
Mohammed Alghobiri, King Khalid University, Abdha, Saudi Arabia
Hikmat Ullah Khan, COMSATS University Islamabad, Pakistan
Ahsan Mahmood, COMSATS University Islamabad, Pakistan

ABSTRACT

The human liver is one of the major organs in the body, and liver disease can cause many problems in human lives. Due to the increase in liver disease, various data mining techniques are proposed by the researchers to predict liver disease. These techniques are improving day by day in order to predict and diagnose liver disease in humans. In this paper, a real-world liver disease dataset is incorporated for diagnosing liver disease in the human body. For this purpose, feature selection models are used to select a number of features that are the most important features to diagnose liver disease. After selecting features and splitting data for training and testing, different classification algorithms in terms of naïve Bayes, supervised vector machine, decision tree, k-nearest neighbor, and logistic regression models to diagnose liver disease in human body are explored. The results are cross validated by tenfold cross-validation methods and achieve an accuracy as good as 93%.

KEYWORDS

Classification, Feature Selection, Liver Disease, Machine Learning, Prediction

1. INTRODUCTION

Liver is the largest organ of the body. Liver helps the body digesting the food, it store energy and removes poisons (Imenda, 2015). The failure of the liver occurs when the major parts of the liver are damaged and liver stop functioning. This condition is life threatening and it requires urgent medical care (Cressman et al., 1996). Although the condition of the liver occurs gradually due to different reasons, a rarer liver condition is called acute liver failure that is very hard to detect on the initial stage (Alonso & Squires, 2017). There are many symptoms of liver failure and liver diseases which can vary although the most common symptoms are legs swelling, easy bruising, change in stool color, change in urine color, and yellowing eyes (Mochida et al., 2018). Although these are the most common symptoms of the liver diseases, there are some cases when there are no symptoms and it is not possible to diagnose the liver disease. Among the other liver diseases, Liver cancer is the second most common cause of cancer death worldwide, preceded only by lung cancer. There were seven hundred thousand deaths due to liver cancer in 2012 (Torre et al., 2015).
The major reason of widespread of the liver infection is because of the deskbound lifestyle, alcoholic consumption and smoking. Due to these things, the liver infections is continuously increasing at a very rapid rate. The types of liver diseases including diseases caused by viruses, such as hepatitis A, hepatitis B, and hepatitis C. Diseases caused by drugs, poisons, or too much alcohol (Cheng et al., 2016). Examples include fatty liver disease and cirrhosis. Inherited diseases, such as hemochromatosis and Wilson disease. Liver function test of LFT are used to test the proper function of the liver. These tests are based on a number of steps that normally identify the reason of liver disease and try to treat the liver infection (Yap & Aw, 2010).

In the recent years, due to the availability of the mobile and other devices, the liver infection is easily detected. These smart devices are able to diagnose the liver diseases very easily through the help of different intelligent models and are able to perform much better. The performance of these devices are accelerated by using different classification systems that distinguish between one type and another. These classifier systems are trained by using machine learning models and are able to perform much better if proper machine learning models are incorporated. As the current age of artificial intelligence make it easy to explore the large amount of data, the integration of the previous data into such devices helps in identifying the liver diseases very easily.

Features selection techniques are usually divided into three categories including filter based techniques, wrapper based techniques and embedded methods of feature selection. While the classification algorithms are supervised, semi supervised or unsupervised. These algorithms are based upon different learning mechanism. In supervised machine learning methods, the labels are defined as the results are known. These methods help in better learning as in this method, the dataset is divided into two parts, for training and testing. In the unsupervised learning models, the label patterns are hidden and there are no available labels. These are more complex data learning models with much complex cases. Many researchers use machine learning techniques to solve other problems (Ayesha, Noor, Ramzan, Khan, & Shoaib, 2017; Khan & Daud, 2017; “Using machine learning techniques for subjectivity analysis based on lexical and non-lexical features,” n.d.) by using feature selection and other methods (Ishfaq, Khan, & Iqbal, 2016) (Khan & Daud, 2017). These tools are also helpful in classification (Khan, 2017) and information retrieval (A. Mahmood, Khan, Zahoor-ur-Rehman, & Khan, 2017; Ahsan Mahmood, Khan, Alarfaj, Ramzan, & Ilyas, 2018).

There are many kinds of liver disease. There are many symptoms of liver diseases and the number of patients of liver disease are continuously increasing due to high consumption of alcohol, harmful gases, and contaminated foods and drugs. In order to properly identify and diagnose the liver disease, real-world dataset is required. In this paper, we used a number of machine learning models incorporated with different feature selection techniques to predict the liver disease based upon real-world dataset. The results show our approach is feasible and that we are able to produce better results in order to predict the liver disease in a patient.

The rest of the paper is divided as follows: in the next section related work is defined followed by the section of research methodology in which we defined the research methodology in detail. In the next section, performance evaluation techniques are discussed followed by the result and discussion section before concluding the paper in the next section.

2. RELATED WORK

Many researchers have used machine learning approaches to predict the liver diseases by using the past data and helps finding the types of disease, causes, and its effects. for this purpose, scientists used machine learning techniques including supervised and unsupervised methods to predict and diagnose the disease in biomedicine (Sajda, 2006). Scientists perform a critical study of a number of classification algorithms for the diagnosis of liver disease in this regard, Ramana BV et al (Ramana, Babu, & Venkateswarlu, 2011), uses different classification algorithms to diagnose the liver disease. they evaluated the performance of their system by using different standard evaluation
measure techniques. Scientists also used digital image processing techniques to classify the liver diseases by using the CT images and support vector machines to correctly predict the liver diseases (Mala, Sadasivam, & Alagappan, 2015). Similarly, Artificial intelligence was also incorporated by the scientists to diagnose the liver disease by proposing an intelligent model which was able to predict the liver disease with good accuracy (Bae et al., 2018).

Kamath PS et al (Kamath et al., 2001), proposed a model to predict if the patient is able to survive with the end-stage liver diseases, the model validly is tested by using different datasets and the overall accuracy of the model was between 78 to 87%. Later, another research was proposed by A Cucchetti et al (Cucchetti et al., 2007), proposes that artificial neural network model works better than the MELD model in predicting the morality of the patients with end-stage liver disease. according to this model, they claim to produce an accuracy upto 96% which is far better than MELD Model. while the scientists proposed methods to predict and diagnose the liver diseases, a hybrid model was proposed to diagnose and determine the types of liver disease. this model was based upon artificial neural networks, analytic hierarchy process and case-based reasoning methods to examine the type of liver disease the patient is suffering from (“A hybrid diagnosis model for determining the types of the liver disease,” 2010). Although the scientists have used artificial intelligence to tackle, predict, diagnose and prevent different sort of diseases, there are only a few models that focuses on liver diseases.

3. RESEARCH METHODOLOGY

The research methodology is based upon a framework that is presented in the Figure 1. The framework consists of mainly 3 parts. In the first paper, dataset is collected from the web source and along with pre-processing and feature selection from the dataset. In the second part, data is divided into two parts for training and testing and different classification algorithms are performed on the training and testing parts of dataset. In the third part, different performance evaluation measure techniques are applied to measure the accuracy of the results along with the cross validation of the results.

The dataset used in this paper is taken from Kaggle.com. This dataset is sample dataset of the people all around the India. The dataset is based upon 550+ instances that are based on 10 different parameters. The number of attributes present in this dataset are 10 while the decisive parameter is described in Boolean value that is either true or false. The detailed description of the dataset is shown in Table 1.

In the first phase, pre-processing is applied on the dataset to normalize the values of dataset. Missing and incomplete values are also corrected in the dataset. This method is carried out to perform the task of feature selection. Initially, the attributes with more than 70% correlation is removed from the dataset while feature selection method was performed by using Rapidminer.
The dataset is then divided into two parts for testing and training. 70% dataset is used for training purpose while 30% dataset is used for testing purposes.

### 3.1. Classification Algorithms

In order to classify the data, a number of classification algorithms are used. After applying the classification algorithms tenfold cross validation is performed for each of the algorithms and results are computed. In this regard, first, naïve Bayes classification algorithms is applied on the dataset. While the naïve Bayes classification performs better in classification, it even performs much better in uncertain dataset. Naïve Bayes is one of the most commonly and most successful classification algorithm (Ren et al., 2009). Second, Decision tree classification is applied on the dataset to classify the dataset the decision tree classification method cover both classification and regression (Xu, Watanachaturaporn, Varshney, & Arora, 2005). In some cases, linear regression also performs well in order to predict on uncertain data (Rathi & Acharjya, 2018a, 2018b). A decision tree is the simplest way of solving the problems by creating a tree and take decision on the basis of its output. Third, K near neighbor or (KNN) is applied on the dataset. KNN tends to product good results in such problems e.g., distance functions. KNN has been used in statistical estimation and pattern recognition already in the beginning of 1970’s as a non-parametric technique (Chen, Pau, & Wang, 1973). Neural network model also performs pretty well in handling accumulated data finding the uncertainties in the data (Anitha & Acharjya, 2015, 2018). Fourth, the logistic regression method is applied for classification. Logistic regression method is binary regression method works better in such environments it is based upon qualitative response/discrete choice model (Harrell, 2015). Researchers also used clustering algorithms for analysis of uncertainty which performs much better (Chowdhary & Acharjya, 2017). Fuzzy models are also interesting while considering the medical data because of their flexibility and perform better as suggested by different research works (Chowdhary & Acharjya, 2016; Das & Acharjya, 2014).
4. PERFORMANCE MEASURE

In order to measure the performance of the system, standard performance measure techniques are used. These performance measure techniques are the standard way of obtaining the performance of the system (Sokolova & Lapalme, 2009). The possible confusion matrix in this case is given in Table 2.

The performance measure techniques used in this paper are given in eq(1), eq(2), eq(3), and eq(4). The precision determines how correctly the model identifies correct values of the class while the recall shows how many correct values have been identified over the total number of relevant instances. Similarly, the specificity or selectivity shows the true negative rate of the overall findings:

\[
\text{Precision} = \frac{tp}{tp + fp} \tag{1}
\]

\[
\text{Recall} = \frac{tp}{tp + fn} \tag{2}
\]

\[
\text{Specificity} = \frac{tn}{tn + fp} \tag{3}
\]

\[
\text{Accuracy} = \frac{tp + tn}{tp + fp + tn + fn} \tag{4}
\]

5. RESULTS AND DISCUSSIONS

This section discusses the results of the experimental setup. The detailed statistics of the dataset are given in Table 3.

Dataset having different charts and graphs with respect to their Attributes. Table 4 shows the distribution of dataset among male and females. The Pie chart describing Group by gender there are 441 are male and 142 are Females. Figure 2 shows the data distribution.

Figure 3 shows distribution of dataset with respect to male and female in detail. In this graph the Blue color shows tot bilirubin Test result, green is for direct bilirubin and red is for total Proteins. At x-axis gender plot.
As overfitting is one of the most common problem with machine learning, tenfold cross validation is performed on the dataset to manage the errors. As overfitting results in bias results, the tenfold cross validation helps in solving the issue. When the data is portioned into two parts, i.e., training and testing part, the training part is used to compute the results that are given in Table 4. According to these results, the highest accuracy is achieved by decision tree as decision tree gains 72.5% overall accuracy. While the other methods are able to attain an accuracy of 72% and 71%. Similarly, the

Table 3. Detailed Dataset Statistics

<table>
<thead>
<tr>
<th>S. No</th>
<th>Attribute name</th>
<th>Missing</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Age</td>
<td>0</td>
<td>4</td>
<td>90</td>
</tr>
<tr>
<td>3.</td>
<td>Total Bilirubin</td>
<td>0</td>
<td>0.400</td>
<td>75</td>
</tr>
<tr>
<td>4.</td>
<td>Direct Bilirubin</td>
<td>0</td>
<td>0.100</td>
<td>19.700</td>
</tr>
<tr>
<td>5.</td>
<td>Alkphos Alkaline Phosphatase</td>
<td>0</td>
<td>0.300</td>
<td>2.800</td>
</tr>
<tr>
<td>6.</td>
<td>Sgpt_Alamine Aminotransferase</td>
<td>0</td>
<td>2.700</td>
<td>9.600</td>
</tr>
<tr>
<td>7.</td>
<td>Sgot_Aspartate Aminotransferase</td>
<td>0</td>
<td>0.900</td>
<td>5.500</td>
</tr>
<tr>
<td>8.</td>
<td>Total Protiens</td>
<td>0</td>
<td>63</td>
<td>2110</td>
</tr>
<tr>
<td>9.</td>
<td>Albumin</td>
<td>0</td>
<td>10</td>
<td>2000</td>
</tr>
<tr>
<td>10.</td>
<td>Albumin_and Globulin Ratio</td>
<td>0</td>
<td>10</td>
<td>4929</td>
</tr>
</tbody>
</table>

Figure 2. Distribution of Dataset

As overfitting is one of the most common problem with machine learning, tenfold cross validation is performed on the dataset to manage the errors. As overfitting results in bias results, the tenfold cross validation helps in solving the issue. When the data is portioned into two parts, i.e., training and testing part, the training part is used to compute the results that are given in Table 4. According to these results, the highest accuracy is achieved by decision tree as decision tree gains 72.5% overall accuracy. While the other methods are able to attain an accuracy of 72% and 71%. Similarly, the
Naïve Bayes algorithm doesn’t perform well in the case as its accuracy is 54%. The precision and recall of decision tree is much better in the form of 100% and 72% which is much better considering the complexity of the data.

Figure 4 shows the comparison between different techniques.

Ten-fold cross validation method is standard evaluation method used to evaluate the performance of machine learning models. In this method, data is first divided into ten parts then in the first phase first of the data is used for testing purpose while all the remaining nine parts are used for training.
purpose. In the second phase, second part is used for testing purpose and remaining nine parts are used for training purpose and so on. After all the ten folds are complete, the average accuracy is said to present the result of ten-fold cross validation.

The accuracy of the classification models is over 70%. The best model being Logistic Regression with 72.00% accuracy. Although Decision tree having 72.57% accuracy but there is misbalancing occur so, we calculate the Sensitivity and Specificity also. Accuracy is 92.29% when no of folds is five but when no of folds is 10 their accuracy is 56.43%.

In order to compute the practical accuracy of the model, cross validation is used. In this manner, cross validation operator is used. The cross validation operator has two sub processes including training sub-process and testing sub-process. These sub-processes are used for training and testing purposes. In the testing phase, the performance of the model is measured. In this phase the number of folds are set to five. The dataset is divided into k subset of equal size. In each of the k subset, the values are computed while the process continuous with k-1 subsets and the cross validation process is then completed this way. The value of k is adjusted in different parameters of tenfold.

As the performance of the model is independent of the test sets, the results yield better output because the data was unseen. It also helps identifying the overfitting of the data. As the model presents the testing data in a better manner, the performance can be less than the actual performance on the training data.

Different Criterion we use for our decision tree includes Gini_index, Accuracy and Least_square. After Applying Pruning and Pre-Pruning Accuracy we got 71.24% and Confidence is 0.25. in machine learning pruning is used to reduce the size of decision tree by removing section of the tree with less importance and usage. This helps reducing the memory requirement and faster process. With pruning and pre-pruning, we got 67.98% accuracy. For KNN classifier the number of folds is five. Accuracy will be 65.09% when no of folds is five but when no. of folds is 10 then their accuracy is 71.41%.

Therefore, above algorithms there is Naïve Bayes Algorithm is best classifier having Accuracy of 92.29%.
6. CONCLUSION

In this paper, different machine learning models are used to perform feature selection and classify the patients of liver disease. The classification models are carefully chosen on the basis of overall performance and are evaluated by using training and testing carefully. The testing part is backed by cross validation by tenfold and other cross validation methods are applied on the dataset to compute the practical performance of the models. According to the results, the decision tree and logistic regression models came as the best classification models as these models gain 72% accuracy while in the cross validation the decision tree gains 71% accuracy. Similarly, although the Naïve Bayes model doesn’t attain good accuracy in training but it gains 92% accuracy in the cross validation phase.

In the future, we aim to use the model on bigger dataset with real-world data available from patients all around the world so the accuracy can be checked from different regions. We also aim to introduce the nonlinear SVM kernel and other classification models to enhance the accuracy of the model. While the number of features available in this dataset are enough to predict the outcome, it is of great importance to know the background data of the patient in order to predict the liver disease as that will help in improving the accuracy of the model.

Table 5. Cross Validation Results

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNN</td>
<td>65.06%</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>92.29%</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>71.24%</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>56.43%</td>
</tr>
</tbody>
</table>
REFERENCES


ENDNOTE

Mohammed Alghobiri is an experienced and highly involved in information systems development and Implementation, especially experimental and participative approaches. Part of his interest is Management and Evaluation of Systems Development, including Software Process Improvement Methods and ERP D&I. Data Base, Data Mining, Decision Support Systems and Electronic Government Concepts are also within his concern. He has very good experience in team leading and IT centres management.

Hikmat Ullah Khan received the master’s degree in computer science and the Ph.D. degree in computer science from International Islamic University, Islamabad. He has been an Active Researcher for the last ten years. He is currently an Assistant Professor with the Department of Computer Science, COMSATS Institute of Information Technology, Wah Cantt, Pakistan. He has authored a number of research articles in top peer-reviewed journals and international conferences. His research interests include Social web mining, Semantic Web, data science, information retrieval, and scientometrics. He is a member of the Editorial board of a number of prestigious Impact Factor Journals.

Ahsan Mahmood received the master’s degree in computer science from the COMSATS University, Attock campus, Pakistan. His research interests include Data Mining, Social Media Analysis, Sentiment Analysis and Machine Learning.