Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2347 – 2357
Received 05 March 2021; Accepted 01 April 2021.

Prediction of Liver Disease Using Machine Learning Algorithm and Genetic Algorithm

B.Poonguzharselvi[1*], Mohammad Mahaboob Ali Ashraf[2], Vadlamani V S S Subhash[3], S.Karunakaran[4]

[1,2,3] Chaitanya Bharathi Institute of Technology, Gandipet, Hyderabad, Telangana-500075, India.
[4]Vardhaman College of Engineering, Shamshabad, Hyderabad, Telangana-501218, India.
[*] poonguzharselvi_cse@cbit.ac.in

ABSTRACT

One of the most vital causes of death worldwide is liver disease. We, humans, have come a long way in the medical field and scientific advancements to treat diseases and it's evident that when these liver diseases are detected early, they can be treated easily. In order to be able to accurately predict if there’s a chance of the liver disease it is imperative to identify the features/symptoms which play a significant role in causing the Liver Disease. In order to improve the performance of the prediction models, it is important to choose the right combination of significant features.

A new system is proposed that identifies the significant features and then predicts whether or not a person may suffer or is suffering from Liver Disease using the identified features. Our system ought to be used as a supplementary tool in diagnosis. Data is essential and we will be using the dataset available on the UIC repository. We will be using genetic algorithms to identify the significant featuresand then use those features to train different classification models like k-Nearest Neighbors, k-means, Random Forest, Support Vector Machines, Naïve Bayes, Logistic Regression, etcetera which will predict if there’s a chance of Liver Disease for a person’s data. We will also be using neural networks with back propagation to perform binary classification.

Ideally our proposed model identified the significant features and finds the best model which predicts the Liver Disease with more accuracy or another statistical measure.

Keywords

Machine Learning; Deep Learning; Neural networks; classification techniques; genetic algorithm; and data mining

INTRODUCTION

Problem Definition including the significance and objective

Liver is the cleaning and detoxification mechanism of our body. If there is any problem with our livers, our bodies cannot properly dispose of its wastes. This can lead to several other problems. Liver diseases are responsible for around 2% of the world’s deaths. Early diagnosis of these diseases helps in preventing from deaths. This project tries to help medical professionals detect liver diseases in its early stages and help reduce the rate of liver diseases.

Liver disease is a category of disease that has the negative effect on the usual working of the liver. The detectionof a liver disease in its early stages is very important to prevent any adverse effects in the future. Thus, the main aim of the proposed model is to detect the liver disease using binary classification. For this purpose we will be using various machine learning classification algorithms. We will also be using genetic algorithms and deep learning techniques to improve our classification accuracy.

Liver disease is a very broad term. When we talk about liver disease there can be many diseases such as fatty liver disease, non-alcoholic fatty liver disease, hepatitis, etc. Liver diseases are responsible for over 2% of the world's deaths. So, detection of the disease in its early stagesis very important. Diagnosing Liver disease involves looking at various statistics about the liver including age, total bilirubin, direct bilirubin, alkaline phosphatase, total protein, albumin and many more [11]. A thorough liver examination is currently required to tell if a person has liver disease or not. This is not accessible to many people around the world. For this reason, we need a cost effective way to tell if a person suffers from liver disease with confidence.

We planned to develop a web application where a medical professional will input various liver functioning data to know whether the person suffers from liver disease or not based on the algorithm used for prediction. Based on the results of various algorithms, the medical professional will determine if the person suffers from liver disease.

Methodologies

There are several steps involved in the process of classification of liver disease. Since we are using machine learning models for the purpose of classification, the steps involved in the process are

1. Data collection
2. Data preparation
3. Choose a model
4. Train the model
5. Evaluate the model
6. Parameter Tuning
7. formulate predictions

Data Collection

Data collection refers to the collection of the required datasets for the purpose of training the models. This can be done from reputed, peer reviewed websites like Kaggle, UCI repository, etc. If the required data is not available, it is our responsibility to collect the required data in sufficient quantity to train the model effectively.

Data Preparation

Data preparation involves in transformation of our raw data into the suitable form that can be used in training our machine learning model. This step is also a good time to perform various visualizations of our data to look for any relevant relationships between our variables. We also perform normalization and scaling in this step to make our data ready for training. We also split the data into training and testing sets in this step.

Choose a model

There are many machine learning models out there. In this step, we identify the machine learning models to train based on the type of problem we are trying to solve and the outcome we are trying to predict. Some of the machine learning models used are SVM, K-Means, Decision tree, Random Forest, etc.

Train the model

After choosing the machine learning model, we move on to training the model. This is the most important step in machine learning classification. In this step, we use our data to improve our model's ability to predict the presence of liver disease. There are several inbuilt classes offered by python for many popular machine learning models. If the data is in the required format, we just pass this data to the objects of various classes and then the prediction score is calculated.

Evaluate the model

After we have trained the model, the performance of our model is evaluated by using the testing set we have obtained in the data preparation phase. We prefer the highest possible score during the process of evaluation.

Parameter Tuning

Once we are done with evaluation, it's possible to boost the performance of our model by fine-tuning our parameters. Few parameters were implicitly assumed during the process of training, we can go back to that step and experiment with various combinations of parameters to see if the prediction score improves.

Formulate Prediction

Machine learning is using data to answer questions. So, prediction is the step where we get answer to the question “Does the person suffer from liver disease?”.

Outline of the results

The results show various machine learning models being used for the prediction of presence of liver disease. We first compared the performance of different models of machine learning algorithms based on the parameters like accuracy, number of information rate, sensitivity, specificity, etc.

Later, we used genetic algorithm to find the right combination of parameters to be used during the process of training various machine learning models and answer the question which is the most effective algorithm to use for the prediction purpose. We also trained a deep learning model and compare its performance with normal machine learning models.

RELATED WORK

Sanjay Kumar and Sarthak Katyal [1] developed a classification model with different data mining algorithms to guess the correct liver diagnosis. Five algorithms were applied to the dataset to give the parameters like precision, recall and accuracy. In this model they used five different

Received 05 March 2021; Accepted 01 April 2021.

algorithms such are K-Means, K-Nearest Neighbour, Naive Bayes, C5.0 and Random Forest. In this model, results showed that Random Forest had the most accuracy of all the algorithms.

Varun Vats, Lining Zhang et.al.[2] proposed a model based on three machine learning techniques including DBSCAN, K-Means, Affinity Propagation. Comparative performance measurement was done based on Silhouette coefficient on previously mentioned three techniques. Finally they concluded that the performance of K-Means technique is optimal.

L. Alice Auxilia [3] used pearson coefficient for feature selection. To predict the liver disease, they used Decision Trees, Naïve Bayes, Random Forest, Support Vector Machine, and Artificial Neural Network. Finally, it was shown that the decision tree performed better than other classification algorithms.

PROPOSED MODEL

A brief workflow of our proposed solution is represented in Figure-1. The steps include preprocessing, Segmentation and combining the results of segmentation. This is followed by contouring to highlight the area of interest.

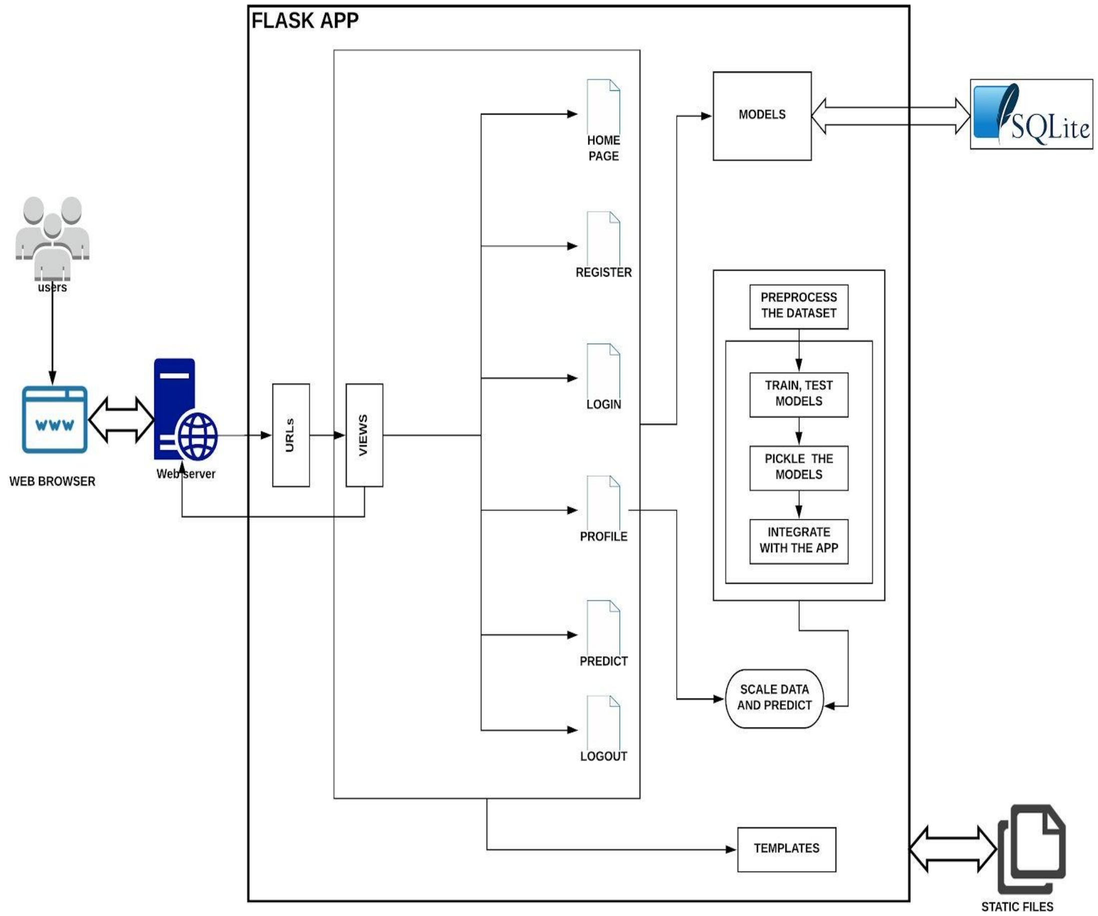

Figure-1: Work flow of the proposed model

Module Description

A. Urls

Flask allows us to have a flexible mapping of URL paths to function calls. This acts analogous to a controller in MVC(Model View Controller) architecture. The app.route function or also known as annotation on other popular platforms maps the urls or the http requests to a function call, which is essentially a view. The function calls in turn decide which view is to be presented/rendered to the end user and also perform operations as needed.

B. Views

It is a Python function. It accepts a web request from the caller and sends the web response back to the caller.

C. Models

Model is one of classes available in the Flask-SQLAlchemy. In any Flask application, the use of SQLAlchemy is done easily by using Flask-SQLAlchemy. The reading, writing, querying and deleting of persistent data from the relational database is achieved through SQLAlchemy.

D. Preprocess

Data preprocessing is a data mining practice which helps in getting the data in valuable and well-organized format by the transformation of the raw data. It includes Data Cleaning, Data Transformation and Data Reduction. Here, we handle missing values, scale the dataset, normalize if necessary and also do feature selection.

E. Train and Test Models

Model is one of classes available in the Flask-SQLAlchemy project. Model fitting is a measure of how fine a machine learning model simplify to related data to that on which it was trained. Outcomes produced by the fine-fitted model are more accurate.

F. Picking the Models

In python, the python objects are serialized and de-serialized using a python pickle module. To save data into disk, the python object must be pickled.

THEORETICAL FOUNDATIONS / ALGORITHMS

A. Adam Optimizer

Adam Optimizer combines the advantages of two other extensions of stochasticgradient descent approaches. They are:

- **Adaptive Gradient Algorithm (AdaGrad)** : It improves the performance with sparse gradients

Received 05 March 2021; Accepted 01 April 2021.

by maintaining the per-parameter learning rate.

• **Root Mean Square Propagation (RMSProp) :** It improves the performance by keeping the per-parameter learning rates that are modified found on the average of recent magnitudes of the gradients for the weight.

Adam understood the advantages of AdaGrad and RMSProp approaches. As an alternative to the parameter learning rates found in RMSProp, Adam used the parameter learning rates based on the average of the second moments of the gradients.

Adam's algorithm calculates an exponential pitiful average of the gradient and the squared gradient,

B. Stochastic Gradient Descent(SGD) optimizer

Stochastic gradient descent optimizer estimates the error gradient using the examples from the training dataset for the current state of the model. Then using the back propagation, weights of the model is updated. The step size or learning rate is the quantity of weights that are updated while training the model. Generally, the step size can be configured while training the neural networks. It takes a positive value from the range 0.0 to 1.0.

C. Momentum(used in SGD)

Including the history to the updated weight makes the training of neural network easier. Incorporating the history with the weight update is called as "momentum". The momentum algorithm builds up an exponentially progressing average of past gradients and carries on to progress in their track. Momentum can speed up learning on problems where the high-dimensional "weightspace" that is being found the way by the optimization procedure has structures that give the wrong impression about the gradient descent algorithm.

D. Dropout

Dropout is a regularization method. It finds the approximation of neural networks with different architectures in parallel while training in large numbers. when training the model, outputs produced by some layers are randomly ignored or "dropped out". Dropping of outputs affect how the layer look-like and how the layer can be treated-like with a diverse kind of nodes and how it is connected to the preceding layer. Each layer is updated with this drop out effects during training phase is done with dissimilar "view" of the configured layer.

Received 05 March 2021; Accepted 01 April 2021.

E. BackPropagation

The backpropagation algorithm is applied to train a neural network using chain rule in an efficient manner. This backpropagation performs a backward pass when adjusting the weight parameter and bias parameter of the model after the completion of each forward pass through a network. In the neural network, for each layer the weight parameter is updated.

F. Genetic Algorithm

The constrained and unconstrained optimization problems that are based on natural selection can be efficiently solved using genetic algorithms. This algorithm follows the process based on the biological progression. The genetic algorithm again and again transforms a population of entity solutions. In every step, from the current population this algorithm randomly picks an entity to be parents. Then, these randomly picked entities are used for producing the children for the subsequent generation. At the end of consecutive generations an optimal solution is found through more evolutions. The genetic algorithm can be used for solving different kind of optimization problems that cannot be solved by standard optimization algorithms. The figure-2 shows the working of a genetic algorithm.

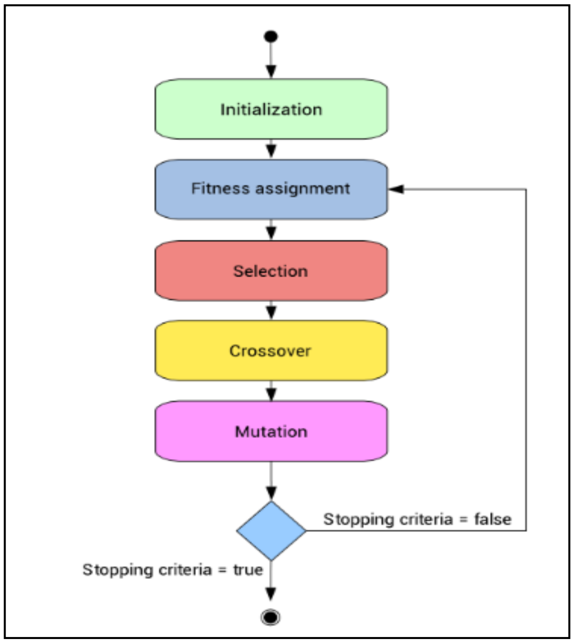

Figure-2: Genetic Algorithm overview

In gentic algorithm, for creating the future generation from the present generation the following three main rules are used.

- *Selection rules:* It selects an entity that becomes as *parents for the next generation.*
- *Crossover rules:* It merges two parent entities to create children for the consecutive generation.
- *Mutation rule:* It randomly picks a parent entity and updates it to produce children.

The following table summarizes the difference between genetic algorithm and classical optimization algorithms.

Classical Algorithm	Genetic Algorithm
In each iteration, single point is generated An optimal solution is obtained at the end of sequence of points.	In each iteration, a population of point is generated. An optimal solution is obtained at the end of sequence best point in the population.
Deterministic computation is used to pick the next point from the sequence.	It randomly picks next population point.

Table-1: Classical algorithm Vs Genetic algorithm

RESULTS AND DISCUSSION

The proposed model is compared with various existing models and the results are shown in figure-3 and table-2.

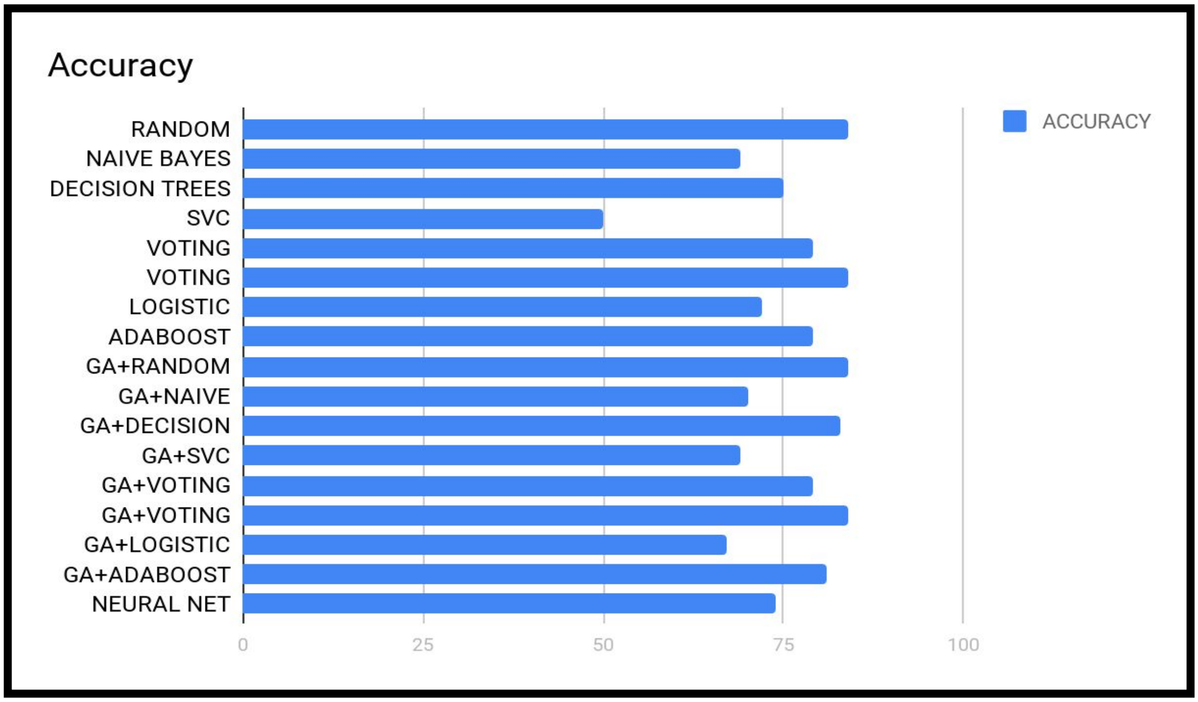

Figure-3: comparison of various algorithms based on Accuracy

ALGORITHM	ACCURACY
RANDOM FOREST	84
NAIVE BAYES	69
DECISION TREES	75
SVC	50
VOTING CLASSIFIER	79
VOTING CLASSIFIER-2	84
LOGISTIC REGRESSION	72
ADABOOST	79
GA+RANDOM FOREST	84
GA+NAIVE BAYES	70
GA+DECISION TREES	83
GA+SVC	69
GA+VOTING CLASSIFIER	79

Table-2: comparison of various algorithms based on Accuracy

From the comparison of various algorithms, we can clearly see that Random Forest performs the best followed by Voting Classifier and Adaboost among the machine learning models with accuracies of 84%, 84 %,79% respectively. Neural Nets gave a validation accuracy of 74 percent. The Genetic Algorithm improved performance of Adaboost by 3 percent, decision trees by 4 percent and SVCby 19 percent.

CONCLUSIONS AND FUTURE WORK

Identifying the liver infectivity at elementary stage is essential in India. The patients must be monitored frequently to identify the preliminary liver disease. So, the patient's can be treated in an efficient manner to avoid any complication related to the patients' life. In this paper, several approaches are investigated to predict the liver infections in patients by using various machine learning techniques. Finally we have found that Random forest performs better than other classification algorithms. We have used oversampling to handle the case of an unbalanced dataset which the base papers didn't take into account. The project also uses Genetic Algorithm for optimization and we were able improve performance of other machine learning models by reducing the features set. We have also used Neural Networks which incorporated regularization techniques like early stopping; drop out to improve the models' performance. The Neural Network, Genetic Algorithm could have performed better if there was more data and a lot more features to train with.

Future Work

A. Incorporate Future Models

We can incorporate more machine learning models which are more efficient and well-suited for our problem in the future.

B. Automating model training and deployment

We can provide a feature which allows users/developers to train a new model and deploy it on the web app. eliminating the process to manually train, pickle and integrate it.

C. Modification of the webapp

In the future, we can make several changes to websites like including a feature to send the performance analysis of a particular model or a group of models to a user's email using email server, scaling up the websites to allow many users to access the prediction mechanism simultaneously, etc.

REFERENCES

[1] Sanjay Kumar and Sarthak Katyal (2018), "Effective Analysis and Diagnosis of Liver Disorder by Data Mining", *International Conference on Inventive Research in Computing Applications (ICIRCA),* ISBN:978-1-5386-2457-9.

[2] Varun Vats, Lining Zhang, Sreejit Chatterjee, Sabbir Ahmed, Elvin Enziama and Kemal Tepe (2018), "A Comparative Analysis of Unsupervised Machine Techniques for Liver Disease Prediction", *IEEE International Symposium on Signal Processing and Information Technology (ISSPIT),* ISSN: 2162-7843.

[3] L. Alice Auxilia (2018), "Accuracy Prediction using Machine Learning Techniques for Indian Patient Liver Disease", *2nd International Conference on Trends in Electronics and Informatics (ICOEI),* ISBN:978-1-5386-3571-1.

[4] Ashwani Kumar, Neelam Sahu (2017), "Categorization of Liver Disease Using Classification Techniques", *International Journal for Research in Applied Science & Engineering Technology (IJRASET),* Volume 5 Issue V.

[5] Chandrasegar Thirumalai and Rashad Manzoor (2017), "Cost Optimization using Normal Linear Regression Method for Breast Cancer Type I Skin", *International Conference on Electronics, Communication and Aerospace Technology ICECA 2017,* ISBN:978-1-5090-5687-3.

[6] Harsha Pakhale, Deepak Kumar Xaxa (2016), "Development of an Efficient Classifier for Classification of Liver Patient with Feature Selection", *International Journal of Computer Science and Information Technologies*, Vol. 7 (3), pp. 1541-1544.

[7] Dr. S. Vijayarani and Mr.S.Dhayanand (2015), "Liver Disease Prediction using SVM and Naïve Bayes Algorithms*", International Journal of Science, Engineering and Technology Research (IJSETR),* Volume 4, Issue 4.

[8] Anju Gulia, Dr. Rajan Vohra , Praveen Rani (2014), "Liver Patient Classification Using Intelligent Techniques", *International Journal of Computer Science and Information Technologies,* Vol. 5 (4), pp. 5110-5115.

[9] Jankisharan Pahareeya, Rajan Vohra, Jagdish Makhijani , Sanjay Patsariya (2014), "Liver Patient Classification using Intelligence Techniques", *International Journal of Advanced Research in Computer Science and Software Engineering*, Volume 4, Issue 2.

[10] Bendi Venkata Ramana, Surendra. Prasad Babu. M, Venkateswarlu. N.B (2011), "A Critical Studyof Selected Classification Algorithms for Liver Disease Diagnosis", *International Journal of Database Management Systems (IJDMS)*, Vol.3, No.2, pp. 101-114.

[11] Silva PRL, Freitas Neto OC, Laurentiz AC, Junqueira OM, Fagliari JJ (2007), "Blood serum components and serum protein test of Hybro-PG broilers of different ages", *Brazilian Journal of Poultry Science*, vol.9,No.4.