# IJARETY


# International Journal of Advanced Research in Education and TechnologY (IJARETY)


## Volume 11, Issue 4, July-August 2024

## Impact Factor: 7.394

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

crossref

निस्केयर NISCAIR

International Journal of Advanced Research in Education and TechnologY(IJARETY)

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 7.394| A Bi-Monthly, Double-Blind Peer Reviewed & Referred Journal |

|| Volume 11, Issue 4, July-August 2024 ||

DOI:10.15680/IJARETY.2024.1104005

# Efficient Machine Learning Techniques for Prediction of Liver Disease

**Mr. Jayanthi Balaji, Mr. Arjun Kurava, Dr.D.William Albert**

M. Tech Student, Department of Computer Science and Engineering, Bheema Institute of Technology and Science, AP, India

Associate Professor, Department of Computer Science and Engineering, Bheema Institute of Technology and Science, AP, India

Associate Professor & Head of the Department, Department of Computer Science and Engineering, Bheema Institute of Technology and Science, AP, India

**ABSTRACT**: The most vital organ in humans, the liver performs a variety of functions including generating bile, eliminating bile and bilirubin, metabolising proteins and carbohydrates, activating enzymes, storing glycogen, vitamins, and minerals, synthesising plasma proteins, and producing clotting factors. The liver is easily impacted by the use of alcohol, prescription drugs, certain dietary items, and a lot of other activities. As of right now, liver function blood test and scan data are analysed to identify disorders connected to the liver. It is more costly and requires more time. One way to speed up the diagnosis of liver illness is to use various data mining techniques to make the procedure easier. The forecast will be more accurate with more data utilised. Cloud storage is utilised to get around the local storage shortage that many healthcare facilities have. Cloud storage might be a suitable option because health care facilities create large amounts of records. This article describes many data mining methods, such as Adaptive Neuro-Fuzzy Inference System (ANFIS), Decision Tree (DT), and K-Nearest Neighbour (KNN), that are utilised to generate a decision support model that might assist a doctor in making predictions about liver illness based on the dataset with the accuracy of 99.794%.. Each algorithm's performance is assessed in terms of its sensitivity, specificity, accuracy, and precision. An overview of these algorithms' efficiency is provided.

**KEY WORDS:** Adaptive Neuro-Fuzzy Inference System (ANFIS), Decision Tree (DT), and K-Nearest Neighbour (KNN).

## I.INTRODUCTION

The most important organ in a human body is the liver. The liver breaks down insulin. By using glucuronidation, the liver breaks down bilirubin, aiding in its excretion into bile [1]. It is also responsible for the decomposition and elimination of a large number of undesirable compounds. It demonstrates a notable function in changing hazardous substances. It plays a significant part in the collapse of pharmaceuticals. We call it drug metabolism. It would weigh 1.3 kilogram's. The liver is divided by two enormous sections: the left estimate and the privileged part. The gallbladder is situated next to the pancreas, beneath the liver. These organs, together with the liver, aid in the consumption and provision of nutrients. Its function is to facilitate the passage of wound materials from the stomach into the bloodstream and then onto whatever remains of the body. Diseases of the liver are brought on by damage to the organ or impairment of its function [2]. Liver problems [3] are complex and multifaceted, resulting from a multitude of factors that define an individual's vulnerability to the illness. Among these are factors related to sex, ethnicity, genetics, body mass index (BMI), concomitant disorders like diabetes, and exposure to the environment (viruses, alcohol, diet, and toxins). Liver disorders are life-threatening conditions that have a high death rate. The initial step in diagnosing liver problems is the standard blood and urine testing. Based on the patient's symptoms, a liver functions test (LFT) is advised [4]. Millions of individuals worldwide are impacted by liver disease, which is a serious health problem. Improved patient outcomes and a lighter load on the medical system can result from early identification and precise categorization of liver illnesses. Non-alcoholic fatty liver disease (NAFLD) is a condition that is becoming more and more common in wealthy countries, affecting one-third of adults and a rising percentage of children [5]. The first indication of the illness is an abnormal accumulation of triglycerides in the liver, which can trigger an inflammatory response in certain individuals and result in cirrhosis and liver cancer. Although obesity, insulin resistance, and non-alcoholic fatty liver disease (NAFLD) are significantly correlated, the path physiology of NAFLD is still not well understood, and there are few therapeutic options available. On the basis of patient data, machine learning approaches have nonetheless shown

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 7.394| A Bi-Monthly, Double-Blind Peer Reviewed & Referred Journal |

|| Volume 11, Issue 4, July-August 2024 ||

DOI:10.15680/IJARETY.2024.1104005

promising results in the prediction and classification of liver disorders. These methods predict results and find patterns in massive datasets by using complex algorithms to analyse and learn from them. Research on the application of machine learning methods for liver disease categorization and prediction is constantly evolving in order to improve accuracy and save medical expenses.

## II. RELATED WORK

Classification algorithms are often intended for use in liver disease forecasting since they may predict a patient's chance of developing liver disease based on certain characteristics. Based on currently available solutions, the F-Tree technique has the highest performance among all the algorithms investigated, making it a viable alternative for liver disease prediction. Feature selection is often used in combination using fuzzy K-means classification approaches for the categorization of liver diseases. These methods can help distinguish between different liver disorders by detecting important features. Liver disease categorization was done using a number of classification methods, such as J48, SVM, RF, and MLP. Because distinct liver illnesses may have the same attribute values, fuzzy-based classification can assist enhance the effectiveness of the method of classification by accounting for the degree of similarity across cases [8]. After assessing the performance of these state-of-the-art algorithms using metrics including data accuracy, data efficiency, and correction rate, the study compared the results. The multilayer perceptron algorithm was found to have the highest accuracy among all the techniques examined in the study. In this paper, researchers utilised Bayesian classification to distinguish between cirrhosis, hepatitis, among non-liver disorders [9]. Both the FT tree and Naive Bayes methods were used to segment the liver patient datasets into a number of disease subgroups. Every method's accuracy and time of execution were evaluated. Their analysis revealed that the Naive Bayes approach outperformed the FT tree strategy in terms of execution time. The effectiveness of several categorization strategies in the diagnosis of liver illnesses was examined by the authors [10]. Among these methods were Naive Bayes, Decision Trees, Multilayer Perceptrons, Random Forests, K-Nearest Neighbours, & Logistic Regression. To evaluate their performance, the writers used metrics like as recall, sensitivity, specificity, and accuracy. The results showed that Naive Bayes yielded the greatest accuracy results out of all the examined algorithms.

## III. METHODOLOGY

The literature reviews indicate that the development of efficient machine learning and deep learning models is necessary in order to detect assaults in datasets. The Indian liver patient dataset was analyzed and trained using four Machine Learning algorithms Random Forest (RF), Decision Tree, Logistics Regression, CNN and KNN. The general layout of the methodology.

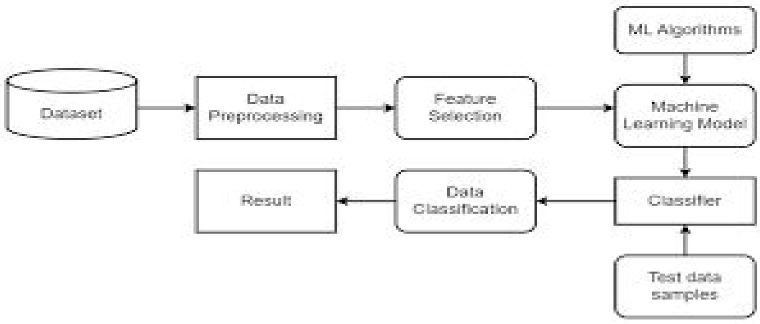



Fig : Processing of ML algorithms

**a. Dataset Collection:**
The original patient liver illness recorded dataset, which was obtained by the Canadian Institute of Cyber security, was condensed into the indian_liver_patient Datasets [21]. It has characteristics with the hospitality sector. There are one class attribute and eleven features per record. Every link is categorised as having a sickness or not. There are 2010 entries in the indian_liver_patient dataset, and each one is categorised as either Normal or Disease Found. In order to train our models, we used 1607 instances. Subsequently, 402 cases were used to assess and test these trained models. Ultimately, validation was performed using the remaining dataset.

**b. Data pre-processing:**
One crucial step in getting the data ready to feed into the algorithm is pre-processing it. Removing uncertainty from the dataset and giving IDS correct data is the aim of data preparation. It integrates normalisation and feature selection. The collection contains a large number of symbolic characteristics with nominal values, such protocol types and flags. To

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 7.394| A Bi-Monthly, Double-Blind Peer Reviewed & Referred Journal |

|| Volume 11, Issue 4, July-August 2024 ||

DOI:10.15680/IJARETY.2024.1104005

improve the dataset's performance, these variables must be transformed to numeric values. Discredited datasets in bin 10 have been used to modify binary classification issues (normal or disease) and multi-class classification problems.

**c. Feature selection:**

Feature Selection removes unnecessary and irrelevant traits to create subsets that are more improved and effective. Assuming that features are conditionally independent given the class, correlation is a well-liked and effective method for determining which features in any dataset are most tightly connected. It describes the strength of the relationship between features. Features that are unrelated and not predicted of one another, yet strongly correlated and predictive of the class, make up a useful feature subset. The CFS Sub Set Eval-Best First result, which was selected for feature selection in WEKA, is displayed in the table.

**Split and discretization**

Because discretized data requires less storage space, the primary goal of discretization is to enhance overall classification performance. Using several classifiers that use discrete data and classifiers that use discrete data discretization is a crucial step before classification. Numerical properties that were discretized utilising unsupervised 10 bin discretization on Weka and a discretization filter are known as discretizations. Additionally, dividing the dataset into testing and training modules is one of the most crucial stages in the construction of any machine learning model. The dataset used in this study was divided into two parts: 80% of the data were used for training, 20% for testing, and the remaining 1% were used for validation.

**e.Classification process**

In this work, we employed supervised machine learning techniques, including KNN, Random Forest (RF), CNN, Decision Tree, and Logistic Regression, to assess the efficacy of NIDS on the indian liver patient dataset. Generally speaking, there are five phases in any machine learning classification process:

- **Logistic Regression Algorithm**

It is a highly popular or extensively used SML model for categorization. For linearly separable classes, the LR model performs admirably and is relatively simple to use. In particular, it is most often utilised in manufacturing. Since LR is a linear model, it is often employed for binary classification; however, by utilising the OvR approach, multiclass classification may be accomplished [9]. To determine if bank cash is real or fake, LR is applied to a dataset and three distinct train test ratios (80:20, 60:40, and 70:30) are taken into account. ROC curves and learning curves are produced for the train-test ratio of 80:20. About 98% of the time, LR is accurate..

- **Decision Tree Algorithm:**

This categorization model has a tree-like topology. The process of developing DT involves segmenting the data collection into smaller sections. There are two kinds of nodes in the DT findings. Leaf nodes and decision nodes. As an illustration, take a look at the Outlook decision node, which has branches named Rainy, Overcast, and Sunny that reflect the tested feature's values. Hours Played, or a leaf node, provides a numerical target value decision. DT is capable of handling both category and numerical data [8]. To determine if bank cash is real or fake, DT is applied to a dataset and three distinct train test ratios (80:20, 60:40, and 70:30) are taken into account. ROC curves and learning curves are produced for the train-test ratio of 80:20.

- **Random Forest Algorithm**

Random Forest being the most often used supervised method. It is most helpful when doing regression and classification tasks. One classifier, called RF, stores numerous decision trees in each subset of an assumed data set and calculates the average value that improves the dataset's prediction accuracy. Decision trees are not necessary for the random forest. Rather, it receives a prediction from each tree, allowing it to estimate the final outcome based on prevalence estimation polls. The accuracy increases and overfitting issues are avoided with a larger number of trees in the forest. The idea of the ensemble approach, which combines several classifiers to solve a complex issue and enhance model performance, is backed by this.

- **K-Nearest Neighbor(KNN) Algorithm**

The supervised learning method is the foundation of K-Nearest Neighbour, one of the most fundamental machine learning algorithms. Based on the premise that the new instance as well as its data have characteristics comparable to the instances that are already available, the K-NN approach assigns the new case to the category that is most equivalent to the existing categories. Once all the pertinent data has been stored, the K-NN approach uses similarity to assess a new data point. This suggests that newly discovered data may be swiftly classified into an appropriate category using

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 7.394| A Bi-Monthly, Double-Blind Peer Reviewed & Referred Journal |

|| Volume 11, Issue 4, July-August 2024 ||

DOI:10.15680/IJARETY.2024.1104005

the K-NN algorithm. While the K-NN approach is mostly applied to classification problems, regression problems may also be addressed using it.

- **Convolutional Neural Network (CNN):**

The most intricate artificial neural network design is known as multi-layer perception. It is largely composed of several levels of perception. This notebook will show you how to create a neural network using the TensorFlow library, a highly well-liked deep learning framework that was published by. To comprehend the meaning of a multi-layer perception, we must utilise Numpy to create one from the beginning. MLP networks are employed in the context of supervised learning. Back propagation's method is another name for a common MLP network learning process. An artificial neural network that creates a set of outputs from a collection of inputs is called a feed forward multilayer perception (MLP).

**f. Evaluation metrics**

An essential step is evaluating the generated categorization models. Additionally, a range of assessment indicators are included throughout the process. The following are applied to metrics for evaluation: • The overall amount of malicious packets that were accurately categorised as True Positives (TP). The Total number of cases accurately identified as normal is known as True Negatives (TN). The total number of malicious packets that were mistakenly classed as assaults is known as False Positives (FP), while the total number of harmful packets that were mistakenly classified as normal is known as False Negatives (FN). The most often used metric for assessing a model is classification accuracy, however it is not a good indicator of how well it will work.

**g. Model validation**

In the last phase, the model is going to be put into practice, trained using the choices made in the earlier stages, and validated to check if it satisfies all requirements and to find out how successful it is at making predictions using fresh data. These evaluations reveal the model's shortcomings and limits, enabling the necessary actions to be made to remedy them. The experiment demonstrates that RF has the best accuracy when compared to other algorithms and that using 13 characteristics for each method results in high accuracy in the binary class. The models' respective closer accuracy percentages are 98.92% and F-measure is 98.9%. It demonstrates that the model is the most effective in identifying liver disease in patients.

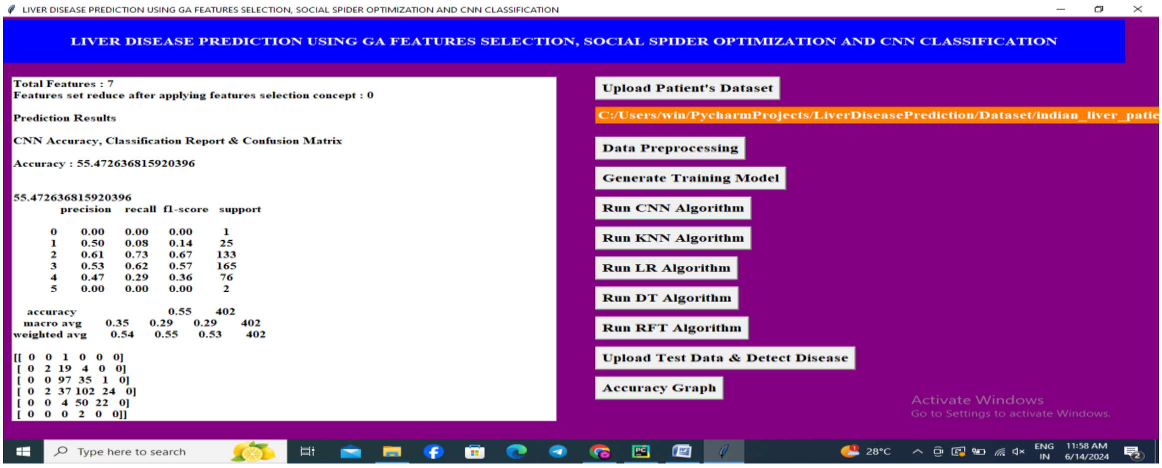
## IV. RESULT AND DISCUSSION

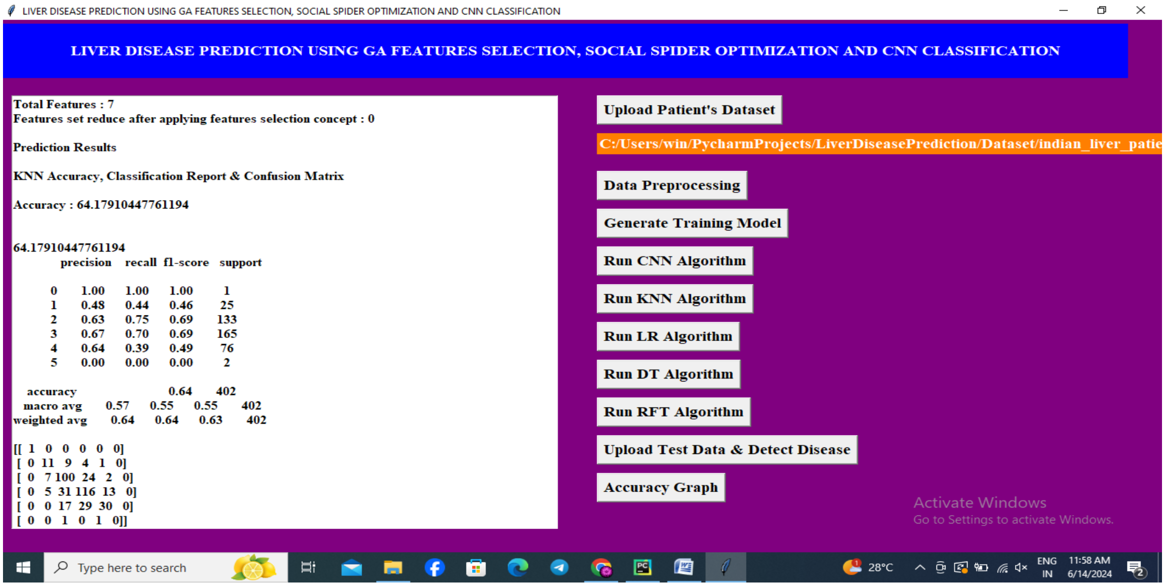
The authors' suggested procedure demonstrates that machine learning is capable of both illness prediction and the detection of subtle patterns that aid in diagnosis and decision-making. Globally, liver problems are thought to be a widespread concern due to the daily increase in instances. This thesis aims to offer proficient results in liver disease identification by classification algorithms [31]. The methods used for this kind of work are Support Vector Machines, K-Nearest Neighbour, and Logistic Regression. Among machine learning algorithms, the classification algorithms [32] are most frequently employed for illness prediction. In the healthcare industry, machine learning techniques have shown to be quite beneficial for predicting illnesses based on medical databases. Researchers and scientists are deliberately enhancing medical diagnostics with machine learning models [33] and classification algorithms practically everywhere on the planet, and the outcomes are improving. This work uses Support Vector Machines, K-Nearest Neighbour, and Logistic Regression to predict liver illness. As everyone knows, the liver is the biggest internal organ in the body and is responsible for several critical processes like the synthesis of glycogen, the generation of bile, triglycerides and cholesterol, and blood clotting factors and proteins. A decline in function often needs to impact far more than 75% of the liver tissue. Therefore, it is essential to detect the condition at an early stage so that it may be treated before it worsens.

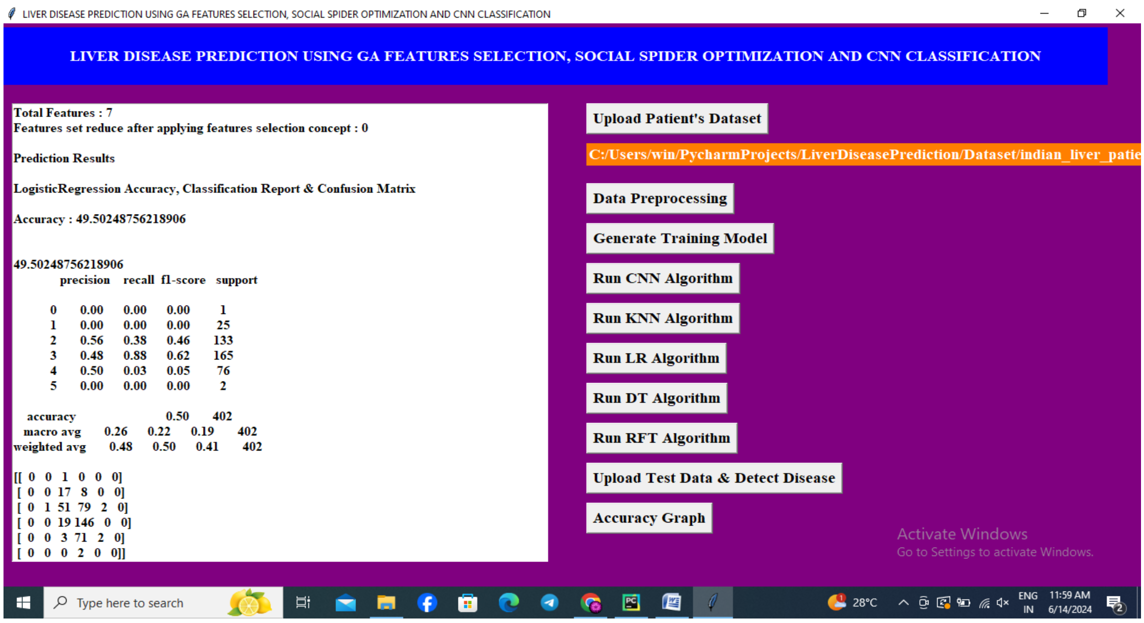
| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 7.394| A Bi-Monthly, Double-Blind Peer Reviewed & Referred Journal |

|| Volume 11, Issue 4, July-August 2024 ||

DOI:10.15680/IJARETY.2024.1104005



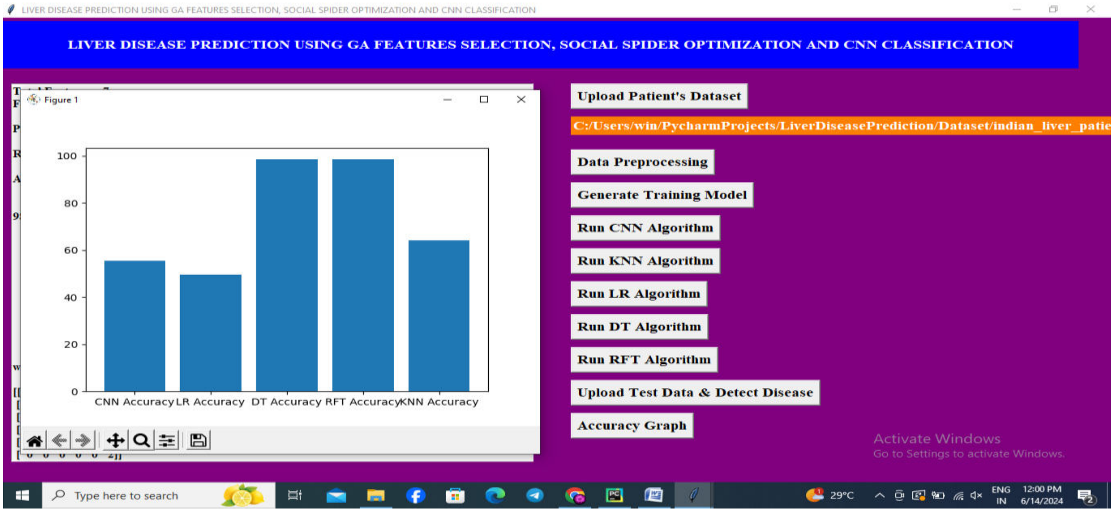



| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 7.394| A Bi-Monthly, Double-Blind Peer Reviewed & Referred Journal |

|| Volume 11, Issue 4, July-August 2024 ||

DOI:10.15680/IJARETY.2024.1104005




## V.CONCLUSION

The main data mining approach utilised in the healthcare industry for illness prediction and medical diagnosis is classification. In order to predict liver illness, this study employed the classification algorithms CNN, MLP, KNN, LR DT, and Random Forest Tree. Based on the performance aspects of execution time and classification accuracy, these algorithms are compared. This paper indicates that based on the experimental data, the DT classifier has the maximum classification accuracy and is thus regarded as the best method. In contrast, the Random Forest Tree requires the least amount of time to execute when comparing execution times.

## REFERENCES

1. .A. Arjmand, C. T. Angelis, A. T. Tzallas, M. G. Tsipouras, E. Glavas, R. Forlano, P. Manousou, and N. Giannakeas, “Deep learning in liver biopsies using convolutional neural networks,” in 2019 42nd International Conference on Telecommunications and Signal Processing (TSP). IEEE, 2019, pp. 496–499.
2. .L. A. Auxilia, “Accuracy prediction using machine learning techniques for indian patient liver disease,” in 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI). IEEE, 2018, pp. 45–50.
3. .A. Spann, A. Yasodhara, J. Kang, K. Watt, B. Wang, A. Goldenberg, and M. Bhat, “Applying machine learning in liver disease and transplantation: a comprehensive review,” Hepatology, vol. 71, no. 3, pp. 1093–1105, 2020.
4. S. Sontakke, J. Lohokare, and R. Dani, “Diagnosis of liver diseases using machine learning,” in 2017 International Conference on Emerging Trends & Innovation in ICT (ICEI). IEEE, 2017, pp. 129–133.
5. .J. C. Cohen, J. D. Horton, and H. H. Hobbs, “Human fatty liver disease: old questions and new insights,” Science, vol. 332, no. 6037, pp. 1519– 1523, 2011.
6. F. Himmah, R. Sigit, and T. Harsono, “Segmentation of liver using abdominal ct scan to detection liver desease area,” in 2018 International Electronics Symposium on Knowledge Creation and Intelligent Computing (IES-KCIC). IEEE, 2018, pp. 225–228.
7. M. B. Priya, P. L. Juliet, and P. Tamilselvi, “Performance analysis of liver disease prediction using machine learning algorithms,” International Research Journal of Engineering and Technology (IRJET), vol. 5, no. 1, pp. 206–211, 2018.
8. T. R. Baitharu and S. K. Pani, “Analysis of data mining techniques for healthcare decision support system using liver disorder dataset,” Procedia Computer Science, vol. 85, pp. 862–870, 2016.
9. U. R. Acharya, S. V. Sree, R. Ribeiro, G. Krishnamurthi, R. T. Marinho, J. Sanches, and J. S. Suri, “Data mining framework for fatty liver disease classification in ultrasound: a hybrid feature extraction paradigm,” Medical physics, vol. 39, no. 7Part1, pp. 4255–4264, 2012.
10. ] N. Nahar and F. Ara, “Liver disease prediction by using different decision tree techniques,” International Journal of Data Mining & Knowledge Management Process, vol. 8, no. 2, pp. 01–09, 2018.
11. A. Naik and L. Samant, “Correlation review of classification algorithm using data mining tool: Weka, rapidminer, tanagra, orange and knime,” Procedia Computer Science, vol. 85, pp. 662–668, 2016.

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 7.394| A Bi-Monthly, Double-Blind Peer Reviewed & Referred Journal |

|| Volume 11, Issue 4, July-August 2024 ||

DOI:10.15680/IJARETY.2024.1104005

12. .A. N. Arbain and B. Y. P. Balakrishnan, "A comparison of data mining algorithms for liver disease prediction on imbalanced data," International Journal of Data Science and Advanced Analytics (ISSN 2563-4429), vol. 1, no. 1, pp. 1–11, 2019.
13. M. A. Kuzhippallil, C. Joseph, and A. Kannan, "Comparative analysis of machine learning techniques for indian liver disease patients," in 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS). IEEE, 2020, pp. 778–782.
14. K. R. Asish, A. Gupta, A. Kumar, A. Mason, M. K. Enduri, and S. Anamalamudi, "A tool for fake news detection using machine learning techniques," in 2022 2nd International Conference on Intelligent Technologies (CONIT). IEEE, 2022, pp. 1–6.
15. M. K. Enduri, A. R. Sangi, S. Anamalamudi, R. C. B. Manikanta, K. Y. Reddy, P. L. Yeswanth, S. K. S. Reddy, and G. A. Karthikeya, "Comparative study on sentimental analysis using machine learning techniques," Mehran University Research Journal of Engineering and Technology, vol. 42, no. 1, pp. 207–215, 2023.
16. M. Islam, C.-C. Wu, T. N. Poly, H.-C. Yang, Y.-C. J. Li et al., "Applications of machine learning in fatty live disease prediction," in Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth. IOS Press, 2018, pp. 166–170.
17. S. Mohanty, P. K. Gantayat, S. Dash, B. P. Mishra, and S. C. Barik, "Liver disease prediction using machine learning algorithm," in Data Engineering and Intelligent Computing: Proceedings of ICICC 2020. Springer, 2021, pp. 589–596.
18. C. Liang and L. Peng, "An automated diagnosis system of liver disease using artificial immune and genetic algorithms," JOURNAL OF MEDICAL SYSTEMS, vol. 37, no. 2, 2013.
19. R. A. Khan, Y. Luo, and F.-X. Wu, "Machine learning based liver disease diagnosis: A systematic review," Neurocomputing, vol. 468, pp. 492–509, 2022.
20. A. S. Abdalrada, O. H. Yahya, A. H. M. Alaidi, N. A. Hussein, H. T. Alrikabi, and T. A.-Q. Al-Quraishi, "A predictive model for liver disease progression based on logistic regression algorithm," Periodicals of Engineering and Natural Sciences (PEN), vol. 7, no. 3, pp. 1255– 1264, 2019.
21. F. E. Harrell, Jr and F. E. Harrell, "Binary logistic regression," Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis, pp. 219–274, 2015.
22. E. M. Hashem and M. S. Mabrouk, "A study of support vector machine algorithm for liver disease diagnosis," American Journal of Intelligent Systems, vol. 4, no. 1, pp. 9–14, 2014.
23. Z. Yao, J. Li, Z. Guan, Y. Ye, and Y. Chen, "Liver disease screening based on densely connected deep neural networks," Neural Networks, vol. 123, pp. 299–304, 2020.
24. M. Abdar, N. Y. Yen, and J. C.-S. Hung, "Improving the diagnosis of liver disease using multilayer perceptron neural network and boosted decision trees," Journal of Medical and Biological Engineering, vol. 38, no. 6, pp. 953–965, 2018.
25. T. Bikku, "Multi-layered deep learning perceptron approach for health risk prediction," Journal of Big Data, vol. 7, no. 1, pp. 1–14, 2020.
26. T. A. Assegie, R. Subhashni, N. K. Kumar, J. P. Manivannan, P. Duraisamy, and M. F. Engidaye, "Random forest and support vector machine based hybrid liver disease detection," Bulletin of Electrical Engineering and Informatics, vol. 11, no. 3, pp. 1650–1656, 2022.
27. J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai et al., "Recent advances in convolutional neural networks," Pattern recognition, vol. 77, pp. 354–377, 2018.
28. J. Murphy, "An overview of convolutional neural network architectures for deep learning," Microway Inc, pp. 1–22, 2016.
29. G. Bonaccorso, Machine learning algorithms. Packt Publishing Ltd, 2017.
30. .I. Cohen, Y. Huang, J. Chen, J. Benesty, J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," Noise reduction in speech processing, pp. 1–4, 2009.
31. M. Ghosh, M. Raihan, M. Sarker, M. Raihan, L. Akter, A. K. Bairagi, S. S. Alshamrani, and M. Masud, "A comparative analysis of machine learning algorithms to predict liver disease." Intelligent Automation & Soft Computing, vol. 30, no. 3, 2021.
32. F. Osisanwo, J. Akinsola, O. Awodele, J. Hinmikaiye, O. Olakanmi, and J. Akinjobi, "Supervised machine learning algorithms: classification and comparison," International Journal of Computer Trends and Technology (IJCTT), vol. 48, no. 3, pp. 128–138, 2017.
33. P. C. Sen, M. Hajra, and M. Ghosh, "Supervised classification algorithms in machine learning: A survey and review," in Emerging technology in modelling and graphics. Springer, 2020, pp. 99–111.

IJARETY

International Journal of Advanced Research in Education and Technology

ISSN: 2394-2975

Impact Factor: 7.394

www.ijarety.in

editor.ijarety@gmail.com