LIVER DISEASE DETECTION AND DIAGNOSIS

[1]G Kiran Kumar, [2]B. Nirupa, [3]K.Preetham Kumar, [4]S. Vaishnavi

[1] Assistant professor, Department of Computer Science and Engineering
Anurag Group of Institutions, Hyderabad, Telangana, India

2,3,4 Department of Computer Science and Engineering
Anurag Group of Institutions, Hyderabad, Telangana, India

ABSTRACT- ***Disease detection & recognition has grabbed a lot of attention & interest from researchers due to the rapid increase of advancement & improvisation in the technology of computer vision stream. Even though there have been a lot of disease detection & recognition systems in the market, projects are still in demand and there is also a need for new, better & accurate recognition Systems. The patient traffic to the diagnosis center is increasing day by day. To reduce the burden for doctors and the time for the patients we introduce a system whose main objective is to detect & recognize if the liver has been damaged or not. We're implementing a machine learning model which states the condition of a liver by considering chemical compounds in the liver such as alkaline, protein, bilirubin, and many others and detects liver condition with high accuracy.***

Keywords: KNN, machine learning, Neural Networks, Logistic Regression, Random forest.

1. INTRODUCTION

The liver is the largest organ of the body and it is essential for digesting food and releasing the toxic element of the body. The viruses and alcohol use lead the liver towards liver damage and lead a human to a life-threatening condition. There are many types of liver diseases whereas hepatitis, cirrhosis, liver tumors, liver cancer, and many more. Among them liver diseases and cirrhosis as the main cause of death. Therefore, liver disease is one of the major health problems in the world. Every year, around 2 million people died worldwide because of liver disease. According to the Global Burden of Disease (GBD) project, published in BMC Medicine, one million people are died in 2010 because of cirrhosis, and millions are suffering from liver cancer. Machine learning has made a significant impact on the biomedical field for liver disease prediction and diagnosis.

The extents of their success ultimately depend on how well the real-time situations go with the kind of learning model the project uses and the factors that it considers in the data set. This field is a subject of a lot of research still, allowing umpteen scope to pick the combinations of various strategies and factors apply them to bring out a model that suits the requirements the best. Data Analysis is a process of inspecting cleansing, modeling data to discover useful information and conclusions. It is a process of analyzing, extracting, and predicting meaningful information from huge data to extract some pattern. This process is used by companies to turn the raw data of their customer into useful information. This analysis can also be used in the field of Medicine.

EXPERMENTAL

Liver Disease detection using machine learning – the proposed system is been developed by K-Nearest Neighbor (KNN), Linear Regression and Random forest algorithms where the data is collected (i.e. Preexisting data) from the agricultural department regarding crops and their significant conditions/requirements which are important to get efficient yield.

The data collected undergoes data cleaning to remove any gaps or redundant data. The data is divided into training and testing data sets, the software used to perform the task is Anaconda Jupiter notebook and the language used is python 3.0. the data set is being imported to the canvas.

KNN is a type of classification where the function is only approximated locally and all computation is deferred until function evaluation. Since this algorithm relies on distance for classification if the features represent different physical units or come in vastly different scales then normalizing the training data can improve its accuracy dramatically.

Both for classification and regression, a useful technique can be to assign weights to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists of giving each neighbor a weight of 1/d, where d is the distance to the neighbor.

The neighbors are taken from a set of objects for which the class (for KNN classification) or the object property value (for KNN regression) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required. A peculiarity of the KNN algorithm is that it is sensitive to the local structure of the data.

LOGISTIC REGRESSION is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (logit regression) is estimating the parameters of a logistic model (a form of binary regression). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labeled "0" and "1". In the logistic model, the log-odds (the logarithm of the odds) for the value labeled "1" is a linear combination of one or more independent variables ("predictors"); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling; the function that converts log-odds to probability is the logistic function, hence the name. The unit of measurement for the log-odds scale is called a logit, from the logistic

unit, hence the alternative names. Analogous models with a different Sigmond function instead of the logistic function can also be used, such as the probit model; the defining characteristic of the logistic model is that increasing one of the independent variables manipulatively scales the odds of the given outcome at a constant rate, with each independent variable having its parameter; for a binary dependent variable, this generalizes the odds ratio.

RANDOM FOREST: builds multiple decision trees and merges them to get a more accurate and stable prediction. The random forest has nearly the same hyperparameters as a decision tree or a bagging classifier. ... Random forest adds additional randomness to the model while growing the trees.

RELATED WORKS

P.Rajeswari, G.SophiaReenahave proposed the data classification is based on liver disorder. The training dataset is developed by collecting data from the UCI repository consists of 345 instances with 7 different attributes. This paper deals with results in the field of data classification obtained with Naïve Bayes algorithms.FT tree algorithms, and KStar algorithms and on the whole performance made know FT Tree algorithm when tested on liver disease datasets, time is taken to run the data for a 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 FT Tree algorithm compare to other algorithms.

Sa'diyah Noor NovitaAlfisahrin, Teddy Montoro, et al., [2013] have proposed to identify if the patients have liver disease 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 study, the performance of the NB Tree algorithm will be the target of improvement of the accuracy by finding the most significant factor in identifying liver disease patients. For future study, the performance of the NB Tree algorithm will be the target of improvement of the accuracy by finding the most significant factor in identifying liver disease patients.

S.Dhamodharan [2014] has proposed many liver disorders require the clinical care of the physician. They predict 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". In this paper, Naïve Bayes and FT tree algorithm accuracy are compared and the result is obtained. The result concludes that the accuracy of the Naïve Bayes algorithm is much better than the other algorithms.

DATA AND METHODS

DATA COLLECTION – The 1st module represents the data collection, the data is collected from the medical department regarding the fields of liver disease (i.e.direct_bilrubin, Total_bilrubin, Alkaline_phosphotase, Alamine_aminotranferas, Aspertate_aminotransferase, Total_proteins, Albumin, etc). The data is stored in the form of tables which is the database for the project.

DATA CLEANING – All the zero values in the data set will be removed and all the multi-valued will be removed.

VISUALIZATION- In this visualization, all the values and attributes will be displayed in a step-wise manner.

DEFINING THE MODEL – This includes data cleaning, data visualization, data-driven modeling, and running algorithms. After the whole process comparison of the models is done.

OUTPUT PREDICTION - The output is predicted which is the estimated output for disease detection

Logistic Function

Logistic regression is named for the function used at the core of the method, the logistic function.

The logistic function, also called the sigma function was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. It's an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.

$$1 / (1 + e^{-value})$$

Where e is the base of the natural logarithms (Euler's number or the EXP() function in your spreadsheet) and value is the actual numerical value that you want to transform. Below is a plot of the numbers between -5 and 5 transformed into the range 0 and 1 using the logistic function. Now that we know what the logistic function is, let's see how it is used in logistic regression.

Representation Used for Logistic Regression

Logistic regression uses an equation as the representation, very much like linear regression.

Input values (x) are combined linearly using weights or coefficient values (referred to as the Greek capital letter Beta) to predict an output value (y). A key difference from linear regression is that the output value being modeled is a binary value (0 or 1) rather than a numeric value. Below is an example logistic regression equation:

$$y = e^{(b0 + b1*x)} / (1 + e^{(b0 + b1*x)})$$

Where y is the predicted output, b0 is the bias or intercept term and b1 is the coefficient for the single input value (x). Each column in your input data has an associated b coefficient (a constant real value) that must be learned from your training data. The actual representation of the model that you would store in memory or a file is the coefficients in the equation (the beta value or b's).

The Random Forest Classifier

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction (see figure below). Visualization of a Random Forest Model Making a Prediction

The fundamental concept behind random forest is a simple but powerful one — the wisdom of crowds. In data science-speak, the reason that the random forest model works so well is:

METRICS - From the above section, it is understood that different machine learning algorithms are being used for the prediction of crop yield. Though different machine learning algorithms are available for use, the selection of a particular algorithm is based on the nature of the application and the accuracy of the prediction algorithm. Prediction accuracy of classifiers is validated by different metrics as Mean Absolute Error (MAE)[8], Root Mean Squared Error(RMSE), Mean Absolute Error(MAE). Here we consider the metric mean squared error. Mean Squared Error(MSE) measures the average squared error of predictions. That is, it calculates the square of the difference between the predicted value and actual value and then averages those values. Computation of MSE is

shown.Here 'y' denotes the expected output and 'yin denotes the predicted output for an ith data and 'i' ranges from 1 to N. Low value of MSE indicates that the accuracy of the classifier is good. For a perfect classifier, the value of MSE would be zero.
The software requirements for the project to be developed are –

Matplotlib This library is used for graph and visualization purposes.

Numpy This library is used to arrange data in one direction i.e. either row or column.

ScipyScipy in python is an open-source library used for solving mathematical, scientific, engineering, and technical problems. It allows users to manipulate the data and visualize the data using a wide range of high-level Python commands.

Seaborn, This is a python library used for data analysis.

Python 3.0 Python programming language is used to develop the code. Python 3.0 was released in 2008. Although this version is supposed to be backward-incompatible, later on, many of its important features have been backported to be compatible with version 2.7 This tutorial gives enough understanding of the Python 3 version programming language.

CONCLUSION

The present project model demonstrates the potential use of machine learning techniques in predicting liver disease based on the required parameters. The accuracy of the prediction of this model is high for given parameters. This model can continue to be an estimate the percentage of the liver that is damaged and help both doctors and patients. In this work, different methods of machine learning were utilized to model liver disease based on a feature such as bilirubin, Total_bilrubin, Alkaline_phosphotase, Alamine_aminotranferase, Aspertate_aminotransferase, Total_proteins, Albumin. In this work, few methods were implemented and compared. The models are sorted from simplest to most complicated and all of them were trained well. The coefficient of the linear model indicates which features have the most effect on the output. For instance, the pesticides squared has the strongest effect, and the region 2 (a dummy variable) intercept has the weakest effect.

REFERENCE

1. https://www.medicinenet.com/liver_disease/articl e.html
2. P. Rajeswari, G. Sophia Reena, Analysis of Liver Disorder Using Data Mining Algorithm, Global Journal of Computer Science and Technology,2010.
3. Sa'sdiyah Noor NovitaAlfisahrin, Teddy Montoro, Data mining Techniques For Optimization of Liver Disease Classification, International conference on advanced Computer Science Application and Technologies,2013.
4. A. Mallikarjuna Reddy, V. Venkata Krishna, L. Sumalatha," Face recognition based on stable uniform patterns" International Journal of Engineering & Technology, Vol.7 ,No.(2),pp.626-634, 2018,doi:10.14419/ijet.v7i2.9922 .
5. Mallikarjuna Reddy, V. Venkata Krishna, L. Sumalatha, "Efficient Face Recognition by Compact Symmetric Elliptical Texture Matrix (CSETM)", Jour of Adv Research in Dynamical & Control Systems, Vol. 10, 4-Regular Issue, 2018.
6. S. Dhamodharan, Liver Disease Prediction Using Bayesian Classification, National Conference on Advanced Computing, Application&Technologies,2014
7. S.E.Sekar, Y.Unal, Z.Erdem, and H.ErdincKocer, Ensembled Correlation Between Liver Analysis Output, International Journal of Biology and Biomedical engineering, ISSN:1998-4150.
8. Ayaluri MR, K. SR, Konda SR, Chidirala SR. 2021. Efficient steganalysis using convolutional auto encoder network to ensure original image quality. PeerJ Computer Science 7:e356 https://doi.org/10.7717/peerj-cs.356
9. A Mallikarjuna Reddy, VakulabharanamVenkata Krishna, LingamguntaSumalatha and AvukuObulesh, "Age Classification Using Motif and Statistical Features Derived On Gradient Facial Images", Recent Advances in Computer Science and Communications (2020) 13: 965. https://doi.org/10.2174/2213275912666190417151247.
10. A.S.Aneesh Kumar, Dr.C.JothiVenkateswaran, A novel approach for Liver disorder Classification using Data Mining Techniques, Engineering, and Scientific International Journal, ISSN 2394- 7179, ISSN 2394-7187,2015.
11. **P.** Thangarajul, R.Mehala, Performance Analysis of PSO-KStar Classifier over Liver Diseases, International Journal of Advanced Research in Computer Engineering, 2015.

12. Swarajya Lakshmi V Papineni, SnigdhaYarlagadda, HaritaAkkineni, A. Mallikarjuna Reddy. Big Data Analytics Applying the Fusion Approach of Multicriteria Decision Making with Deep Learning Algorithms *International Journal of Engineering Trends and Technology*, 69(1), 24-28, doi: 10.14445/22315381/IJETT-V69I1P204.
13. Srinivasa Reddy, K., Suneela, B., Inthiyaz, S.,Kumar, G.N.S., Mallikarjuna Reddy, A." Texture filtration module under stabilization via random forest optimization methodology "International Journal of Advanced Trends in Computer Science and Engineering, Volume 8, No.3, May - June 2019.
14. Onwodi Gregory, Prediction of Liver Disease (Biliary Cirrhosis) Using Data Mining Technique, International Journal of EmergingTechnology&Research, ISSN (E):2347-5900, ISSN (P):2347-6079, 2015.