ZKG INTERNATIONAL

ISSN: 2366-1313

LIVER DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS

[1]Dr.M. DHASARATHAM, [2]VITTEDI VARSHINI, [3]SOMISHETTY SAIKRUPA, [4]AMBALA PAVAN

[1]Professor, Dept.of IT, TKR College of Engineering & Technology, Meerpet, Hyderabad, dasarath.m@gmail.com

[2]BTech student, Dept.of IT, TKR College of Engineering & Technology, Meerpet, Hyderabad, vittedivarshinireddy756@gmail.com

[3]BTech student, Dept.of IT, TKR College of Engineering & Technology, Meerpet, Hyderabad, somishettysaikrupa@gmail.com

[4]BTech student, Dept.of IT, TKR College of Engineering & Technology, Meerpet, Hyderabad, amabalapavn123@gmail.com

Abstract: *Liver is the second largest organ in human bodies and located in the upper right-hand portion of the abdominal cavity. Viral infections, immune system disorders, hereditary diseases, cancer and additionally take up of too many toxics, can lead to various liver diseases. Symptoms of different types of liver diseases differ from each other. If they are not detected at early stage, then it might be difficult to treat. As a result, liver health problem needs to be detected timely and treated appropriately.*

Keywords: *Liver disease detection, machine learning*, SVM, WKNN, RF

I. INTRODUCTION

Liver is one of the largest organs that is present in the upper right part of abdominal cavity, and it is also the second largest organ after skin. It is wedging shape. And it is also the largest gland of the body which secretes chemical substances called hormones. Liver performs more than 500 functions in human body and also supports most of the organ which is vital for our survival [1]. In adults, it is observed that the liver weighs about 2% of body weight, in Males the liver weighs about 1.4 – 1.8 kgs, in females the liver weighs about 1.2 – 1.4 kgs and in new born it weighs 150 g.

Liver disease is the swelling of the liver caused by toxic substances, bacteria or inherited disease which causes the liver not to function properly as it is essential for digestion and get rid of bacteria. Liver diseases are commonly found in people

ISSN: 2366-1313

around the age group of 40-60 years and it is found mostly in men. There is around 10 lakh people diagnosed with liver disease every year and a total of 1.4 lakh deaths in a year in India. Machine Learning is a part of Artificial intelligence (AI) which simulates human intelligence into machines that is used to program to think like humans and mimic their actions. In other words, ML helps the system to gain knowledge without any specific knowledge [2]. In Supervised algorithm, the user inputs and the outputs are used for training process and accuracy prediction. Machine learning has extended its space to health care as well. One of the problems faced by health care the increase in the number of patients. Applications of Machine Learning can be potentially boosting accuracy for treatment accuracy. Various automatic medical diagnostic methods use classification techniques. The symptoms of liver disease difficult to detect early on since the organ works properly despite being partially destroyed. Patients' survival rate will increase on early diagnosis liver problem. The presence of enzymes in the blood can be used to identify liver disease. In this paper we are using liver patient dataset to predict whether the patient is having liver disease or not [3].

The liver is the largest solid organ in the body. It filters gastrointestinal blood, which contains a lot of toxins and antigens produced by the body. The liver is also an essential organ because it detoxifies toxins, metabolizes drugs, and generates proteins that are required for blood coagulation and other biological functions. The risks of liver disease are serious, and organ failure is unavoidable unless the condition is discovered early. Liver illness can have a variety of symptoms, making it difficult to diagnose promptly and accurately. The patient's problems are difficult to notice because the liver operates normally even when partially impaired. To solve this challenge, machine learning techniques can be utilized [4].

The purpose of "WEB BASED FRAMEWORK FOR LIVER DISEASE

DIAGNOSIS USING MACHINE LEARNING MODELS" is to improve liver disease diagnosis using machine learning approaches.

The need of this project is to serve the medicinal community, for the diagnosis of liver disease among patients, through a graphical user interface. The GUI (Graphical User Interface) can be readily utilized by doctors and medical

ISSN: 2366-1313

practitioners as a screening tool for the liver disease.

The main objective of this project is to use classification algorithms to identify the liver patients from healthy individuals.

The aim of this project is to develop a model with best accuracy and a web application that collects information and predict the results

Fig.1 Liver disease – stages

Liver harm is the one of the best deadliest ailments on the planet. The fundamental driver of liver harm are Fatty liver, Liver Fibrosis, Cirrhosis, hepatitis and diseases. Fig 1. Demonstrates the phases of liver harm, in the principal arrange solid liver will end up greasy liver because of gathering of cholesterol and triglycerides, following couple of months to years greasy liver will ends up liver fibrosis, later it prompts last phase of liver harm known as cirrhosis. In the beginning times of the liver ailment, it is exceptionally hard to identify despite the fact that liver tissue has been harm decently, it sources numerous restorative specialists over and over neglect to analyse the sickness. This can twist to wrong pharmaceutical and treatment, so early location is essential and important to spare the patient.

II. LITERATURE SURVEY

Machine learning calculations are exceptionally useful in giving essential measurements, continuous information, and progressed examination regarding the patient's illness, lab test results, circulatory strain, family history, clinical preliminary information, and more to specialists. Presently a day's Machine learning calculations are exceptionally valuable for removing and looking at the therapeutic information with the end goal to manufacture certain expectation models to rise the precision of analysis in a particular malady. Be that as it may, just couple of works in machine learning explore liver issue, in spite of the fact that this infection is forcefully expanding and getting to be one of the deadliest infections in a few nations.

Different classification techniques, such as Logistic Regression, Support Vector Machine, and K-Nearest Neighbor, were utilised by Thirunavukkarasu K,Ajay S. Singh,Md Irfan,Abhishek Chowdhur in their study to predict liver illness. All of these algorithms were compared based on

ISSN: 2366-1313

classification accuracy, which was determined using a confusion matrix. Logistic Regression and K-Nearest Neighbour have the highest accuracy, but logistic regression has the highest sensitivity, according to the experiment. As a result, we can conclude that Logistic Regression is a good way to predict liver illness

In[5] The paper "Comparative study of different classification algorithms on ILDP dataset to predict liver disorder" was published by Ayesha Pathan, Diksha Mhaske, Shruti Ka Jadhav, Rupali Bhondave, Dr. K. Rajeswari in 2018. ILDP is the dataset used in this study (Indian Liver Patient Dataset). Feature selection is carried out on the dataset. In order to pre-process and cluster the data, the k means clustering technique is utilized. The clustered data is then fed into various classification algorithms like Naive Bayes, Ada Boost, J48, Bagging and Random Forest. The performance of each algorithm is evaluated, and a comparative study has been carried out. Based on the performance comparison, Random Forest algorithm provides better performance as compared to Naive Bayes, AdaBoost, J48 and Bagging.

In [6] The paper "Early Detection of the Liver Disorder from Imbalance Liver Function Test Datasets" was published by Pushpendra Kumar, Ram Jeevan Singh Thakur in 2019. One of the datasets used in this investigation is ILDP (Indian Liver Patient Dataset) dataset. While the second dataset used is MPRLPD (Madhya Pradesh Region Liver Patient Dataset) dataset. The unbiased result is obtained using 10-fold cross-validation. To develop the system, we used support vector machine and k-nearest neighbor algorithms, as well as synthetic minority oversampling techniques to balance the datasets. On both the ILDP and MPRLPD imbalance and balance datasets, SVM (Support Vector Machine) and KNN (K-Nearest Neighbor) algorithms are used. For both the imbalanced and balanced datasets, we compared the results of both algorithms on various parameters. SVM improves accuracy, specificity, precision, and false positive rate (FPR) parameters on balanced datasets, whereas KNN improves accuracy, specificity, sensitivity, FPR, and false negative rate (FNR) parameters on balanced datasets. On majority of the parameters, the suggested system improves the results on the balance dataset. For balanced datasets, this method achieves an accuracy of 73.96 percent with SVM and 74.67 percent with KNN.

ISSN: 2366-1313

In[7] The paper "A critical study of selected classification algorithms for liver disease diagnosis" was published by Bendi Venkata Ramana, Prof. M. Surendra Prasad Babu, Prof.

N. B. Venkateswarlu. One of the datasets used in this study is Andhra Pradesh State of India. While the Bupa Liver Disorders datasets were used as second dataset. In this work, the performance of five classification methods was compared using data from liver patients: Naive Bayes classification (NBC), C 4.5 Decision Tree, Back Propagation, K- Nearest Neighbor (KNN), and Support Vector Machines (SVM).

On both the datasets, KNN, Back propagation and SVM are giving better results with all the feature set combinations.

In[8]The paper "Liver Patient Classification Using Intelligent Techniques" was published by Anju Gulia, Dr. Rajan Vohra, Praveen Rani. In this study, J-48, Multilayer perceptron, Support Vector Machine, Random Forest and Bayesian network are used. In three steps, this research uses hybrid model development and comparison analysis to improve prediction accuracy of liver patients. The classification phase is the initial step. On the original liver patient datasets, algorithms are applied. In the second phase, by utilizing feature selection a subset of liver patient from the entire liver patient dataset is achieved as it consists of only significant attribute. On a significant subset of the data obtained, classification algorithms were used. The outcomes of the third phase are presented. In third phase, the results of classification algorithms with and without feature selection are compared with each other.

Using feature selection, the Random Forest (RF) algorithm surpassed all other strategies with an accuracy of 71.8696 percent, according to the findings of our study.

Reshul Dani, proposed two techniques with the end goal to order the ceaseless liver illness, one strategy is a symptomatic way to deal with finding, and second one includes a hereditary way to deal with the finding. Proposed approach is the utilization of Artificial Neural Networks and Multi-Layer Perceptron to Miniaturized scale Array Analysis.

III. EXISTING SYSTEM

Indians are at a higher risk of liver failure. Existing system used feature selection and pre-processing to cluster data. Various techniques like Ada Boost, Bagging,

ISSN: 2366-1313

Random Forest were applied on the modified data. But these approaches are inefficient to classify liver data and is also time consuming.

There we have limitations in the existing work

- The performance in the training and testing of the liver disorder dataset is poor.
- It requires high computation time for prediction of liver disease.

IV. PROPOSED SYSTEM

To solve the problems that are associated with those research articles, in this paper we are using data preprocessing technique and machine learning algorithms like Weighted K-Nearest Neighbor algorithm, SVM and Random Forest.

How does Machine Learning Work?

A model is created by training a machine learning algorithm on a training data set. When new input data is presented to the ML method, the model is used to produce a prediction.

The accuracy of the prediction is tested, and if it is acceptable, the Machine Learning method is used. If the accuracy isn't good enough, the Machine Learning algorithm is retrained with a new batch of training data.

Building a Predictive model that can be used to discover a solution to a Problem Statement is part of the Machine Learning process. Assume you've been given a challenge to tackle using Machine Learning in order to grasp the Machine Learning process.

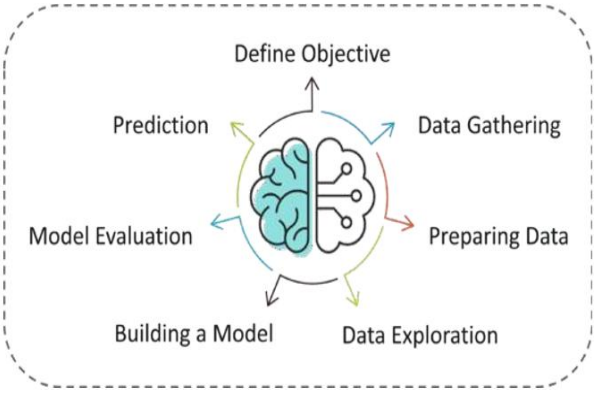

Fig.1 Working of machine learning algorithms

The following steps are followed in a Machine Learning process:

Step 1: Define the objective of the Problem Statement

We must first figure out what actually needs to be determined. The goal in this scenario is to use weather conditions to estimate the likelihood of rain. It's also important to make detailed notes at this point about the data types that can be used to address the issue or the strategy you'll need to take to arrive at a solution.

ISSN: 2366-1313

Step 2: Data Gathering

Here, you must be asking questions such as,

- What type of data is needed to solve this problem?
- Is the data available?
- How can I get the data?

After you've figured out what kind of data you'll need, you'll need to figure out how to get it. Data can be collected manually or by web scraping. If you're a newbie trying to learn Machine Learning, though, you won't have to worry about acquiring data. Because on the web there are already 1000s of data resources which are available, so you can just download the data set and keep going.

Returning to the issue at hand, the data required for weather forecasting comprises factors like humidity, temperature, pressure, location, whether you reside in a hill station, and so on. Such information should be gathered and preserved in order to be analysed.

Step 3: Data Preparation

Is often not the data you acquire in the correct format. Missing values, redundant variables, duplicate values, and other irregularities will be found in the data collection. It's critical to eliminate such inconsistencies because they can lead to incorrect estimations and forecasts. As a result, at this point, you check the data set for any errors and correct them right away.

Data Processing

Data processing is the process of transforming data from one format to another that is more useable and desirable, i.e., rendering it more relevant and useful. This entire process is automated using Machine Learning algorithms, mathematical modelling, and statistical expertise. Graphs, videos, charts, tables, photos, and a variety of other formats can be produced as a result of this entire process, based on the task given and the machine's requirements. This may appear straightforward, but when it comes to extremely large businesses such as Twitter and Facebook, as well as administrative bodies such as Parliament and UNESCO, and health-care organizations, the entire procedure must be carried out in a very systematic manner.

Machine Learning Algorithms Weighted K-Nearest Neighbor algorithm:

Weighted KNN is a modified version of K-Nearest Neighbor algorithm. One of the main issues that affect the performance of

ISSN: 2366-1313

the KNN algorithm is the choice of the hyperparameter k. If k is too small, the algorithm would be more sensitive to outliers. If k is too large, then the neighborhood may include too many points from other classes.

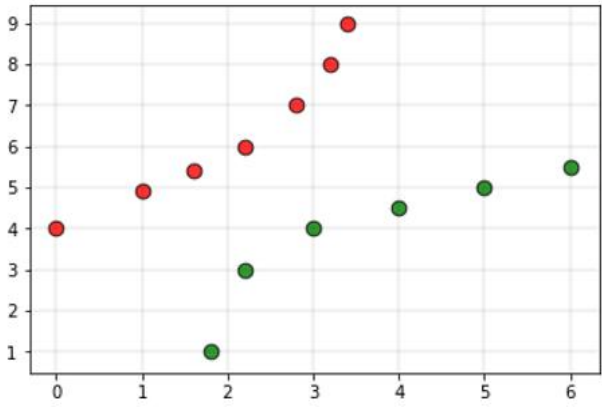

Fig. 2 KNN algorithm

The red labels indicate the class 0 points, and the green labels indicate class 1 points. Consider the white point as the query point

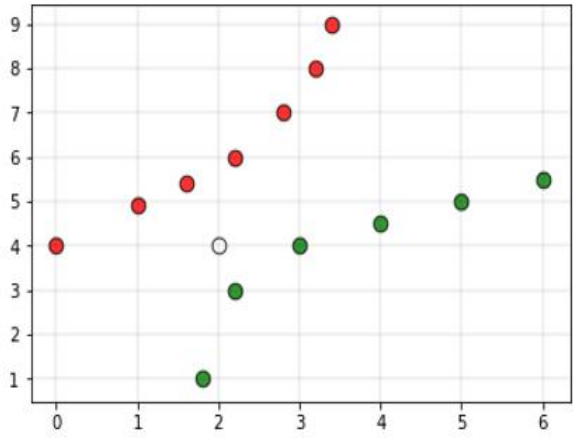

Fig.3 Weighted K-Nearest Neighbor algorithm

If we give the above dataset to a KNN based classifier, then the classifier would declare the query point to belong to the class 0. But in the plot, the point is closer to the class 1 points compared to the class 0 points. To overcome this disadvantage, weighted KNN is used. In weighted KNN, the nearest k points are given a weight using a function called as the kernel function.

The intuition behind weighted KNN, is to give more weight to the points which are nearby and less weight to the points which are farther away. Any function can be used as a kernel function for the weighted KNN classifier whose value decreases as the distance increases. The simple function which is used is the inverse distance function.

Support vector machine (SVM):

The separation hyper plane in the SVM classifier is designed to minimize the expected classification error of the unseen test patterns. SVM is a powerful classifier that can distinguish between two groups. SVM assigns the test picture to the class with the greatest distance to the training's nearest point. The SVM training process created a model that can predict whether a test image belongs to this or another class. Even if we limit ourselves to single pose (frontal) detection, SVM requires a large quantity of training data to identify an

ISSN: 2366-1313

affective decision border, and the computational cost is very high.

The SVM is a categorization learning method. It seeks to locate the best separation hyper plane for unseen patterns such that the expected classification error is as low as possible. The input is transferred to a high-dimensional feature space where they can be separated by a hyper plane for linearly non-separable data. Kernels are used to efficiently accomplish this projection into a high-dimensional feature space. The SVM seeks to discover the best separating hyper plane given a set of training samples and the accompanying decision values -1, 1.

Random Forest:

Random Forest is a well-known and very effective machine learning technique. Bagging, also known as Bootstrap Aggregation, is a type of machine learning technique. The bootstrap is a very powerful statistical approach for estimating a value from a data sample, such as mean. Many data samples are obtained, the mean is computed, and then all of the mean values are averaged to provide a better estimate of the true mean value.

The similar procedure is employed in bagging, although decision trees are commonly utilized instead of estimating the mean of each data sample. Several samples of the training data are considered, and models are built for each sample. While a prediction for any data is needed, each model gives a prediction, and these predictions are then averaged to get a better estimation of the real output value.

Data collection

The data sets were obtained from the Kaggle machine learning repository, which is an open data resource for machine learning and predictive analytics.

Data Preprocessing

The dataset that is required for the study is selected. If the dataset contains any null values for attributes, then those values are to be deleted and replaced with mean, mode or medium of a particular column or with zero or one. To deal with the character data One Hot Encoding method is needed. This method converts the character data into binary values. If there is any imbalance in the dataset, then it should be fixed by means of data duplication.

Extraction of the features would reduce the data size involved to characterize a wide collection of data. This approach chooses features based on their performance evaluation, hence features with a better

ISSN: 2366-1313

correlation value are chosen as a significant prediction function for liver.

Data Prediction

The data should be divided into two datasets: the train dataset and the test dataset. The data is trained with accessible input and output data using the machine learning algorithm. The data is checked in the test phase to see if the e-accuracy model is satisfactory. The machine learning module then predicts the new data.

V. RESULTS

SYSTEM ARCHITECTURE

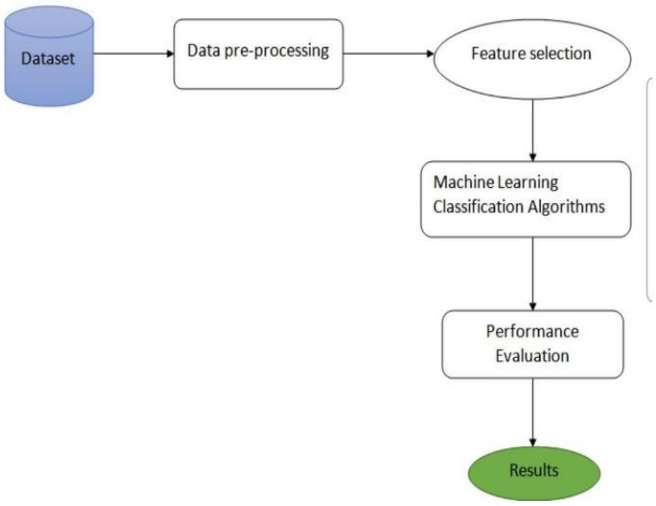

Fig.4 System architecture

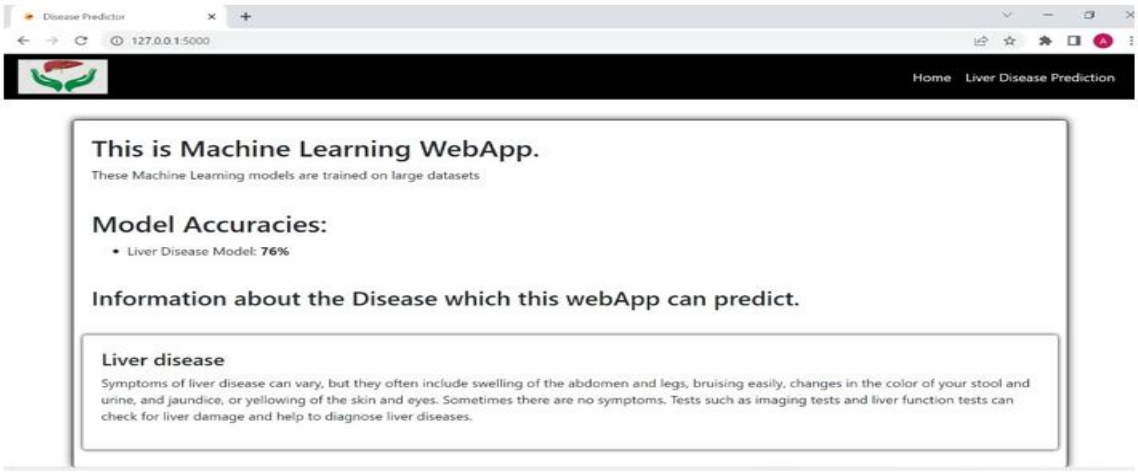

Fig.5 Home page

ISSN: 2366-1313

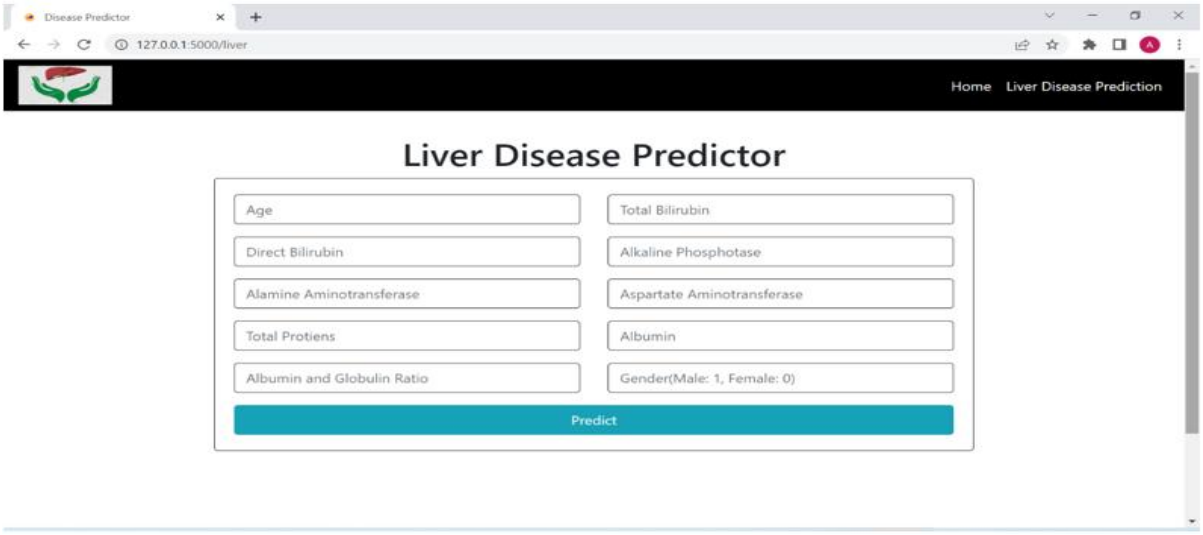

Fig.6 GUI developed using python

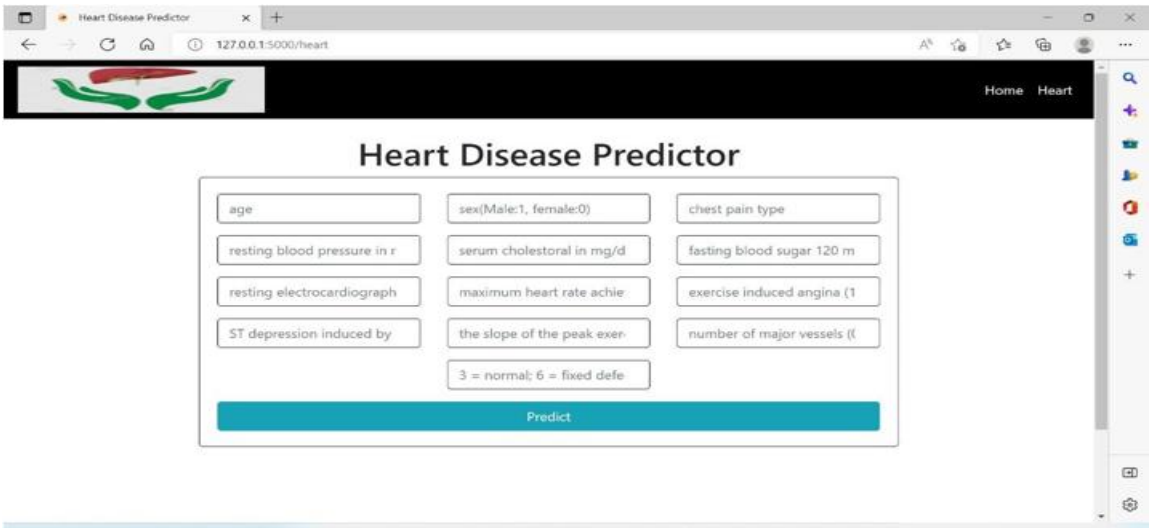

Fig.7 Input by User

ISSN: 2366-1313

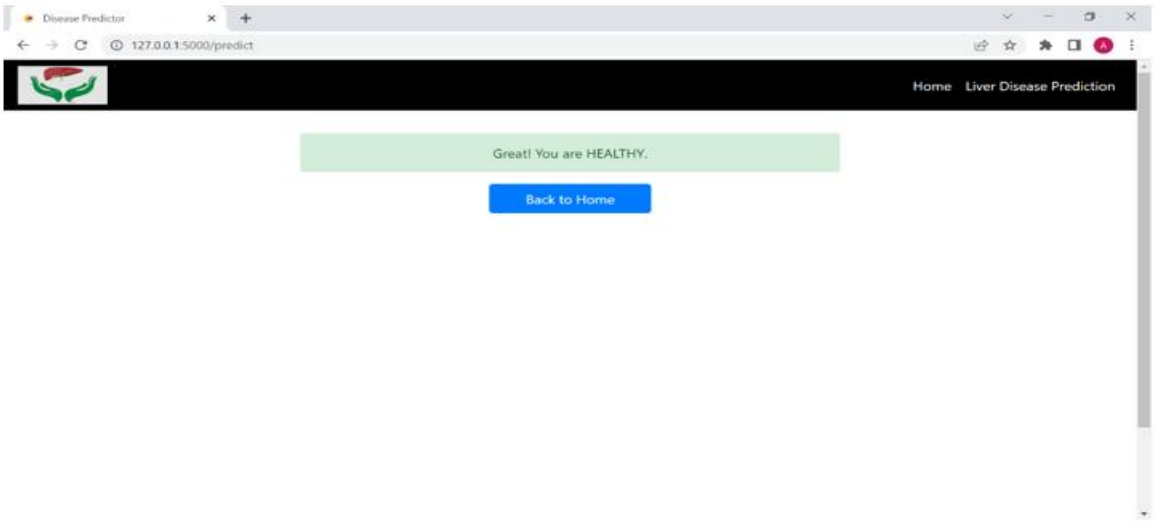

Fig.8 Result page

VI. CONCLUSION

We proposed approaches for identifying liver illness in patients using machine learning techniques in this study. Random Forest, KNN, and SVM were among the three machine learning algorithms tested. All the models were used to implement the system, and their performance was analyzed. Certain performance metrics were used to evaluate performance. With a 76 percent accuracy, SVM was the model that produced the best results. When comparing this work to earlier research, it was determined that SVM was extremely effective. SVM was utilized to create a graphical user interface that would be used as a medical tool by hospitals and medical personnel.

The proposed system can be developed in a variety of ways, and there is a lot of room for development. These include: 1. Improving the algorithms' accuracy. 2. Improving the algorithms to improve the system's efficiency and functionality. 3. Working on certain additional characteristics to counter liver disease even more effectively.

REFERENCES

1. Pushpendra Kumar, Ramjeevan Singh Thakur (2019), “Early Detection of the Liver Disorder from Imbalance Liver Function Test Datasets”.
2. Bendi Venkata Ramana, Prof. Surendra Prasad Babu, Prof. N. B. Venkateswarlu (2011), “A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis”.
3. M. Banu Priya, P. Laura Juliet, P.R. Tamilselvi (2018), “Performance

ISSN: 2366-1313

Analysis of Liver Disease Prediction Using Machine Learning Algorithms".

4. Ayesha Pathan, Diksha Mhaske, Shrutika Jadhav, Rupali Bhondave, Dr.K. Rajeswari "Comparative Study of Different Classification Algorithms on ILPD Dataset to Predict Liver Disorder".
5. Anju Gulia, Dr. Rajan Vohra, Praveen Rani (2014), "Liver Patient Classification using Intelligence Techniques".
6. M. Banu Priya, P. Laura Juliet, P.R. Tamilselvi (2018), "Performance Analysis of Liver Disease Prediction Using Machine Learning Algorithms,".
7. A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis " Bendi Venkata Ramana1, Prof. M.Surendra Prasad Babu2, Prof. N. B. Venkateswarlu3 - (IJDMS), Vol.3, No.2, May 2012
8. S. B. Kotsiantis, Supervised Machine Learning: A Review of Classification Techniques, Informatica (2007) 249-268 249.
9. A.S.Aneeshkumar and C.Jothi Venkateswaran, "Estimating the Surveillance of Liver Disorder using Classification Algorithms",
10. International Journal of Computer Applications (095-8887), Volume 57-No.6, November 2012
11. T. Mitchell, Machine Learning. McGrawHill ,1997 Saranya, A., and G. Seenuvasan. "A COMPARATIVE STUDY OF DIAGNOSING LIVER DISORDER DISEASE USING CLASSIFICATION ALGORITHM." (2017).
12. Bendi Venkata Ramanaland Prof.M.Surendra Prasad Babu, " Liver Classification Using Modified Rotation Forest", Internationa Journal of Engineering Research and Development ISSN: 2278-067X, Volume 1, Issue 6 (June 2012), PP.17-24.