J Ramesh ,INDIA / International Journal of Research and Computational Technology
Vol.14 Issue.1 Free Journal Publication Pages: 27- 35
ISSN: 0975-5662, June, 2022 www.ijrct.com

A WEB BASED FRAMEWORK FOR LIVER DISEASE DIAGNOSIS USING MACHINE LEARNING MODELS

#J.RAMESH[1],Email:rameshjarapala@gmail.com
#A.Sushmitha[2],Email:sushmithar640@gmail.com
A.Abhinay kumar[3],Email:ammulaabhinay8888@gmail.com
#CH. Bala siva sai[4] Asst. Professor, TKR College of Engineering and Technology, Hyderabad,INDIA.
#BTech Student, TKR College of Engineering and Technology, Hyderabad,INDIA.

Abstract: The liver is that the second-largest organ in human bodies. It's regarding the scale of a football game and sits just under your ribs on the right side. The burden of the liver is regarding one. The weight of the liver is 36 kilogram and reddish brown in colour. Liver health problem might find yourself in liver failure or malignancy. Fully completely different reasons finish in different types of illness. Viral infections, immune system disorders, hereditary diseases, cancer, and additionally the uptake of too many toxics, among various things, can cause illness. Symptoms of some forms of liver disease are uncommon. The foremost current symptom is jaundice, that's characterised by a yellowing of the skin. Abdominal pain, bruising merely, changes in excretion colour, exhaustion, nausea, and swelling inside the arms or legs are all indicators of illness. Certain forms of illness can raise your possibilities of getting liver disease. Others will still hurt the liver if left untreated. As a result, liver health problem got to be detected timely and treated appropriately.

Liver disorders are considered as 2nd leading cause of death among all digestive disorders in the US and the 5th most common cause of death in the UK. Liver diseases accounts for approximately 2 million deaths annually and cirrhosis is currently the 11th most common cause of death globally.

1. INTRODUCTION

The liver is the body's largest solid organ. It serves as a filter for gastrointestinal blood, which has a high concentration of toxins and antigens produced by the body. The liver also detoxifies toxins, metabolizes medications, and produces proteins that are necessary for blood clotting and other bodily activities, making it a vital organ. The dangers of liver disease are significant, and organ failure is unavoidable unless it is detected early on. Liver disease may associate with several symptoms, making it challenging to identify it quickly and accurately. However, because the liver functions normally even when partially injured, patient's issues are difficult to detect. Machine learning approaches can be used to overcome this problem.

The purpose of "WEB BASED FRAMEWORK FOR LIVER DISEASE DIAGNOSIS USING MACHINE LEARNING MODELS" is to improve liver disease diagnosis using machine learning approaches.

2. LITERATURE SURVEY

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, it is clear that Random Forest algorithm provides better performance as compared to Naive Bayes, AdaBoost, J48 and Bagging.

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 and KNN 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 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 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.

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.

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

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 algorithm surpassed all other strategies with an accuracy of 71.8696 percent, according to the findings of our study.

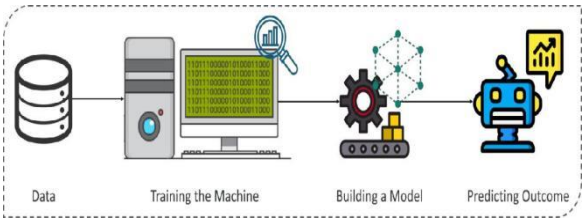
III. PROPOSED SYSTEM

In the existing system, Liver failures are at high rate of risk among Indians. Existing system used feature selection and preprocessing to cluster data. Various techniques like Ada Boost, Bagging, Random Forest were applied on the modified data. But these approaches to classify liver data is time consuming and is inefficient. The following are the drawbacks

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

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.

Machine learning is a subset of artificial intelligence. Machine learning is a growing technology which enables computers to learn automatically from past data. Machine learning uses various algorithms for building mathematical models and making predictions using historical data. The detailed process is shown in the following figure



Machine Learning algorithm is trained using a training data set to create a model. When new input data is introduced to the ML algorithm, it makes a prediction on the basis of the model.

The prediction is evaluated for accuracy and if the accuracy is acceptable, the Machine Learning algorithm is deployed. If the accuracy is not acceptable, the Machine Learning algorithm is trained again and again with an augmented training data set.

The Machine Learning process involves building a Predictive model that can be used to find a solution for a Problem Statement. To understand the Machine Learning process let's assume that you have been given a problem that needs to be solved by using Machine Learning.

The below steps are followed in a Machine Learning process:

Step 1: Define the objective of the Problem Statement

At this step, we must understand what exactly needs to be predicted. In our case, the objective is to predict the possibility of rain by studying weather conditions. At this stage, it is also essential to take mental notes on what kind of data can be used to solve this problem or the type of approach you must follow to get to the solution.

Step 2: Data Gathering
At this stage, you must be asking questions such as,

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

Once you know the types of data that is required, you must understand how you can derive this data. Data collection can be done manually or by web scraping. However, if you're a beginner and you're just looking to learn Machine Learning you don't have to worry about getting the data. There are 1000s of data resources on the web, you can just download the data set and get going.

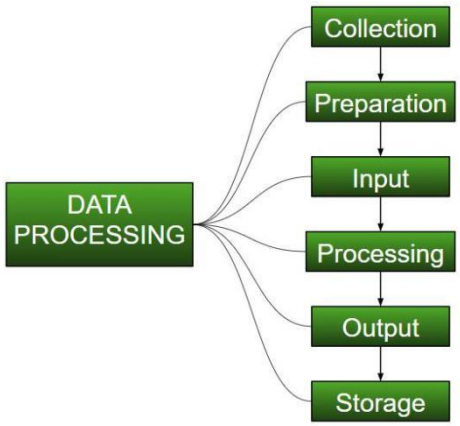
Coming back to the problem at hand, the data needed for weather forecasting includes measures such as humidity level, temperature, pressure, locality, whether or not you live in a hill station, etc. Such data must be collected and stored for analysis.

Step 3: Data Preparation

The data you collected is almost never in the right format. You will encounter a lot of inconsistencies in the data set such as missing values, redundant variables, duplicate values, etc. Removing such inconsistencies is very essential because they might lead to wrongful computations and predictions. Therefore, at this stage, you scan the data set for any inconsistencies and you fix them then and there.

Data Processing:

Data Processing is a task of converting data from a given form to a much more usable and desired form i.e., making it more meaningful and informative. Using Machine Learning algorithms, mathematical modeling and statistical knowledge, this entire process can be automated. The output of this complete process can be in any desired form like graphs, videos, charts, tables, images and many more, depending on the task we are performing and the requirements of the machine. This might seem to be simple but when it comes to really big organizations like Twitter, Facebook, Administrative bodies like Parliament, UNESCO and health sector organizations, this entire process needs to be performed in a very structured manner.



Collection:

The most crucial step when starting with ML is to have data of good quality and accuracy. Data can be collected from any authenticated source like data.gov.in, Kaggle or UCI dataset repository. For example, while preparing for a competitive exam, student's study from the best study material that they can access so that they learn the best to obtain the best results. In the same way, high-quality and accurate data will make the learning process of the model easier and better and at the time of testing,the model would yield state of the art results.

A huge amount of capital, time and resources are consumed in collecting data. Organizations or researchers have to decide what kind of data they need to execute their tasks or research. Example: Working on the Facial Expression Recognizer, needs a large number of images having a variety of human expressions. Good data ensures that the results of the model are valid and can be trusted upon.

Preparation

The collected data can be in a raw form which can't be directly fed to the machine. So, this is a process of collecting datasets from different sources, analyzing these datasets and then constructing a new dataset for further processing and exploration. This preparation can be performed either manually or from the automatic approach. Data can also be prepared in numeric forms also which would fasten the model's learning. **Example:** An image can be converted to a matrix of N X N dimensions; the value of each cell will indicate image pixel

Input:

Now the prepared data can be in the form that may not be machine-readable, so to convert this data to readable form, some conversion algorithms are needed. For this task to be executed, high computation and accuracy is needed. Example: Data can be collected through the sources like MNIST Digit data (images), twitter comments, audio files, video clips.

Processing:

This is the stage where algorithms and ML techniques are required to perform the instructions provided over a large volume of data with accuracy and optimal computation.

Output:

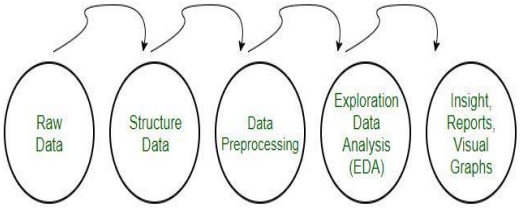
In this stage, results are procured by the machine in a meaningful manner which can be inferred easily by the user. Output can be in the form of reports, graphs, videos, etc.

Storage:

This is the final step in which the obtained output and the data model data and all the useful information are saved for the future use.

Data Preprocessing for Machine learning in Python

- Pre-processing refers to the transformations applied to our data before feeding it to the algorithm.
- Data Preprocessing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis.



Need of Data Preprocessing

• For achieving better results from the applied model in Machine Learning projects the format of the data has to be in a proper manner. Some specified Machine Learning model needs information in a specified format, for example, Random Forest algorithm does not support null values, therefore to execute random forest algorithm null values have to be managed from the original raw data set.

- Another aspect is that data set should be formatted in such a way that more than one Machine

Learning and Deep Learning algorithms are executed in one data set, and best out of them is chosen. Rescale Data

- When our data is comprised of attributes with varying scales, many machine learning algorithms can benefit from rescaling the attributes to all have the same scale.
- This is useful for optimization algorithms in used in the core of machine learning algorithms like gradient descent.
- It is also useful for algorithms that weight inputs like regression and neural networks and algorithms that use distance measures like K-Nearest Neighbors.
- We can rescale your data using scikit-learn using the Min Max Scaler class.

Binarize Data

- We can transform our data using a binary threshold. All values above the threshold are marked 1 and all equal to or below are marked as 0.
- This is called binarizing your data or threshold your data. It can be useful when you have probabilities that you want to make crisp values. It is also useful when feature engineering and you want to add new features that indicate something meaningful.
- We can create new binary attributes in Python using scikit-learn with the binarized class.

Standardize Data

- Standardization is a useful technique to transform attributes with a Gaussian distribution and differing means and standard deviations to a standard Gaussian distribution with a mean of 0 and a standard deviation of 1.
- We can standardize data using scikit-learn with the standard scaler class.

Data Cleansing

Introduction:

Data cleaning is one of the important parts of machine learning. It plays a significant part in building a model. Data Cleaning is one of those things that everyone does but no one really talks about. It surely isn't the fanciest part of machine learning and at the same time, there aren't any hidden tricks or secrets to uncover. However, proper data cleaning can make or break your project. Professional data scientists

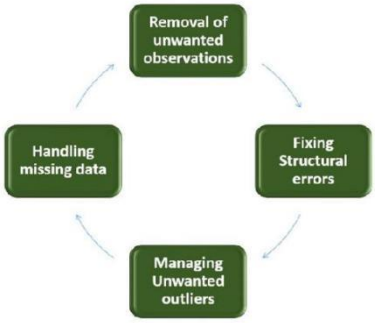
usually spend a very large portion of their time on this step.

Because of the belief that, “Better data beats fancier algorithms”.

If we have a well-cleaned dataset, we can get desired results even with a very simple algorithm, which can prove very beneficial at times.

Obviously, different types of data will require different types of cleaning. However, this systematic approach can always serve as a good starting point.

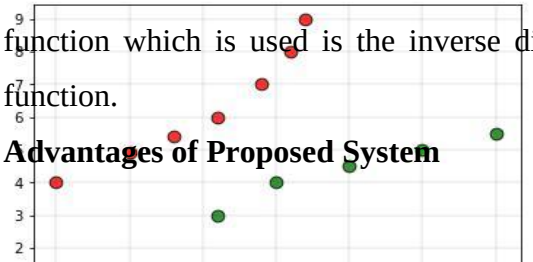
Steps involved in Data Cleaning



1. Removal of unwanted observations
2. Fixing Structural errors
3. Managing Unwanted outliers
4. Handling missing data

Weighted K-Nearest Neighbor algorithm:

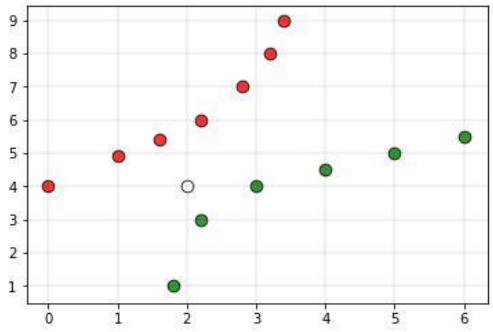
Weighted KNN is a modified version of K-Nearest Neighbor algorithm. One of the main issues that affect the performance of the KNN algorithm is the choice of the hyperparameter k. If k is too small, the algorithm would be more sensitive to outliers. If k is too large, then the neighborhood may include too many points from other classes.



function which is used is the inverse distance function.

Advantages of Proposed System

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



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, it is clear that 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

- The performance classification of liver diseases is further improved.
- The system is fast and efficient.

IV. Implementation and Testing

from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test=train_test_split(X, y ,test_size = 0.1 random_state = 42)

print("Train Set: ", X_train.shape, y_train.shape)

curacy is {round(accuracy_score(y_test, neigh.predict(X_test))*100,2)}")
from sklearn import svm
clf = svm.SVC()clf.fit(X_train, y_train)print(f"Accuracy is {round(accuracy_score(y_test, X_test))*100,2)}")
estimate says that 50% of whole software development process should be tested. Errors may ruin
the software from critical level to its own removal. Software testing is done while coding by the developers and thorough testing is conducted by testing experts at various levels of code such as module testing, program testing, product testing, in-house testing and testing the product at user's end. Early discovery of errors and their remedy is the key to reliable software. There are three types of possible test outcomes:

- OK – This means that all the tests are passed
- FAIL – This means that the test did not pass and an Assertion Error exception is raised.
- ERROR – This means that the test raises an exception other than Assertion Error.

10. Conclusion:

In this project, we have proposed methods for diagnosing liver disease in patients using machine learning techniques. The four machine learning techniques that were used include Random Forest, KNN and SVM. The system was implemented using all the models and their performance was evaluated. Performance evaluation was based on certain performance metrics. SVM was the model that resulted in the highest accuracy with an accuracy of 76%. Comparing this work with the previous research works, it was discovered that SVM proved highly efficient. A GUI, which can be used as a medical tool by hospitals and medical staff was implemented using SVM

10. Future Scope

The proposed system can be developed in many different directions which have vast scope for improvements in the system. These includes:

1. Increase the accuracy of the algorithms.
2. Improvising the algorithms to add more efficiency of the system and enhance its working.
3. Working on some more attributes so to tackle liver disease even more.

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 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,