WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Machine Learning Based Liver Disease Prediction System

B. Thilagavathi[1]*, Mr. S. Barath*[2]

[1]Master of Computer Application, krishnasamy college of engineering &Technology, Cuddalore
[2]MCA., M. Phil., NET., Assistant Professor, Master of Computer Application, krishnasamy college of engineering & Technology, Cuddalore.

ABSTRACT

In recent decades, liver diseases have risen to prominence as one of the world's top causes of death and a condition that can be fatal. According to the WHO, chronic diseases are responsible for over 59 percent of global mortality and 46% of conditions, and they claim the lives of almost 35 million people worldwide. As the liver continues to operate even when partially wounded, problems with the liver are typically not identified until it is too late. Potentially, early discovery can save a person's life. This project's need is to outline a framework for an iterative method of finding high-risk patients' events that is based on a clinical data repository and machine learning algorithm. In order for the prediction to be accurate, the article is also open to new difficulties and potential adjustments to other cutting-edge technology. Damage to the liver, which performs a vital function for the body, could have catastrophic repercussions. Early liver disease detection is crucial for this reason. This work attempted to predict liver disease using an ensemble technique (Gradient Boosting Classifier + AdaBoost Classifier) based on various clinical data collected from healthy blood donors and liver patients.

I. INTRODUCTION

Liver is one of the largest organ that is present in the upper right part of abdominal cavity, and it is also the second largest organ after skin. It is wedge 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.

The liver disease which has a type called fatty liver disease, which is caused by the accumulation of fat in the liver, is a very common condition in India, with over 10 million cases reported each year. Testing is required for diagnosis due to the lack of adverse effects. Any problem with the liver's function that creates illness is known as liver disease The liver is in charge of many dangerous tasks in the body, and if it becomes diseased or injured, such functions will be lost, causing serious harm to the body. Hepatic disease is another term for liver disease. Liver disease is a broad phrase that encompasses all of the issues that might cause the liver to fail to execute its tasks.

Before a decline in function occurs, more than 75 percent or three-quarters of the liver tissue must be affected. Artificial Intelligence (AI) is a type of computerized reasoning that is enabled by the ability of computer programs to learn, acquire knowledge, and then use that knowledge in various fields. Man-made awareness is now used in almost every sphere of application, and it is made up of various components, such as Deep Learning and Machine Learning. Obesity, inhalation of toxic gases, consumption of polluted food, excessive use of foods and medications, and alcohol are all major causes of liver disease. The goal of this research is to present machine learning approaches based on the Classification of Liver Disorders to relieve clinicians of their workload. The, implementation involves splitting the dataset into training and testing sets. The model will be trained and applied on the testing set and the performance will be measured based on performance on the parameters such as data collection, data Processing, classification, decision tree algorithm, random forest algorithm, support vector machine etc. The aim that has been incorporated by the author is to use the different classification techniques of machine learning and take the dataset to train the model and then evaluate the performance based on the features that are being used. The area which is one of the most significant where the concept of data mining is being used is Biomedical science. It is a field of science which mainly works with the lives of humans and is considered very highly sensitive. After many years, the researchers from the community have done research on various diseases using the concept of data mining. As we go deeper into this concept the research that has been done in the years by the researchers in this field many works that are being done using data mining have been performed such as forecasting, prevention etc.

II. RELATED WORKS

Auxilia et al[1] made an accurate prediction for liver disease using different ML methods, including SVM, Random Forest, Decision Trees, Artificial Intelligence and Naïve Bayes. The research was conducted using R on the Indian Liver Patient Records dataset, with 583 instances and 11 attributes. The accuracies were obtained from SVM (77%).Wu et al.[2] did a prediction analysis on patients having Fatty Liver Disease (FLD).The research collected 700 patient records from New Taipei Hospital, which had screening tests for fatty liver disease; out of 700 patients, 577 records were considered depending on the patient's age and sufficient data. Of those 577 patients, 377 had fatty liver disease, and the remaining had no fatty liver disease. The dataset contains patient health details of age, gender, systolic and diastolic blood pressure, abdominal girth, glucose level, triglyceride, HDL-C, SGOT-AST, and SGPT-ALT. Synthetic Minority Over-Sampling Technique (SMOTE) was applied at the data preprocessing stage, and normalisation was done. Singh et al.[3] focused their research on predicting liver disease using different classification methods with feature selection and implementing software for easy prediction. The study was conducted on the Indian Liver Patient Records dataset. Some attributes were removed during the feature selection phase using the Correlation-based Feature Selection Subset Evaluator with the Greedy Stepwise search method in WEKA. Only five attributes were selected through this method: Total Bilirubin, Direct Bilirubin, Alkaline Phosphatase, Alamine Aminotransferase, and Aspartate Aminotransferase. With this, six different classification methods were applied: Logistic Regression, Naïve Bayes, Sequential Minimal Optimization (SMO), Random Forest, Instant based Classification (IBk), and Logistic Regression has provided the highest accuracy with 74.36%. The least accuracy was produced by Naïve Bayes (55.9%).

III. METHODOLOGY AND IMPLEMENTATION

In the proposed system we implement Liver disease prediction using Ensemble Technique. This system helps to reduce the burden on the doctor by analyzing patient's conditions using machine learning techniques. Ensemble approaches provide more accurate results than a single model.In our proposed system we use Gradient Boosting Classifier + AdaBoost Classifier (Ensemble Technique). Our goal is to get prediction on the basis of given datasets of people whether the person is having the liver disease symptoms or not.

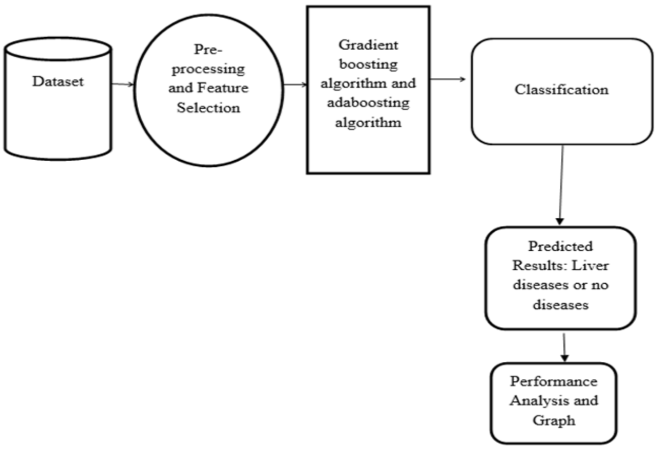

Dataset
Pre-processing and Feature Selection
Gradient boosting algorithm and adaboosting algorithm
Classification
Predicted Results: Liver diseases or no diseases
Performance Analysis and Graph

Figure 1 Proposed model

1.Data Preparation

Wrangle data and prepare it for training. Clean that which may require it (remove duplicates, correct errors, deal with missing values, normalization, data type conversions, etc.)Randomize data, which erases the effects of the particular order in which we collected and/or otherwise prepared our data Visualize data to help detect relevant relationships between variables or class imbalances (bias alert!), or perform other exploratory analysis Split into training and evaluation sets

2.Model Selection

We used Gradient Boosting Classifier + Ada Boost Classifier (Ensemble Technique) using machine learning algorithm; we got a accuracy of 92.1% on test set so we implemented this algorithm.

Ensemble Technique

Ensemble methods are techniques that aim at improving the accuracy of results in models by combining multiple models instead of using a single model. The combined models increase the accuracy of the results significantly. This has boosted the popularity of ensemble methods in machine learning.

Categories of Ensemble Methods

Ensemble methods fall into two broad categories, i.e., sequential ensemble techniques and parallel ensemble techniques. Sequential ensemble techniques generate base learners in a sequence, e.g., Adaptive Boosting (AdaBoost). The sequential generation of base learners promotes the dependence between the base learners. The performance of the model is then improved by assigning higher weights to previously misrepresented learners.The majority of ensemble techniques apply a single algorithm in base learning, which results in homogeneity in all base learners. Homogenous base learners refer to base learners of the same type, with similar qualities. Other methods apply heterogeneous base learners, giving rise to heterogeneous ensembles. Heterogeneous base learners are learners of distinct types.

Main Types of Ensemble Methods

1. Bagging

Bagging, the short form for bootstrap aggregating, is mainly applied in classification and regression. It increases the accuracy of models through decision trees, which reduces variance to a large extent. The reduction of variance increases accuracy, eliminating overfitting, which is a challenge to many predictive models.

Aggregation in bagging is done to incorporate all possible outcomes of the prediction and randomize the outcome. Without aggregation, predictions will not be accurate because all outcomes are not put into consideration. Therefore, the aggregation is based on the probability bootstrapping procedures or on the basis of all outcomes of the predictive models.

2. Boosting

Boosting takes many forms, including gradient boosting, Adaptive Boosting (AdaBoost), and XGBoost (Extreme Gradient Boosting). AdaBoost uses weak learners in the form of decision trees, which mostly include one split that is popularly known as decision stumps. AdaBoost's main decision stump comprises observations carrying similar weights.

Gradient boosting adds predictors sequentially to the ensemble, where preceding predictors correct their successors, thereby increasing the model's accuracy. New predictors are fit to counter the effects of errors in the previous predictors. The gradient of descent helps the gradient booster identify problems in learners' predictions and counter them accordingly.

XGBoost makes use of decision trees with boosted gradient, providing improved speed and performance. It relies heavily on the computational speed and the performance of the target model. Model training should follow a sequence, thus making the implementation of gradient boosted machines slow.

3. Stacking

Stacking, another ensemble method, is often referred to as stacked generalization. This technique works by allowing a training algorithm to ensemble several other similar learning algorithm predictions. Stacking has been successfully implemented in regression, density estimations, distance learning, and classifications. It can also be used to measure the error rate involved during bagging.

3. Analyze and Prediction

- In the actual Dataset, we chose only 10 features:
- Age: Age of the patient
- Female: Gender of the patient (1 if Female, 0 if Male)
- TB: Total Bilirubin
- DB: Direct Bilirubin
- Alkphos: Alkaline Phosphotase
- Sgpt: Alamine Aminotransferase
- Sgot: Aspartate Aminotransferase
- TP: Total Protiens
- ALB: Albumin
- A/R: Albumin and Globulin Ratio
- class: 1 Liver diseases and 0 no diseases

4.Accuracy on test set

We got an accuracy of 92.1% on test set.

5. Saving the Trained Model

Once you're confident enough to take your trained and tested model into the production-ready environment, the first step is to save it into an .h5 or .pkl file using a library like pickle.

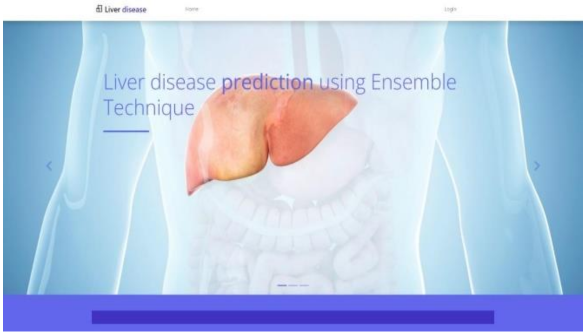
Make sure you have pickle installed in your environment.

Next, let's import the module and dump the model into .pkl file

IV. RESULTS AND DISCUSSION

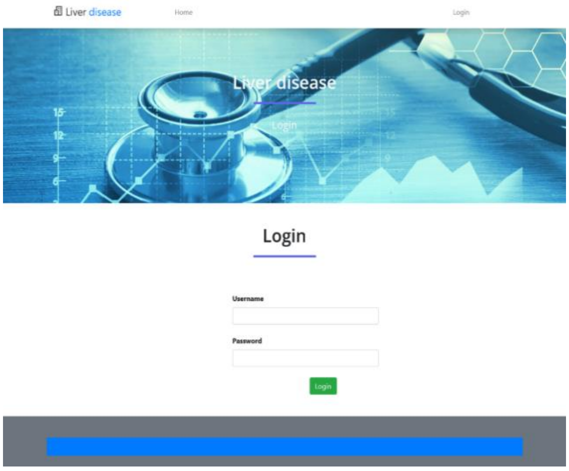
Liver diseases are becoming one of the most fatal diseases in several countries. Patients with Liver disease have been continuously increasing because of excessive consumption of alcohol, inhale of harmful gases, intake of contaminated food, pickles and drugs. liver patient datasets are investigate for building classification models in order to predict liver disease. This dataset was used to evaluate prediction algorithms in an effort to reduce burden on doctors. In that paper, we proposed as checking the whole patient Liver Disease using Machine learning algorithms

Home page



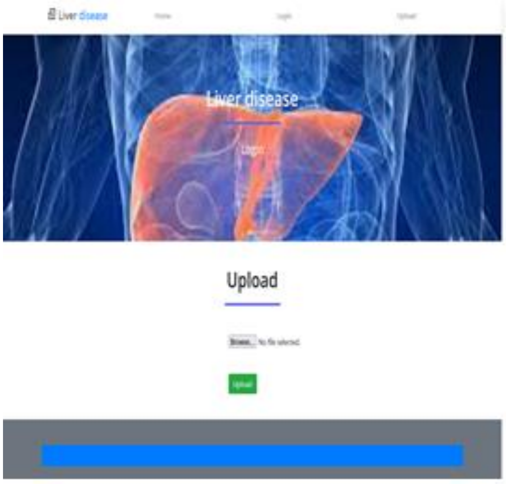
This is Home page from

Login page



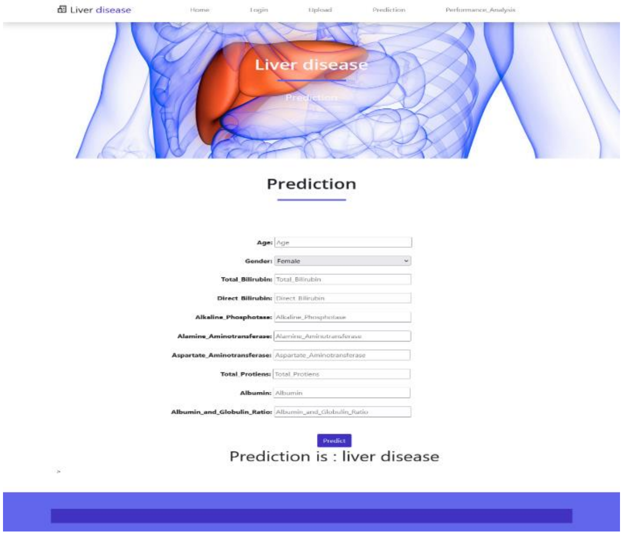
This is Login page

Upload page



This is upload page where liver disease dataset is upload.

Prediction Result



Prediction

Age: Age

Gender: Female

Total_Bilirubin: Total_Bilirubin

Direct_Bilirubin: Direct_Bilirubin

Alkaline_Phosphotase: Alkaline_Phosphotase

Alamine_Aminotransferase: Alamine_Aminotransferase

Aspartate_Aminotransferase: Aspartate_Aminotransferase

Total_Protiens: Total_Protiens

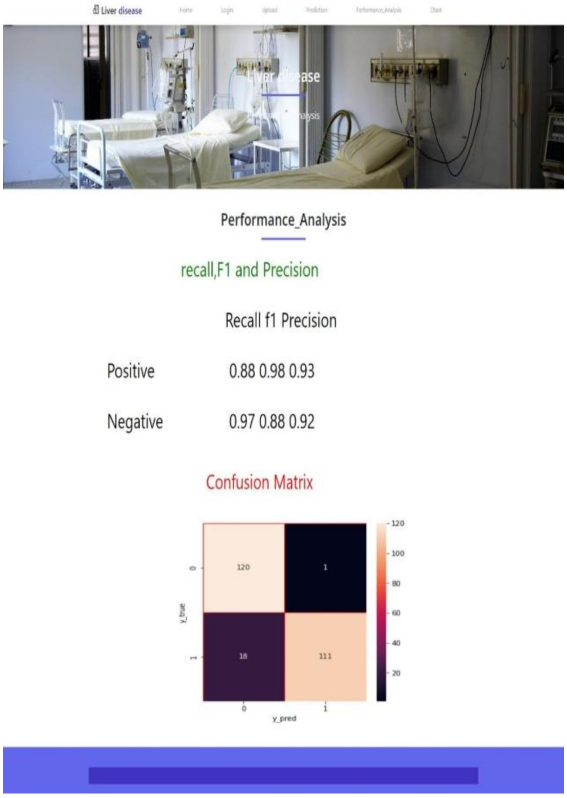
Albumin: Albumin

Albumin_and_Globulin_Ratio: Albumin_and_Globulin_Ratio

Prediction is : liver disease

This is page for Predict the Liver disease result

Analysis Page



Performance_Analysis

recall,F1 and Precision

	Recall	f1	Precision
Positive	0.88	0.98	0.93
Negative	0.97	0.88	0.92

Confusion Matrix

This is page show the total number of liver static

V.CONCLUSION

After the authors performed the entire process of getting the novel output than all the related work starting from the data preparation to preprocessing to selecting the different models and then combining it. The output received would be used by the medical industry in getting proper predications for the liver disease which would prove really beneficial to save lives of the people suffering from this disease. The proposed model achieved an accuracy of Train Accuracy: 100% and Test Accuracy of 92%.

REFERENCES

[1] AgroConsultant: Intelligent Crop Recommendation System Using Machine Learning Algorithms. Zeel Doshi , Subhash Nadkarni , Rashi Agrawal, Prof. Neepa Shah.

[2] Kalimuthu M, Vaishnavi P and Kishore M 2020 Crop Prediction Using Machine Learning 2020 Third International Conference on Smart Systems and Inventive Teachnology (ICSSIT) pp 926-3210.1109/ICSSIT

10.1109/ICSSIT48917.2020.9214190.

[3] Prof. D.S. Zingade ,Omkar Buchade ,Nilesh Mehta ,Shubham Ghodekar ,Chandan Mehta "Crop Prediction System Using Machine Learning".

[4]M. Sameer, A. K. Gupta, C. Chakraborty, and B. Gupta, "Epileptical Seizure Detection: Performance analysis of gamma band in EEG signal Using Short-Time Fourier Transform," in 2019 22nd International Symposium on Wireless Personal Multimedia Communications (WPMC), 2019, pp. 1–6, doi: 10.1109/WPMC48795.2019.9096119.

[5] A. Mahajan, K. Somaraj, and M. Sameer, "Adopting Artificial Intelligence Powered ConvNet To Detect Epileptic Seizures," in 2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES), 2021, pp. 427–432, doi: 10.1109/IECBES48179.2021.9398832.

[6] N. Nasir, N. Afreen, R. Patel, S. Kaur, and M. Sameer, "A Transfer Learning Approach for Diabetic Retinopathy and Diabetic Macular Edema Severity Grading," Rev. d'Intelligence Artif., vol. 35, pp. 497–502, Dec. 2021, doi: 10.18280/ria.350608.

[7] M. Sameer and B. Gupta, "ROC Analysis of EEG Subbands for Epileptic Seizure Detection using Naive Bayes Classifier," J. Mob. Multimed., pp. 299–310, 2021.

[8] M. Sameer and B. Gupta, "Time–Frequency Statistical Features of Delta Band for Detection of Epileptic Seizures," Wirel. Pers. Commun., 2021, doi: 10.1007/s11277-021-08909-y.

[9] S. M. Beeraka, A. Kumar, M. Sameer, S. Ghosh, and B. Gupta, "Accuracy Enhancement of Epileptic Seizure Detection: A Deep Learning Approach with Hardware Realization of STFT," Circuits, Syst. Signal Process., 2021, doi: 10.1007/s00034-021-01789-4.