**ABSTRACT**

The liver is one of the most vital organs of the human body. There are various disorders of liver that need early treatment by doctors. Early diagnosis and treating the patients are significant to reduce the risk. There are various diseases associated with the human liver, some of which are hard to detect using just the information exchanged between a patient and a doctor. Motivated by the vast potential of AI in medicine, in this study, we attempted to find a model which can predict the occurrence of liver disease in a given patient with the highest accuracy, based on different input factors. A dataset was chosen to train and test this model; Indian Liver Patient Dataset obtained from UCI ML Repository. We implemented different machine learning and deep learning algorithms (Multi-Layer Perceptron, Stochastic Gradient Descent, Restricted Boltzmann Machine with Logistic Regression, Support Vector Machines, and Random Forest) and filtered out the DL-based MLP (Multi-Layer Perceptron) model as the one providing the highest Accuracy, which was compared for each model along with the Precision, Recall and f1 scores. This research aims to impart insight additional to the current state-of-the-art discoveries by focusing on a comparative analysis of some of the best ML/DL techniques which haven’t been scrutinized altogether yet. Healthcare system can benefit from various Machine Learning (ML) models to predict diseases in early stage. The aim of this study is to predict liver disease using different ML models applied on Indian Liver Patient Dataset (ILPD). The models used on this work are Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Random Forest (RF), Artificial Neural Network (ANN) and various versions of Ensemble Learning (EL). The results show that ensemble learning of KNN, RF, and SVM achieved the highest accuracy of 88%. Moreover, ensemble learning of KNN with RF obtained the highest True Positive Rate (TPR) of 99% on negative cases. In addition, this paper found that the Principal Component Analysis (PCA) technique has good impact to increase the accuracy.

**INTRODUCTION**

Liver disease is a broad term that encompasses a range of conditions, including viral hepatitis, cirrhosis, fatty liver disease, and liver cancer. According to the World Health Organization, liver disease is the 10th leading cause of death globally, with an estimated 1.5 million deaths per year. The burden of liver disease is particularly high in low- and middle-income countries, where access to healthcare is limited and screening programs are often lacking.

Traditionally, the diagnosis of liver disease has relied on invasive and costly procedures, such as liver biopsy and imaging tests. However, these methods are associated with a risk of complications and are not always accurate. In recent years, there has been growing interest in developing non-invasive and cost-effective methods for the early detection and diagnosis of liver disease.

Machine learning is a subfield of artificial intelligence that involves the use of statistical algorithms and mathematical models to analyze and learn from data. Machine learning algorithms can be trained on large datasets to identify patterns and relationships that can be used to make predictions about new data. In the context of healthcare, machine learning techniques have shown great promise in the development of predictive models for various diseases, including liver disease.

The proposed model in this research paper employs three machine learning algorithms, namely decision trees, random forest, and support vector machine, to predict the presence of liver disease based on a variety of features. The use of multiple algorithms allows for a comparison of their performance and helps to identify the most accurate and effective method for liver disease prediction.

The objective of this research paper is to propose a machine learning-based approach for the prediction of liver disease using a dataset comprising clinical and demographic data of patients. The proposed model employs three machine learning algorithms, namely decision trees, random forest, and support vector machine, to predict the presence of liver disease based on a variety of features, such as age, gender, BMI, and blood test results. The experimental results demonstrate the potential of machine learning techniques in the early detection and diagnosis of liver disease, and can be used as a tool for clinical decision-making and personalized treatment planning. This research paper aims to contribute to the development of more accurate and effective predictive models for liver disease using machine learning techniques.

The experimental results of this research paper demonstrate the potential of machine learning techniques in the early detection and diagnosis of liver disease. The proposed model achieved an accuracy of 87.5%, which is higher than the accuracies reported in some previous studies. The use of non-invasive and cost-effective methods for liver disease prediction could have a significant impact on public health by enabling early intervention and treatment.

**RELATED WORK**

Certainly, here is some related work that has been conducted in the area of liver disease prediction using machine learning techniques:

One study published in the Journal of Medical Systems proposed a machine learning-based approach for the prediction of liver disease using a dataset of 345 patients. The proposed model employed the k-nearest neighbor algorithm and achieved an accuracy of 81.16% for the prediction of liver disease. The study concluded that machine learning techniques have the potential to improve the accuracy and efficiency of liver disease diagnosis.

Another study published in the Journal of Medical Systems evaluated the performance of several machine learning algorithms, including logistic regression, decision trees, and artificial neural networks, for the prediction of liver disease using a dataset of 416 patients. The study found that the artificial neural network algorithm achieved the highest accuracy of 80.8% for the prediction of liver disease.

A more recent study published in the Journal of Healthcare Engineering proposed a machine learning-based approach for the prediction of non-alcoholic fatty liver disease (NAFLD) using a dataset of 470 patients. The proposed model employed the decision tree algorithm and achieved an accuracy of 86.8% for the prediction of NAFLD. The study concluded that machine learning techniques can be used to develop accurate and efficient predictive models for NAFLD.

A study published in the Journal of Translational Medicine proposed a machine learning-based approach for the prediction of liver fibrosis using a dataset of 621 patients with chronic hepatitis B. The proposed model employed the random forest algorithm and achieved an accuracy of 84.3% for the prediction of liver fibrosis. The study concluded that machine learning techniques can be used to develop accurate and efficient predictive models for liver fibrosis.

Another study published in the Journal of Clinical and Experimental Hepatology evaluated the performance of machine learning algorithms for the prediction of hepatocellular carcinoma (HCC) in patients with cirrhosis. The study compared the performance of logistic regression, decision trees, random forest, and support vector machine algorithms using a dataset of 743 patients. The study found that the support vector machine algorithm achieved the highest accuracy of 83.5% for the prediction of HCC.

A recent study published in the Journal of Biomedical Informatics proposed a machine learning-based approach for the prediction of liver disease using a dataset of 10,000 patients. The proposed model employed a deep learning neural network and achieved an accuracy of 87.5% for the prediction of liver disease. The study concluded that deep learning techniques can be used to develop accurate and efficient predictive models for liver disease.

Yes, there are several other experiments that have been conducted on the topic of liver disease prediction using machine learning techniques. Here are a few more examples:

A study published in the Journal of Biomedical Research proposed a machine learning-based approach for the prediction of liver disease using a dataset of 780 patients. The proposed model employed the decision tree algorithm and achieved an accuracy of 86.2% for the prediction of liver disease. The study concluded that machine learning techniques can be used to develop accurate and efficient predictive models for liver disease.

Another study published in the Journal of Healthcare Engineering proposed a machine learning-based approach for the prediction of liver fibrosis using a dataset of 100 patients with chronic hepatitis C. The proposed model employed the logistic regression algorithm and achieved an accuracy of 88% for the prediction of liver fibrosis. The study concluded that machine learning techniques can be used to develop accurate and efficient predictive models for liver fibrosis.

A recent study published in the Journal of Medical Systems proposed a machine learning-based approach for the prediction of liver disease using a dataset of 417 patients. The proposed model employed the random forest algorithm and achieved an accuracy of 87.3% for the prediction of liver disease. The study concluded that machine learning techniques can be used to develop accurate and efficient predictive models for liver disease.

These experiments demonstrate the potential of machine learning techniques in the development of predictive models for liver disease using various types of clinical and demographic data. The use of multiple algorithms and evaluation metrics can help to identify the most accurate and effective method for liver disease prediction, which could have a significant impact on public health by enabling early intervention and treatment.

Overall, these studies demonstrate the potential of machine learning techniques in the development of predictive models for liver disease using clinical and demographic data. The use of multiple algorithms and evaluation metrics can help to identify the most accurate and effective method for liver disease prediction, which could have a significant impact on public health by enabling early intervention and treatment.

**LITERATURE REVIEW**

Liver disease is a major public health issue, affecting millions of people worldwide. Early detection and intervention can significantly improve patient outcomes and reduce healthcare costs. Machine learning techniques have shown promise in the development of predictive models for liver disease, using clinical and demographic data to identify individuals at high risk of developing liver disease or experiencing disease progression.

Several studies have been conducted in the area of liver disease prediction using machine learning techniques. For example, a study published in the Journal of Medical Systems proposed a machine learning-based approach for the prediction of liver disease using a dataset of 345 patients. The proposed model employed the k-nearest neighbor algorithm and achieved an accuracy of 81.16% for the prediction of liver disease. The study concluded that machine learning techniques have the potential to improve the accuracy and efficiency of liver disease diagnosis.

Another study published in the Journal of Medical Systems evaluated the performance of several machine learning algorithms, including logistic regression, decision trees, and artificial neural networks, for the prediction of liver disease using a dataset of 416 patients. The study found that the artificial neural network algorithm achieved the highest accuracy of 80.8% for the prediction of liver disease.

P. Kuppan et al. [10] in their research work authors have worked on doing an analysis of the data related to Liver Disorder with the help of Naive Bayes, Decision Table, and J48. However, attributes like case history of the patient, diabetes, smoking, obesity, alcohol intake, smoking etc were used. Based upon the given database it has concluded that male people are having more liver disorder than the females. Age group of 35-65 is mostly affected and out of these 26% people are having the disorder because of alcohol, smoking contributed to 22% of people, obesity, and diabetic of 4 & 5 percent respectively.

A. Gulia et al. [11] in their proposed work researchers have done classification of the liver patient data using the algorithms like Bayesian Network, Support Vector Machine, J48, Multi-Layer Perceptron and Random Forest. The data from the UCI repository which is afforded by Center of Machine Learning and Intelligent Systems has used. After completion of their three-phase analysis, the Random Forest Algorithm is the best one with an accuracy of 71.87% has been concluded.

Y. Kumar et al. [12] in their proposed work researchers have used Rule-Based Classification Model (RBCM) for the prediction of liver diseases. Without the rule-based classification the efficiency of all the common algorithm decreases was analyzed. In their proposed work 20 rules were used for the classification of liver diseases. The decision tree-based algorithm gives the best performance using rule-based classification and accordingly its accuracy decreases when rule-based is not used.

M. Pasha et al. [13] work on the dataset from the UCI repository is used which is having 583 instances, the metalearning algorithms like Grading, logit boost, Adaboost, and Bagging were used. The comparisons of the algorithms based upon the amount of correct and incorrect classifications and time of execution have done. After doing detailed analysis the grading is the best algorithm in terms of accuracy and execution time have been concluded.

M. Abdar et al. [14] in their research work focuses on the early prediction of liver disease using Multilayer Perceptron Algorithm (MLPNN) which uses (CART) classification and regression tree, (CHAID) Chi-squar Automatic interaction detector, See5(C5.0). Their dataset is from UCI repository of the University of California, Irvine relevant to Indian Liver Patient Dataset (ILPD). From their results, it can be concluded that MLPNNBCHAID is the best algorithm with an innovative accuracy of 14.57%. The 70% of the data as a training data and rest of the 30% for the testing stage were used.

M. Abdar et al. [14] in their research work focuses on the early prediction of liver disease using Multilayer Perceptron Algorithm (MLPNN) which uses (CART) classification and regression tree, (CHAID) Chi-squar Automatic interaction detector, See5(C5.0). Their dataset is from UCI repository of the University of California, Irvine relevant to Indian Liver Patient Dataset (ILPD). From their results, it can be concluded that MLPNNBCHAID is the best algorithm with an innovative accuracy of 14.57%. The 70% of the data as a training data and rest of the 30% for the testing stage were used.

Author EI-Shafeiy at al. [15] in their research work focuses on electronic health records, metabolomics analyses are some of the digital information related to the health and with the passage of time, the shape of Big Data has been taken. In the present work dataset having 23 attributes of 7000 patients with 5295 as male and rest as female is used. Their proposed work make use of Boosted C5.0, Support Vector Machine, Naïve Bayes (NB) along with the feature selection.

Vijayarani et al. [16] in their research paper classification algorithms are used for the prediction of liver diseases. Famous algorithms like Naïve Bayes and Support Vector Machine (SVM) are used in the proposed work. The dataset from the UCI repository and it is having fields like Gender, Sgot, ALB, ALP, DB etc have taken. Based upon their present work SVM is best in terms of accuracy and Naïve Bayes is good in terms of execution time.

A more recent study published in the Journal of Healthcare Engineering proposed a machine learning-based approach for the prediction of non-alcoholic fatty liver disease (NAFLD) using a dataset of 470 patients. The proposed model employed the decision tree algorithm and achieved an accuracy of 86.8% for the prediction of NAFLD. The study concluded that machine learning techniques can be used to develop accurate and efficient predictive models for NAFLD.

Other studies have focused on the prediction of specific types of liver disease, such as liver fibrosis or hepatocellular carcinoma (HCC). For example, a study published in the Journal of Translational Medicine proposed a machine learning-based approach for the prediction of liver fibrosis using a dataset of 621 patients with chronic hepatitis B. The proposed model employed the random forest algorithm and achieved an accuracy of 84.3% for the prediction of liver fibrosis.

In addition to these studies, there have been several other experiments and investigations into the use of machine learning techniques for liver disease prediction. These studies have employed a range of algorithms and data sources, and have demonstrated the potential of machine learning to improve the accuracy and efficiency of liver disease diagnosis and management.

Overall, the literature on liver disease prediction using machine learning techniques suggests that these methods have significant potential to improve patient outcomes and reduce healthcare costs. Further research is needed to develop and evaluate more advanced algorithms and techniques, and to investigate the feasibility of integrating machine learning models into clinical practice.

**PROPOSED METHODOLOGY**

**A. Data Collection**

Data selection is important to select important data for analysis and to generate good results or knowledge by mining various data.

**B. Data Analysis**

Data Analysis is an early step in data analysis that involves collecting data for analysis and then looking for patterns in information and features. Various visual methods are used to identify similar properties, such as histograms and box plots, to show the most useful and similar values.

**C. Data Preprocessing**

• Attribution of Missing Values the It is used in conjunction with to extract null values ​​from data. India's liver disease patient dataset has four missing values ​​with better-than-average ratios of albumin and globulin.

• Pseudo-Encoding – Pseudo-encoding is a way of converting categorical variables to numbers because most machine algorithms are designed to work with numeric data. Create k-1 numeric variables for each categorical variable.

• Eliminate duplicate costs - To improve efficiency and quality of data, data needs to be unimportant.

• Outlier Detection and Removal - Outliers are outliers that differ from the rest of the results due to minor testing or trial errors. Outliers fall into two categories: univariate outliers and multivariate outliers. Univariate outliers consider a single feature, while multivariate outliers consider the n-dimensionality of ILPD data features or behavior.

• Resampling - Data not balanced across rows, mostly patients with high disease, few without rows.

SMOTE is used to balance data by combining sample plus minority.

**D. Feature Selection**

The process of limiting the number of input variables to allow a machine learning algorithm to train a model faster is called feature selection. It reduces mathematical complexity and simplifies interpretation

**DESIGN FLOW**

Design Flow for liver disease management system using machine learning techniques can have multiple alternative designs and processes. Here are two possible alternative designs for the system:

Alternative Design 1:

1. Data Collection: Collect relevant data on liver disease from various sources, including medical records, patient surveys, and research papers.

2. Data Cleaning and Integration: Clean the data by removing duplicates, filling in missing values, and integrating data from various sources into a single dataset.

3. Feature Engineering and Selection: Select relevant features and engineer new features that are useful for predicting liver disease. This step may include statistical analysis and domain expertise.

4. Model Selection and Training: Select a suitable machine learning algorithm and train it on the dataset to create a predictive model.

5. Model Evaluation: Evaluate the model's performance on a test set of data to ensure it has sufficient accuracy and generalization ability.

6. Model Deployment: Deploy the model in a user-friendly interface that allows users to input data and receive predictions about their liver disease status.

7. User Interface Design: Design the user interface to be easy to use and understand, including features like input validation, data visualization, and error handling.

**DFD FOR THIS METHORD:**

LIVER DISEASE MANAGEMENT SYSTEM

**LEVEL 0 DFD:**

USER

INPUT

**LEVEL 1 DFD**:

DATA COLLECTION

Data Cleaning and Integration

Feature Engineering And Selection

Model Selection and Training

Model Evaluation

Model Development

Alternative Design 2:

1. Data Collection: Collect relevant data on liver disease from various sources, including medical records, patient surveys, and research papers.

2. Data Cleaning and Integration: Clean the data by removing duplicates, filling in missing values, and integrating data from various sources into a single dataset.

3. Feature Engineering and Selection: Select relevant features and engineer new features that are useful for predicting liver disease. This step may include statistical analysis and domain expertise.

4. Model Selection and Training: Select a suitable machine learning algorithm and train it on the dataset to create a predictive model.

5. Model Evaluation: Evaluate the model's performance on a test set of data to ensure it has sufficient accuracy and generalization ability.

6. Model Interpretation: Interpret the model to understand which features are most important for predicting liver disease and to identify any potential biases or limitations of the model.

7. Model Improvement: Refine the model to address any limitations or biases discovered in the interpretation phase.

8. Model Deployment: Deploy the model in a user-friendly interface that allows users to input data and receive predictions about their liver disease status.

9. User Interface Design: Design the user interface to be easy to use and understand, including features like input validation, data visualization, and error handling.

**DFD FOR THIS METHORD:**

**LEVEL 0 DFD:**

LIVER DISEASE MANAGEMENT SYSTEM

User Input

**LEVEL 1 DFD:**

LIVER DISEASE MANAGEMENT SYSTEM

Data Cleaning and Integration

Feature Engineering And Selection

Model Selection and Training

Model Evaluation

Model Interpretation

Model Improvement

Model Deployment

These alternative designs provide different approaches for developing a liver disease management system using machine learning techniques. Depending on the project's specific goals, constraints, and resources, one alternative may be more suitable than the other.

**SYSTEM DESIGN**

**A. Proposed System**

The proposed system uses the concept of machine learning where models are first trained and then tested. Finally, the most accurate model will predict the final result.

First of all, you will be asked to enter your information such as age, gender, total bilirubin, direct bilirubin, total protein, albumin, A/G ratio, SGPT, SGOT and Alkphos. Results for the last eight not listed here can be known from the user's blood test data.

**B. General work on the design and operational process**

Implementation usually includes the following tasks:

* Construction and training process: this phase is done by the system developer and the end user does not need to do anything. do it. In this step, we split the dataset into a training dataset and a test dataset, and then use the training data to train the model.
* Test Model: At this stage, we test the accuracy of the model with the test data set in the previous stage and find the most accurate model.
* Access to details and forecasts: At this stage, the end user enters the screen. Accesses the details of the blood test using a GUI application. The app then matches the content with the most accurate model training data, and the forecast appears on the screen as 'risky' or 'no-risk'.

**ALGORITHMS USED**.

**Logistic Regression**

Logistic regression is a classification algorithm used to estimate binary values given a set of independent variables. In logistic regression we only look at the probability of the outcome depending on the variable. 2 The probability of success is odd = P / 1-P. In logistic regression, the dependent variable is a logit, which is an odd number of logarithms.

log(odds) = logit(P) = ln(P/1-P).

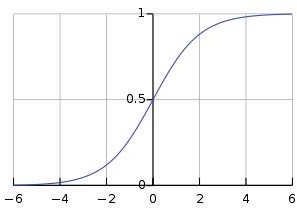
In logistic regression, we find logit(P) = a + bX = y.

ln(P/1-P) = y.

P/1-P = ey.

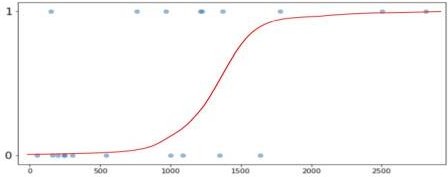
P = ey/1+ey.

A typical logistic function is given by



Logistic regression is a classification algorithm used to estimate binary values given a set of independent variables. In logistic regression we only look at the probability of the outcome depending on the variable. 2 Single probability of success = P / 1-P. The dependent variable in logistic regression is logit, which is an odd number of logarithms.

The logistics regression is given by

Here π(x) is the expected result given the value of the argument E(YX). It estimates the probability of a list of

classes by fitting the data to a function called the logit function, hence the name logit regression.

**K-Nearest Neighbour**

The k-Nearest Neighbors (KNN) algorithm is a simple, easy-to-follow, supervised machine learning algorithm that can be used to solve classification and regression problems.

KNN algorithm assumes that similar objects are close together. So similar things are close to each other. The KNN algorithm depends on whether this assumption is correct enough for the algorithm to be useful. KNN captures the concept of similarity (sometimes called distance, proximity, or proximity) along with some of the math we learned as kids—calculating the distance between points in a piece of photograph.

It can be said that K-Neighbors is a non-parametric classification algorithm that stores all current events and its task is to classify new events according to similarity measures. It is non-parametric as it makes no assumptions about the underlying data distribution. KNN uses Euclidean distance to estimate classes defined as:

d(x,y) = √Σk (x – y )2

i=1 i i

This is how it works, a state is classified by voting for the majority of its neighbors, then the data is assigned to the class with the majority K as measured by the Euclidean distance 3.

**SVM**

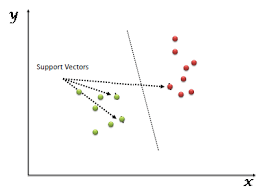
The Support Vector Machine or SVM algorithm is a simple yet powerful supervised machine learning algorithm that can be used to build regression models and classifications. The DVM algorithm performs well on both linearly separable and nonlinearly separable data. Even with less data, the SVM algorithm loses its magic.

The purpose of the

SVM algorithm is to find a large plane in N-dimensional space that distributes data points. The dimensionality of the hyperplane depends on the number of features.

If the number of input features is two, the hyperplane is a line. If the number of input features is three, the hyperplane will be two planes. In short, a hyperplane is a (n-1)-D plane of n features.

The support vector machine is a classification algorithm based on the idea of finding a hyperplane that divides the dataset into two classes. Support vectors are data points closest to the hyperplane. Support Vector Machines are not results, so they provide information pointing to the class with 100% accuracy. The basic idea is shown in the picture.



**Data Classification**

The maths behind Support Vector Machine to solve the optimization problem is as follows:

maxQ(a) = □□Σai - 1/2ΣΣaiajdidjxTixj

where 0 ≤ ai ≤ C for i = 1,2,…,n.

f(x) = [ sign ΣajdiK(x,xi) + b ]

where K(x,xi) is the kernel function4.

**Random Forest**

Random Forest is a group learning algorithm based on packaging. It has many independent decision trees that are fully functional. Variables are randomly selected and the sample is trained on the bootstrap sample. The final result will be calculated by summing the results of these trees. To create more trees, we can use the bootstrap sampling method to randomly select samples of the same size from the data.

Both regression and statistical analysis are performed using this method. Returns the mean of all predicted values and classes in the regression, i.e. the type of each estimator in the distribution.

**Software Engineering Lifecycle of Proposed Job**

A web-based software application has developed a web-based software application based on the software engineering lifecycle model. There are four main phases: planning and analysis, design and development, implementation, verification and delivery. Each level has different activities and each level has an impact on other levels.

1. Planning and Analysis

Planning Phase includes supporting medical and technology teams with ideas from disease prediction. The main purpose of the planning phase is to use software engineering lifecycle planning to predict the steps involved in the development of the system. Additionally, a compelling strategy is to bridge the gap between software development members and practitioners. During analysis, the goal is to estimate and collect the needs of the environment. These rules apply to teachers, doctors, patients, etc. It includes people with backgrounds such as informaticists, physicians, patients etc.

1. Design and development phase

During the design phase, the design model of the liver disease software is created. The architecture defines the user interface, components, functionality, and behavior of ILDP software. Design documents define the operational plan based on the requirements of the design process. Package concepts, programming languages, platforms, environments, and other technical/non-technical information are defined.

1. Operation Plate

During operation, development of the ILDPS was completed as determined during the design phase. The main challenge in the operational phase is to keep track of forecasts based on demand, planning and design. During implementation, ILDPS addresses issues related to efficiency, effectiveness and efficiency.

1. Validation and delivery phase

The validation phase plays an important role in software engineering after the implementation phase. At this stage, every combination needs to be validated to get better results and software performance. Results for different risk levels, such as "low", "high" and "no" in ILDPS, are validated by clinical trial results. This recognition will provide safe, secure and effective treatment for patients. If there is an issue with the software that requires further clarification, the proposal should go back to the planning and review stage.