Comparative Study of Data Mining And Neural Techniques For An Automated Skin Disease Diagnostic System

Rashika Singh[1*], Kashish Desai[2], Tanvi Ruparel[3] and Prof. Mitchell D'silva[4]

[1-3] Students, [4]Assistant Professor
[1-4] Information Technology
[1-4] Dwarkadas J. Sanghvi College of Engineering, Mumbai, India

[1]singhrashika05@gmail.com, [2]kashishdesai97@yahoo.com,
[3]tanviruparel97@gmail.com, [4]mitchell.dsilva@djsce.ac.in

Abstract

Skin disease can be seen among people on a daily basis with tremendous suffering and hardships involved. Dermatologists with high-level of expertise are required for efficient diagnosis of the skin disease due to the variety in visual aspects involved. The diagnosis also may be subjective. There arises a need to develop a computer aided system that can give a reliable and objective diagnosis. This paper proposes the development of a system that makes use of not only images but also symptoms provided by the patient. It compares six techniques that can be used to predict the diseases namely K-nearest neighbour (K-NN), Naive Bayes Classifier, Random Forest, Support Vector Machine (SVM), Convolutional Neural Network (CNN), and Artificial Neural Network (ANN). Based on the performance of all the algorithms, CNN emerges as the most suitable algorithm for the proposed system. The paper also includes a development model for the proposed system.

Keywords: Diagnosis of skin disease, K-nearest neighbour(KNN), Naive Bayes Classifier, Random Forest, Support Vector Machine(SVM), Convolutional Neural Networks(CNN), Artificial Neural Network (ANN).

1. Introduction

Skin diseases like psoriasis, lichen planus, seborrheic dermatitis are some of the most common infections seen among people. Due to the disfigurement and associated hardships, skin disorders cause lots of trouble to the sufferers. Diagnosis of skin diseases is not easy as it requires high-level of expertise [1]. This diagnosis may be subjective because it depends on the medical practitioner's expertise and level of understanding. This diagnosis may vary from one expert to the other. To avoid this subjectivity and give a more accurate prediction, there occurs the need of a computer aided system. A computer system which makes use of images to classify and predict skin diseases shall result in more objective and reliable diagnosis. To further increase the accuracy, this paper proposes the development of a system that makes use of not only images but also symptoms provided by the patient. There occurs a need of such a system in today's day and age because a user with a minute skin disease does not have time to go to the dermatologist. The users can be assured by the help of this system whether their disease is serious or benign and if a visit to the dermatologist is necessary. Thus by using the system

the users can assess their condition rather than ignoring it and making it worse. The proposed system uses about 23,000 images collected from Dermnet (www.dermnet.com) and calls it the Dermnet dataset along with a Erythemato-Squamous Disease (ESD) dataset taken from UCI Machine Learning Repository database, to train the CNN architecture. The ESD dataset has 34 clinical and histopathological attributes used to classify the diseases. This paper presents a comparative study of the algorithms that can be used to implement the proposed system. Section 2 describes the literature review of various existing systems for identifying and predicting skin diseases. Sections 3 and 4 provide the description and comparison of the various algorithms and models that can be used for skin disease prediction. Section 5 presents the design of the proposed system which uses the technique or model that is suitable and efficient for the selected data set and section 6 contains conclusion of the paper.

2. Literature review

This section discusses the various data mining techniques and neural network models that are used for detecting and predicting the various skin diseases.

Several techniques are used to improve the efficiency of identifying and predicting skin diseases. The feasibility of skin disease classification system using deep CNN is examined in [1]. The classification is done by fine-tuning ImageNet pretrained models-VGG16, VGG19, GoogleNet with the Dermnet dataset. The same models are also applied to the dataset OLE which gives a comparatively lower accuracy as there is not a broad variance in the training set. This shows that by increasing the variance in the training set better accuracy can be achieved.

A system that makes use of Multi SVM (Support Vector Machine) classifier, K-NN and Naïve Bayesian classifier is proposed in [2]. The input is images from the AOCD unit database and the MIT unit database which are pre-processed using Gaussian Filter technique. K-means clustering algorithm is then applied to partition the disease affected area and the non-affected area followed by feature extraction using Grey Level Co-occurrence Matrix (GLCM) for examining texture which gave the statistical parameters, for better classification efficiency. The combination of the algorithms used has provided the highest accuracy for the selected dataset.

The techniques presented in [3] present a 2-stage process for prediction of skin diseases, where the disease region is converted into a feature vector and then used for training the network. This system is capable of detecting 5 main skin disorders namely psoriasis, melanoma, scleroderma, eczema, impetigo.

A combination of Image Processing and Neural Networks such as Artificial Neural Network (ANN) is discussed in [4]. The proposed system is able to successfully detect the dermatological disease present in the image. This system is used for detecting diseases such as Nevus and Ringworms. The dataset is relatively small hence, the error in prediction was less.

3. Study of Algorithms

This section provides a brief description about the various data mining techniques and neural network models that are used by the existing systems.

3.1. K-Nearest Neighbour (KNN)

The k-Nearest Neighbour is one of the simplest supervised machine learning algorithm used for classification. It requires little or no knowledge to classify the data. K-NN classifies new cases based on similarity while storing the available cases. In this approach, the data classification is done by estimating how likely a data point will belong to one class or the other, depending on what class of data point nearest to it are in [5]. K-NN is a

type of instance-based learning or lazy learning, in which it does not built the model until a query of data set is performed on the training set. The only calculation it makes is when it is asked to vote the data point's neighbour. This makes it easy to implement for data mining. The distance between the two data points is calculated using Euclidean distance which is the most widely used distance function. The value of "k" in K-NN stands for number of nearest neighbours we use to account for assigning a class to current testing data point. The efficiency or the performance of this algorithm is primarily based on the selection of the parameter k. Since K-NN is based on the feature similarity; choosing the right value of k is the process of parameter tuning and is important for better accuracy. If the value of k is small, then noise will have higher influence on the results and might get skewed results. So the output obtained is noisy. However, if the value of k is large then it degrades the classification performance. It introduces outliers from other classes. It even makes it computationally expensive. A general rule of thumb can be used to determine the value of k which is equal to the square root of N, where N stands for the number of samples in the training data set. Another way is to select odd value of k to avoid confusion between the two classes of data. Sometimes value of k depends on different individual cases, so the best way is to run through each possible value of k and choose the best value suited for the selected dataset.

3.2. Naive Bayes Classifier

Naive Bayes Classifier is a simple classification technique based on Bayes' Theorem with the assumption that the predictors or events used are independent of each other. This assumption is called as class conditional independence. Naive Bayes uses a conditional probability model that estimates the likelihood of a property given the set of input or training data [9]. It requires initial data to estimate parameters such as mean and variance that are necessary for classification and the instance is classified to a class with the highest conditional probability. Naive Bayes Classifier assumes the events or predictors are independent of each other. This assumption is called as class conditional independence [10]. Bayes Theorem provides a way of calculating posterior probability, P(C|X), from $P(C)$, $P(X)$, and $P(X|C)$.[10]

$$P(C|X) = P(C)\ P(X|C)\ P(X) \qquad \text{[10]…….. (Equation 1)}$$

$$P(C|X) = P(X_1|C)\ .\ P(X_2|C)\ ...\ P(X_n|C)\ .\ P(C) \qquad \text{[10]…….. (Equation 2)}$$

Since features are independent of each other in Naive Bayes Classifier, denominator predictor prior probability, P(X), is effectively constant as it does not depend on C and the values of the feature X_i. A Naive Bayes model is easy to build due to its simplicity as no complicated iterative parameter estimation is required. It requires a small amount of training data to estimate or classify the parameters. Despite its simplicity, it outperforms many sophisticated classification techniques and good results are obtained in most of the cases [11]. Its disadvantage lies in the assumption of class conditional independence. Practically, dependencies exist among variables which cannot be modelled by Naive Bayesian Classifier [11].

3.3. Random Forest

Random Forest or Random Decision Forest is one of the most accurate and simplified classification technique that constructs a multitude of decision trees at training time and outputs the class that is the mode of the classes of the individual trees [13]. It is one of the most accurate classification algorithms. It runs efficiently on large data sets and can handle thousands of variables without variable deletion. Since it makes use of multiple

decision trees it reduces the risk of overfitting and the training time required is less [14]. When large portion of data is missing it maintains its accuracy by estimating the missing data. However, in few cases Random Forest is observed to overfit for some data set with noisy classification tasks. They are not easily interpretable, so basically Random Forest is used when high performance is required with less need of interpretation [15].

3.4. Support Vector Machine (SVM)

Support vector machines (SVMs, also support vector networks) are supervised learning models that construct a hyperplane (decision plane) or set of hyperplanes in an infinite-dimensional space, which can be used for classification, regression, or other tasks like outliers detection [16]. A hyperplane is a linear classifier for a set of training points in a dataset. The vectors defining the hyperplane are known as the support vectors. SVM aims at minimizing an upper bound of the generalization error through maximizing the margin between the separating hyperplane and the data [17]. SVMs can be used to analyse a dataset of images for classification analysis. The goal of SVM (Support Vector Machine) is to produce a model that can perform accurate classification for a given labelled training dataset. Pre-decided labels help in indicating the correct working. There are two known approaches for multi-class SVM. One is to construct and combine binary classifiers while the other is to consider all of the data directly in one optimization formulation. In order to solve multi-class SVM problems in a single step, the variables should be proportional to the number of classes. Therefore, for multi-class SVM methods, either several binary classifiers have to be constructed or a larger optimization problem is needed. Hence in general it is computationally more expensive to solve a multiclass problem [18].

3.5. Convolutional Neural Networks (CNN)

Neural networks process similar to a human brain. The network consists of parallel processing elements in a large number which are working in parallel for solving a specific problem. Each neuron generates a output value by certain functions applied to the input layer. The function which is applied comprises of a bias and a vector of weights which are adjusted when learning progresses in the network. Convolutional Neural Networks are a class of forward feed neural network which implies that there are no cycles in the network and flow of information is only in one direction. CNN are primarily used for image classification and are gaining prominence due to their high accuracy. The architecture combines benefits of standard network training with convolutional operation for image classification. A Convolutional layer will apply the convolution operator to a single layer and subsequently pass it to the next layers. The CNN architecture comprises of input layer, output layer and hidden layer which again consists of the pooling layer, convolutional layer, fully connected layers and normalization layer. It has features such as local connectivity where neurons are connected to a subset of input image and parameter sharing which is sharing of neurons. The accuracy of CNN's is deemed to be higher than humans in certain cases. They see images piece by piece and extract the useful features layer wise. The most important feature is "pooling" in which the system takes an image and shrinks it down to include the most important features.

3.6. Artificial Neural Network (ANN)

ANN systems "learn" to perform tasks by considering examples, generally without being programmed with any task-specific rules [22]. It is used for image classification where it classifies object based on the label and this learning can be applied to other tasks. Every non-looping computation can be expressed as Artificial Neural Network so it refers to a

general implementation in computability terms. ANN is used for effectively classifying images with hidden layers.

4. Comparison of data mining and neural network techniques

Table 1: Comparison of various techniques

Criteria	**Data Mining algorithms**	**Convolutional Neural Network**	**Artificial Neural Network**
Efficiency	Works well for few small datasets.	Works well for small and large datasets.	Works well for small and large datasets.
Overall Performance	Depends on the training dataset values	Depends on the architectural parameters	Depends on the architectural parameters
Advantages	1.Simplicity of the algorithm 2.Building of the model is cheap 3.Little or no knowledge is required for classification 4.Small amount of dataset required for classification 5.Less computing time required	1.Transfer learning is possible 2.Efficient zooming features 3. Suitable for large data sets 4. Less pre-processing is required as compared to others	1. Suitable for hidden layers in images 2. Ability to work with incomplete knowledge. 3. Fault tolerance
Disadvantages	1.No transfer learning 2.Accuracy degrades in presence of noisy features 3.Variable interdependencies are ignored 4.Computation is highly expensive 5.. Large number of binary classifiers are required	1. Computation time depends on the filter size and the number of fully connected layer units [23]	1. Rarely used for predictive modelling 2. Difficult for the network to interpret the problem

5. Design of the proposed system

Figure 1 shows the design of the Proposed System Architecture for the proposed system

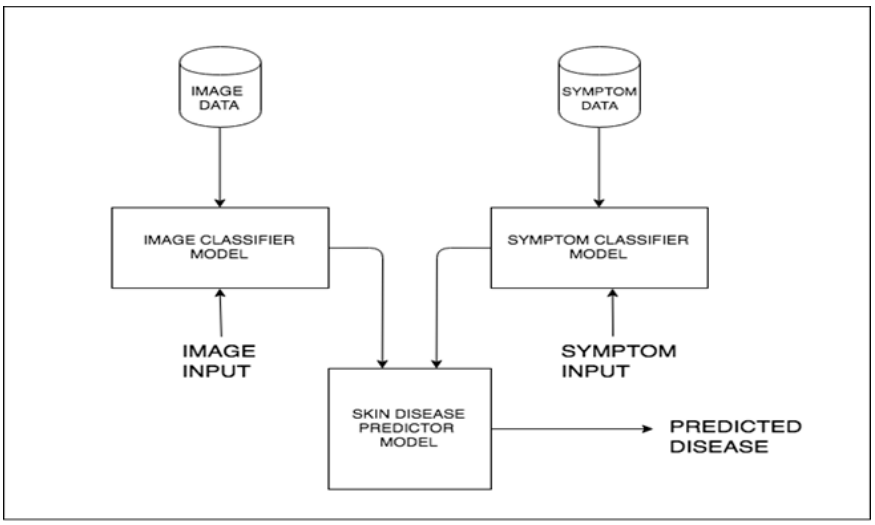

Figure 1. Proposed System Architecture

Figure 2 describes the implementation flow of the proposed system for diagnosis of skin diseases.

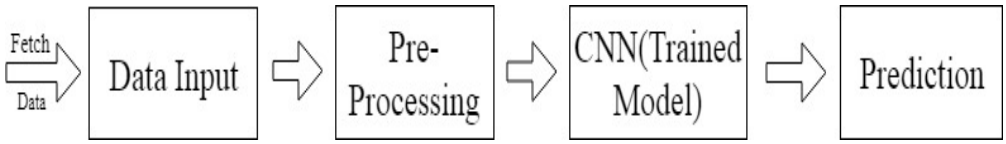

Figure 2. Implementation Flow of the Proposed System

A. Input Image:

The input for this system are the digital images of skin disease. The dataset we use comprises of 23,000 images collected from Dermnet (www.dermnet.com) and call it the Dermnet dataset along with a Erythemato-Squamous Disease (ESD) dataset that was taken from UCI Machine Learning Repository database, to train the CNN model. The ESD dataset has 34 clinical and histopathological attributes used to classify the diseases.

B. Pre-Processing:

Image preprocessing is one of the basic step required for image enhancement before applying machine learning algorithms in order to highlight interesting details and remove noise from the image. The main purpose of this step is to remove the unwanted or surplus region in the background of the image in order to increase the accuracy of the system. It is used to perform image augmentation which has includes image scaling, shearing, rotation, etc. It further converts the image to a vector consisting of RGB values of the image. The objective of this

stage can be achieved by image enhancement techniques like sharpening filters, smoothening filters and histogram processing.

C. Classification:

The vector generated in the pre-processing stage is applied as input to the Convolutional Neural Network. CNN model will extract the features from the vector and uses these features as different layers to train the model which will predict the skin diseases.

6. Conclusion

Many learning algorithms like K-NN, SVM, Naive Bayesian, and Random Forest have been used for image classification. However, CNN has emerged as the model of choice owing to certain disadvantages of the other systems and certain unique features of CNN. One primary reason is due to the fact that transfer learning is not possible in machine learning algorithms and thus using CNN which comes under deep learning provides a mean for transfer learning in which a base model is developed for a task and this model acts as a starting point for the next task. CNN proves to be the most scalable for a large dataset and hence is chosen to classify the skin diseases for the dataset which uses about 23,000 images. CNN provides an efficient way to zoom in, and zoom out of the images that we use along with making predictions even for the small areas. Thus the paper proposes the use of Convolutional Neural Network (CNN) to make the proposed skin disease predictor accurate and efficient.

References

1. Liao, Haofu. “A Deep Learning Approach to Universal Skin Disease Classification.” (2015).
2. S. Reena Parvin, O.A. Mohamed Jafar “Prediction of Skin Diseases using Data Mining Techniques-Multi-SVM classifier, K-NN and Naïve Bayesian classifier” International Journal of Advanced Research in Computer and Communication Engineering-2017.
3. Lakshay Bajaj,Himanshu Kumar,Yasha Hasija “Automated System for Prediction of Skin Disease using Image Processing and Machine Learning” International Journal of Computer Applications-2018
4. R. Yasir, M. A. Rahman and N. Ahmed, "Dermatological disease detection using image processing and artificial neural network," *8th International Conference on Electrical and Computer Engineering*, Dhaka, 2014, pp. 687-690.doi: 10.1109/ICECE.2014.7026918
5. What is K-Nearest Neighbor (K-NN)? - Definition from Techopedia (2018,August,20) Techopedia.com. [Online]. Available: http://www.techopedia.com/definition/32066/k-nearest-neighbor-k-nn.
6. S. B. Imandoust and M. Bolandraftar, “Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background,” International Journal of Engineering Research and Applications, Vol. 3, Issue 5, Sep-Oct 2013 .
7. Value of k in k Nearest Neighbor Algorithm (2018,August,25) Stack Overflow (2018,August,20) stackoverflow.com/questions/11568897/value-of-k-in-k-nearest-neighbor-algorithm.
8. KNN Algorithm - How KNN Algorithm Works With Example | Data Science For Beginners | Simplilearn (2018,September,06) YouTube [Online] Available: https://www.youtube.com/watch?v=4HKqjENq9OU.
9. U. K. Pandey, “Data Mining : A prediction of performer or underperformer using classification,”*(IJCSIT) International Journal of Computer Science and Information Technologies*, 2011.

10. Naive Bayesian (2018,September,12) saedsayad [Online] Available: https://www.saedsayad.com/naive_bayesian.htm.
11. Naive bayes (2018,September,15)slideshare [Online] Available: https://www.slideshare.net/ashrafmath/naive-bayes-15644818.
12. Naive Bayes Classifier(2018, September,06) wikipedia[Online]Available: https://en.m.wikipedia.org/wiki/Naive_Bayes_classifier.
13. Random forest (2018,September,15) Wikipedia [Online] Available: https://en.wikipedia.org/wiki/Random_forest.
14. Random Forest Algorithm - Random Forest Explained | Random Forest in Machine Learning | Simplilearn (2018,September,15) YouTube [Online] Available: https://www.youtube.com/watch?v=eM4uJ6XGnSM.
15. Machine Learning Algorithms: Introduction to Random Forests (2018,September,20) DATAVERSITY [Online] Available: http://www.dataversity.net/machine-learning-algorithms-introduction-random-forests/.
16. Support Vector Machine (2018, September, 26) wikipedia[Online]Available: https://en.wikipedia.org/wiki/Support_vector_machine.
17. Hyeran Byun and Seong-Whan Lee. "Applications of Support Vector Machines for Pattern Recognition: A Survey"(2002)
18. Chih-Wei Hsu and Chih-Jen Lin."A Comparison of Methods for Multi-class Support Vector Machines"(2011).
19. Overview of Convolutional Neural Network in Image Classification (2018,August,28) analyticsindiamagazine[Online] Available: https://www.analyticsindiamag.com/convolutional-neural-network-image-classification-overview/
20. A Gentle Introduction to Transfer Learning for Deep Learning (2018,September,05) machinelearningmastery[Online]Available: https://machinelearningmastery.com/transfer-learning-for-deep-learning/
21. Convolutional Neural Network (2018,September,10) wikipedia[Online]Available: https://en.wikipedia.org/wiki/Convolutional_neural_network
22. Artificial Neural Network(2018,September,15)wikipedia[Online]Available: https://en.wikipedia.org/wiki/Artificial_neural_network
23. Dependence of CNN computation time on the filter size and the number of fully connected layer units(2018,September,22)hackerearth[Online]Available: https://www.hackerearth.com/practice/notes/dependence-of-cnn-computation-time-on-the-filter-size-and-the-number-of-fully-connected-layer-units/
24. What are the advantages/disadvantages of Artificial Neural networks(2018,September,25) quora[Online]Available: https://www.quora.com/What-are-the-advantages-disadvantages-of-Artificial-Neural-networks