2022 OPJU International Technology Conference on Emerging Technologies for Sustainable Development (OTCON) | 978-1-6654-9294-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/OTCON56053.2023.10113966

Prediction of Liver Abnormality using Machine Learning

Lavanya Gottemukkala
Assistant Professor, Department of Information Technology
Gokaraju Rangaraju Institute of Engineering and Technology,JNTUH
Hyderabad, India
lavanya1202@grietcollege.com

Jeevan Nagendra Kumar Y
Professor and Dean TIC Department of Information Technology
Gokaraju Rangaraju Institute of Engineering and Technology,JNTUH
Hyderabad, India
jeevannagendra@griet.ac.in

Sai Manikanta Phani Teja U
Department of Information Technology
Gokaraju Rangaraju Institute of Engineering and Technology,JNTUH
Hyderabad, India
phanitejau@gmail.com

Tanishq Dhanraj N
Department of Information Technology
Gokaraju Rangaraju Institute of Engineering and Technology,JNTUH
Hyderabad, India
tanishqraj26@gmail.com

Nitish Y
Department of Information Technology
Gokaraju Rangaraju Institute of Engineering and Technology,JNTUH
Hyderabad, India
nitishyamsani@gmail.com

***Abstract*—Cancer according to the American Society of Clinical Oncology journal was first described back in 1600 B.C. and has been prevalent ever since. The technological advancement in the field of medical sciences aided with advancement in the field of machine learning and deep learning has brought us to a situation today, where a subject can be informed of the dangers or the possibility of possessing an infected liver. In the prediction model, different enzymes were studied, and appropriate ratios were to determine the stability of the hepatocytes in the liver. The data was employed by different Machine Learning algorithms and based on their accuracy levels the final prediction has been made using the most appropriate algorithm. In an attempt to take the model to the next level, a few more algorithms were employed and explored the dataset even more. The results of each algorithm are compared using ROC graphs and ROC AUC SCORE to achieve a better model for this prediction model. Each algorithm is given by certain hyper-parameters which would increase the fitting nature more towards the best. The most important features calculated by each algorithm are mentioned and used accordingly to calculate the results.**

Keywords— Liver Disease, Machine Learning

I. INTRODUCTION

The main function of the liver in the body is to provide, regulate and maintain the metabolism rate in the body. This metabolism depends on the cells that make up the liver. These cells are called hepatocytes. The extent to which these hepatocytes have been damaged determines the level at which the liver has been misfunctioning. There are broadly three stages that lead to the failure or infection of the liver namely, liver inflammation, Liver Cirrhosis, and the last stage Liver Carcinoma or Liver Cancer. According to the study Hepatic disease is another name for liver disease. Symptoms of hepatic disorders include nausea, vomiting, exhaustion, stomach pain and swelling, back pain, lethargy, and weight loss. Certain patients have been observed to have jaundice (yellowing of the skin and eyes), fluid in an atypical cavity, pale feces, and, in particular, an enlarged spleen and gallbladder. Many scientists have tried building prediction models using intelligent algorithms for predicting FLD (Fat Liver Disease). For example, Italian researcher Giorgio Bedogni gathered data on, age, alanine aminotransferase, aspartate aminotransferase, gender, body mass index (BMI), alcohol consumption, waist circumference, the total of four skinfolds, and other factors to develop a NAFLD prediction model. However, the majority of the models are based on medical tests and Surveys. They employ several features that are difficult to gather in big amounts. These models are challenging to generalize due to data quantity constraints and feature complexity. The main objective of this work is to build a convenient, effective FLD prediction model by using machine learning algorithms that can be used by clinicians to identify people who need liver tests more and for epidemiological screening on a big scale. A few enzymes that are secreted by the liver are taken into account to forecast their functionality. The data is trained on eight machine learning algorithms- Decision Tree, Support Vector Machine, Logistic Regression, K-Nearest Neighbors, XG Boost, Gradient Boosting, Random Forest, and Neural Networks. After it is trained, the accuracy of these respective algorithms is shown and hence any algorithm can be chosen to predict the result based on the values given by the user.

II. LITERATURE SURVEY

Rebecca L. Siegel, Kimberly D. Miller, Hannah E. Fuchs, Ahmedin Jemal worked on the types of cancer that affect men and women in the population of USA [1]. This work involves several estimates of cancer cases and cancer-related deaths in the United States based on various surveys. A model was then developed to predict the number of new cancer cases as well as cancer-related deaths.

Dong Jin Park, Min Woo Park, Homin Lee, Young-Jin Kim, Yeongsic Kim & Young Hoon Park built a machine-learning model on the test reports from various laboratories [2]. The machine learning models like LightBGM(light gradient boosting machine) and XGBoost(extreme gradient boost) are built for predicting various diseases. The F1-score and prediction accuracy of the model is 81% and 92% respectively.

Konstantina Kourou Themis P, Exarchos Konstantinos P.Exarchos Michalis V.Karamouzis Dimitrios I.Fotiadis: developed a model for classification of cancer-based on the risk factor [3]. The model is developed with the help of

978-1-6654-9294-2/23/$31.00 ©2023 IEEE

supervised ML techniques like Support vector machines (SVMs) decision Trees Artificial Neural Networks and Bayesian Networks. The research contains various results which are obtained from various datasets and in the classificationof various cancers.

Nazmun Nahar, Ferdous Ara worked on developing a Machine Learning model for predicting the cancer of a patient [4]. The model is mainly developed on a supervised learning technique which is a decision tree. Which is used in the classification of cancer. The dataset used is the liver disease dataset, the dataset is trained on various decision tree algorithms like Random Forest, J48, Random tree, LMT, Decision Stump, REPTree, and, Hoeffding Tree. Decision Stump provides better accuracy when compared with others.

Dr. S. Vijayarani, and S.Dhayanand worked with data mining tools for disease prediction on the voluminous medical database [5]. Through the use of SVM(support vector machines) and Naïve Bayes algorithms, they developed a model for prediction of the liver disease. The F-Score of SVM and Naive Bayes are 0.331 and 0.251 respectively.

Shambel Kefelegn, Pooja Kamat Liver disease prediction with the help of various data mining classification algorithms like SVM, C4.5, NBC [6]. The model's performance is compared with its accuracy.

III. Architecture

The system architecture is shown in Figure 1.

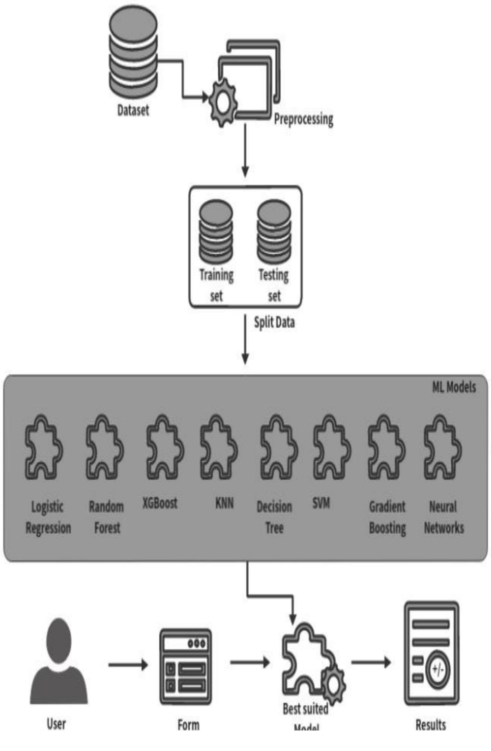


Fig. 1. Model Architecture

A. Developing the Model:

To begin with, the dataset is pre-processed, after which it is fed into the model for training and testing. Various algorithms are trained and tested on the above-split dataset. These models are compared based on accuracy and ROC-AUC score. The best-performing model is saved.

B. Developing GUI:

A Graphical user interface is developed with the help of the python library Tkinter which consists of various user inputs that have to be provided by the user.

C. Deploying the Model:

The model saved in the developing phase is deployed onto the GUI. The inputs given by the user are used by the model in predicting Liver disease.

IV. Methodology

The occurrence of abnormality in the functioning of the liver depends on the secretion of a few enzymes. Hence, the occurrence of these enzymes has been considered to build the ml model. This methodology aims to develop a model that can predict whether the liver of an individual is affected, or whether it has a chance of being affected before physical symptoms arise.

For this reason, biological enzymes and other biological attributes which facilitate the detection of the abnormality without being dependent on any physical symptom such as fever, alcohol intake, etc. have been considered. The model will enable an individual to know whether he is at risk or not during any regular check-up and will be able to detect asymptomatic liver dysfunction as well.

A. Dataset

The dataset used for training and testing the models is taken from Kaggle(Indian Liver Patient) and consists of attributes like Aspartate Aminotransferase, Alamine Aminotransferase, age, direct bilirubin, alkaline phosphatase, Total Proteins, Albumin, Albumin, Globulin Ratio, gender, and total bilirubin. It consists of 584 instances.

B. Splitting the Dataset

The dataset is split into two sections: 80% for training and 20% for testing. In other words, the training set consists of 451 instances, and testing consists of 113 instances.

C. Algorithms

1) Logistic Regression: The Logistic Regression algorithm is used to estimate the likelihood of a specific result based on the parametric values supplied. Hence, Logistic Regression is considered to not only approach prediction in a categorical format, but to approach it from a continuous data format as well, this particular algorithm might yield a less accuracy percentage, but provides a needed overall review and performance of the data concerning his problem, as the goal is to highlight what would not work or give an average result a much as something that works and gives the best result [7]. This increases the awareness and transparency of the project being conducted.

2) Random Forest: The Random Forest algorithm is considered to be a more accurate version of the decision tree algorithm [8]. The "Forest" here refers to the group of different decision trees that together culminate to represent a forest. The general logic of this algorithm lies in performing different sub-algorithms and combing the result to improve precision and thus make the result more efficient. The reason to include a Random Forest, as well as a decision tree in the project is to judge the versatility of the data and show the results of an individual process and the same process is done multiple times to increase the efficiency [21]

3) XG Boost: XG Boost is a distributed gradient boosting library that has been developed to be very efficient, portable, and, adaptable. It is used to solve issues requiring supervised learning. The XG Boost approach produces decision trees sequentially. All the weighted variables are independent [9]. These weights are used by the decision tree to forecast outcomes. The weights of the tree's incorrectly predicted variables are increased, and these variables are subsequently put into the second decision tree. After that, all of the individual predictions are integrated to produce a more powerful and exact model. It can handle problems like user-defined prediction, classification regression, and ranking. [20]

4) K-Nearest Neighbors: The K-Nearest Neighbors or KNN algorithm can be used to handle both regression and classification issues [10]. During the training of the model, the machine learns how the data is classified into infected or not infected based on the values of the parametric features. This algorithm has been chosen as the KNN algorithm gives not only the visual understanding to the person viewing the result but also gives precision concerning every individual value [19].

5) Support Vector Machine: Support Vector Machine or SVM is an algorithm, that has proven to give the best results and most accurate values compared to the other 17 algorithms that have been used to make the predictions [11]. SVM is mostly used for classification problems, hence making it the most appropriate algorithm for giving most the accurate results

6) Decision Tree: The Decision Tree algorithm works based on the continuous splitting of the data based on a certain parameter during the training phase. The attributes used in the experiment were: Alkaline Phosphatase, Aspartate Aminotransferase, Alamino Aminotransferase, and Age of a person [12]. A classification tree is used here to categorize based on the parameters if the subject is affected or has the chance of being affected by an ailment or not.

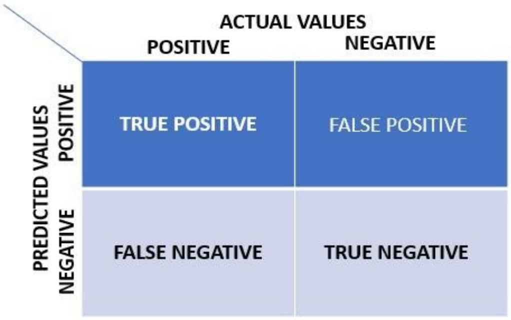
7) Gradient Boosting: Gradient Boosting may be used to forecast both long-term objective variables (as a Regressor) and short-term objective variables (as a Classifier). The cost work is Mean Square Error (MSE) when the calculation is used as a regressor, and Log loss when it is used as a classifier. [13] As it is observed, errors in AI computations are broadly classified into two types, fr example, Inclination Error, and Variance Error. As one of the assisting computations for reducing the model's predisposition blunder, gradient boosting is used.

8) Neural Networks: Neural Network is a proficient registering framework whose focal topic is acquired from the similarity of organic brain networks additionally called Artificial Neural Networks Brain networks are just a collection of computations that are freely demonstrated after the human mind and are meant to recognize designs [14]. These computations translate palpable data using machine discernment, marking, or bunching input. They detect mathematical examples stored in vectors into which all certifiable information, whether sound, text, or time series, should be decrypted. Brain networks assist in grouping and sorting. They may be described as a layer of bunching and grouping on top of the information you save and use [18]. They assist in grouping unlabeled information based on similarities across model sources of information, and they organize information when they have a named dataset to prepare.

D. Evaluation

To evaluate the model, the Confusion matrix, accuracy, ROC graph, ROC AUC Score, and AUC Test score were used.

1) Confusion Matrix: Fig 2 is a representation of the actual and expected values. It's a table-like layout that measures how effective the machine learning classification model works [22].



		ACTUAL VALUES	
		POSITIVE	NEGATIVE
PREDICTED VALUES	POSITIVE	TRUE POSITIVE	FALSE POSITIVE
	NEGATIVE	FALSE NEGATIVE	TRUE NEGATIVE

Fig. 2. Confusion Matrix

2) Accuracy Score: It is defined as the proportion of right predictions made to total predictions made as shown in Figure 3

3) Receiver Characteristic Curve: The Receiver Operating Characteristic Curve displays the performance of a classification model [15]. It compares the TPR (True Positive Rate) and FPR (False Positive Rate) at various categorization levels.

4) ROC AUC Score: A key aspect of machine learning is the performance of the models. An AUC (Area Under the Curve) Score and ROC (Reciever Operating Characteristics) Score are used in machine learning to test and visualize the performance of classification algorithms [16-17]. It is the most important evaluation metric for the performance check of any classification algorithm.

5) AUC Score: The fraction of the ROC plot that is below the curve is known as AUC. It displays the performance across all available categorization criteria. AUC is defined as the probability that the model evaluates an example that is random higher than an example that is random lower.

V. RESULTS

Algorithm	Accuracy	True Positive	True Negative	False Positive	False Negative
Random Forest	64.60%	5	68	13	27
Logistic Regression	69%	1	77	4	31
KNN	69.91%	4	75	6	28
SVM	70.79%	3	77	4	29
Decision Tree	69.915%	1	78	3	31
Gradient Boosting	69.92%	13	65	16	19
XG Boost	71.68%	0	81	0	32
sNeural Networks	69.9%	4	75	6	28

Table . 1. Accuracy and Confusion Matrix of various Algorithms

Table 1 displays various evaluation metrics like Accuracy, False Positive, True Negative, False Negative, and True Positive of the algorithms[23]

A. Random Forest

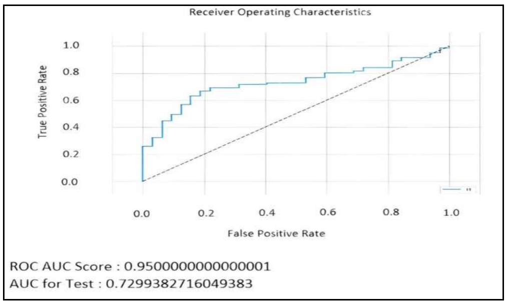


Fig. 3. ROC Graph of Random Forest

The ROC AUC Score and AUC for Test for Random Forest are depicted in Figure 3 where True Positive Rate is represented on the x-axis and the y-axis depicts the False Positive Rate[24].

B. Logistic Regression

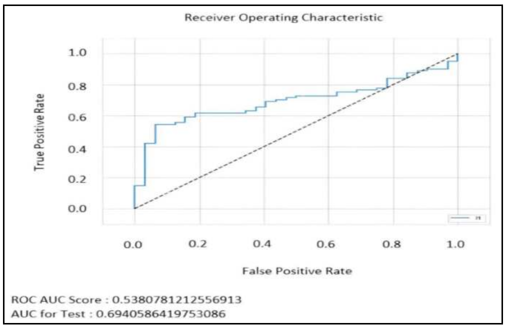


Fig. 4. ROC Graph of Logistic Regression

The ROC AUC Score and AUC for Test for Logistic Regression are depicted in Figure 4 where the True Positive Rate is represented on the x-axis and the y-axis depicts False Positive Rate [25]

C. K Nearest Neighbors

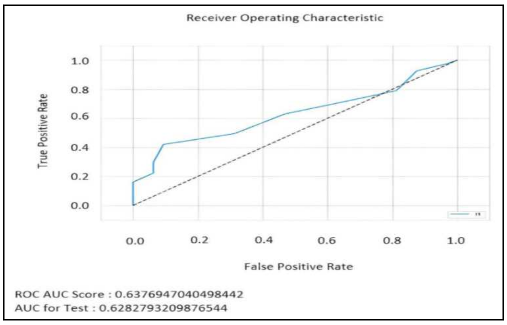


Fig. 5. ROC Graph of K Nearest Neighbors

The ROC AUC Score and AUC for Test for K-Nearest Neighbors are depicted in Figure 5 where True Positive Rate is represented on the x-axis and the y-axis depicts False Positive Rate[26].

D. Support Vector Machine

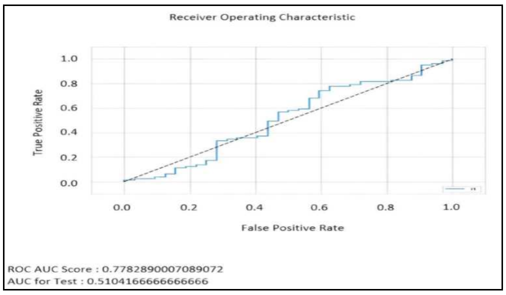


Fig. 6. ROC Graph of Support Vector Machine

The ROC AUC Score and AUC for the Test for Support Vector Machine are depicted in Figure 6 where True Positive Rate is represented on the x-axis and y-axis depicts False Positive Rate [27].

E. Decision Trees

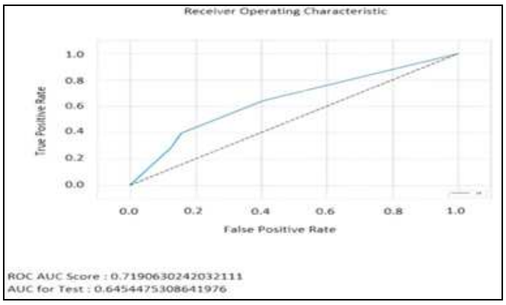


Fig. 7. ROC Graph of Decision Trees

The ROC AUC Score and AUC for Test for Decision Tree are depicted in Figure 7 where True Positive Rate is represented on the x-axis and the y-axis depicts False Positive Rate[28].

F. XG Boost

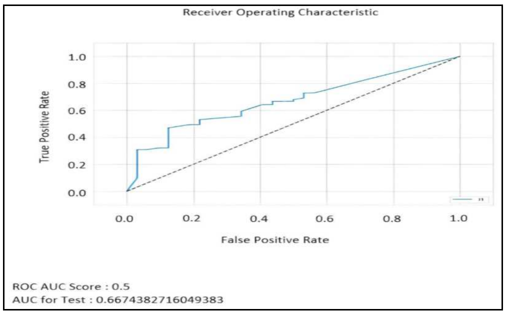


Fig. 8. ROC Graph of XG Boost

The ROC AUC Score and AUC for Test for XG Boost are depicted in Figure 8 where True Positive Rate is represented on the x-axis and the y-axis depicts the False Positive Rate [29].

G. Gradient Boosting

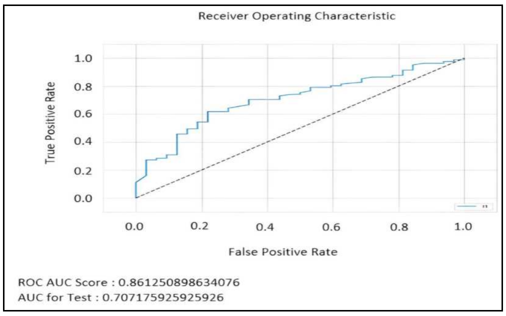


Fig. 9. ROC Graph of Gradient Boosting

The ROC AUC Score and AUC for Test for Gradient Boosting are depicted in Figure 9 where the True Positive Rate is represented on the x-axis and the y-axis depicts the False Positive Rate [30]

H. Neural Networks

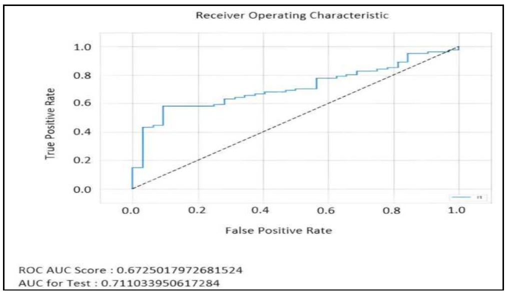


Fig. 10. ROC Graph of Neural Networks

The ROC AUC Score and AUC for Test for Neural Network are depicted in Figure 10 where True Positive Rate is represented on the x-axis and the y-axis depicts the False Positive Rate [31, 32]

VI. CONCLUSION

To conclude, based on various evaluation metrics it is observed that the XG Boost algorithm is more fitting to this model. Thus, the XG Boost algorithm is used to make predictions and is deployed onto a GUI [33]. All in all an attempt has been made to identify and alert the subject as well as their doctor of any chronic disease that the subject might bare the risk of having. The features that were selected were done so by reading multiple persons and in-person interaction with various students and practicing doctors. The features on their own are not sufficient to classify the results and give the right prediction that might be useful. The main features that have been selected are namely, Alkaline Phosphatase, Aspartate Aminotransferase, Alamino Aminotransferase, and the age of a person are the features that are directly linked with damaging the regeneration of the hepatocytes that in turn cause liver inflammation leading to Hepatitis, Cirrhosis, and the final stage, Liver Carcinoma. This project is an attempt to find the underlying asymptomatic conditions that prevail on the inside of the body based on the enzyme readings and the Albumin-Globulin ratio, Bilirubin levels, etc. On a positive note, concluding the research project by alerting and cautioning the subject and doctor of any liver abnormalities, hoping this to be a stepping stone to numerous future possibilities [24].

VII. FUTURE WORK

The research project's main objective is to predict the presence of any functionality of the liver in the body. In the future, this method can further be enhanced to narrow down the result to the kind of ailment that the liver is suffering from and the type of disease it may be at risk of getting infected with. This can further be enhanced in a situation where every person going for a regular check-up is made aware of the functionality of different organs of the subject's body and the possible risk that their organs bare. However, the methodology presented in this paper is limited to informing about the liver exclusively, but with a different set of attributes appropriate to the task in hand, this project can further be expanded to different parts of the body and thus making the machine and its predictions smarter and more useful

REFERENCES

[1] Kimberly D. Miller: A Cancer Journal for Clinicians January 2021

[2] Dong Jin Park: Development of Machine Learning Mode for Diagnostic Disease Prediction based on Laboratory tests April 2021

[3] Konstantina Kourou: Machine Learning in Cancer prognosis and prediction

[4] Nazmun Nahar, Ferdous Ara , ―Liver Disease Predictionusing by using different decision tree techniques‖ InternationalJournal of Data Mining & Knowledge Management Process (IJDKP) Vol.8, No.2, March 2018

[5] Dr. S. Vijayarani, S.Dhayanand , ―Liver DiseasePrediction using SVM and Naïve Bayes Algorithms‖,International Journal of Science, Engineering and Technology Research (IJSETR) Volume 4,Issue 4 ,April 2015,816-820.

[6] Shambel Kefelegn, Pooja Kamat, ―Prediction and Analysis of Liver Disorder Diseases by using Data Mining Technique: survey ‖, International Journal of pure and applied mathematics ,volume 118,No 9,765-770,2018

[7] https://www.analyticsvidhya.com/blog/2017/08/skilltest-logisticregression/#:~:text=True%2C%20Logistic%20regression%20is%20a,when%20you%20train%20the%20model%20.

[8] https://builtin.com/data-science/random-forest-algorithm

[9] https://www.geeksforgeeks.org/xgboost/

[10] https://becominghuman.ai/comprehending-k-means-and-knn algorithmsc791be90883d

[11] https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vectormachine/

[12] https://www.xoriant.com/blog/product-engineering/decision-trees-machinelearning-algorithm.html

[13] https://machinelearningmastery.com/gentle-introduction-gradient-boostingalgorithm-machine-learning/

[14] https://wiki.pathmind.com/neural-network

[15] https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

[16] https://towardsdatascience.com/understanding-auc-roc-curve68b2303cc9c5#:~:text=ROC%20is%20a%20probability %20curve,and%201%20classes%20as%201.

[17] https://www.researchgate.net/publication/334763090_Prediction_of_Liver_Disease_using_Classification_Algorithms https://www.frontiersin.org/articles/10.3389/fpubh.2021.6 68351/full

[18] https://www.researchgate.net/publication/334763090_Prediction_of_Liver_Disease_using_Classification_Algorithms https://www.frontiersin.org/articles/10.3389/fpubh.2021.6 68351/full

[19] https://www.ijert.org/liver-disease-prediction-system-using-machine-learningtechniques

[20] https://www.sciencedirect.com/science/article/pii/S187705092030692X

[21] Atef Zaguia, Vikram Raju, Y. Jeevan Nagendra Kumar, Umashankar Rawat,"Secure Vertical Handover to NEMO using Hybrid Cryptosystem" Hindawi, Article ID 6751423, Hindawi.

[22] Y. Jeevan Nagendra Kumar, Dr. T. V. Rajini Kanth, "GIS-MAP Based Spatial Analysis of Rainfall Data of Andhra Pradesh and Telangana States Using R", International Journal of Electrical and Computer Engineering (IJECE), Vol 7, No 1, February 2017, Scopus Indexed Journal, ISSN: 2088-8708

[23] B Sankara Babu, A Suneetha, G. Charles Babu, Y. Jeevan Nagendra Kumar, G. Karuna, "Medical Disease Prediction using Grey Wolf optimization and Auto Encoder based Recurrent Neural Network" Periodicals of Engineering and Natural Sciences, Vol 6 Issue 1 Pg: 229-240 ISSN 2303-4521 Jun 2018

[24] Dr. Y. Jeevan Nagendra Kumar, Guntreddi Sai Kiran, Partapu Preetham, Chila Lohith, Guntha Sai Roshik, G. Vijendar Reddy, "A Data Science View on Effects of Agriculture & Industry Sector on the GDP of India" International Journal of Recent Technology and Engineering, Volume-8, Issue-1, May 2019, ISBN: 2277-3878.

[25] Dr. Y. Jeevan Nagendra Kumar, Guntreddi Sai Kiran, Partapu Preetham, Chila Lohith, Guntha Sai Roshik, G. Vijendar Reddy, "A Data Science View on Effects of Agriculture & Industry Sector on the GDP of India" International Journal of Recent Technology and Engineering, Volume-8, Issue-1, May 2019, ISSN: 2277-3878

[26] Y. Jeevan Nagendra Kumar, N. Kameswari Shalini, P.K. Abhilash, K. Sandeep, D. Indira, "Prediction of Diabetes using Machine Learning" International Journal of Innovative Technology and Exploring Engineering, ISSN: 2278-3075, Volume-8 Issue-7 May 2019

[27] Y. Jeevan Nagendra Kumar, B. Mani Sai, Varagiri Shailaja, Singanamalli Renuka, Bharathi Panduri, "Python NLTK Sentiment Inspection using Naïve Bayes Classifier" International Journal of Recent Technology and Engineering, ISSN: 2277-3878, Volume-8, Issue-2S11, Sep 2019

[28] Srikanth Bethu, V Sowmya, B Sankara Babu, G Charles Babu, Y. Jeevan Nagendra Kumar, "Data Science: Identifying influencers in Social Networks", Periodicals of Engineering and Natural Sciences, ISSN 2303-4521 Vol.6, No.1, pp. 215~228

[29] Y Jeevan Nagendra Kumar, V Spandana, VS Vaishnavi, K Neha, VGRR Devi, "Supervised Machine Learning approach for Crop Prediction in Agriculture Sector", IEEE - 5th International Conference on Communication and Electronics Systems (ICCES), ISBN: 978-1-7281-5370-4 pg: 736-741

[30] D. Srinivasa Rao, Ch. Ramesh Babu, Y. J. Nagendra Kumar, N. Rajasekhar, T. Ravi, "Medical Image Fusion Using Transform Based Fusion Techniques", International Journal of Recent Technology and Engineering, Volume-8 Issue-2 ISSN: 2277-3878

[31] Atef Zaguia, Vikram Raju, Y. Jeevan Nagendra Kumar, Umashankar Rawat,"Secure Vertical Handover to NEMO using Hybrid Cryptosystem" Hindawi, Article ID 6751423, Hindawi

[32] Tiwari, V., Thakur, R. S., & Tiwari, B. "Optimization of EHR Data Flow Toward Healthcare Analytics." *In Proceedings of International Conference on Recent Advancement on Computer and Communication: ICRAC* 2017, pp. 637-643. Springer Singapore, 2018.

[33] Y Jeevan Nagendra Kumar, Partapu Preetham, P Kiran Varma, P Rohith, P Dilip Kumar, "Crude Oil Price Prediction Using Deep Learning" IEEE Second International Conference on Inventive Research in Computing Applications (ICIRCA-2020) ISBN: 978-1-7281-5373-5 Pg:119-124