Liver Disease Analysis Using Machine Learning

Trupti Kherde[1], Supriya Jadhav[2], Yash Chougule[3], Rakshita Khidbide[4], Riya Mohite[5]

[1]Professor, Dept. of Computer Engineering, Pimpri Chinchwad College Of Engineering And Research, Maharashtra, India
[2]Student, Dept. of Computer Engineering, Pimpri Chinchwad College Of Engineering And Research, Maharashtra, India
[3]Student, Dept. of Computer Engineering, Pimpri Chinchwad College Of Engineering And Research, Maharashtra, India
[4]Student, Dept. of Computer Engineering, Pimpri Chinchwad College Of Engineering And Research, Maharashtra, India
[5]Student, Dept. of Computer Engineering, Pimpri Chinchwad College Of Engineering And Research, Maharashtra, India

---***---

Abstract - *The increasing ubiquity of multimedia data, spanning text, audio, and video, has necessitated the development of effective summarization techniques. In our digital era, where information overload is prevalent, the ability to generate concise and coherent summaries is paramount. Deep Learning (DL) techniques have emerged as potent tools for addressing the multifaceted challenges inherent in multimedia summarization. This study focuses on summarization methods that leverage DL to extract meaningful content from diverse multimedia formats. The objective is to provide a comprehensive overview of the state-of-the-art DL techniques employed in summarizing multimedia data. By exploring various multimedia formats and their associated challenges, this research contributes to the evolving landscape of multimedia summarization, offering insights into its applications and future potential.*

***Key Words*: Machine Learning, Classification, Random Forest, Naive Bayes, Regression**

1.INTRODUCTION

The liver, an indispensable organ in the human body, plays a pivotal role in various critical processes such as metabolism, detoxification, and nutrient storage. However, liver illnesses pose significant threats to overall health and well-being. Early detection and intervention are paramount for effectively managing these conditions. In recent years, artificial intelligence (AI), particularly its subset of machine learning, has emerged as a potent tool in the medical field, offering promising avenues for analyzing complex medical data and predicting disease outcomes, particularly those associated with liver health.

This research project delves into the application of machine learning algorithms in analyzing medical data related to liver diseases, aiming to improve early detection and prognosis. By leveraging cutting-edge computational methods and predictive modeling, we aim to enhance our understanding of liver illnesses, including cirrhosis, hepatitis, and liver cancer. The utilization of machine learning allows for the analysis of diverse patient data sets, encompassing genetic data, laboratory test results, and patient records.

The objectives of this study are manifold. Firstly, we aim to explore how machine learning algorithms can detect early symptoms and risk factors for liver disorders, enabling timely interventions and improving patient outcomes. Additionally, we seek to investigate how machine learning techniques can aid in the classification of different liver diseases based on distinct patterns in patient data. This classification is crucial for tailoring treatment plans and predicting disease progression accurately. Moreover, our research endeavors to develop predictive models that assess an individual's risk of developing or exacerbating liver disease, informing personalized preventive measures and therapies. Furthermore, we aim to optimize treatment strategies by evaluating the efficacy of various therapies across diverse patient cohorts, leading to more targeted and efficient interventions.

By addressing these objectives, our research aims to contribute to the advancement of medical practices in liver disease management, ultimately leading to improved patient care and outcomes.

1.1 Methodology

In this study, we employ a Random Forest classifier to analyze a dataset comprising clinical features, laboratory test results, and patient demographics. We preprocess the data to handle missing values, normalize features, and ensure compatibility with the algorithm. Hyperparameters such as the number of trees, maximum tree depth, and maximum features considered for splitting are optimized using grid search cross-validation. The trained model is evaluated using metrics such as accuracy, precision, recall, and F1-score to assess its performance in predicting liver disease outcomes.

1.2 Algorithms And Techniques

Three supervised learning approaches are selected for this problem. Care is taken that all these approaches are fundamentally different from each other, so that we can cover as wide an umbrella as possible in term of possible approaches. For example- We will not select Random Forest and Ada Boost together as they come from the same family of 'ensemble' approaches: For each algorithm, we will try out different values of a few hyperparameters to arrive at the best possible classifier. This will be carried out with the help of grid search cross validation technique. The algorithms are described below:

1.2.1.Random Forest Classifier:

The Random Forest algorithm constructs decision trees using sample data and aggregates predictions from all the trees to determine the best overall prediction through a voting mechanism. The dataset is divided into training and testing sets, with 20% reserved for testing and 80% for training. The algorithm then partitions the data into multiple groups and subgroups, forming a tree-like structure. Each group is separated by hyperplanes that maximize the distance to the nearest data point in the training set, ensuring effective classification between classes.

a) n_estimators(number of trees in a forest)

b) max_depth(maximum depth of one single tree)

c) max_features(decides how many features are to be used)

d)oob_score(decides whether to include out-of-bag or prediction error)

Accuracy scored: 0.68

1.2.2. Gaussian Naive Bayes Classifier :

Naive Bayes is a classification algorithm for binary (two-class) and multi-class classification problems. The technique is easiest to understand when described using binary or categorical input values. The representation for naive Bayes is probabilities. A list of probabilities are stored to file for a learned naive Bayes model. This includes:

Class Probabilities: The probabilities of each class in the training dataset.

Conditional Probabilities: The conditional probabilities of each input value given each class value.

Accuracy scored: 0.5613

1.2.3. Logistic Regression:

Since the outcome is binary and we have a reasonable number of examples at our disposal compared to number of features, this approach seems suitable. At the core of this quantifies the difference between each prediction and its corresponding true value. When presented with a number of inputs, it assigns different weights to features (based on their relative importance).

Since for this data it already knows the output beforehand, it continuously adjusts the weights such that when these weights summed up with their features are introduced in the logistic function, the results are as near as possible to the actual ones. Once presented with a test value, it again inserts the value into our logistic function and returns the output as a number between 0 and 1, which represents the probability of that test value being in a particular class.

Accuracy scored: 0.7143

2. FUNCTIONAL MODEL AND DESCRIPTION

A data flow diagram is a graphical representation of the "flow" of data through an information system, modeling its process aspects. Often they are a preliminary step used to create an overview of the system which can later be elaborated. Data Flow Diagrams can also be used for the visualization of data processing (structured design).

The Data Flow Diagram Level 0 identifies external entities and processes of the system. Level 0 explains the architecture that would be used for developing a software product.

Data Flow Diagram Level 1 shows the main processes in the work and the entities involved in it.

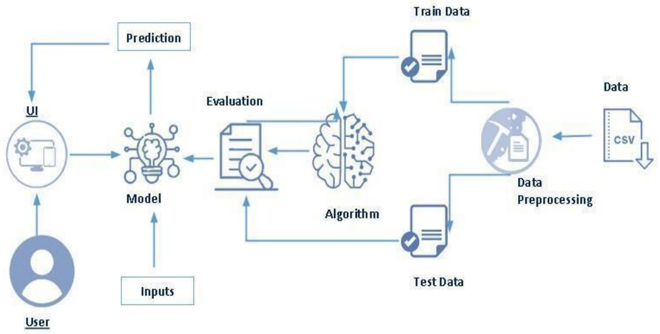

Fig -1: Data Flow Diagram

3. EXPLORATORY DATA ANALYSIS

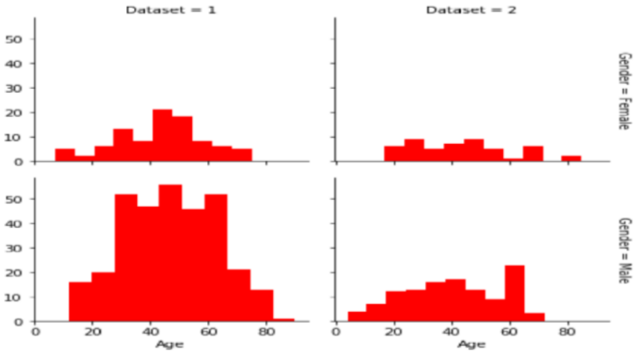

Fig -2: Facetgrid On Disease By Gender And Age

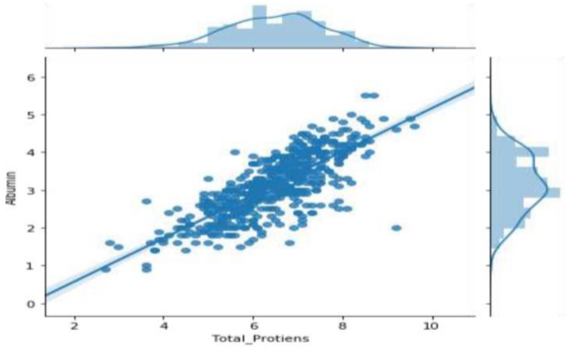

Fig -3: Jointplot On Direct_Bilirubin And Total Bilirubin

4. PROPOSED SYSTEM

In the proposed system, we begin by importing the liver patient dataset in CSV format. We then proceed to preprocess the dataset, eliminating anomalies and addressing empty cells to enhance the accuracy of liver disease prediction. Next, we construct a Confusion matrix to offer a comprehensive view of correct and incorrect predictions. Following this, we implement various classification and prediction techniques, potentially combining different algorithms, to evaluate accuracy. Our main goal is to develop a code that achieves a precision level of 90%. The anticipated advantages of the system include improved classification, early risk detection, and enhanced prediction accuracy.

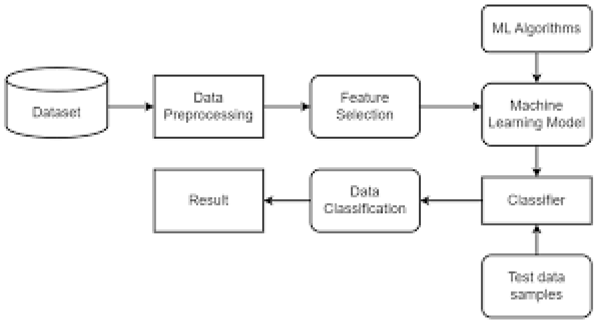

Fig -4:Proposed System

5. CONCLUSION

Utilizing the liver patient dataset, we implemented prediction and classification algorithms to alleviate the workload on medical practitioners. Our proposal advocates for the application of machine learning techniques to assess the overall liver condition of patients. Chronic liver conditions, persisting for at least six months, are considered, with both positive and negative instances utilized in our analysis. Employing a confusion matrix, we represent the classifier's outcomes in processing percentages of liver disease. The incorporation of training datasets enhances the performance of our proposed classification methods significantly. By employing a machine learning classifier, we distinguish between favorable and unfavorable values, demonstrating the accuracy of our classification model outputs.

Our research extends to the application of deep learning techniques for liver disease prediction. Future endeavors aim to enhance prediction and classification accuracy by incorporating more diverse data sources and combining multiple machine learning techniques. Additionally, machine learning models could be trained to predict the likelihood of liver disease based on individual characteristics. An essential aspect of our work involves developing explainable models for liver disease prediction and classification. These models should offer transparent insights into the factors contributing to liver disease, empowering healthcare professionals to make informed decisions and provide optimal patient care.

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.

[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, 2009.30, no. 3, 2021.