iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** VIII **Month of publication:** August 2024

DOI: https://doi.org/10.22214/ijraset.2024.63879

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue VIII Aug 2024- Available at www.ijraset.com

Liver Disease Detection using Machine Learning

Mrs. MD. Bushra[1], Saipavan Chittiprolu[2], Kandru Paramjyothi[3], Shaik Thabassum[4]
[1]*Assistant Professor, Department of Computer Science and Engineering, PSCMR College of Engineering & Technology*
[2, 3, 4]*Student, Department of Computer Science and Engineering, PSCMR College of Engineering & Technology*

Abstract: ***Liver disease is a serious worldwide health issue, and prompt diagnosis and treatment are essential for successful outcomes. Traditional diagnostic techniques, however, may be expensive and time-consuming and can require intrusive procedures. In this work, we suggest a machine learning-based method for liver disease identification that makes use of the Support Vector Machine (SVM) and Random Forest Decision Tree algorithms. Our approach uses a large dataset with pertinent clinical characteristics including biochemical signs and patient demographics to categorize people into liver disease-positive or -negative groups. Furthermore, we incorporate an intuitive UI with Streamlit, improving accessibility and usability for end users and healthcare providers alike. By means of meticulous testing and assessment, we exhibit the efficiency and functionality of our suggested framework. In terms of precision, specificity, and sensitivity, our model performs admirably, demonstrating its promise as a trustworthy instrument for the identification of liver illness. Moreover, the incorporation of Streamlit improves the system's practicality and usability, making it easier to install and use in actual healthcare environments.***
Keywords: ***Detection, Disease, Healthcare, Liver, Machine Learning, Prediction, Streamlit and SVM.***

I. INTRODUCTION

India is facing a serious public health crisis as a result of the country's rising liver disease burden over time. Recent data from the Indian Council of Medical Research (ICMR) indicates [1] that a significant share of morbidity and death in the country are caused by liver illnesses. In a population over one billion, liver problems are quite prevalent and might include non-alcoholic fatty liver disease (NAFLD), alcoholic liver disease, and viral hepatitis. These illnesses have an impact on the quality of life of those who are afflicted and place a significant financial strain on healthcare systems. In this work, we primarily address the issue of the Indian population's lack of effective, non-invasive diagnostic methods for liver disease early identification and intervention. Conventional diagnostic techniques frequently depend on invasive procedures like liver biopsies, which are expensive, fraught with danger, and may not be available to all patients, especially those who reside in rural or underdeveloped areas [2]. Furthermore, depending too much on subjective interpretation of test results might result in delayed or incorrect diagnoses, which can worsen health outcomes by preventing prompt treatment.

We suggest a novel machine learning-based method for liver disease diagnosis that is suited to the Indian setting in order to overcome these difficulties. Modern algorithms that have proven effective in medical diagnostics, such as Random Forest Decision Tree and Support Vector Machine (SVM), are incorporated into our technique [3]. Our approach uses a large dataset that includes clinical history, demographic data, and pertinent biomarkers to reliably categorize people into liver disease-positive or -negative groups. We want to create a reliable and accurate prediction tool that may help medical practitioners identify liver disorders early on and stratify patients based on their risk by utilizing machine learning. This study's approach consists of a few essential components. First, we gather a representative and varied dataset of anonymized patient information from Indian healthcare facilities. This dataset includes a broad variety of clinical factors, such as liver function tests, imaging results, and comorbidities, in addition to demographic information including age, gender, and geographic region [4]. The quality and integrity of the dataset are then ensured by preprocessing the data to manage missing values, normalize features, and solve class imbalance concerns.

We next use feature selection approaches to determine which of the significant variables are linked to the outcomes of liver disease. This is a critical step in lowering dimensionality and increasing interpretability of the model, which will improve our prediction algorithms' performance and generalizability. After the feature set has been established, we use cross-validation to train a variety of machine learning models, such as Random Forest Decision Trees and Support Vector Machines [5], in order to minimize overfitting and optimize hyperparameters. Concurrently, we utilize Streamlit, a well-liked Python package for creating interactive web apps, to create an intuitive user experience. Through the usage of this interface, which acts as a platform for the deployment of our predictive models, end users and healthcare providers may input patient data and receive real-time predictions about the risk of liver disease. Streamlit's integration [6] improves our prediction tool's usability and accessibility, making it easier to incorporate into clinical processes and encouraging its wider use in a variety of healthcare settings.

©IJRASET: All Rights are Reserved

II. LITERATURE SURVEY

Chen et al. proposed a machine learning algorithm was created to forecast individuals' chances of developing fatty liver disease (FLD). Tests on 577 patients were conducted using the model, which includes logistic regression, random forest, Naïve Bayes, and artificial neural networks. With an area underneath the receiver's operating characteristic of 0.925, the random forest model had the best accuracy among the models, according to the results [7]. Physicians may find it easier to stratify patients with fatty liver for prevention, early therapy, monitoring, and management if the random forest model is used. Strict attention to feature selection and model refining methods has been a recurrent subject in previous studies. Scholars have discerned that the crucial function of choosing relevant clinical features and refining algorithm parameters is to guarantee the resilience and applicability of prediction models. Previous research have attempted to improve the accuracy and reliability of liver disease prediction by carefully selecting input variables and fine-tuning model parameters, taking into account the complexities of various patient groups and disease presentations. Azam et al. this paper addresses the use of computed tomography, magnetic resonance imaging, and ultrasound in computer-aided identification of hepatic lesions in diffuse- and localized liver illnesses. With an emphasis on denoising, deblurring, segmentation, texture characteristics, and support vector machines, it contrasts preprocessing, attribute analysis, and classification approaches [8]. Convolutional neural networks with a deep learning foundation perform best. Future developments in machine learning algorithms that take pathological variables and biopsy samples into account might enhance prediction performance. Additionally, previous studies have emphasized the need of incorporating diverse data sources, such as biochemical markers, lifestyle variables, and demographic features, into prediction models. These studies have highlighted the need of an integrated approach to integrating data, seeking to capture the subtle interaction of numerous risk factors and disease drivers, in light of the complex character of liver illnesses. Prior studies have attempted to create more contextually relevant and sophisticated prediction models that can offer practical insights for medical decisions by integrating a wide range of clinical characteristics.

Rahman et al. proposed that Chronic liver disease, a serious worldwide health concern, is brought on by a number of conditions, including obesity, hepatitis, alcohol abuse, and renal failure [9]. This condition is expensive and difficult to diagnose. The purpose of this study is to assess how well various machine learning algorithms work in order to lower the expense of diagnosing chronic liver disease. Six algorithms were employed by the researchers: Random Forest, K Nearest Neighbors, Decision Tree, Support Vector Machine, Naïve Bayes, and Logistic Regression. The outcomes demonstrated that the LR algorithm had the best accuracy. The study also looks at various data representation techniques and the use of clinical data for predicting liver disease. Furthermore, the transfer of predictive algorithms from studies to clinical practice has been made easier by recent developments in computational methodologies and user interface design. Researchers have attempted to close the gap among data analytics and practical healthcare applications by utilizing cutting-edge technology like intuitive interfaces and decision assistance tools. Previous research has attempted to democratization access to statistical analysis by creating user-friendly and easily navigable platforms for medical practitioners. This has given doctors the tools they need to direct patient treatment and enhance health outcomes.

Ghazal et al. proposed that Liver Disease (LD) is the most common cause of mortality globally, impacting a significant population. It is costly and time-consuming to diagnose LD. Algorithms for machine learning (ML) [10] have the ability to diagnose diseases automatically. The purpose of this study is to evaluate whether or not machine learning algorithms can lower the cost of liver disease detection by using prediction. The prevalence of liver problems is rising, making early identification essential. A suggested intelligent model employing machine learning techniques has an accuracy of 0.884 and a miss-rate of 0.116. Prior studies pertaining to the identification of liver illness have made a deliberate attempt to leverage machine learning techniques in order to improve prediction powers. Numerous algorithms have been examined in these investigations, from more complex methods like random forests, decision trees, and support vector machines to more conventional logistic regression. The investigation of such a wide range of computational strategies highlights the understanding of the complexity involved in diagnosing liver illness and the need for advanced modeling methods to successfully address these issues.

III. METHODOLOGY

The first stage in every data analysis endeavor is data collecting. It entails obtaining pertinent data from a variety of sources, including online scraping, databases, surveys, and APIs. After the data is gathered, it must be prepared so that it may be used for analysis. For example, in order to obtain insights into consumer behavior, data may be gathered for a marketing analysis project via social media platforms, website analytics tools, and customer surveys. Cleaning and getting ready the data for research is known as data preprocessing, and it's an important step. Taking care of duplicates and missing data is frequently part of this phase. Using statistical techniques like mean, median, or mode, null values can be eliminated or imputed.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

If there are any duplicate values, they are found and eliminated to guarantee the accuracy of the data. For example, before continuing with the analysis, it is necessary to manage the values that are missing in the shape of nulls or NaN in the dataset.

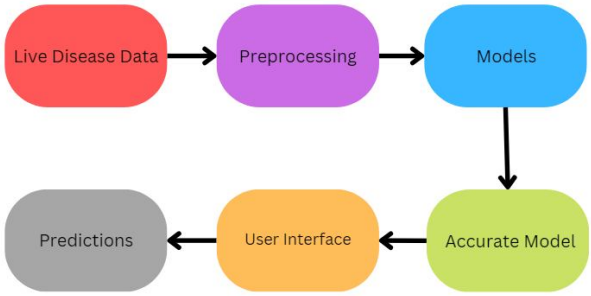


Fig.1 Working Methodology

One preprocessing method for turning data that is categorical into numerical form is label encoding. Label encoding gives each category a distinct number value since numerous algorithms for machine learning are unable to work with categorical information directly. This is especially helpful for decision trees, logistic regression, and support vector machine (SVM) algorithms. When a dataset has categorized variables like "red," "blue," which and "green," for instance, label encoding would translate these groups into integers like 0, 1, and 2, accordingly. Another preprocessing method for scaling numerical characteristics to a range is called min-max scaling. This range is usually between 0 and 1. This is significant because similar-scale features yield superior results for many machine learning techniques. Min-max scaling guarantees that every characteristic have the same scale while maintaining the original distribution's form. The algorithm could assign greater weight to the data with the broader range if, for example, a particular attribute range from 0 to 100 and another from 0 to 100,000. This could lead to biased findings. Several models based on machine learning can be used on the preprocessed data when it has been finished. Random forests, logistic regression, and support vector machines (SVM) are examples of common models. SVM is a potent classification algorithm which can be utilized for both non-linear and linear data, random forest is a method of ensemble learning that can be used for regression as well as classification applications, and logistics regression is a linear model that is utilized for binary classification issues. To generate predictions or glean insights from the data, such algorithms can be taught on the preprocessed material.

IV. MODELS

A. Logistic Regression

When there are only two potential outcomes for a categorical outcome variable [11], such as in binary classification tasks, logistics regression is a statistical approach that is employed. Logical regression is a categorization technique, not a regression procedure, despite its name. It uses the logistic function, which is also referred to as the sigmoid function, to describe the likelihood that a given input falls into a particular category. Upon training on a dataset, logistic regression [12] may attain a high degree of accuracy, frequently above 98%. Using a logistic function, the process in logistic regression determines the likelihood that a given input falls into a specific category. The likelihood is then multiplied by a threshold number, usually 0.5, to get a binary result. The input is categorized as pertaining to one group if the computed probability is higher than the threshold; if not, it is categorized as pertaining to the other group. The excellent precision with which logistic regression was able to differentiate between the two groups within the dataset is demonstrated by its high accuracy.

Because of its high accuracy, logistic regression is used extensively in a variety of industries, such as marketing, banking, and healthcare, to predict things like whether an individual has a certain condition, whether a buyer will purchase a product, and whether or not an email is spam. Even though it's straightforward, logistic regression may be rather useful if the data behaves well and there is a roughly linear connection between the input and result variables. On substantially non-linear data, however, it might not perform well; in this scenario, more sophisticated models [13], such decision trees or support vector machines, would be more suited. Predictions and their probability are mapped using logistic regression using a logistic function known as the sigmoid function. An S-shaped curves that transforms any real number into an interval between 0 and 1 is known as the sigmoid function. Moreover, the framework predicts this the particular instance corresponds to that class if the estimated probability produced by the sigmoid function exceeds a predetermined threshold on the graph.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

The model anticipates that an instance does not belong in the class if its calculated likelihood is less than the predetermined threshold. For logistic regression, a sigmoid function [14] is known as a function of activation and is described as follows:

$$f(x) = \frac{1}{1 + e^{-x}}$$

Where e is the base of natural algorithms and value is numerical value that needs to transform.

B. Support Vector Machine

In machine learning, support vector machine (SVM) is a popular and effective classification technique. It works especially well for situations involving the categorization of both non-linear and linear data. SVM operates by identifying the ideal hyperplane for dividing the dataset's various classes. The margin, or the separation between the hyper plane and the closest point of data from each class, is maximized by selecting this particular hyperplane. SVM seeks to identify the most reliable decision boundary that performs well when applied to previously unknown data by optimizing the margin. With a dataset as its training set, SVM may get an outstanding precision of up to 71%.

SVM's [15] 71% accuracy rate practically indicates that it is good at categorizing data into relevant groups. SVM is often utilized in many different disciplines, including the classification of images, text classification, and bioinformatics, because to its capacity to handle complicated data distribution and non-linear decision boundaries. However, selecting the right kernel function and fine-tuning model parameters are necessary to get best performance using SVM. Furthermore, even while SVM works well on a wide range of datasets, it may not function as well on very big or highly noisy datasets. Nonetheless, SVM may be a very useful tool for machine learning classification jobs with the right optimization and parameter [16] adjustment.

C. Random Forest

A potent ensemble learning technique for both regression and classification applications is called Random Forest [17]. In order to function, it builds a large number of choice trees throughout training and outputs the class that represents the mean prediction (regression) or method for the classes (classification) of the individual trees [18]. A random portion of the data used for training and a selected number of the characteristics are used to construct each tree in the forest. Random Forest may attain an astounding level of precision up to 100% when trained on a dataset. The fact that Random Forest has a 100% accuracy rate shows how well it can categorize the information into the right categories. By combining the predictions from several decision trees, the Random Forest lessens overfitting and raises the model's overall accuracy. Random Forest is a popular option for a variety of classification applications due to its robustness against noise and outliers.

The Random Forest technique has been effectively applied in a number of fields, including marketing, finance, and healthcare. It is a flexible and dependable technique for classification problems due to its capacity to handle missing values, non-linear correlations, and high-dimensional data [19]. The effectiveness of Random Forest is dependent on a number of variables, including feature selection, dataset characteristics, and parameter adjustment, much like any other machine learning technique. In spite of this, Random Forest continues to provide consistently high levels of accuracy [20], making it one of the most popular and successful algorithms for classification jobs.

V. RESULTS

The Random Forest technique has been effectively applied in a number of fields, including marketing, finance, and healthcare. It is a flexible and dependable technique for classification problems due to its capacity to handle missing values, non-linear correlations, and high-dimensional data. The effectiveness of Random Forest is dependent on a number of variables, including feature selection, dataset characteristics, and parameter adjustment, much like any other machine learning technique. In spite of this, Random Forest continues to provide consistently high levels of accuracy, making it one of the most popular and successful algorithms for classification jobs. Determining any age-related patterns in liver disease requires an understanding of the age distribution across patients. The distribution of frequencies of ages may be displayed on a histogram, which helps us determine which age categories liver disease patients most frequently fall into. Understanding the socioeconomic makeup of liver disease patients and maybe identifying potential age-related risk factors can both be aided by this knowledge.

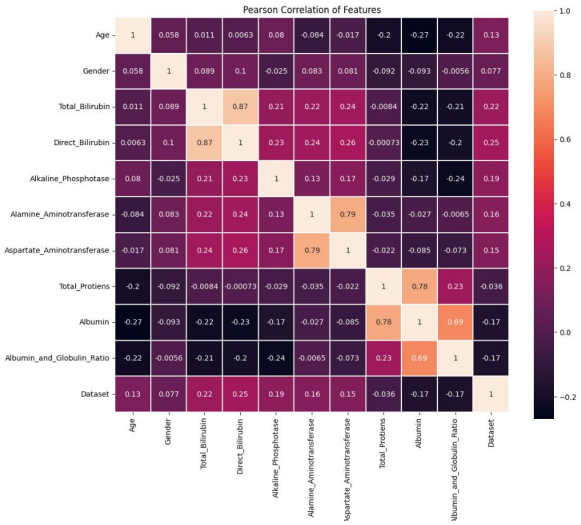


Fig.2 Pearson Correlation of Features

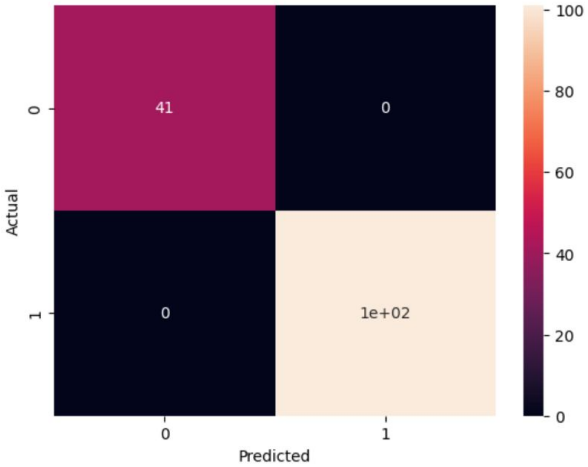


Fig.3 Confusion Matrix for Logistic Regression

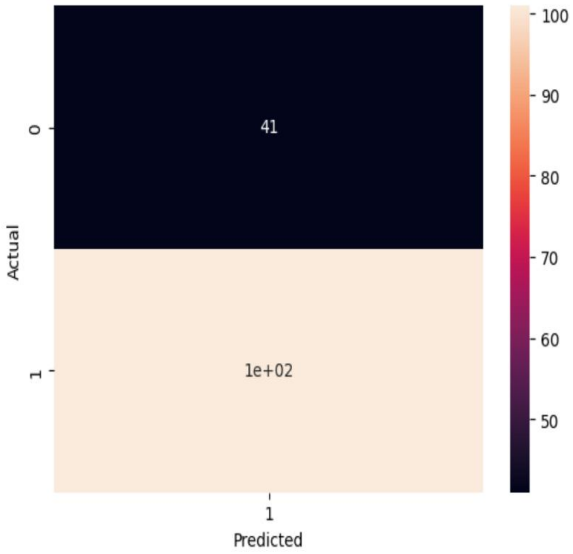


Fig.4 Confusion Matrix for SVM

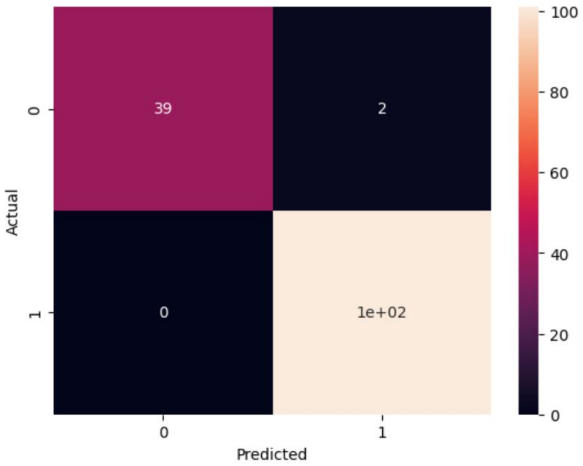


Fig.5 Confusion Matrix for Random Forest

The magnitude and axis of the linear link between several characteristics may be determined by computing the coefficient of Pearson correlation between them. For example, to determine the relationship between these properties, we may compute the correlation between Alkaline Phosphotase, Total Bilirubin, and Direct Bilirubin. These associations may be graphically represented by a correlation matrix, which enables us to determine which traits have the strongest correlations with one another. A table called a confusion matrix is frequently used to explain how well a classification model performs when applied to a set of test data. The confusion matrix indicates the proportion of true positives, true negatives, false positives, and fake negatives for each class in this example, liver disease and no liver illness. We may assess each model's performance by contrasting the expected and actual classes. The confusion matrix gives us important information about how well the models are doing, enabling us to spot any misclassifications and evaluate the models' overall accuracy.

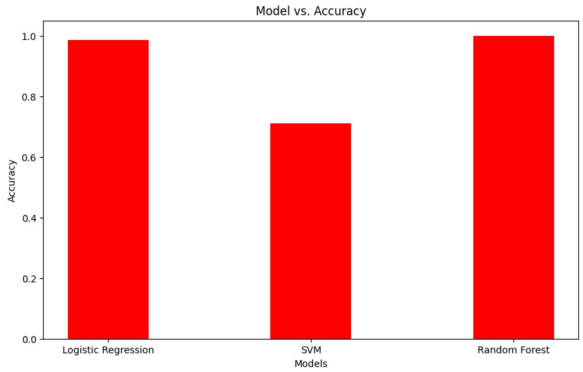


Fig.6 Comparision of Model Accuracies

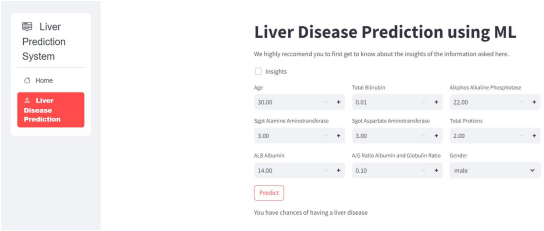


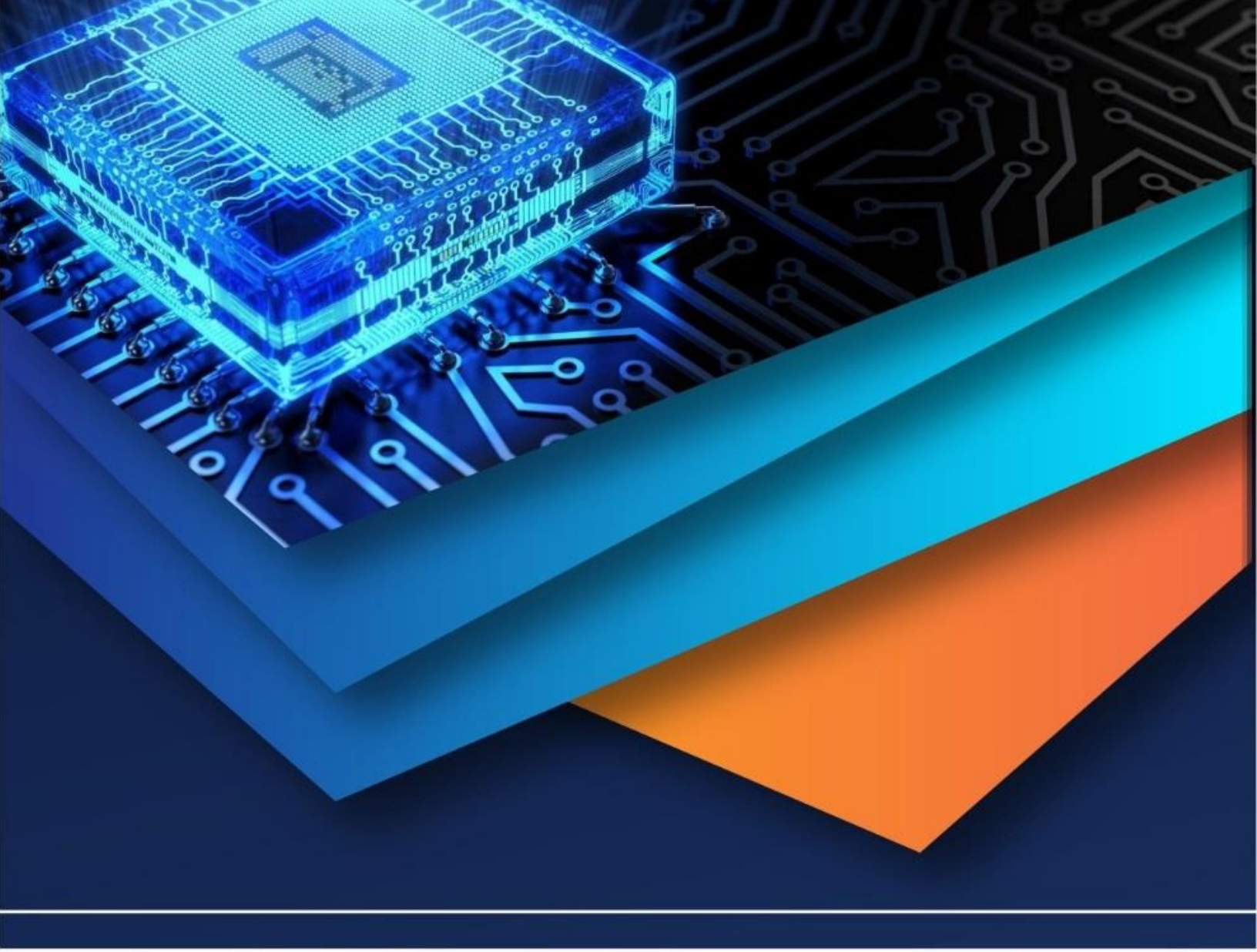
Fig.7 User Interface

The optimal model for liver disease real-time detection must be chosen by comparing the accuracy of several models. We can ascertain which model works best on the provided dataset by comparing each model's precision, recall, precision, accuracy, and F1-score. In addition, to make sure the outcomes are solid and dependable, we might employ strategies like cross-validation. The most effective model may then be used to identify liver illness in real time.

Designing an easy-to-use interface that enables users to enter data and get immediate information on their liver condition is a crucial step in developing a real-time liver disease detection user interface. The interface may provide numerical fields for lab test results, such as Total Bilirubin, Direct Bilirubin, and Alkaline Phosphotase levels, in addition to fields for inputting demographic data, such as age and gender. The interface can show the estimated likelihood of liver illness and any suggestions for additional testing or therapy when the user enters their data. Users with different degrees of technical skill should be able to easily utilize and access the interface. To further safeguard user data, it must be secure and adhere to any privacy standards.

REFERENCES

[1] Acharya, Subrat K. "Epidemiology of hepatocellular carcinoma in India." Journal of clinical and experimental hepatology 4 (2014): S27-S33.
[2] PAUL, Salisu Ojonemi, Michael Sunday Agba, and DC Jr CHUKWURAH. "Rural development programmes and rural Underdevelopment in Nigeria: A rethink." International journal of public administration and management research 2.4 (2014): 1-14.
[3] Heumann, Benjamin W. "An object-based classification of mangroves using a hybrid decision tree—Support vector machine approach." Remote Sensing 3.11 (2011): 2440-2460.
[4] Lee, Chuan-Mo, et al. "Age, gender, and local geographic variations of viral etiology of hepatocellular carcinoma in a hyperendemic area for hepatitis B virus infection." Cancer: Interdisciplinary International Journal of the American Cancer Society 86.7 (1999): 1143-1150.
[5] Pranckevičius, Tomas, and Virginijus Marcinkevičius. "Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification." Baltic Journal of Modern Computing 5.2 (2017): 221.
[6] Boddapati, Mohan Sai Dinesh, et al. "YouTube Comment Analysis Using Lexicon Based Techniques." International Conference on Cognitive Computing and Cyber Physical Systems. Cham: Springer Nature Switzerland, 2022.
[7] Wu, Chieh-Chen, et al. "Prediction of fatty liver disease using machine learning algorithms." Computer methods and programs in biomedicine 170 (2019): 23-29.
[8] Khan, Rayyan Azam, Yigang Luo, and Fang-Xiang Wu. "Machine learning based liver disease diagnosis: A systematic review." Neurocomputing 468 (2022): 492-509.
[9] Rahman, A. Sazzadur, et al. "A comparative study on liver disease prediction using supervised machine learning algorithms." International Journal of Scientific & Technology Research 8.11 (2019): 419-422.
[10] Ghazal, Taher M., et al. "Intelligent model to predict early liver disease using machine learning technique." 2022 International Conference on Business Analytics for Technology and Security (ICBATS). IEEE, 2022.
[11] Long, J. Scott, and Simon Cheng. "Regression models for categorical outcomes." Handbook of data analysis (2004): 259-284.
[12] Tsangaratos, Paraskevas, and Ioanna Ilia. "Comparison of a logistic regression and Naïve Bayes classifier in landslide susceptibility assessments: The influence of models complexity and training dataset size." Catena 145 (2016): 164-179.
[13] Franses, Philip Hans, and Dick Van Dijk. Non-linear time series models in empirical finance. Cambridge university press, 2000.
[14] Perlich, Claudia, Foster Provost, and Jeffrey Simonoff. "Tree induction vs. logistic regression: A learning-curve analysis." (2003).
[15] Wang, Haifeng, and Dejin Hu. "Comparison of SVM and LS-SVM for regression." 2005 International conference on neural networks and brain. Vol. 1. IEEE, 2005.
[16] Syarif, Iwan, Adam Prugel-Bennett, and Gary Wills. "SVM parameter optimization using grid search and genetic algorithm to improve classification performance." TELKOMNIKA (Telecommunication Computing Electronics and Control) 14.4 (2016): 1502-1509.
[17] Rigatti, Steven J. "Random forest." Journal of Insurance Medicine 47.1 (2017): 31-39.
[18] Prasad, Anantha M., Louis R. Iverson, and Andy Liaw. "Newer classification and regression tree techniques: bagging and random forests for ecological prediction." Ecosystems 9 (2006): 181-199.
[19] Ver Steeg, Greg, and Aram Galstyan. "Discovering structure in high-dimensional data through correlation explanation." Advances in Neural Information Processing Systems 27 (2014).
[20] Belgiu, Mariana, and Lucian Drăguț. "Random forest in remote sensing: A review of applications and future directions." ISPRS journal of photogrammetry and remote sensing 114 (2016): 24-31.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)