PREDICTION OF HEPATITIS DISEASE USING MACHINE LEARNING TECHNIQUE

Meesala Durga Prasad, G.Sasibhushana rao Dept. of Electronics and Communications Engineering Andhra University college of engineering ,

***Abstract*— The objective of this work is to choose the best tool for diagnosis and detection of Hepatitis as well as for the prediction of life expectancy of Hepatitis patients. In this work, a comparative study between various machine learning tools and neural networks were carried out. The performance metric is based on the accuracy rate and the mean square error. The Machine Learning (ML) algorithms such as Support Vector Machines (SVM), K Nearest Neighbor (KNN) and Artificial Neural Network (ANN) were considered as the classification and prediction tools for diagnosing Hepatitis disease. A brief study on the above algorithms were performed based on the prediction accuracy of disease diagnosis. All the ML algorithms were implemented and validated using MATLAB software.**

Keywords—Hepatitis, SVM, ANN, KNN, Confusion matrix, mean square error.

I. INTRODUCTION

Medical diagnosis is an important and a quite complex task which requires accurate identification. It is important to diagnose the disease at proper time and to be cured at the earliest. Liver is the vital part of a human body. One of the severe diseases that affect the functionality of liver is hepatitis, which causes inflammation of the liver. The main factor for Hepatitis disease is the presence of virus in liver [1]. Hepatitis is a worldwide disease with high mortality rate. If accurate measures are not taken in proper time, it may affect the vital functions of the body and may cause to cirrhosis, severe scarring and increase the risk of liver cancer [2]. Early detection through proper diagnosis and proper medication can cure the disease. For diagnosis of any disease, the two important things are: (i) The selection of right parameters of diagnosis and (ii) proper analysis of the data with an experienced expertise.

Machine Learning (ML) is the tool which could make a system to learn by itself by detecting different patterns and different relationships for the given data using different algorithms [5]. This would enable automatic diagnosis of any diseases, where the two important things considered with utmost care are: selection of parameters and the tool used for analyzing these parameters.

In this work, a study of three different tools that are used for Hepatitis prediction namely: KNN, SVM and ANN are carried out.

Different researches have undergone for the diagnosis and the prediction of diseases using machine learning techniques [14]. Somaya et al. evaluated different machine learning techniques in the prediction of advanced fibrosis that incorporates serum biomarkers [7]. Haydon et al. used artificial neural networks for the prediction of cirrhosis in patients using routine clinical host and viral parameters [4]. An automatic diagnosis system was proposed by Jiaxin et al. using extreme learning machine on serum indices data of patients to predict the fibrosis stage and inflammatory activity grade of chronic hepatitis C [8]. Sushrutha et al. proposed a hybrid model for the prediction of hepatitis [9]. They have developed a combination of genetic search algorithm and multilayer perceptron technique. The paper
[10] investigates the impact of applying varied different fold for cross validation on missing values using PCA-MLP imputation method. Janhel et al. investigated different ANN models for the prediction of Hepatiti[11]. Cai et al. proposed a classification method using serum biochemistry data of patients in which a collaborative representation model is used with smoothly clipped absolute deviation to diagnosis chronic hepatitis C [3]. Rouhani et al designed various neural network such as RBF, GRNN, PNN, LVQ and SVM to diagnosis hepatitis disease and compared the performance [12].The main objective of this work is to perform a comparative study for a specific dataset by training the same dataset using different ML tool and neural network architecture, and choosing those best tool for diagnosis of Hepatitis disease. Comparative study of various technique is performed based on the accuracy. In this work, we used SVM and KNN ML algorithms and neural network. The rest of the paper is arranged as follows. Section 2 describes the system architecture. In section 3 methodology and the techniques adopted are explained. In section 4 experimental results are discussed in comparison with ML tool and neural network. Some concluding studies and future works are given in section 5.

II. SYSTEM ARCHITECTURE

In this work, the required data set is chosen from UCI repository, considering different clinical cases. This dataset consists of 155 instances with 20 attributes, one among the same attributes is the class to decide the life expectancy of a hepatitis patient. The 19 attributes for classification are shown in table I. Machine learning algorithm such as SVM and KNN were applied to the dataset for training and testing. Followed by this, neural network approach was

performed on the same dataset for performance analysis. Comparison was evaluated based on the prediction accuracy of the tool used as well as the mean square error. Lower the mean square error, better the performance.

TABLE I. ATTRIBUTES IN DATASET

Attributes	Value
Age	Numerical value
Sex	male(1), female(2)
Steroid	no(1), yes(2)
Liver Big	no(1), yes(2)
Liver Firm	no(1), yes(2)
Spiders	no(1), yes(2)
Antivirals	no(1), yes(2)
Fatigue	no(1), yes(2)
Malaise	no(1), yes(2)
Spleen Palpable	no(1), yes(2)
Ascites	no(1), yes(2)
Varices	no(1), yes(2)
Varices	no(1), yes(2)
Bilirubin	0.39, 0.80, 1.20, 2.00, 3.00, 4.00
Alkaline Phosphate	33, 80, 120, 160, 200, 250
Aspartate transaminase	13, 100, 200, 300, 400, 500
Albumin	2.1, 3.0, 3.8, 4.5, 5.0, 6.0
Pro-time	10, 20, 30, 40, 50, 60, 70, 80, 90
Histology	no(1), yes(2)

The system model for hepatitis disease diagnosis is shown in fig.1. The main processes involved in this model can be described as: loading the data, imputing the data and data preprocessing, classifying the data, applying ML technique and diagnose the disease. Steps involved in the process are described in detail in the following sections:

A. Loading the data.

The data set is extracted from UCI repository which consists of 155 instances with 20 attributes. Since ML learns from examples, sufficient and smoothened data have to be given to the network model. To get sufficient data, data imputation was performed on the available dataset.

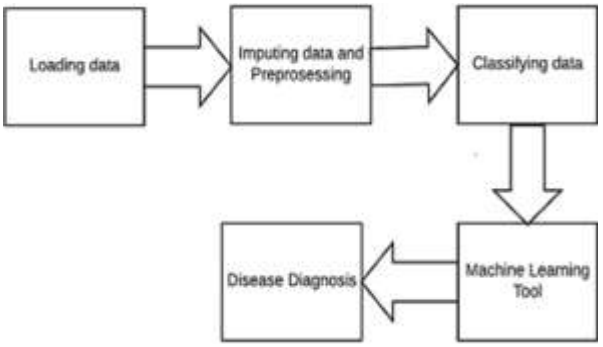

Fig. 1. System Model

B. Imputting the data and data preprocessing.

The presence of missing data in any field, if not handled wisely may result in incorrect prediction and affects the quality of the result. In this hepatitis database, out of 155 instances 75 of them have missing values. To get sufficient data for training, validation and testing, the data augmentation technique was used. Data augmentation was performed thrice to increase the precision of the results. Out of all the instances of the dataset, the missing values corresponding to the instances are removed and the imputation process was performed on the basis of remaining data. The performance of the imputation method was evaluated using the metric error rate, in which the imputed data is compared with the data with non-missing values in the attribute.

C. Classifying the data.

After loading the data and preprocessing phase, classification of the data is carried out. Data classification is carried out in two phases, namely training phase and classification phase. In the training phase, the data is classified into training set and validation set. After that, the classifier algorithm builds the classifier with the training dataset. In the second classification phase, the trained model is used for disease classification and the life expectancy of the Hepatitis person. The data is divided into training (60%), testing (20%) and validation (20%) in a stratified manner.

D. Applying ML tool and diagnose the disease

There are mainly three stages in implementing the machine learning code as well as neural network namely; training, validating and testing. In our research, with the given hepatitis dataset, initially the data was split in to these three categories using stratified splitting. After this, we train the data using suitable machine learning tool and then the data for validation is given to the network. Using the trained network, the testing data is validated, and this is the phase which gives the accuracy of prediction of life expectancy of patients with Hepatitis. SVM, KNN and neural network are the ML tools that we have used in this work.

III. METHODOLOGY

We have used three different algorithms for the classification of hepatitis, which are mentioned below:

A. Support Vector Machine

SVM technique is one of the popular and efficient machine learning techniques. SVM maps the input variable to an n-dimensional feature space. For a given labeled training data, SVM creates a hyperplane that separates the feature space by their class, thereby it avoids the overfitting. Given a sample dataset, it assigns a new category to any of the labelled classes [13].

B. K-Nearest Neighbor

K-nearest neighbor popularly known as KNN is one

of the simplest algorithm for classification. This algorithm mainly works by properly choosing the optimum *k* nearest neighbors [6]. This is a popular machine learning algorithm used for data sets because of its ability to select the neighbor. By selecting the very low and high values of *k* we may not end up with correct results. So we should choose an optimum *k* value for the algorithm for precise result.

C. Artificial neural networks

Artificial neural networks is a non-linear machine learning technique. Linear regression and logistic regression aren't much effective to handle large amounts of data like an image data and other data, so we use the concept of neutral network. It is one of the popular machine learning algorithms, by this method we were able to reach a good accuracy for prediction.

IV. RESULTS AND ANALYSIS

In this section, results are analyzed using different classification techniques such as SVM, KNN and ANN. All the classification technique implementation were performed in MATLAB. Here the dataset contains various attributes of health related data and the class name, which is the life expectancy of patients. Here we divided the data into two parts which consists of training data and other is validating data.

With the given training data, we trained an SVM model. Then using validating data we validated the model and plotted a confusion matrix and also found out the mean square error. By using SVM we achieved a result of 89.58
%. The Mean Square Error (MSE) obtained using this method is 0.1042. These two results are obtained from MATLAB and the obtained results are shown in fig.2. Plot of confusion matrix is given in fig.3. Confusion matrix is used to indicate the quality of the classifier for correct prediction. The count value in the confusion matrix shows the number of correct and incorrect classifier predictions. The top row of the confusion matrix gives the predicted positive events with true positive and bottom row corresponds to no events with true negatives. In other words, the diagonal elements represent the number of predicted target classes equal to the true target class. And off diagonal elements corresponds to the misclassified or incorrectly predicted target class.

```
accuracy = 0.8958
mean_square_error =  0.1042
>>
```

Fig.2 Mean Square Error and accuracy of SVM.

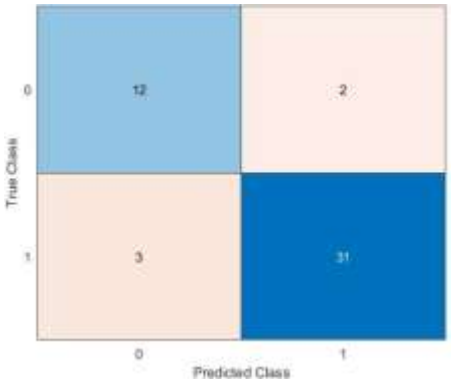

Fig.3. Confusion matrix for SVM.

The same data set was used for classification using KNN model. The data is divided into training data and the validating data by stratified splitting method. Also using the algorithm, we selected different neighbors and implemented the KNN algorithm using optimum *k* value to get good result. Based on the accuracy as well as mean square error (MSE), we have chosen *k* as 7 which gives an accuracy of 85.29 percentage. Also the MSE obtained in MATLAB using KNN method is 0.1471. We plotted the accuracy graph for different *k* values. These two results are shown in fig.4. For a *k* value of 5 and 3 accuracy is good, but mean square error is large which doesn't gives a good result. Also, plot of confusion matrix is given in fig.5 for predicted classes against true classes. From the said methods, SVM gives good accuracy and less mean square error than KNN. Even though, ANN outperforms all these methods.

```
accuracy =  85.2941
mean_square_error =  0.1471
>>
```

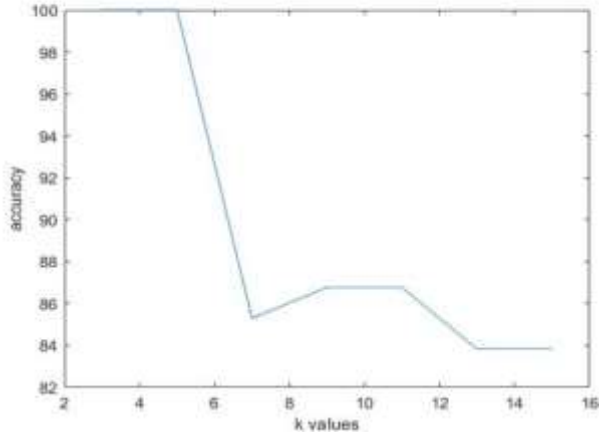

Fig.4. MSE and accuracy of KNN, and accuracy plot for various *k* values.

Fig.5. Confusion matrix for KNN.

ANN model is performed on the same dataset containing various attributes, and life expectancy of patients was predicted more accurately in this model than SVM and KNN. Here we divided the data into three parts which consists of training data, validating data and testing data. With the given training data, we trained the neural network. Then using validating data we validated the network and also tested the network's accuracy and plotted a confusion matrix for the complete dataset, which is for training data, validating data and testing data. By using ANN accuracy is increased to 96.15 percentage which results in a good prediction for disease diagnosis. Using this data the missed classification is 3.84 percentage only. The result obtained from MATLAB is given in fig.6.

Fig.7, shows the plot of confusion matrix separately for training, validation and prediction. In the confusion matrix the rows correspond to output or predicted class and the column corresponds to actual target class. The diagonal elements correspond to correctly classified output observations. Whereas, off diagonal elements correspond observations that are incorrectly classified. From the final confusion matrix it is shown that out of 93 hepatitis prediction, 90 are correctly predicted as hepatitis, which contributes to 96.8% of all true classes and 3 are misclassified, which corresponds to 3.2%. After the performance analysis of the techniques described earlier, ANN was found to be the best fit for classification. To understand the performance of ANN network further in detail, the validation performance of the network was also taken into account for different epochs. Fig.8. shows the validation performance of the neural network. It shows that after 6 epochs the validation performance came to a minimum level and the training stopped at that point. Beyond 6 epochs, the network starts overfitting the data and may degrade the performance of the network.

```
Percentage Correct Classification   : 96.153846%
Percentage Incorrect Classification : 3.846154%
>>
```

Fig.6. Accuracy for ANN

Neural Network (ANN) and K Nearest Neighbor (KNN) to get the accurate prediction of the disease.

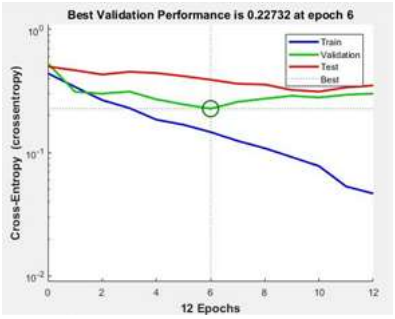

Fig.8. Validation performance of ANN model.

With this study, it is inferred that out of all models considered and its performance, ANN is most accurate that gives a good prediction accuracy of 96 percentage and a minimum mean square error. As a future work the same will be implemented using RNN for the prediction of occurrence of other diseases.

Fig.7. Confusion matrix for ANN.

Neural networks is one of the famous machine learning algorithms and by using this algorithm, the predicting

accuracy came up to a good level.

V. CONCLUSION

In this work, different machine learning techniques and neural networks were used for the diagnosis of hepatitis. A comparison on the accuracy for a particular data set was performed by using various ML and ANN techniques, for identifying the best tool for Hepatitis disease diagnosis. We have used Support Vector Machine (SVM), Artificial

REFERENCES

[1] Ghumbre S. U.; Ghalot A.A, "Hepatitis B Diagnosis using Logical Inference And Self Organizing Map", 2008 ; *Journal of Computer Science* ISSN 1549-3636.

[2] M. A. Chinnaratha, G. P. Jeffrey, G. Macquillan, E. Rossi, B. W. D. Boer, D. J. Speers, and L. A. Adams, "Prediction of morbidity and mortality in patients with chronic hepatitis c by non-invasive liver fibrosis models," *Liver International*, vol. 34, no. 5, pp. 720–727, 2014.

[3] Roslina, A. H., and A. Noraziah. "Prediction of hepatitis prognosis using Support Vector Machines and Wrapper Method." In *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery*, vol. 5, pp. 2209-2211. IEEE, 2010.

[4] G. H. Haydon, R. Jalan, M. Alakorpela, Y. Hiltunen, J. Hanley, L. M. Jarvis, C. A. Ludlum, and P. C. Hayes, "Prediction of cirrhosis in patients with chronic hepatitis c infection by artificial neural network analysis of virus and clinical factors," *Journal of Viral Hepatitis*, vol. 5, no. 4, pp. 255–264, 2010.

[5] Atif Khan, John A. Doucette, Robin Cohen, "Integrating Machine Learning into a Medical Decision Support System to Address the Problem of Missing Patient Data", 2012 *IEEE DOI* 10.1109/ICMLA.2012.82.

[6] Uhmn, Saangyong, Dong-Hoi Kim, Young-Woong Ko, Sungwon Cho, Jaeyoun Cheong, and Jin Kim. "A study on application of single nucleotide polymorphism and machine learning techniques to diagnosis of chronic hepatitis." *Expert Systems* 26, no. 1 (2009): 60- 69.

[7] KayvanJoo, Amir Hossein, Mansour Ebrahimi, and Gholamreza Haqshenas. "Prediction of hepatitis C virus interferon/ribavirin therapy outcome based on viral nucleotide attributes using machine learning algorithms." *BMC research notes* 7, no. 1 (2014): 565.

[8] Vijayarani, S., and S. Dhayanand. "Liver disease prediction using SVM and Naïve Bayes algorithms." *International Journal of Science, Engineering and Technology Research (IJSETR)* 4, no. 4 (2015): 816- 820.

[9] Sartakhti, Javad Salimi, Mohammad Hossein Zangooei, and Kourosh Mozafari. "Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM- SA)." *Computer methods and programs in biomedicine* 108, no. 2 (2012): 570-579.