2013 Jun 30;5(2).
11. 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.
12. Feng-feng SH. Building Diagnostic Model for Dermatology Disease by Using Data Mining Methods. Science & Technology Vision. 2015(12):215.
13. Leopord H, Cheruiyot WK, Kimani S. A survey and analysis on classification and regression data mining techniques for diseases outbreak prediction in datasets. Int. J. Eng. Sci. 2016;5(9):1-1.
14. Manjusha KK, Sankaranarayanan K, Seena P. Prediction of different dermatological conditions using naive bayesian classification. International Journal of Advanced Research in Computer Science and Software Engineering. 2014 Jan;4(1)
15. Varghese DP, Tintu PB. A survey on health data using data mining techniques. International Research Journal of Engineering and Technology (IRJET). 2015 Oct;2(07):2395-0056.
16. Witten, I. H. and E. Frank ,Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, San Francisco. 2005, P. 411
17. Velickovski, F., Ceccaroni, L., Roca, J., Burgos, F., Galdiz, J.B.,Marina, N., and Lluch-Ariet, M., Clinical Decision Support Systems (CDSS) for preventive management of COPD patients. Journal of translational medicine, 201412(2):28.
18. Chaurasia, V., and Pal, S., Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability. International Journal of Computer Science and Mobile Computing, 2014 3:10–22.
19. Heydari, M., Teimouri, M., Heshmati, Z., and Alavinia, S.M.,Comparison of various classification algorithms in the diagnosis of type 2 diabetes in Iran. International Journal of Diabetes in Developing Countries, 201636(2):167–173.
20. Patil, B.M., Joshi, R.C., and Toshniwal, D., Hybrid prediction model for type-2 diabetic patients. Expert systems with applications, , 2010 37(12):8102–8108.
21. Ram S, Gupta S. Building Machine Learning Based Diseases Diagnosis System Considering Various Features of Datasets. InEmerging Trends in Expert Applications and Security 2019 (pp. 147-155). Springer, Singapore.