PREDICTION OF SKIN DISEASE USING DECISION TREE AND ARTIFICIAL NEURAL NETWORK (ANN)

Tinuke Omolewa Oladele, Dorcas Romoke Olarinoye, Samuel Segun Adebisi

Department of Computer Science, University of Ilorin, Ilorin, Nigeria

Corresponding Author: Tinuke Omolewa Oladele, tinuoladele@gmail.com

ABSTRACT: Skin diseases are common diseases that exist between the children and adults in the society. The issue of finding and proffering a better skin disease predictive model in the health care system has been identified to be major problem. Thus, this study provides a comparative evaluation on two data mining classification techniques; Decision Tree and Multi-layer Neural Network apply for the prediction of skin diseases. All experimental analysis were carried out in WEKA data mining tool environment. Each individual classifier was put through training and testing using the N-fold cross validation technique (N value was set to 10). The two classifiers are Decision Tree and Neural Network family respectively. The predictive model obtained from the J48 and Multi-layer Perceptron (MLP) was measured and evaluated accordingly with the use of basic parameters such as accuracy, kappa statistics, TP Rate, FP Rate, Precision, Recall, ROC area. Multi-layer neural network presented accuracy of 96.9945 % while J48 gave an accuracy of 93.9891%.

KEYWORDS: Skin Disease, Prediction, Decision Tree, Artificial Neural Network, Multi-layer Perceptron (MLP).

1. INTRODUCTION

The human skin represents one of the largest organs of the body that helps to protect against infection, heat, injury, and any form of damage which can be caused by ultraviolet (UV) radiation to body ([PJ17]). Diseases of the skin are common to everyone and different types of reactions are becoming more popular ([A+15]). Many of these diseases are very dangerous, particularly if not treated at an early stage. The technique of data mining involves a method of obtaining hidden knowledge from data ([MSS15]). The approaches and techniques available in data mining aid the transformation of large data into relevant data for knowledge discovery, prediction and intelligent decision making ([A+17]) The field of medicine is the significant field for the application of data mining and analysis technique as it concerns human life which is the most important thing in the society ([SS13]).

The predictive analytic is a branch of data mining that automatically produce model for classification from training dataset and employ such model to predict further classes of datasets that are not classified ([VVG12]). Predictive data mining approach handles learning algorithms to help clinicians in monitoring tasks, diagnostics and therapeutics. In predictive modeling, data is collected, a statistical model is formulated, predictions are made and the model is validated or revised as additional data becomes available. Clinical data mining is based on strategic research to retrieve, analyze and interpret both qualitative and quantitative information available from medical datasets or records.

There are several classification techniques in data mining such as Naive Bayes, K-nearest neighborhood, neural network, logistic regression, support vector machine (SVM) and decision tree ([AQR15]). Among these techniques, decision tree is one of the most powerful classification approaches used in decision support system and machine learning process ([SRM12]). A decision tree is a predictive modeling technique that is used in classification, clustering and predictive task. This technique applies a divide-conquer approach to split the problem search space into subsets. The decision tree classification is easy to understand approach, which especially used when there is a need to understand the structure of trained knowledge models.

In healthcare system, there is a lot of significant information connected to the patient's environment for the diagnosis, but this information is often difficult to be synthesized by a human brain. Thus, employing the computation speed of and the strength of data mining algorithms is of most relevance as it saves time and many other resources. In addition, this gave rise to where the artificial neural networks, a mathematical model of how the brain work, prove their power. In this paper, two machine learning algorithms; Decision Tree (DT) and Artificial Neural Networks (ANN) were used to develop a Predictive Model for the prediction of skin diseases.

2. RELATED WORK

([Kad17]) developed classification system for skin diseases using data mining approach. The study used a system based on the decision tree for mining and processing image data. Decision tree and digital

image processing principals were applied to detect skin diseases using some features found in a digital image for a skin. The major steps involved in the system are: pre-processing, features extraction and decision tree classifier. This system enhanced the classification process to be more accurate. The system was designed and implemented in MATLAB environment and also tested with the images of various databases.

([G+17]) considered a train of images that obtained from a given data set and preprocessing and segmentation was performed on each image. After the image segmentation, the skin was properly checked whether it has been affected by any disease or not. Three diseases were taken into consideration; psoriasis, vitiligo and skin cancer. Once the presence of disease is detected the portion of area affected by the disease will be highlighted indicating the exact location of the disease on the skin. From the affected area the classification of disease through data mining will be performed. The segmentation of image is done using a tolerance value, this value is calculated through histogram.

([AD17]) implemented a system for real time detection and classification of skin diseases. The skin diseases were pre-processed before it was passed to classification stage. The detection of pigmented skin lesion acquired from the patient was performed and a database prepared. The captured images are in RGB format, thus colour to monochrome image conversion was achieved. The gray scale image was used to perform image binarization using thresholding. Finally, identification of border and automating of the original image was done. This helped in detecting the infected skin fast and accurately. Feature extraction is performed for efficient classification of skin diseases.

([MSS15]) studied different dermatological diseases with like symptoms which may even prove deadly if not properly attended at exact time. The medical dataset used for the work contained 230 instances with 22 attributes. The study experimented on data gathered from the southern part of Kerala, India. For better prediction calculation of accuracy of mining algorithm is important. Two data mining classification algorithm Naive Bayes' (NB) and J48 were used for data analysis. WEKA Open Source data mining tool was used to carry out analysis and also to reveal the chances of different dermatological disease and also finds out the probabilities of occurrence of each disease.

([OOO13]) focus was on the designing and modeling of a system that will collate past Pigmented Skin Lesion (PSL) image results. Analysis, corresponding observations and conclusions by medical experts were done using prototyping approach. A part of the system used computational intelligence technique to analyze, process, and classify the image library data based on texture and possibly morphological features of the images. Trained medical personnel in a remote location can use mobile data acquisition devices (such as cell phone) to generate images of PSL, supply such images as input to the proposed system, which in turns would intelligently be able to specify the malignancy (life-threatening) or benign (non-threatening) status of the imaged PSL.

3. METHODOLOGY

The proposed system is a predictive data mining model aimed at providing a better prediction by applying Decision Tree and Neural Networks. WEKA data mining tool was used for carrying out this proposed study. The J48 and Multi-layered perceptron algorithm was applied on the dataset. The J48 algorithm was selected under the decision tree category while Multi-layered perceptron was chosen from the Neural Network category as both algorithms have complementary advantages and disadvantages. Knowledge representation of Decision Tree is easily understood by humans, which is not the case for Neural Networks and also Decision Trees have trouble dealing with noise in training data, which is again not the case for Neural Networks. Decision Trees learn fast and Neural Networks learn relatively slowly. This experiment was carried out step-by-step with focus on applying the selected classification algorithms on the Dermatology dataset. The procedures included environment setup, data preprocessing (putting together of clinical and histopathological attributes of patients and collating all of them together in a single database and formatting the database into usable dataset), choosing the data mining software, running the simulation and evaluation and measurement of the performance of the classifiers.

3.1 DATASET

This study was developed on the basis of the information gotten from the online dataset for skin diseases from Dermatology Database of Gazi University, School of Medicine. After filtering and correcting missing values, 366 skin disease data were obtained which was used for prediction of six skin diseases which showed similar symptoms. Among all, 112 pieces of data were psoriasis, 61 seboreic dermatitis, 72 lichen planus, 49 pityriasis rosea, 52 cronic dermatitis and 20 pityriasis rubra pilaris as shown Table 1.It was also discovered that the database contains 34 attributes, 33 of which are linear valued and one of them is nominal. They all share the clinical features of erythema and scaling, with very little differences. The diseases in this group are psoriasis, seboreic dermatitis, lichen planus, pityriasis

rosea, cronic dermatitis, and pityriasis rubra pilaris. Usually a biopsy is necessary for the diagnosis but unfortunately these diseases share many histopathological features as well.

Table 1: Skin Disease Records in Database

CLASS CODE	SKIN DISEASE CLASS	NUMBER OF INSTANCES
1	Psoriasis	112
2	Seboraic dermatitis	61
3	Lichen planus	72
4	Pityriasisrosea	49
5	Cronic dermatitis	52
6	Pityririasisrubra pilaris	20

4. RESULTS AND DISCUSSION

The skin disease dataset was passed as input to each base classifier. The multilayer perceptron and J48 machine learning algorithm and their individual performance were recorded respectively and the output measured and evaluated as follows:

4. 1 Experimental Result of Multilayer Perceptron Algorithm for Classification

Results of multilayer perceptron neural network is represented in Figure 1, with the parameters for performance measurement evaluation of MLP on the size of skin disease predicted and their weighted average obtained for more accurate results as shown in Table 2 and Table 3.

4.2 Experimental Result of J48 Algorithm for Classification

Results of J48 Decision tree is represented in Figure 2, with the parameters for performance measurement evaluation of J48 on the size of skin disease predicted and their performance evaluation results are represented in Table 4 and Table 5.

```
Class 6
    Input
    Node 5


Time taken to build model: 83.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         355               96.9945 %
Incorrectly Classified Instances        11                3.0055 %
Kappa statistic                          0.9624
Mean absolute error                      0.0411
Root mean squared error                  0.0927
Relative absolute error                 15.4167 %
Root relative squared error             25.4006 %
Total Number of Instances              366

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0.004      0.991      1         0.996      1         1
                 0.918     0.016      0.918      0.918     0.918      0.996     2
                 0.986     0          1          0.986     0.993      1         3
                 0.898     0.016      0.898      0.898     0.898      0.996     4
                 1         0          1          1         1          1         5
                 1         0          1          1         1          1         6
Weighted Avg.    0.97      0.006      0.97       0.97      0.97       0.999

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
 112   0   0   0   0   0 |   a = 1
   1  56   0   4   0   0 |   b = 2
   0   0  71   1   0   0 |   c = 3
   0   5   0  44   0   0 |   d = 4
   0   0   0   0  52   0 |   e = 5
   0   0   0   0   0  20 |   f = 6
```

Figure 1: Result of MLP for Skin Disease Predictive Model

Table 2: Performance Evaluation of MLP

PARAMETER	Result
CORRECTLY CLASSIFIED INSTANCES (%)	96.9945
INCORRECTLY CLASSIFIED INSTANCES (%)	3.0055
KAPPA STATISTICS	0.9624
MEAN ABSOLUTE ERROR	0.0411
ROOT MEAN SQUARED ERROR	0.0927
RELATIVE ABSOLUTE ERROR	15.4167
ROOT RELATIVE SQUARED ERROR	25.4006

Table 3: Performance Evaluation of MLP

PARAMETER	Result
TP RATE	0.97
FP RATE	0.006
PRECISION	0.97
RECALL	0.97
F-MEASURE	0.97
ROC AREA	0.999

Table 4: Performance Evaluation of J48

PARAMETER	RESULT
CORRECTLY CLASSIFIED INSTANCES (%)	93.9891
INCORRECTLY CLASSIFIED INSTANCES (%)	6.0109
KAPPA STATISTICS	0.9246
MEAN ABSOLUTE ERROR	0.0264
ROOT MEAN SQUARED ERROR	0.1365
RELATIVE ABSOLUTE ERROR	9.9147
ROOT RELATIVE SQUARED ERROR	37.397

Table 5: Performance Evaluation of J48

PARAMETER	RESULT
TP RATE	0.94
FP RATE	0.015
PRECISION	0.94
RECALL	0.94
F-MEASURE	0.94
ROC AREA	0.977

```
ives   :      30

:ree :       40

) build model: 0.01 seconds

:d cross-validation ===
===

issified Instances          344               93.9891 %
Classified Instances         22                6.0109 %
:ic                           0.9246
: error                       0.0264
ared error                    0.1365
olute error                   9.9147 %
: squared error              37.397  %
 of Instances                366

 Accuracy By Class ===

   TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
     0.964     0.024       0.947     0.964     0.956      0.983    1
     0.885     0.023       0.885     0.885     0.885      0.958    2
     0.972     0.01        0.959     0.972     0.966      0.983    3
     0.878     0.009       0.935     0.878     0.905      0.967    4
     0.981     0.006       0.962     0.981     0.971      0.989    5
     0.9       0.003       0.947     0.9       0.923      0.973    6
     0.94      0.015       0.94      0.94      0.94       0.977

 Matrix ===

   d    e    f   <-- classified as
   0    1    0 |   a = 1
   3    0    1 |   b = 2
   0    1    0 |   c = 3
  43    0    0 |   d = 4
   0   51    0 |   e = 5
   0    0   18 |   f = 6
```

Figure 2: Result of J48 for Skin Disease Classification

5. SUMMARY AND DISCUSSION

The prediction of skin diseases using two data mining classification techniques gives different results for each technique used. It was observed that the accuracy of 96.9945% was obtained in multilayer perceptron neural network (MLP), while 93.9891% was recorded in J48 decision tree. The misclassification rate of 3.0055% was obtained in MLP, while the J48 presented misclassification of 6.0109%. The MLP got 0.97 True Positive Rate (TP RATE), while J48 got 0.94 True Positive (TP RATE). For the False Positive Rate (FP RATE), MLP recorded 0.006, while J48 obtained 0.015 for False Positive Rate (FP RATE). For precision, MLP gave 0,97, while the J48 gave 0.94. The MLP gave recall rate of 0.97, while the J48 recorded 0.94. Finally, from the overall results obtained from the skin diseases predictive system, it was revealed that MLP approach outperformed the J48 decision tree approach.

6. CONCLUSION

The diagnosis process in medicine is an essential but also intricate task that should be done with great accuracy and efficiency. In dermatology, there are many diseases that show similar pattern in appearance and symptoms. Diagnosis of a single patient can differ significantly, if the patient is examined by different physicians. There is need for proper measures for predicting such a disease. This paper applies two data mining techniques; Decision Tree and Multi-layer Neural Network for the prediction of skin diseases. Each classification approach was put through training and testing using the N-fold cross validation technique (N value was set to 10). The predictive model obtained from the J48 and Multi-layer Perceptron (MLP) was measured and evaluated accordingly with the use of basic evaluation metrics which include; accuracy, kappa statistics, TP Rate, FP Rate, Precision, Recall, ROC area. The experimental results from Multi-layer neural network showed that MLP peformed better than the J48 decision tree.

REFERENCES

[AD17] **K. Abhijith, L. K. Deepa** - *Detection and Classification of Skin Diseases by Image Analysis Using MATLAB*, Int. J. Emerg. Res. Manag. Technol., vol. 9359, no. 5, pp. 779–784, 2017.

[AQR15] **S. Q. Ahmad, P. Qamar, S. Rizvi** - *Techniques of Data Mining In Healthcare: A Review*, Int. J. Comput. Appl., vol. 120, no. 15, pp. 38–50, 2015.

[A+15] **A. A. L. C. Amarathunga, E. P. W. C. Ellawala, G. N. Abeysekara, C. R. J. Amalraj** - *Expert System For Diagnosis Of Skin Diseases*, Int. J. Sci. Technol. Res., vol. 4, no. 1, pp. 174–178, 2015.

[A+17] **T. O. Aro, B. J. Muhammed, O. B. Ayoade, I. D. Oladipo** - *A Review on Data Mining Techniques for Heart Disease Prediction*, Anale. Ser. Inform., vol. XV, no. 1, pp. 99–103, 2017.

[G+17] **S. Gindhi, S., Nausheen, A., Zoya, A., Ruhin** - *An Innovative Approach for Skin Disease Detection Using Image Processing and Data*, vol. 5, no. 4, pp. 8135–8141, 2017.

[Kad17] **Q. K. Kadhim** - *Classification of Human Skin Diseases using Data Mining*, Int. J. Adv. Eng. Res. Sci., vol. 4, no. 1, pp. 159–163, 2017.

[MSS15] **P. Manjusha, K K, Sankaranaayana, K. Seena** - *Data Mining in Dermatological Diagnosis : A Method for Severity Prediction*, Int. J. Comput. Appl., vol. 117, no. 11, pp. 11–14, 2015.

[OOO13] **D. A. Okuboyejo, O. O. Olugbara, S. A. Odunaike** - *Image Classification*, World Congress on Engineering and Computer Science, 2013, vol. II, pp. 23–25.

[PJ17] **S. R. Parvin, O. A. M. Jafar** - *Prediction of Skin Diseases using Data Mining Techniques*. Int. J. Adv. Res. Comput. Commun. Eng., vol. 6, no. 7, pp. 313–318, 2017.

[SS13] **D. Singh, H. Naveen, J. Samota** - *Analysis of Data Mining Classification with Decision Tree Technique*, Glob. J. Comput. Sci. Technol., vol. 13, no. 13, pp. 1–6, 2013.

[SRM12] **Seema, M. Rathi, Mamta** - *Decision Tree: Data Mining Techniques*, Int. J. Latest Trends Eng. Technol., vol. 1, no. 3, pp. 150–155, 2012.

[VVG12] **D. B. Varun, K. M. Vijay, S. V. Gayathri** - *Hepatitis Prediction Model based on Data Mining Algorithm and Optimal Feature Selection to Improve Predictive Accuracy*, Int. J. Comput. Appl., vol. 51, no. 19, pp. 13–16, 2012.