Machine Learning Approach to Predict the Presence of Liver Disease

Kiran More, Shantanu Sontakke, Atharva Tayade, Prajwal Surve, Utsav Tayde, Dnyaneshwari Shriram, Shashank Joshi

Department of Engineering, Sciences and Humanities (DESH)

***Abstract* —** Kaggle dataset containing the data of 196 Indian liver patients was taken, consisting of the attributes related to liver as in the liver function tests. 10 features were chosen and the Decision Tree algorithm and Random Forest algorithm were applied to them. The corresponding labels consisted of whether the patients had a liver disease or not. The data was also visualised by reducing the 10 dimensional graph to 2 dimensions using the t-SNE algorithm and the accuracy of the predictions was calculated. Also, GUI programming was done in order to allow the patients to enter their own data in order to check whether they may have a liver related problem or not.

Keywords* —** ***machine learning, classification, liver disease, python, patients

I. INTRODUCTION

Machine Learning and artificial intelligence has completely revolutionized the world. They have their applications in almost every domain and are very promising in solving the tiniest of problems at high speeds and great accuracies. Machine learning has had a huge impact on the healthcare sector, and have gone beyond expectations in discovering patterns, detecting diseases, discovering drugs and medicines, using computer vision and deep learning on images for disease diagnosis, prediction of potential diseases, etc. Machine learning is basically a branch of Artificial Intelligence that provides systems the ability to automatically learn and improve from experience using mathematical algorithms without being explicitly programmed.It also focuses on the development of computer programs that can access data and themselves use it for learning.

By analyzing health records, Sutter Health and Georgia Institute of Technology researchers showed that they could predict heart failure as much as nine months before doctors using traditional means. Researchers and startups are using GPU-accelerated deep learning to automate analysis and increase the accuracy of diagnosticians. Behold.ai is a New York startup working to reduce the number of incorrect diagnoses by making it easier for healthcare practitioners to identify diseases from ordinary radiology image data. Arterys, a San Francisco-based startup, provides technology to visualize and quantify heart flow in the body using any MRI machine to help speed diagnosis. San Francisco startup Enlitic analyzes medical images to identify tumors, nearly invisible fractures, and other medical conditions.

Machine learning is broadly classified into three types- Supervised learning, Unsupervised learning and Reinforcement learning. Supervised learning deals with the type of machine learning in which the data is labelled whereas unsupervised learning does not have labelled data and is a bit more challenging. Supervised learning is divided into two types- classification and regression. Classification has a categorical or discrete values associated with it whereas regression has a continuos output. Companies use it to classify whether a Unsupervised learnin is g broadly divided into clustering and association. In clustering, similar groups are discovered in the data whereas association is done to discover rules that describe a large portions of data. Rinforcement learning is about taking suitable action to maximize reward in a particular situation. It is employed by

softwares and machines to find the best possible behavior or path it should take in a specific situation.

It is estimated that there are about 2 million deaths worldwide due to liver disease. About 1 million die due to complications of cirrhosis. Cirrhosis is the 11th most common cause of death globally and liver cancer is the 16th most common cause of death.About 2 billion people consume alcohol worldwide and nearly 75 million of them are at the risk of alcohol-related liver disease. The global prevalence of viral hepatitis is high, and drug-induced liver injury is still increasing to be a major cause for acute hepatitis. Alarmingly, less than 10% of global liver transplantation needs are met at current rstes.They are very important to address quickly to improve public health as most of the causes of liver diseases are preventable.

II. LITERATURE REVIEW

P. shimpi, S. Shah, M. Shroff and A. Godbole (2017), studied the classification of cardiac arrythmia patients using Random Forest, SVM, Logistic Regression and KNN classifiers and achieved 91.2% accuracy by SVM classifier[1].

S Tharaha and K Rashika (2017), studied the Hybrid Artificial Neural network and Decision Tree algorithm for Disease Recognition and Prediction in Human Blood Cells and concluded that the performance level of the hybrid algorithm of decision tree and artificial neural network had better performance than the individual algorithms[2].

M. Somvanshi, S. Tambade, P.Chavan and S.V. Shinde (2016), reviewed the machine learning techniques Decision Tree and Support Vector Machines[3].

H. Ayeldeen, O. Shaker, G. Ayeldeen and K.M. Anwar (2015) studied the prediction of Liver Fibrosis stages by Machine Learning using Decision Tree algorithm and achieved 93.7% accuracy in its prediction[4].

J.H. Oh, R. Al-Lozi and I. El Naqa (2009) studied the application of machine learning techniques for the prediction of radiation Pneumonitis in Lung Cancer patients[5].

S.K. Asrani and J.Hepatol (2019), how the massive burden of liver diseases is affecting the world[8].

III. TOOLS USED FOR THE STUDY

Python and machine learning were used to examine the Kaggle dataset of 196 liver patients. Decision Tree and Random Forest Algorithms were primarily used to train the machine on the data

IV. METHODOLOGY/EXPERIMENTAL

A. Decision Tree Algorithm

Decision tree algorithm uses a decision tree to go from observations about an item that are represented in the branches, to conclusions about the item's target value that are represented in the leaves. It is one of the predictive modeling approaches used in statistics and machine learning. Tree models where the target variable can take a discrete set of values are called classification trees; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels.

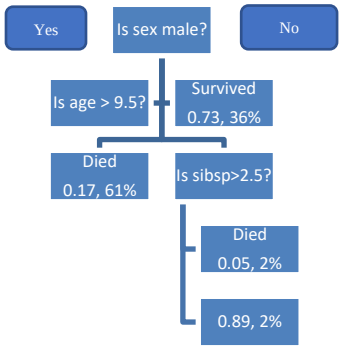


Fig. 1: Example of decision tree (The figures in leaves show the probability of survival and the percentage of observations in the leaf)

The decision tree uses a top-down approach asking at each node "Which feature or attribute should be tested at each node?" To answer this, each feature is evaluated using a statistical test to determine how well it alone classifies the training examples. A descendant of the root node is then created for each possible value of this attribute, and the training examples are sorted to the appropriate descendant node.

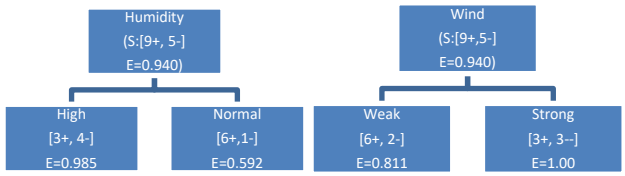
A statistical property called Information gain is used to select which attributes or features to test at each node in the tree. It is simply the expected reduction in entropy caused by partitioning the examples according to this feature. The information gain, Gain(S, A) of a feature A, relative to a collection of examples S, is defined as-

$$Gain(S,A) \equiv Entropy(S) \sum_{v \in Values(A)} \frac{|Sv|}{|S|} Entropy(Sv) \quad (1)$$

Given a collection S, containing positive and negative examples of some target concept, the entropy of S relative to this boolean classification is defined as-

$$Entropy(s) \equiv -p_{\oplus} log_2 p_{\oplus} - p_{\ominus} log_2 p_{\ominus} \quad (2)$$

The entropy is 1 when the collection contains an equal number of positive and negative examples. If the collection contains unequal numbers of positive and negative examples, the entropy is between 0 and 1.



Gain(S,Humidity)=0.940-(7/14).985-(7/14).592 =0.151

Gain(S, Wind)=0.940-(8/14).811-(6/14)1.0 =0.048

Fig. 2: Selecting best feature at node

This algorithm has various advantages-

- Simple to understand and interpret
- Able to handle both numerical and categorical data
- Requires little data preparation. Other techniques often require data normalization
- Performs well with large datasets
- Mirrors human decision making more closely than other approaches

It also has some disadvantages-

- Trees can be very non-robust. A small change in the training data can result in a large change in the tree
- Decision-tree learners can create over-complex trees that do not generalize well from the training data. This is called overfitting
- Mechanisms such as pruning are necessary to avoid the problem of overfitting

B. Random Forest

Random forest or random decision forests are an ensemble learning method for classification and regression. It works by constructing multiple decision trees at the training time and outputting the class that is mode of classes for classification i.e. the majority of what each decision trees predict by summing the number of labels predicted by them. Random decision forests correct for decision trees' habit of overfitting to their training set.

The training algorithm for random forests applies the general technique of bagging, to tree learners. Given a training set $X = x_1, ..., x_n$ with responses $Y = y_1, ..., y_n$, bagging repeatedly (B times) selects a random sample with replacement of the training set and fits trees to these samples: For $b = 1, ..., B$:

1) Sample, with replacement, n training examples from X, Y; call these X_b, Y_b.
2) Train a classification or regression tree f_b on X_b, Y_b.

Then the output in classification is given by just taking a majority vote among them. This leads to better model performance because it decreases the variance of the model, without increasing the bias. This means that while the predictions of a single tree are highly sensitive to noise in its training set, the average of many trees is not, as long as the trees are not correlated.

C. Synthesis/Algorithm/Design/Method

Ten features were selected from the dataset as follows-

1. Age of the patient
2. Gender of the patient
3. Total Bilirubin mg/dL
4. Direct Bilirubin mg/dL
5. Alkaline Phosphotase IU/L
6. Alamine Aminotransferase(SGPT) IU/L
7. Aspartate Aminotransferase (SGOT) IU/L
8. Total Protiens g/dL

9. Albumin g/dL
10. Albumin and Globulin Ratio A/G ratio

After preprocessing the data, the selected features were trained with the label 'Dataset' of which 1 meant presence of a liver disease and 2 meant absence of liver disease.
The data was then randomly split into 80% of training data and 20% testing data. Decision Tree classifier and Random Forest classifier was used to train the model. For this csv, pandas and sklearn libraries were used.

Using tkinter package, GUI programming was done to create the following window in which the patient can enter the values of their liver function tests.

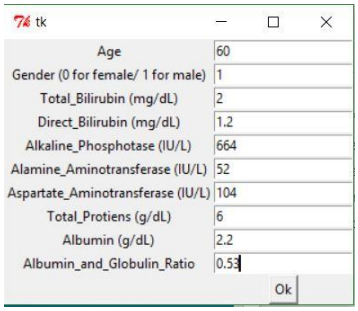


Fig. 3: Pop-up window 1

The following window pops up if the predicted output is 1:

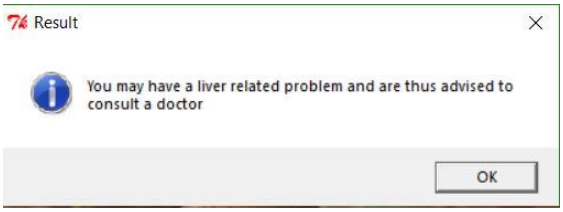


Fig. 4: Pop-up window 2

And this window pops up if the output is 2:

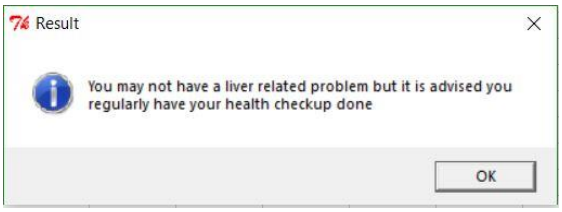


Fig. 5: Pop-up window 3

To visualize the data, the 10-dimensional data is projected onto the plane of Direct-Bilirubin and Age using the matplotlib library-

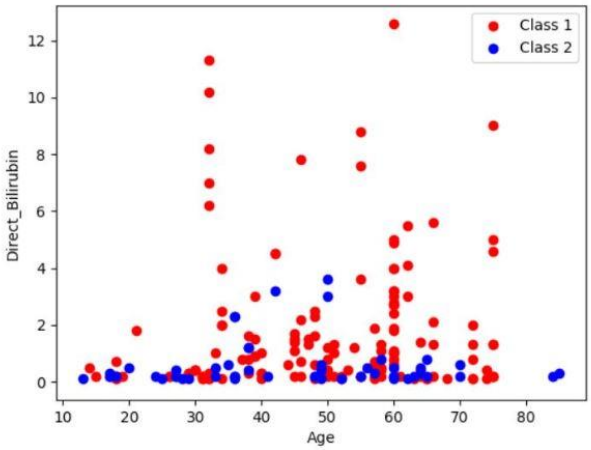


Fig. 6: Graph of Direct Bilirubin V.s. Age

T-distributed Stochastic Neighbor Embedding (t-SNE) is amachine learning algorithm that is a nonlinear dimensionality reduction technique well-suited for embedding high-dimensional data for visualization in a low-dimensional space of two or three dimensions. Specifically, it models each high-dimensional object by a two- or three-dimensional point in such a way that similar objects are modeled by nearby points and dissimilar objects are modeled by distant points with high probability.

The t-SNE algorithm comprises two main stages. First, t-SNE constructs a probability distribution over pairs of high-dimensional objects in such a way that similar objects have a high probability of being picked while dissimilar points have an extremely small probability of being picked. Second, t-SNE defines a similar probability distribution over the points in the low-dimensional map, and it minimizes the Kullback–Leibler divergence between the two distributions with respect to the locations of the points in the map. a heavy-tailed Student-t distribution (with one-degree of freedom, which is the same as a Cauchy distribution) The degrees of freedom refers to the number of independent observations in a set of data. is used to measure similarities between low-dimensional points in order to allow dissimilar objects to be modeled far apart in the map.

The minimization of the Kullback–Leibler divergence with respect to the other points is performed using gradient descent. The result of this optimization is a map that reflects the similarities between the high-dimensional inputs well.

Hence the 10 Dimensional data is visualized in two dimension as follows:

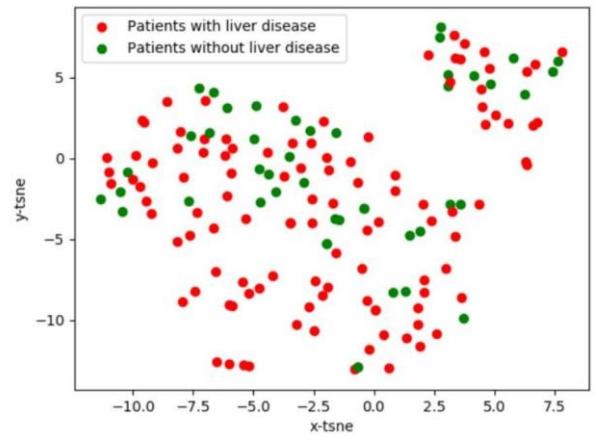


Fig. 7: Visualising 10-dimensional data with the using t-SNE

V. RESULTS

The predictions achieved by the Decision tree Algorithm was about 70% whereas the Random Forest Algorithm achived near about 83.75% accuracy in its prediction.

VI. FUTURE SCOPE

Machine learning and AI are expanding like never before and the research papers published solely on machine learning has more then doubled in the past few years. There is a huge scope in solving problems in not only healthcare but also any imaginable field.

VII. CONCLUSION

Thus, the results suggest that machines can assist doctors to learn from large sets of data as well as the patients can use them to verify their health test results in a trusted way. Also, algorithms like Random Forest and Decision Trees in Machine Learning can help achieve high accuracy rates and help transform the healthcare domain.

ACKNOWLEDGMENT

We thank our guide and our college Vishwakarma Institute of Technology for guidance and support and for giving us the opportunity to conduct this study.

REFERENCES

[1] P. shimpi, S. Shah, M. Shroff and A. Godbole (2017), "A Machine Learning Approach for the Classification of Cardiac Arrhythmia", IEEE 2017 International Conference on Computing Methodologies and Communication (ICCMC), Mumbai, India.

[2] S Tharaha and K Rashika (2017), "Hybrid Artificial Neural network and Decision Tree algorithm for Disease Recognition and Prediction in Human Blood Cells", 2017 International Conference on Innovations in information Embedded and Communication Systems (ICIIECS), Sriperumbudur, India.

[3] M. Somvanshi, S. Tambade, P.Chavan and S.V. Shinde (2016), "A Review of Machine Learning Techniques using Decision Tree and Support Vector Machine", Pune, India.

[4] H. Ayeldeen, O. Shaker, G. Ayeldeen and K.M. Anwar (2015), "Prediction of Liver Fibrosis stages by Machine Learning model: A Decision Tree Approach",

[5] J.H. Oh, R. Al-Lozi and I. El Naqa (2009), "Application of Machine Learning Techniques for Prediction of Radiation Pneumonitis in Lung Cancer Patients", 2009 International Conference on Machine Learning and Applications, USA.

[6] https://www.nvidia.com/object/deep-learning-in-medicine.html

[7] https://www.geekforgeeks/reinforcement-learning

[8] S.K. Asrani and J.Hepatol(2019), "Burden of liver diseases in the world",European Association for the Study of Liver

[9] https://en.wikipedia.org/wiki/Decision_tree_learning

[10] Tom M. Mitchell, Machine Learning, McGraw-Hill Science, 1997

[11] https://en.wikipedia.org/wiki/T-distributed stochastic_neighbor_embedding

[12] www.kaggle.com