International Journal of Computer Trends and Technology
ISSN: 2231-2803 / https://doi.org/10.14445/22312803/IJCTT-V68I8P107

Volume 68 Issue 8, 44-52, August 2020

Original Article

Prediction and Diagnosis of Liver Disease in Human Using Machine Learning

Adekola Olubukola Daniel[1] Ekanem Edikan Uwem[2], Omidiran Daniel Tolulope[3], Owoade Samuel Jesupelumi[4]

[1,2,3,4]*Computer Science Department, School of Computing and Engineering Sciences, Babcock University, Ilishan Remo, Ogun State, Nigeria.*

Received Date: 03 August 2020
Revised Date: 14 August 2020
Accepted Date: 31 August 2020

Abstract *- Disease diagnosis is the most vital task in medicine, and this mostly depends on the doctor's intuition based on experiences in the past. Unfortunately, the difficulties in recognizing correct symptoms result in a misdiagnosis. To avoid such medical misdiagnosis, this study utilized a dataset to intelligently detect liver disease in humans. This study aimed to implement an effective data mining method and algorithm to predict and diagnose the occurrence of liver diseases in humans in order to eliminate the use of manual methods of analysis relating to liver diseases. The study embodies case studies, systematic literature reviews and surveys. Important requirements were also identified in related papers. The relevant documents obtained were qualitatively analyzed for convergence, and relevant details were extracted using an inductive approach. Subsequently, a liver disease diagnosis system (LDDS) was developed to tackle the problem of early detection of the disease in humans. LDDS is a web application created to ease the prediction of the occurrence of liver disease in humans.*

Keywords *- Misdiagnosis; Dataset; Machine-Learning algorithms; Data mining; Liver Disease Diagnosis System*

I. INTRODUCTION

Diagnosis of a disease is based on a doctor's knowledge and experience. However, under some circumstances, the prediction can be wrong, which leads to incorrect treatment to the patient. Hence to confirm if the diagnosis is right, it would be useful if there could be a tool that uses data from several patients to diagnose the disease and its probability of occurrence. In this research, liver disease is automated for diagnosis. The dataset is pre-processed, and patterns of disease are analyzed using several machine-learning algorithms. A User Interface (UI) is designed to collect the symptoms of patients that are to be diagnosed. The machine learning algorithm is selected to diagnose the symptoms of the disease entered by the user in the UI, and the result of the diagnosis is displayed to the user.

The liver is one of the biggest and strongest organs in the human body. The location of the liver in the upper right portion of the abdomen, surrounded by the rib cage. The gallbladder, along with parts of the pancreas and intestines, sits below the liver. The liver performs many complex tasks in the body. The liver is a large solid organ that sits on the right side of the belly, weighing about 3 pounds (the largest gland in the human body). The two large sections that the liver has is the right and the left lobes. The liver works together with various organs to digest, absorb, and process food. The liver also makes proteins important for blood clotting and other functions [1].

Liver disease is a large term that covers all the potential problems that cause the liver to fail or to perform its designated functions. Liver disease can also be referred to as Hepatic disease. Usually, more than 75% or three-quarters of liver tissue needs to be affected before a decrease in function occurs [1].

[1] attempted to detect heart disease using Coronary Heart Disease Dataset from Cleveland Clinical Foundation in two ways. The first approach is using a single data-mining algorithm for the heart disease dataset and benchmark the baseline accuracy. Further, in the second approach, they tried using hybrid algorithms such as the J4.8 decision tree and bagging algorithms. It was proposed that the application of voting two techniques, different data discretization levels and reduced error, can increase the accuracy of data mining algorithms. They achieved a maximum of 84.1% accuracy using their techniques and claimed that further research is in progress. Also, in [2], the heart disease dataset from the University of California Irvine (UCI) machine learning repository was analyzed. The authors used an ensemble learning method – adaptive boosting algorithm for classification on the Hungarian Institute of Cardiology (HIC) dataset, and the highest accuracy obtained after various experiments done with an ensemble of classifiers is 96%. The researchers in [3] used real-world datasets such as the UCI repository and SMBBMU (Shaheed Mohtarma Benazir Bhutto Medical University) for diagnosing various thyroid diseases such as Euthyroid, Hypothyroid, Hyperthyroid, Sub-clinical-hypothyroid and subclinical-Hyperthyroid. Two classifiers, such as multi and binary SVM, are used. The accuracy of the system is 95.7% with 10-k fold cross-validation. Apart from heart disease and thyroid disease, there has been working on other diseases such as diabetes, breast cancer, dengue, hepatitis, liver disorders, and more. To identify liver diseases, in [4], SVM and Naïve Bayes algorithms are used individually, and the highest accuracy of 79% is

obtained. In [5], the same classification is performed using various algorithms such as J48, MLP, SVM, Bayesian network and random forest classifiers individually. The highest accuracy obtained is 71% for the same. For the detection of diabetes in [6], the authors use the decision tree and Naïve Bayes algorithm individually and reach a maximum of 79.68% accuracy. In [7], the CART algorithm, adaboosting and logiboosting techniques are used, and the accuracy result is only 3 78.65% maximum. However, in [8], the authors use a hybrid-decision tree algorithm with Naïve Bayes algorithm for heart disease prediction and achieve 95% accuracy. In the same work, the hybrid algorithm is tested for different datasets such as contact lens, iris plant, breast cancer, soybean, glass, image segmentation, and tic-tac-toe win. The results of hybrid algorithms have better performance than using an individual decision tree or Naïve Bayes algorithm. When comparing work in [8] with other related research, the usage of the hybrid algorithm has better accuracy than the usage of individual algorithms. From the literature, we know that the Naïve Bayes algorithm is computationally expensive if used with many attributes. Hence the fewer the attributes, the better is its performance. However, the attributes to be considered for Naïve Bayes could be a question. As an answer to this question, a hybrid algorithm of the Naïve Bayes algorithm is used with a decision tree in [8]. The top K-most important attributes involved in classification for a decision tree can be used for the Naïve Bayes algorithm, thereby reducing the cost of computation.

II. REVIEW OF CLOSELY RELATED WORKS

- [9] proposed data classification based on liver disorder. The training dataset is developed by collecting data from the UCI repository consisting of 345 instances with seven different attributes. This paper deals with results in the field of data classification obtained with Naïve Bayes algorithms, FT tree algorithms and K-Star algorithms. Overall performance made known, FT Tree algorithm when tested on liver disease datasets, time taken to run the data for the result is fast when compared to another algorithm with an accuracy of 97.10%. Based on the experimental results, the classification accuracy is found to be better using the FT Tree algorithm compared to other algorithms.
- [10] worked on identifying liver disease in patients based on the 10 important attributes of liver disease using a Decision Tree, Naive Bayes, and NB Tree algorithms. The result shows NB Tree algorithm has the highest accuracy; however, the Naïve Bayes algorithm gives the fastest computation time. For future studies, the performance of the NB Tree algorithm would be the target of accuracy improvement by finding the most significant factor in identifying liver disease in patients.
- [11] stated that there are many liver disorders that require the clinical care of the physician. The study predicts three major liver diseases such as liver cancer, cirrhosis, hepatitis with the help of distinct symptoms. The primary goal is to predict the class types from classes such as liver cancer, cirrhosis, hepatitis and “no diseases”. This study compares the accuracy of Naïve Bayes and the FT tree algorithm, and the result concludes that the accuracy of the Naïve Bayes algorithm is much better.
- [12] analyzed liver diseases data using particle swarm optimization algorithm (PSO) with K Star classification. The proposed algorithm enhanced the performance of accuracy when compared to existing classification algorithms. The PSO-Kstar algorithm is the best suitable algorithm for the classification of liver disorders as it improved the performance in prediction accuracy, as earlier mentioned. The PSO-KStar algorithm is considered one of the good data mining algorithms with respect to understandability, transformability and accuracy, giving 100%.
- The work of [4] is to predict liver diseases using classification algorithms. The algorithms used in this work are Naïve Bayes and support vector machine (SVM). Comparisons of these algorithms are made, and it is based on the performance factors classification accuracy and execution time. From the results, this work concludes the SVM classifier is considered the best classification algorithm because of its highest classification accuracy values. On the other hand, while comparing the execution time, the Naïve Bayes classifier needs minimum execution time from the implementation results. It is observed that the SVM is a better Classifier for predicting liver diseases and comparing the execution time; the Naïve Bayes classifier needs minimum execution time.

A. Types of Liver Infections

a) Hepatitis

Hepatitis refers to an inflammatory condition of the liver. It is commonly caused by a viral infection, but there are other possible causes of hepatitis. These include autoimmune hepatitis and hepatitis that occurs as a secondary result of medications, drugs, toxins, and alcohol.

b) Cirrhosis

Cirrhosis, also known as liver cirrhosis or hepatic cirrhosis, is a condition in which the liver does not function properly due to long-term damage. This damage is characterized by the replacement of normal liver tissue by scar tissue. Typically, the disease develops slowly over months or years. Early on, there are often no symptoms. As the disease worsens, a person may become tired, weak, itchy, have swelling in the lower legs, develop yellow skin, bruise easily, have fluid built up in the abdomen, or develop spider-like blood vessels on the skin. The fluid build-up in the abdomen may become spontaneously infected. Other serious complications include hepatic encephalopathy, bleeding from dilated veins in the oesophagus or dilated stomach veins, and liver cancer. Hepatic encephalopathy results in confusion and may lead to unconsciousness.

a) Liver Cancer: Liver cancer, also known as hepatic cancer and primary hepatic cancer, is cancer that starts

in the liver. Other causes include aflatoxin, non-alcoholic fatty liver disease and liver flukes.

b) Ascites: Ascites is the abnormal build-up of fluid in the abdomen. Technically, it is more than 25 ml of fluid in the peritoneal cavity. Symptoms may include increased abdominal size, increased weight, abdominal discomfort, and shortness of breath. Complications can include spontaneous bacterial peritonitis.

c) Gallstones: Gallstones are hardened deposits of digestive fluid that can form in the gallbladder. The gallbladder is a small, pear-shaped organ on the right side of your abdomen, just beneath your liver. The gallbladder holds a digestive fluid called bile that is released into the small intestine.

d) Hemochromatosis: An inherited disorder characterized by abnormally high absorption of iron by the intestinal tract, resulting in excessive storage of iron in the body.

e) PSC: Primary Sclerosing Cholangitis (PSC) is a chronic liver disease characterized by a progressive course of cholestasis with inflammation and fibrosis of the intrahepatic and extrahepatic bile ducts. The underlying cause of the inflammation is believed to be autoimmunity.

f) PBC: Primary biliary cholangitis (PBC), previously known as primary biliary cirrhosis, is an autoimmune disease of the liver. It results from slow, progressive destruction of the small bile ducts of the liver, causing bile and other toxins to build up in the liver, a condition called cholestasis.

B. Liver Tests

The following blood tests are performed for diagnosis of liver disease:

a) Liver function panels: This panel consists of many different blood tests and checks on how properly the liver is functioning.

b) ALT (Alanine Aminotransferase): The higher level of ALT assists in recognition of diseases of the liver such as hepatitis.

c) AST (Aspartate Aminotransferase): In addition, with increased ALT, Aspartate Aminotransferase AST also analyses the causes of liver destruction.

d) Alkaline phosphates: In bones as well as bile secreting cells in the liver, alkaline phosphates are found. High level often means bile flow out of the liver is blocked.

e) Bilirubin: a difficulty with the liver is occurred by elevated bilirubin quantity.

f) Albumin: Albumin analyses how perfect the liver is functioning, as it is a component of protein.

g) Ammonia: when the liver is not working accurately or perfectly, the level of ammonia is increased.

h) Hepatitis A test: The doctors will test the working of the liver in addition to antibodies to identify the hepatitis A virus if hepatitis A is diagnosed.

i) Hepatitis B test: If hepatitis B is suspected, doctors will determine the level of antibodies to analyze whether the person is suffering from hepatitis B virus or not.

j) Hepatitis C test: Its blood test analyses if a person has been suffering from hepatitis C and examines the working of the liver.

k) Prothrombin Time (PT): This test is accomplished to check if someone is taking the adequate dosage of medicine of blood-thinning warfarin. It can check problems of blood clotting.

l) Partial Thromboplastin Time (PTT): This test is done to determine blood-clotting trouble.

In numerous automatic medical diagnoses, classification techniques are generally used. At the early stage, problems with liver patients are not simply revealed. Even if the liver is partially damaged, it will perform all its functions normally [13].

Knowledge is the most important strength of any organization in an Information Technology (IT) driven society. People are expecting better and reasonable healthcare due to the ongoing development in IT applications in the area of healthcare. Valuable medical information's are now easily accessible due to computerized hospital information systems (HIS). The use of sophisticated equipment in the practice of modern medicine produces a huge amount of data. After storing this data in digital form, to extract knowledge by automated methods of data analysis, a significant amount of effort is being made for this purpose. Knowledge can be used for well speedy clinical decision-making. The area that consists of data mining and knowledge discovery tools are very helpful in achieving these goals. Including data mining, knowledge discovery is a distinct process that consists of different steps. Usually, the healthcare environment is very informative but lacks knowledge. However, for better knowledge in the healthcare environment, data mining techniques can be applied for this purpose. As a large amount of medical data is being produced, it is need of the hour to improve methods of data analysis and knowledge discovery using appropriate data mining techniques. A convenient resource for data mining and knowledge discovery provides a proper medical database that was created with intention mining. Knowledge discovery in databases (KDD) and data mining are the two terms that are interchangeable. The process of finding useful information and extracting knowledge is called KDD and data mining, respectively [14]. Learning from the learned knowledge is roughly called Meta-learning. It is a modern technology used to compute high-level models. In principled fashion, Meta classifiers are integrated. Separately learned classifier cleans the information to enhance the predictive performance. In the process of Meta-learning, the number of the learning program is applied on the number of data subsets, parallel. Then in the form of a classifier, the combined outcome is gathered [7]. For detecting liver diseases, the recommendation is a computerized method that is based on previous data.

III. DISCUSSION

A. *Problem statement*

The diagnosis of a disease is the most critical and vital task in medicine, and this mostly depends on a doctor's intuition based on experiences in the past. **Unfortunately, the difficulties in recognizing correct symptoms result in a misdiagnosis.** To avoid such medical misdiagnoses, this study utilizes large datasets collected by healthcare industries to automate the diagnosis of diseases. The need to develop a tool that could aid doctors and prevent them from unwarranted errors and unwanted biases in diagnosis is established in this research. Therefore, an automated medical diagnosing system to tackle the problem of correct **early detection of liver disease** is considered as one of the outputs of this study.

B. *Aim and Objectives*

The main aim of this study is to implement an effective data mining method and algorithm to predict and diagnose the occurrence of liver diseases in humans in order to eliminate the use of manual methods of analysis relating to liver diseases.

The objectives of the study include:

1. To obtain data on signs and symptoms of liver diseases in humans and to transform the stated factors into a suitable form for system coding.

2. To model a neural network that can predict the occurrence of liver disease based on the data set provided.

3. To implement an automated medical diagnosing system for early detection of liver diseases.

C. *Significance*

The output from this study would mitigate incorrect prediction and diagnosis, which would save more lives. There would be less need for a patient referral form from one hospital to another, provided this tool is available.

D. *Methodology*

This study implements an effective liver disease diagnosis system using the Random Forest classification model. The algorithms act on the input taken. Data sets are accessed, and effective liver disease diagnosis is carried out.

a) Use Case Diagram

The use case diagrams represent the initial phase graphical model analyzing the interaction between the system and its external environment. It shows the summary of what the proposed system must do. The following unified modelling language (UML) use case diagrams describe the interaction between users and the developed system:

Table 1. Data set attributes (International Journal of Research In Advent Technology, 2019)

S/NO	ATTRIBUTES	ATTRIBUTE INFORMATION	ATTRIBUTE TYPE
1	AGE	AGE OF PATIENT	NUMERIC
2	GENDER	GENDER OF PATIENT	NOMINAL
3	TB	TOTAL BILIRUBIN	NUMERIC
4	DB	DIRECT BILIRUBIN	NUMERIC
5	ALKPHOS	ALKALINE PHOSPHATE	NUMERIC
6	SGPT	ALAMINE AMINOTRANSFERASE	NUMERIC
7	SGOT	ASPARTATE AMINOTRANSFERASE	NUMERIC
8	TP	TOTAL PROTEINS	NUMERIC
9	ALB	ALBUMIN	NUMERIC
10	A/G RATIO	ALBUMIN AND GLOBULIN RATIO	NUMERIC
11	SELECTOR FIELD	LABELLED BY EXPERTS	BINOMINAL

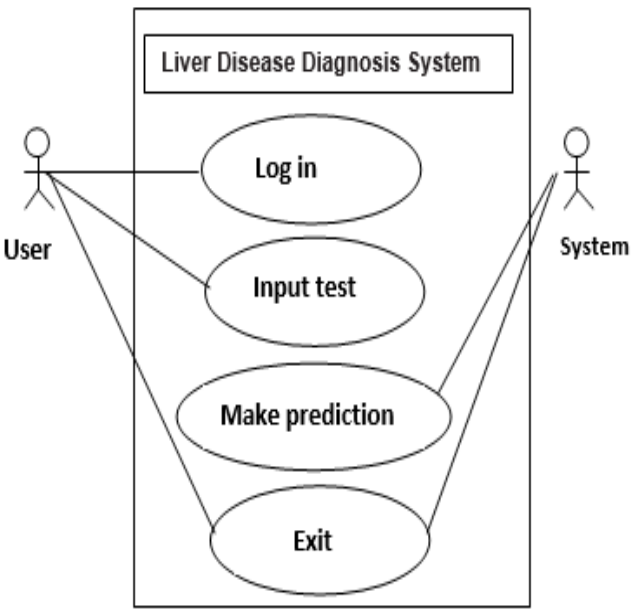


Fig.1 Liver disease diagnosis system use cases

b) Data set Description

Indian Liver patient data set is obtained from the UCI machine learning repository. This data set was gathered from Andhra Pradesh, India. The data set consists of 583 instances in which disease is present in 416 patients, and 167 are healthy patients. Data set contains 11 attributes that are age, Gender, TB (total Bilirubin), DB (direct Bilirubin), Alkphos (Alkaline Phosphatase), SGPT Alamine Aminotransferase, SGOT Aspartate Aminotransferase, TP (total proteins), ALB (albumin), A/G ratio (Albumin and Globulin Ratio) and Selector field. The selector is a class that divides the whole data set into two parts, "Liver patients or non-Liver Patients".

Table 1. Data Set Attributes

In this research, the system analysis stages are;

1. Train the model; the dataset
2. Test it against overfitting and underfitting
3. Implement chosen machine learning algorithm
4. Evaluate the model
5. Predict liver disease

c) Project Design

The project involves a frontend GUI that collects and displays data and a backend machine-learning model that analyses data from the frontend. The front is a GUI that contains three sets of screens. The first screen, also known as the welcome screen, allows the user to prompt to the diagnosis page. The second screen, which is the data collector screen, prompts the user to enter the test values of the selected disease. After submitting the values, the third screen pops up and displays the diagnosis made by the machine-learning model, the accuracy of the machine learning model used and the prediction text. The model that generalizes well by performing classification with the highest accuracy is selected to diagnose the disease symptoms entered by the user. Upon classification, the diagnosis, probability of diagnosis, and the accuracy of the model are passed to the frontend to display the results to the user. The block diagram below gives a diagrammatic representation of the project design.

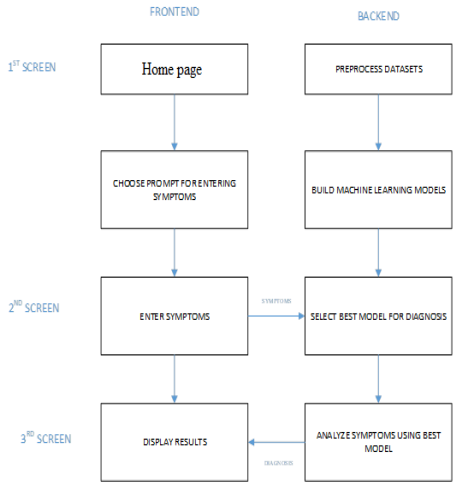


Fig. 2 Block diagram representing project design

IV. SYSTEM DESIGN VIEW

When a user opens the webpage, the user should be able to view a welcome page, which is just a landing page where the user will click on a button that proceeds to the diagnosis page. After diagnosis, there is a landing page where the result is displayed.

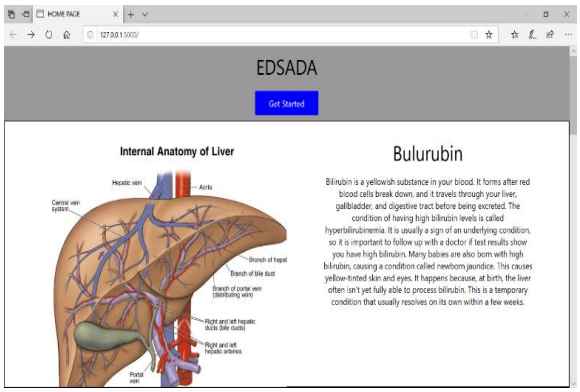


Fig. 3 Home page I

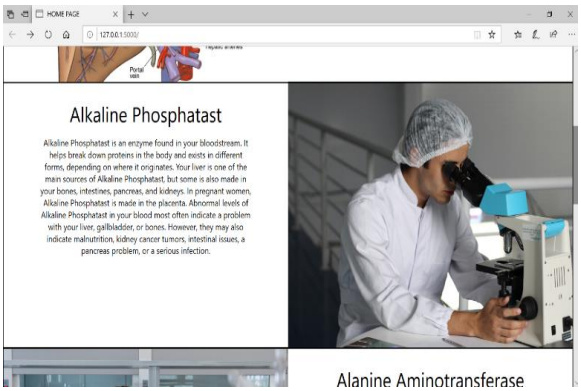


Fig. 4 Home page II

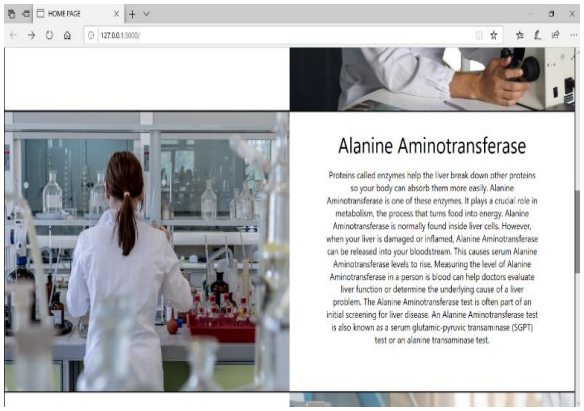


Fig. 5 Home page III

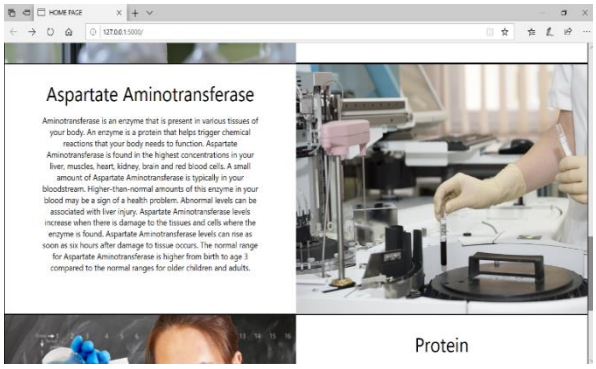


Fig. 6 Home page IV

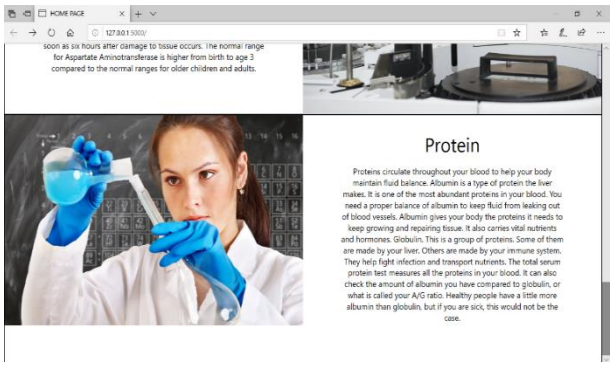


Fig. 7 Home page V

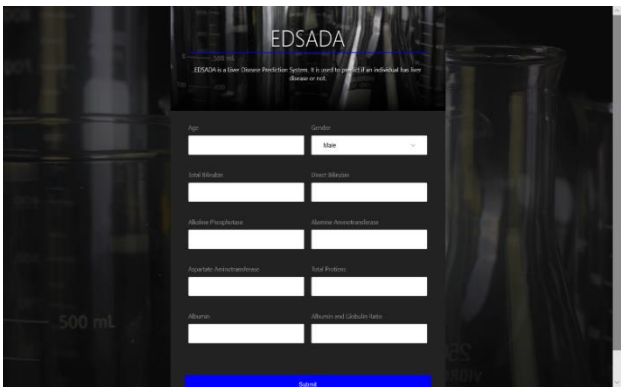


Fig. 8 Prediction page

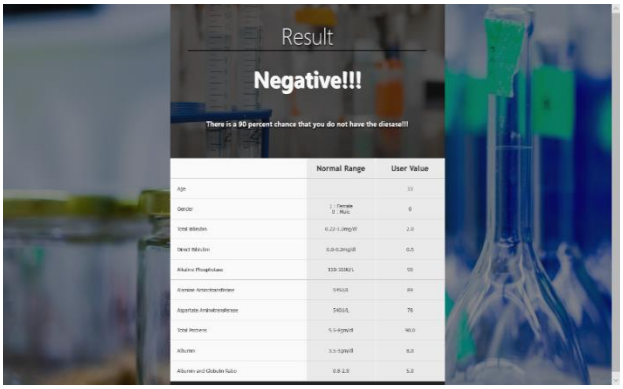


Fig. 9 Result page

V. DATA FORMATTING AND ANALYSIS

The data used for the project was obtained from UCI and Kaggle repositories. Analysis was required to be carried out on the data so as to be familiar with the dataset, to choose the important features to input into the model for prediction and also to make sure the data is well-formatted for use.

a) Data Analysis

For the analysis of data, the following libraries were imported and used to carry out statistical and analytical operations:

- Pandas
- NumPy
- Seaborn
- Matplotlib

```
In [1]: #import libraries

        import pandas as pd
        import numpy as np
        import seaborn as sns
        import matplotlib.pyplot as plt


        %matplotlib inline
```

Fig. 10 Import libraries for analysis

The data set which was gotten from Kaggle and UCI repositories was then imported and was viewed.

```
In [2]: #import dataset

        # data = pd.read_csv('Indian Liver Patient Dataset (ILPD).csv')
        data= pd.read_csv('indian_liver_patient.csv')
```

```
In [3]: #perform basic analysis on the dataset to be familiar with the dataset

        data.head()
```

Out[3]:

	Age	Gender	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alanine_Aminotransferase	Aspartate_Aminotransferase	Total_Protiens	Albumin	Al
0	65	Female	0.7	0.1	187	16	18	6.8	3.3	
1	62	Male	10.9	5.5	699	64	100	7.5	3.2	
2	62	Male	7.3	4.1	490	60	68	7.0	3.3	
3	58	Male	1.0	0.4	182	14	20	6.8	3.4	
4	72	Male	3.9	2.0	195	27	59	7.3	2.4	

Fig. 11 Import dataset for analysis

After importing and viewing the dataset, the dataset was examined to be sure it was in the correct format suitable for the experiment. It was observed noticed that the dataset consisted of integers, floating numbers and objects (Strings). But for having to work with mathematical models, the object data would have to be transformed to numbers (integers) so as to be able to perform both statistical analyses and use the values in performing the prediction.

```
In [10]: M data.dtypes
Out[10]: Age                         int64
         Gender                     object
         Total_Bilirubin           float64
         Direct_Bilirubin          float64
         Alkaline_Phosphotase        int64
         Alamine_Aminotransferase    int64
         Aspartate_Aminotransferase  int64
         Total_Proteins            float64
         Albumin                   float64
         Albumin_and_Globulin_Ratio float64
         Dataset                     int64
         dtype: object
```

Fig. 12 Checking dataset for correct format

Statistical analysis was then carried out to better understand the dataset. Some analysis carried out includes checking the maximum and minimum values of each column checking the mean, standard deviation and quartile ranges.

```
In [7]: M data.describe()
Out[7]:
```

	Age	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Proteins	Albumin
count	583.000000	583.000000	583.000000	583.000000	583.000000	583.000000	583.000000	583.000000
mean	44.746141	3.298799	1.486106	290.576329	80.713551	109.910806	6.483190	3.141852
std	16.189833	6.209522	2.808498	242.937989	182.620356	288.918529	1.085451	0.795519
min	4.000000	0.400000	0.100000	63.000000	10.000000	10.000000	2.700000	0.900000
25%	33.000000	0.800000	0.200000	175.500000	23.000000	25.000000	5.800000	2.600000
50%	45.000000	1.000000	0.300000	208.000000	35.000000	42.000000	6.600000	3.100000
75%	58.000000	2.600000	1.300000	298.000000	60.500000	87.000000	7.200000	3.800000
max	90.000000	75.000000	19.700000	2110.000000	2000.000000	4929.000000	9.600000	5.500000

Fig. 13 Checking the max and min of data

After performing basic statistical analysis on the dataset, the correlation among each column of the dataset was examined.

```
In [16]: M plt.figure(figsize=(12,8))
         sns.heatmap(data.corr(), annot=True)
Out[16]: <matplotlib.axes._subplots.AxesSubplot at 0x1317fcb8>
```

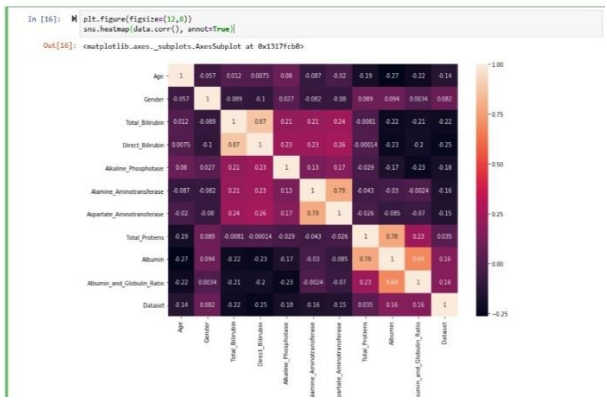


Fig. 14 Checking correlation of data

b) Analysis of Categorical Data

Analysis was made on Gender being the only categorical data. Count means the number of rows in the column, unique means unique values in the column, which is 2 because there are only Male and Females, top means the unique value with the highest frequency, which is Male, Freq means how many times Male appeared.

```
In [8]: M #make analysis of categorial data
        data.describe(include=['object'])
Out[8]:
```

	Gender
count	583
unique	2
top	Male
freq	441

Fig. 15. Analysis of gender in data

A graph that shows age against Gender using the frequency of the people who had the disease and people who did not.

```
In [11]: M import warnings
         warnings.filterwarnings("ignore")
         sns.factorplot (x="Age", y="Gender", hue="Dataset", data=data);
```

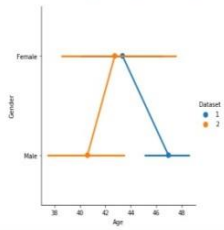


Fig. 16 Graph of age against gender in data

A graph that shows Gender against count to show the number of males and females in the dataset.

```
In [12]: M sns.countplot(data=data, x = 'Gender', label='Count')
         M, F = data['Gender'].value_counts()
         print('Number of patients that are male: ',M)
         print('Number of patients that are female: ',F)
         Number of patients that are male:  441
         Number of patients that are female:  142
```

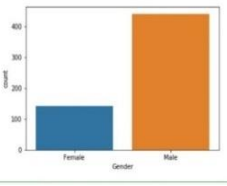
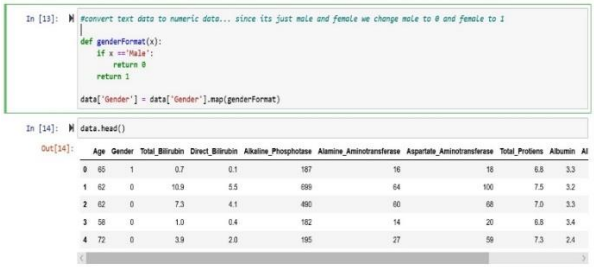


Fig. 17 Graph of gender against count in data

The categorical data (the column with the datatype float) was then changed to integers so that it could be inputted into the model. Since the data was for Gender, there are only have two categories (Male and female), then the data transformation was performed manually, that is, value Male was changed to Zeros (0's) and the values Female to Ones (1's).



```
In [13]: #convert text data to numeric data... since its just male and female we change male to 0 and female to 1

         def genderFormat(x):
             if x =='Male':
                 return 0
             return 1

         data['Gender'] = data['Gender'].map(genderFormat)

In [14]: data.head()

Out[14]:
```

	Age	Gender	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Protiens	Albumin	Al
0	65	1	0.7	0.1	187	16	18	6.8	3.3	
1	62	0	10.9	5.5	699	64	100	7.5	3.2	
2	62	0	7.3	4.1	490	60	68	7.0	3.3	
3	58	0	1.0	0.4	182	14	20	6.8	3.4	
4	72	0	3.9	2.0	195	27	59	7.3	2.4	

Fig. 18 Converting Text Data to Numeric Data

b) Model Selection, Implementation and Evaluation

After the analysis was performed on the dataset, the model used for the prediction system was selected, evaluation was carried out, and then the model was saved for post-production purposes.

```
In [22]: #start machine learning on the data

         from sklearn.model_selection import train_test_split

         #select your features and your target
         variables = ['Age', 'Gender', 'Total_Bilirubin', 'Direct_Bilirubin',
                      'Alkaline_Phosphotase', 'Alamine_Aminotransferase',
                      'Aspartate_Aminotransferase', 'Total_Protiens', 'Albumin',
                      'Albumin_and_Globulin_Ratio']

         X = data[variables]
         y = data.Dataset

In [23]: #split data into training and testing

         X_train, X_test, y_train, y_test = train_test_split(X, y,
                                                             shuffle=True, stratify=data.Dataset4t)

In [24]: # Print number of observations in X_train, X_test, y_train, and y_test

         print(X_train.shape, X_test.shape, y_train.shape, y_test.shape)

         (424, 10) (142, 10) (424,) (142,)
```

Fig. 19 Training and Testing the Model

```
In [30]: #implement chosen machine learning algorithm

         from sklearn.ensemble import RandomForestClassifier

         # model = RandomForestClassifier()
         # model = RandomForestClassifier(n_estimators=500, min_samples_split=2, min_samples_leaf=4)
         model = RandomForestClassifier(n_estimators=100, max_depth=None, min_samples_split=2,  min_samples_leaf=4)

         model.fit(X_train, y_train)

Out[30]: RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
                                max_depth=None, max_features='auto', max_leaf_nodes=None,
                                min_impurity_decrease=0.0, min_impurity_split=None,
                                min_samples_leaf=4, min_samples_split=2,
                                min_weight_fraction_leaf=0.0, n_estimators=100,
                                n_jobs=None, oob_score=False, random_state=None,
                                verbose=0, warm_start=False)
```

Fig. 20 Implementation of chosen algorithm

```
In [31]: #evaluate our model using confusion matrix

         from sklearn.metrics import confusion_matrix

In [32]: y_pred = model.predict(X_test)
         confusion_matrix(y_test, y_pred)

Out[32]: array([[86, 15],
                [26, 15]], dtype=int64)

In [33]: confusion_matrix(y_test, y_pred).T

Out[33]: array([[86, 26],
                [15, 15]], dtype=int64)

In [34]: #extracting true_positives, false_positives, true_negatives, false_negatives
         TN, FP, FN, TP = confusion_matrix(y_test, model.predict(X_test)).ravel()

         print('Testing Accuracy = "{} !"'.format((TP + TN) / (TP + TN + FN + FP)))

         Testing Accuracy = "0.7112676056338029 !"

In [35]: print('Random Forest Classifier Training Accuracy:', model.score(X_train, y_train))

         Random Forest Classifier Training Accuracy: 0.950471698113207S

In [37]: from sklearn.metrics import accuracy_score
         accuracy = accuracy_score(y_test, y_pred)
         accuracy

Out[37]: 0.7112676056338029
```

Fig. 21 Evaluation of the model

```
In [53]: from sklearn.externals import joblib
         joblib.dump(model, 'C:/Users/HP/Documents/Clean Start/Omidiran/model.pkl')
         print("Model dumped!")

         # Load the model that you just saved
         model = joblib.load('C:/Users/HP/Documents/Clean Start/Omidiran/model.pkl')

         Model dumped!
```

Fig. 22 Saving of the model

VI. LIMITATIONS

A supervised learning algorithm was used, and one has to manually retrain the model every time. There was no room for users of the system to input values with results to get new data set for the system. Because of the specifications of the system used in the implementation of the work, a deep learning and neural network approach could not be used.

VII. RECOMMENDATIONS

This system can be extended to include modification or enhancements such as:

1. Increasing the number of liver diseases covered.

2. Providing the system on other platforms.

3. A deep learning approach could be explored as a future direction.

VIII. CONCLUSION

This study implemented an effective algorithm to predict and diagnose the occurrence of liver diseases in humans, obtained information relating to the signs and symptoms that are associated with liver diseases in humans, transformed the stated factors and information into code and finally modelled a neural network that could predict the occurrence of liver disease based on the data set provided.

IX. REFERENCES

[1] Mai, S., Tim, T., & Rob, S., Using Data Mining Techniques in Heart Disease Diagnosis and Treatment. Northcott Drive: University of New South Wales at the Australian Defense Force Academy, (2012).

[2] Kathleen, M. H., Julia, M. H., & George, M. J., Diagnosing Coronary Heart Disease Using Essemble Machine Learning. International Journal of Adanced Computer Science and Application, (2016).

[3] Soomrani, R., Adul, M., & Jamil, A., Thyroid Disease Type Diagnostics. Sukkur: Sukker Institute of Bussiness Administeration, (2016).

[4] Vijayarami, S., & Dhayanaand, S., Liver Disease Prediction Using SVM and Naive Bayes Algorithms. International Journal of Sciences, Engineering and Technology Research, (2015).

[5] Anju, G., Rajan, V., & Praveen, R., "Liver Patient Classification Using Intelligent Techniques". International Journal of Computer Science and Information Technology, (2014).

[6] Aiswarya, L., Jeyalatha, S., & Ronak, S., Diagnosis of Diabetis Using Classification Mining Techniques. International Journal of Data Mining & Knowledge Management Process, (2015).

[7] Sen, S. K., & Dash, S., Application of Meta-Learning Algorithims for the prediction of Diabetes Disease. International Journal of Advanced Research in Computing, (2014).

[8] Dewan, F. M., Chowdhury, Mofizur, R., Hossain, M. A., Strachan, R., & Zhang, L., Hybrid Decision Tree and Naive Bayes classifiers for multi-class classification tasks. Expert Systems with Applications, (2014).

[9] PapeTyari, Retrieved from PaperTyari: www.papertyari.com/general-awareness/it-knowledge/bcnf-4nf-5nf/, (2014).

[10] L. Rothkrantz, Dynamic routing using maximal road capacity. Paper presented at the Proceedings of the 16th International Conference on Computer Systems and Technologies, (2015).

[11] Sadiyah, N., Novita A., & Teddy, M., Data mining Techniques for optimization of Liver Disease Classification. International Conference on Advanced Computer Science Applications, (2013).

[12] Dhamodharan, S. (2014). Liver Disease Prediction Using Bayesian Classification. COMPUSOFT, (2014).

[13] Thangarajul, P., & Mehala, R. Analysis of PSO-KStar Classifier over Liver Disease. International Journal of Advanced Research in Computer Engineering, (2015).

[14] Jankisharan, P., Rajan, V., Jagdish, M., & Sanjay, P. Liver Patient Classification using Intelligence Techniques. International Journal of Advanced Research in Computer Science and Software Engineering, (2014).

[15] Wasan, S. K., Bhatnagar, V., & Kaur, H. The Impact of Data Mining Techniques on Medical Diagnostics. Data Science Journal, (2006).