International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Liver Disease Prediction Using Machine Learing

K K Archana[1], Mahalakshmi R S[2], Narmatha Bharathi A[3], Reem A[4]

[1]Assistant Professor, Department Of Computer Science and Engineering, Sri Ramakrishna Institute of Technology, Coimbatore, Tamil Nadu, India.

[2,3,4]Student,Department Of Computer Science and Engineering, Sri Ramakrishna Institute of Technology, Coimbatore, Tamil Nadu, India.

Abstract:
This study focuses on predicting liver disease using machine learning, highlighting the importance of accurate diagnosis and early detection. Through thorough data preprocessing—such as handling missing values, addressing class imbalance with SMOTE, and applying iterative imputation—a high-quality dataset is created. The performance of SVM, KNN, HVC, and Random Forest algorithms is evaluated, with Random Forest achieving an accuracy of 85%. Ultimately, this research makes a significant contribution to healthcare analytics by providing a systematic, data-driven framework for the diagnosis and management of liver disease. This research enhances data-driven decision support for timely and accurate liver disease diagnosis, ultimately improving patient outcomes.

1. INTRODUCTION

Liver sickness is a great sized and developing worldwide fitness challenge, affecting tens of hundreds of thousands of humans international and posing essential dangers to public health and affected person well-being. Early and correct diagnosis is crucial for improving clinical intervention, patient care, and ordinary fitness effects. Modern day upgrades in tool studying and data analytics have created new possibilities to decorate the predictive accuracy of liver disorder diagnosis. This project investigates the creation and assessment of predictive fashions for liver sickness, making use of a complete dataset of patient statistics. Through emphasizing early diagnosis and nicely timed scientific intervention, this check presents a detailed angle on healthcare analytics inside the control of liver disease. The proposed approach starts with thorough facts preprocessing, addressing worrying conditions like missing values, specific feature encoding, and the usually encountered elegance imbalance, all essential for boosting dataset great. To evaluate and compare the effectiveness of various system studying models, we make use of a spread of algorithms, in conjunction with aid Vector tool (SVM), ok-Nearest buddies (KNN), hard voting Classifier (HVC), and Random wooded place. Extensively, our Random wooded location Classifier completed an splendid accuracy of 85%, particularly even as trained on decided on capabilities.

1.1.General Introduction

Liver ailment represents a large and developing worldwide fitness issue affecting thousands and thousands of humans worldwide and posing severe risks to public fitness and man or woman nicely-being early and correct diagnosis is critical to enhancing medical effects permitting well timed clinical intervention and improving patient care with latest improvements in machine mastering and statistics analytics there is a promising opportunity to decorate the predictive accuracy of liver disorder prognosis.

This challenge explores the improvement and evaluation of predictive fashions for liver sickness using a complete dataset of affected person facts with the aid of focusing on early detection and facts-driven

selection-making this take a look at highlights the capability of healthcare analytics in managing liver-related conditions the proposed method starts offevolved with robust facts preprocessing addressing key demanding situations along with missing values specific function encoding and sophistication imbalancecrucial steps.

To ensure the high-quality and reliability of the dataset to evaluate the overall performance of numerous machine gaining knowledge of processes more than one algorithms are applied consisting of assist vector gadget svm ok-nearest buddies knn tough voting classifier hvc and random wooded area among these the random forest classifier tested the best accuracy of 85 especially whilst trained on a carefully decided on set of functions.

1.3.Problem Statement

Liver ailment represents a widespread and growing worldwide health project affecting tens of millions of individuals around the arena and posing a extreme danger to public health and patient nicely-being well timed and correct analysis is essential for enhancing clinical interventions improving patient care and attaining higher common fitness outcomes cutting-edge improvements in system getting to know and statistics analytics have created new possibilities to enhance the predictive accuracy of liver ailment analysis.

One of the number one worrying conditions in diagnosing liver disease is the complexity of its clinical manifestations and the related risk elements liver illnesses can present in numerous strategies ranging from asymptomatic times to severe symptoms and symptoms which makes early identification difficult the overlap of signs among one in every of a type liver sicknesses in addition complicates the diagnostic manner additionally there's an urgent want to cope with the hassle of sophistication imbalance in liver sickness datasets in which the general public of sufferers do no longer have the circumstance this skewed instance can negatively have an impact on the overall performance of diagnostic fashions by using the usage of tackling.

Those demanding situations we purpose to extend particular and green device mastering models for predicting liver sicknesses our primary consciousness is on facilitating early and reliable diagnoses in the end we need to provide healthcare experts with a tool that complements the diagnostic manner and improves patient care associated with liver ailments

1.4. Algorithm

1. 1.Support Vector Machine
2. k-nearest neighbors
3. Hard Voting Classifier
4. Random Forest

2. LITERATURE SURVEY

Elias Dritsas et al.,(2013) included statistics preprocessing techniques particularly employing the synthetic minority oversampling approach smote to address the imbalance in instance distribution within the dataset this meticulous step is deemed vital for the improvement of powerful class fashions especially in predicting the prevalence of liver sickness Moreover the authors done an extensive evaluation of function significance the usage of several strategies together with pearson correlation gain ratio and random forest inside the studies a complete comparative evaluation of the general performance of more than one gadget getting to know models modified into finished assessment metrics in conjunction with

precision don't forget f-diploma accuracy and the place beneath the curve auc were applied to gauge version effectiveness.

Farkhondeh Rampur et al.,(2014) investigated 3 goal label variables fatty liver degree no steatosis and fibrosis tiers their examine employed eight super gadget gaining knowledge of techniques in particular okay-nearest neighbor KNN assist vector device svm radial basis characteristic rob SVM gaussian manner gp random wooded location rf neural network in ad boost and nave bayes to decide the handiest modeling approach the authors conducted 50 iterations of training and evaluation recognizing the want for more than one runs to account for potential variations in classifier outputs because of differing initial situations for a strong evaluation of classifiers model accuracy and the vicinity below the curve auc have been documented for each device mastering approach information preprocessing steps covered normalization and segmentation addressing lacking values thru linear interpolation and making use of crucial component assessment pca to extract statistics attributes the dataset became then separated into check and training sets accompanied through in addition preprocessing regarding function preference and class using the maximum applicable capabilities in the version processing diploma various models were performed and the nice-appearing version have become decided on for the final assessment.

Ilkyu Park et al., (2016) investigated the improvement of non-alcoholic fatty liver ailment nafld with the aid of the use of classifying every ailment type primarily based on a histopathological algorithm the evaluation focused entirely on controls with out steatosis ballooning and infection for records preprocessing of the proper rna-seq dataset sickle v133 was used to eliminate the adapter sequences in the long run the readings corresponded to the human genome assembly grch38104 using superstar v278 with an average of ninety four09 of the reads efficaciously mapped to the reference genome in every pattern the usage of stringtie v215 transcripts had been quantified in fragments in keeping with kilobase million fpkm layout calculating expression values and normalizing countsfrom the overall huge sort of genes protein-coding genes in keeping with the ensemble database had been determined on data normalization involved reworking the whole dataset the use of the log2pkm 1 method genes were excluded if their fpkm rate in the patient corporation surpassed 60 or if the maximum fpkm in every sample have become a good deal less than 1 the evaluation targeted on 16512 genesto end up privy to differentially expressed genes degs numerous statistical tests had been employed which include the t-check wilcoxon rank-sum test and fold trade the normality of each gene was assessed the use of the shapiro-wilk take a look at with genes displaying an equal or less than 005 p-value present process the t-take a look at p zero05 for genes not displaying a normal distribution p zero05 the wilcoxon rank-sum check come to be implemented the fold change technique was applied to choose the most great genes emphasizing smooth variations in expression

Anil Utku et al., (2016) explores the multilayer perceptron mlp a neural network primarily based with a couple of interconnected neurons hierarchically incorporating non-linear activation functions the mlp consists of an enter layer one or more intermediate hidden layers and an output layer the enter layer gets facts for processing this is then forwarded via the community the use of weights connecting it to the hidden layers activation capabilities which encompass relu sigmoid and tanh are applied in the hidden layers to procedure input sequences with iterations based mostly on the range of hidden layers the output layer answerable for duties like regression or elegance employs activation talents decided on based totally completely at the particular problem typesigmoid for binary kind and softmax for multi-elegance magnificenceschooling of mlp neurons is performed the usage of the backpropagation set of policies utkus developed mlp-based version utilizes affected character information in conjunction with demographic information and blood take a look at results as input to expect the presence of cirrhosis the version features

an enter layer that receives demographic and blood take a look at records incorporating two hidden layers for computational capabilities hyperparameter evaluation carried out thru gridsearchcv determines the most superb kind of neurons and epochs in the hidden layers the relu activation characteristic is applied in the input layer and within the hidden layers to prepare and perform non-linear calculations given the binary magnificence nature of the mission the output layer uses the sigmoid activation characteristic.

Shibam Ch Karmakar et al., (2017) hired a system getting to know technique to forecast the superiority of liver sickness and have a look at predictive accuracy the preliminary step worried the development of a logistic regression algorithm geared closer to predicting the early stages of liver disorder with the objective of improving precision in forecasting more than one logistic regression a device gaining knowledge of technique become finally applied to are awaiting a binary outcome via the use of thinking about one or more input variables this approach proved instrumental in organising quantitative institutions a number of the variable sets.

3. EXISTING METHODOLOGY

Traditional methods for diagnosing liver disease primarily rely on clinical evaluation, blood tests, and imaging techniques such as ultrasound, CT scans, and MRI. Physicians analyze liver function test results, including bilirubin levels, enzyme activity, and protein concentrations, to assess liver health. Statistical techniques and conventional classification algorithms, such as logistic regression and decision trees, have been used to predict liver disease based on structured medical data. However, these approaches often struggle with handling missing data, categorical variables, and class imbalances in medical datasets. They also tend to rely on manually selected features, which may not capture the full complexity of the disease, leading to reduced predictive accuracy.

3.1.DISADVANTAGE

Despite its widespread use, the traditional methodology for liver disease prediction has several limitations. One major drawback is its limited accuracy, as conventional statistical models and early machine learning techniques often fail to capture the complex relationships within medical data. Additionally, handling missing data remains a significant challenge, with many approaches simply discarding incomplete records, which can reduce the dataset's effectiveness. The imbalance in medical datasets further complicates predictive modeling, as standard methods struggle to learn from underrepresented cases. Moreover, traditional techniques lack advanced feature engineering, relying on predefined variables that may not fully capture the intricacies of liver disease. Lastly, the diagnostic process can be time-consuming, requiring multiple medical tests and expert interpretation, leading to delays in disease detection and treatment initiation.

4. PROPOSED METHODOLOGY

The proposed device employs several crucial strategies to enhance liver sickness prediction precision and regular overall performance facts cleaning and preprocessing ensure that the affected character information is correct and reliable addressing issues like lacking values and outliers this easily records bureaucracy as the inspiration for feature preference wherein the most informative variables are decided on optimizing the tool reading models ordinary general performance and decreasing computational complexity.

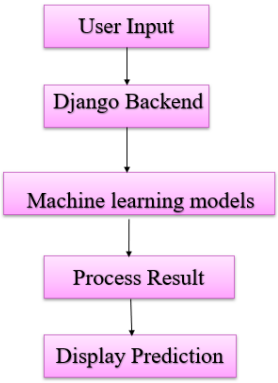
To cope with the challenge of imbalanced information the system uses the artificial minority over-sampling approach smote and an opportunity method the usage of the iterative imputer from sci-kit-studies

smote generates artificial data elements for the minority beauty generating a dataset that is balanced the coronary heart of the systems operation lies within the education of tool reading fashions algorithms like assist vector machines hard voting classifier properly sufficient-nearest buddies random forests are done to create predictive fashions

4.1.ADVANTAGE

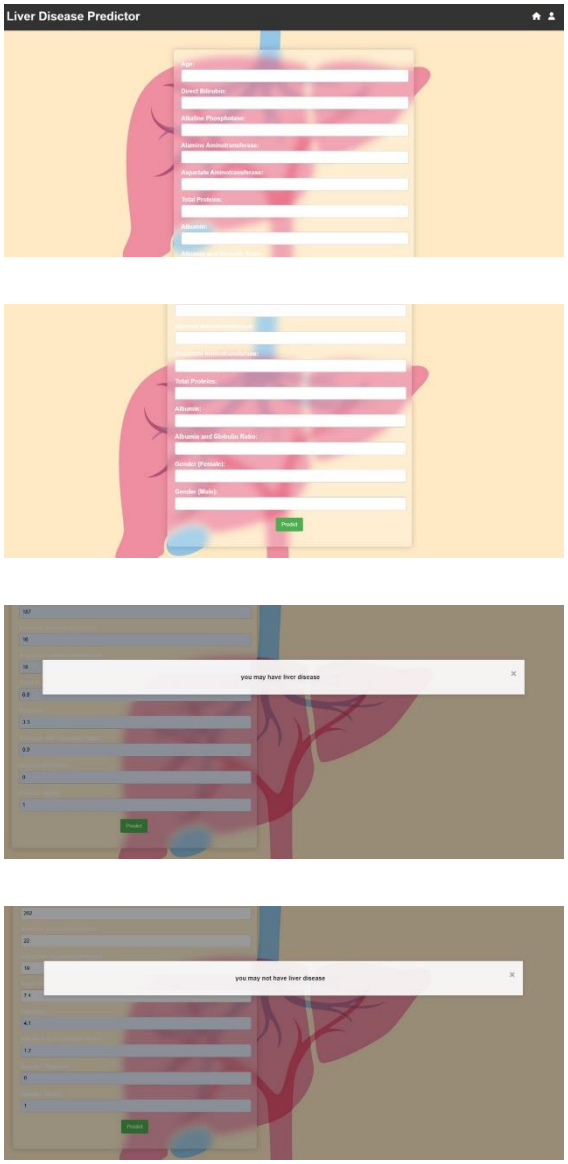
The proposed gadget leverages superior device learning techniques and strong statistics preprocessing techniques to enhance the accuracy and performance of liver ailment prediction by successfully managing missing statistics and categorical variables the model guarantees that no precious facts is misplaced leading to extra reliable predictions the use of hyperparameter tuning particularly within the random wooded area version drastically improves overall performance achieving an impressive 85 accuracy additionally by addressing elegance imbalance the device enhances its capability to come across liver disease even in underrepresented cases the practical application of this approach in healthcare analytics facilitates early prognosis customized remedy hints and higher aid allocation ultimately improving patient consequences moreover the system contributes to scientific research by using figuring out patient cohorts for medical trials and assisting the assessment of recent remedies for this reason advancing the sector of liver disorder management.

4.2. BLOCK DIAGRAM



5. RESULTS

The results of the liver disease prediction system demonstrate the effectiveness of the implemented machine learning approach through both quantitative metrics and visual analysis. Models such as Random Forest were evaluated, with Random Forest delivering the highest accuracy and overall performance. Key evaluation metrics including precision, recall, and F1-score confirmed the model's reliability. Visualizations such as feature importance graphs and confusion matrices provided insights into which medical features had the most influence on predictions, enhancing the model's interpretability. Overall, the system effectively supports early detection of liver disease and shows strong potential as a decision-support tool in the medical field.



6. CONCLUSION

This project harnesses the power of machine learning and advanced data preprocessing techniques to enhance liver disease prediction. By effectively handling missing data, categorical variables, and class imbalances, we have significantly improved the reliability of predictive models. Our approach ensures that the dataset is optimized for analysis, leading to more accurate and meaningful results in diagnosing liver disease.

Through extensive experimentation with various machine learning models, we identified the Random Forest algorithm as the most effective, achieving an impressive **85% accuracy** after hyperparameter tuning. By selecting the most relevant features and refining the model's parameters, we have maximized its predictive capabilities. This level of accuracy demonstrates the potential of AI-driven analytics in revolutionizing disease detection and healthcare decision-making.

Beyond its technical success, this project has meaningful implications for real-world healthcare applications. Faster and more precise liver disease diagnosis can lead to early interventions, personalized treatment recommendations, and better resource management in medical facilities. Additionally, the insights gained from this research can aid in identifying patient cohorts for clinical trials and evaluating

new treatments, ultimately contributing to advancements in medical research and improved patient outcomes.

REFERENCES

1. P.J. Thuluvath, A. Saraya, and M. Rela, ‘‘An introduction to liver disease in India,’’ Clin. Liver Disease, vol. 18, no. 3, pp. 105–107, Sep. 2021, doi: 10.1002/cld.1149.
2. G.Shaheamlung, H. Kaur, and M. Kaur, ‘‘A survey on machine learning techniques for the diagnosis of liver disease,’’ in Proc. Int. Conf. Intell. Eng. Manag., 2020, pp. 337–341, doi:10.1109/ICIEM48762.2020.9160097.
3. Acharya U. R., Faust O., Molinari F., Sree S. V., Junnarkar S. P., Sudarshan V. (2015). Ultrasound-based tissue characterization and classification of fatty liver disease: A screening and diagnostic paradigm. Knowledge-Based Systems, 75, 66-77
4. Shaheamlung, G., Kaur, H., & Kaur, M. (2020, June). A Survey on machine learning techniques for the diagnosis of liver disease. In 2020 International Conference on Intelligent Engineering and Management (ICIEM) (pp. 337-341). IEEE.
5. J. Zhang et al., "A Novel Approach for Early Detection of Liver Disease Using Machine Learning Algorithms," IEEE Transactions on Biomedical Engineering, vol. 67, no. 5, pp. 1356-1364, May 2020.
6. Deivendran.P, S. Selvakanmani, S. Jegadeesan, V. Vinoth Kumar "Liver Infection Prediction Analysis using Machine Learning to Evaluate Analytical Performance in Neural Networks by Optimization Techniques" Volume 71 Issue 3(Feb 2021), 377-384
7. Farkhondeh Razmpour, Reza Daryabeygi-Khotbehsara, Davood Soleimani, HamzehAsgharnezhad, Afshar Shamsi, Ghasem Sadeghi Bajestani, Mohsen Islam"Application of machine learning in predicting non-alcoholic fatty liver disease using anthropometric and body composition indices" volume 53 Issue (July 2020)
8. Abdul Quadir Md, Sanika Kulkarni ,Christy Jackson Joshua, Tejas Vaichole, Senthilkumar Mohan and Celestine Iwendi"Enhanced Preprocessing Approach Using Ensemble Machine Learning Algorithms for Detecting Liver Disease" volume 12, issue(Jan 2021).