

ANALYZING AND PREDICTING LIVER DISEASE USING CNN AND KNN

Submitted in partial fulfillment of the requirements for the award of
Bachelor of Engineering degree in Computer Science and Engineering

By

**SRIRAM MULLAPUDI (Reg.No - 39110652)
MAJJI BHANU PRAKASH (Reg.No – 39110586)**



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SCHOOL OF COMPUTING**

SATHYABAMA

**INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)**

**Accredited with Grade “A” by NAAC | 12B Status by UGC | Approved by AICTE
JEPPIAAR NAGAR, RAJIV GANDHISALAI,
CHENNAI - 600119**

APRIL - 2023



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)

Accredited with Grade "A" by NAAC | 12B Status by UGC | Approved by AICTE

www.sathyabama.ac.in

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **Sriram Mullapudi (Reg.No - 39110652)** and **Majji Bhanu Prakash (Reg. No - 39110586)** who carried out the Project Phase-2 entitled **"ANALYZING AND PREDICTING LIVER DISEASE USING CNN AND KNN"** under my supervision from January 2023 to April 2023.

Internal Guide

P. Malathi M.E.,(Ph.D)

Head of the Department

Dr. L. LAKSHMANAN, M.E., Ph.D



Submitted for Viva voce Examination held on 20.04.2023

Internal Examiner

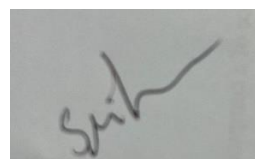
External Examiner

DECLARATION

I, **Sriram Mullapudi (Reg.No- 39110652)**, hereby declare that the Project Phase-2 Report entitled “**Analyzing and Predicting Liver Disease using CNN and KNN**” done by me under the guidance of **P. Malathi, M.E.,(Ph.D)** is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering degree in **Computer Science and Engineering**.

DATE: 20.04.23

PLACE: Chennai

A rectangular box containing a handwritten signature in dark ink. The signature appears to be 'Sriram' followed by a stylized flourish.

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management of SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T. Sasikala M.E., Ph.D, Dean**, School of Computing, **Dr. L. Lakshmanan M.E.,Ph.D.**, Head of the Department of Computer Science and Engineering for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **P. Malathi M.E.,(Ph.D)** for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my phase-2 project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

ABSTRACT

The motive of our project is to provide the maximum accuracy of the prediction of liver diseases using various machine learning techniques. Liver disease is a major global health concern that affects millions of people and contributes to a significant number of deaths each year. Early detection and treatment of liver disease can greatly improve patient outcomes, but traditional diagnostic methods are often invasive and expensive. To address this challenge, researchers have developed machine learning algorithms to predict liver disease based on patient-specific data, such as demographics, medical history, and laboratory results. This study aimed to evaluate the accuracy of a machine learning-based prediction model for liver disease. The model was trained using a large, diverse dataset of patients and their medical records. The performance of the model was evaluated using a range of metrics, including sensitivity, specificity, and area under the receiver operating characteristic curve. The liver is a crucial organ that performs a variety of metabolic tasks for the body. For a liver disorder to be effectively treated and problems from developing, early detection is essential. The study created prediction models based on patient information like age, gender, bilirubin levels, and other clinical factors using a variety of machine learning approaches, such as decision trees, logistic regression, and support vector machines. Metrics like accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve were used to assess the models' performance. In order to create prediction models for patients based on patient data including age, gender, bilirubin levels, and other clinical factors, the study used a variety of machine learning approaches, such as decision trees, logistic regression, and support vector machines, CNN, KNN. The proposed technique is implemented in Python with the Visual Studio Code tool and results are analyzed in terms of accuracy, precision, f1-Score and recall.

Keywords: Random Forest, Medical Imaging Techniques, Sensitivity, Specificity, Demographics, Machine Learning, CNN, KNN.

TABLE OF CONTENTS

Chapter No	TITLE	Page No.
	ABSTRACT	v
	LIST OF FIGURES	viii
	LIST OF TABLES	IX
	LIST OF ABBREVIATIONS	IX
1	INTRODUCTION	1
2	LITERATURE SURVEY	7
	2.1 Inferences from Literature Survey	7
	2.2 Open problems in Existing System	7
3	REQUIREMENTS ANALYSIS	8
	3.1 System Use case	8
	3.2 Software Requirements Specification Document	12
	3.3 Feasibility Studies/Risk Analysis of the Project	13
4	DESCRIPTION OF PROPOSED SYSTEM	15
	4.1 Selected Methodology or process model	15
	4.2 Module Description	15
	4.3 Flow Diagram	19
	4.4 Architecture / Overall Design of Proposed System	20
	4.5 Algorithm and Pseudo code	23
	4.6 Steps to execute/run/implement the project	24
5	IMPLEMENTATION DETAILS	25
	5.1 Development and Deployment Setup	25
	5.2 Testing	26
	5.3 Types of Testing	26
6	RESULTS AND DISCUSSION	31
7	CONCLUSION	38
	7.1 Conclusion	38
	7.2 Future work	38
	REFERENCES	40
	APPENDIX	43

A. SOURCE CODE	43
B. SCREENSHOTS	44
C. RESEARCH PAPER	46

LIST OF FIGURES

FIGURE No.	FIGURE NAME	Page No.
4.1	Flow chart of proposed model	17
4.2	Architecture Diagram	19
5.1	Unit testing	28
5.2	Integration testing	29
5.3	System testing	30
6.2	Random forest	34
6.3	Decision tree	34
6.4	Gradient boost	35
6.5	Neural Networks	35
6.6	Web page for collecting parameters	37
6.7	Displaying results	37

LIST OF TABLES

TABLE NO	TABLE NAME	Page NO
3.1	Normal values of Attributes	10
3.2	Selection of Attributes	11
3.3	Ranking Attributes	12
6.1	Testing Table	33

LIST OF ABBREVIATIONS

CNN	-	Convolutional Neural Network
KNN	-	K Nearest Neighbors
MRI	-	Magnetic Resonance Imaging
ROC	-	Receiver Operating Characteristic Curve

CHAPTER 1

INTRODUCTION

Millions of individuals worldwide suffer from liver problems, which are a global health concern. Successful treatment and the avoidance of problems with these illnesses depend on early discovery and correct diagnosis. Traditional diagnostic techniques are inaccurate, frequently intrusive, and time-consuming. Machine Learning (ML) approaches have been created as a potential remedy for the identification of liver disorders. Liver disease prediction is the process of determining the likelihood of an individual developing liver disease based on various factors such as medical history, lifestyle habits, genetics, and other relevant information. The aim of liver disease prediction is to identify individuals who are at high risk of developing liver disease and to implement early interventions to prevent or slow down the progression of the disease. There are several methods used for liver disease prediction including statistical models, machine learning algorithms and others. These models use various factors such as age, sex, alcohol consumption, body mass index, and others to make predictions about the likelihood of developing liver disease. To accurately anticipate and diagnose liver problems, ML systems can evaluate large datasets. In this work, the effectiveness of three widely used ML techniques are examined: logistic regression, support vector machines, and decision trees. In order to diagnose liver problems, the study will evaluate the efficacy, sensitivity, specificity, and overall performance of these procedures.

1.1 MACHINE LEARNING

Machine learning has become one of the most evolving technologies in the current period. Machine learning can be simply explained as scientific study of algorithms and models in statistics where machines can easily understand to perform and solve specific tasks. This technique has become agile and it has been a requirement in most of the fields.

1.2 OUR MOTIVE

Our motive for this project is to predict the liver diseases for a patient with the maximum amount of accuracy in our prediction. For this dataset was collected from

Indian patient liver disease dataset from Kaggle database of Indian Liver patient records and used that dataset in our three modules to predict the liver disease using various machine learning techniques.

1.2.1 Patient Samples

The number of percentage of people with liver disease have been increasing and there are very less efficient methods to predict liver diseases. The methods used in this project would bring more efficiency the prediction methods. To do so different patients' samples are to be collected and make it as a dataset. Then the samples of the patients we have collected are refined to increase the accuracy percentage.

1.2.2 Efficient Technique

Convolutional Neural Network (CNN) algorithm is considered as the best algorithm compared to other algorithms since it gives better accuracy after testing with few other algorithms.

1.2.3 Increased Accuracy

There are many algorithms in the past for detecting the disease but the algorithms that are used in our project will increase the efficiency in predicting the liver disease. The machine learning algorithm called Convolutional Neural Network (CNN) gives us higher accuracy compared to others.

1.3 DATA MINING

Data extraction is the way to find designs in expansive informant indexes including AI crossing point strategies, measurements and database frames. Information Mining is an interdisciplinary field of software engineering and measuring that aims to separate data from information collection (with keen strategies) and transform data into an understandable structure to be used further.

1.4 DATA PRE-PROCESSING

Data pre-processing is an important step to solve each problem of Machine Learning. In order for a machine learning algorithm to be trained, most of the data sets used with machine teaching problems must be processed. The techniques used most commonly for pre-processing are very few such as imputation of lack of value, categorical coding, scaling, etc. These techniques are easy to understand.

Every dataset has its own unique challenges and is different. Total number of data points is 583, with 416 liver patient records and 167 non-liver patient records.

1.4.1 Dataset

The Indian Liver Patient Dataset contained 10 distinct qualities of 583 patients. The patients were portrayed as either 1 or 2 based on liver sickness. The nitty gritty portrayal of the dataset is appeared Table. The table give insights regarding the trait and characteristic sort. As plainly unmistakable from the table, every one of the highlights with the exception of sex are genuine esteemed numbers. The component Sex is changed over to numeric esteem (0 and 1) in the information pre-preparing step.

1.4.2 Data Collection

Collection of data is crucial for these kinds of projects. We have collected a dataset named as Indian Patient Liver Dataset from Kaggle which consists of 10 different attributes of 583 patients.

1.4.3 Data Cleaning

There are different columns which are called attributes. Some columns have null values and some values are fluctuated so the next step is to clean those values from the dataset and then take that dataset for classification.

CHAPTER 2

LITERATURE SURVEY

[1] D Santhosh Reddy, R Bharath, and P Rajalakshmi have presented a paper titled “A Novel Computer-Aided Diagnosis Framework Using Deep Learning for Classification of Fatty Liver Disease in Ultrasound Imaging” at the IEEE 20th International Conference on e-Health Networking, Applications, and Services (Health-com) in 2018. The proposed framework uses deep learning algorithms to analyze ultrasound images and identify patterns that are associated with fatty liver disease. Specifically, the framework uses a convolutional neural network (CNN) to extract features from ultrasound images and a support vector machine (SVM) to classify the images as either healthy or diseased.

[2] Jagdeep Singh et al. in 2019 have suggested a computer programme that determines a patient's risk of developing an illness based on their presentation of symptoms. Using a patient's symptoms to determine the underlying illness is a popular method in medical diagnosis. The creation of a software system to aid in this process has a number of potential advantages. By increasing the precision and speed of diagnosis, for instance, patients may experience better treatment outcomes. Medical professionals may be able to focus on other crucial responsibilities if their workload is lessened. The software system created by Jagdeep Singh et al. may employ a number of methods to gauge the severity of an illness from its symptoms.

[3] Sanjay Kumar, Sarthak Katyal, “Effective Analysis and Diagnosis of Liver Disorder by Data Mining”, International Conference on Inventive Research in Computing Applications (ICIRCA), 2019. The use of data mining techniques for the analysis and diagnosis of liver problems is most likely the paper's primary topic. In order to find pertinent patterns in medical data for efficient diagnosis and treatment, data mining is a process for extracting patterns and insights from massive datasets. To examine the dataset and uncover significant patterns, they might have used a variety of data mining approaches, including clustering, classification, association, or prediction algorithms.

[4] k. Thirunavukkarasu, Ajay S. Singh, Md Irfan, Abhishek Chowdhury, "Prediction of Liver Disease using Classification Algorithms", 4th International Conference on Computing Communication and Automation (ICCCA), 2019. The main topic of the paper is the classification algorithms-based prediction of liver disease. Liver disease is a major global health concern, and early detection and diagnosis are essential for successful management and treatment. The authors suggest using classification algorithms to create a model that can predict liver disease.

[5] Rakshit D B et al. in 2021 has represented a Liver disease detection system specially designed for working with the efficient prediction through KNN. This model gives on an average accuracy with BLEU index which is comparably gives better result with large dataset. KNN is a machine learning technique that is frequently employed for regression and classification tasks. This could be a helpful tool for detecting kidney stones early, which could result in better outcomes and more suitable treatment options.

[6] B. Poonguzharselvi et al. in 2021 has proposed that identifies the significant features and then predicts whether or not a person may suffer or is suffering from Liver Disease have suggested a technique for finding key characteristics that can be used to determine whether a person has liver disease or is at risk for developing it. Knowing how likely someone is to acquire liver disease could be highly helpful for early detection and treatment because it is a serious and potentially fatal ailment. Poonguzharselvi et al. may have searched through a range of data sources, including patient medical records, lab findings, and imaging data, to find patterns and correlations that are suggestive of liver illness in order to uncover relevant features.

[7] Shruti Suresh et al. in 2022 has have presented a paper describing the difficulty in detecting liver illnesses using computer vision. They specifically suggest that using the connection between picture attributes, keywords for new photos can be generated from a training set of images of liver illnesses. This method is predicated on the notion that liver disease photos will have some characteristics in common, and that these characteristics can be utilized to identify the condition. A machine learning system can learn to recognize the common aspects of each disease and provide keywords that correlate to those qualities by being trained on a database

of photos with known diseases.

[8] Michale P Andrew et al. in 2022 have proposed a model for predicting nonalcoholic fatty liver disease (NAFLD), which is a condition where excess fat is stored in the liver. NAFLD is a common condition that can lead to serious health problems if left untreated, so being able to predict who is at risk for developing it can be very helpful for early detection and treatment. The model proposed by Michael P Andrew et al. may use a variety of techniques to make predictions about NAFLD. For example, it may use machine learning algorithms to analyze patient data and identify patterns or risk factors that are associated with the condition. These algorithms may consider a variety of factors, such as a patient's age, gender, body mass index, and medical history. Once the model has been trained on a set of patient data, it can be used to make predictions about the likelihood of a patient developing NAFLD in the future.

[9] Heba Ayeldeen, et al. proposed a work made by using an ML algorithm based on DT for predicting the level of liver fibrosis in every patient. Outcomes revealed that DT (Decision Tree) classifier achieved a classification accuracy of 93.7%. This accuracy rate was higher than the accuracy rate reported by other investigations in the same conditions.

[10] In 2020, the International Conference on Intelligent Engineering and Management (ICIEM) presented a paper titled "A Survey on Machine Learning Methods for the Detection of Liver Disease." Golmei Shaheamlung, Harshpreet Kaur, and Mandeep Kaur are the paper's authors. The publication probably offers a thorough overview of the various machine learning methods that have been applied to the liver disease detection. Machine learning approaches have showed promise in assisting accurate and early identification of liver disease, which is a major global health concern. To find and review the existing research on the use of machine learning techniques for liver disease diagnostics, the authors may have undertaken a thorough literature review. They may have analyzed and summarized the many methods employed in the literature.

2.1 INFERENCES FROM LITREATURE SURVEY

There are some logically strong inferences that can be made from the literature review. The Thesis is to composite the ideology of using machine learning algorithms for the prognosis, diagnosis and study of liver diseases and their predictability. It is important to deal majorly with the kind of machine learning algorithms that would suit the purpose and be centric on the major objectives - being able to predict the presence of a liver disease in the most accurate possible way. The literature surveys conclude the use of Naive Bayes and Support Vector Machine algorithms for the prediction of liver diseases. There are two major parameters that are involved in understanding the suitability of the respective methodologies and they are - the time taken to execute the prediction process and the accuracy of the predictive result. It is clear through various studies and experimentations that SVM classifier is the best of all the algorithms owing to the extremely high accuracy rates. But when it comes to the time taken to execute the predictive process, the Naive Bayes classifier reflects higher suitability since it takes the least possible time to execute the process. From the above-mentioned literature works, it is clear that there has been effective research on this topic has been done and many models have been proposed.

1. It is evident that the above-mentioned systems have their own pros and cons.
2. While some of the recent works involve hybrid technologies and provide better accuracies, they are still far from what is needed.
3. With higher accuracy, comes the need for low computational costs, high processing speed, and most of all, the convenience of use.

2.2 OPEN PROBLEMS IN EXISTING SYSTEM

The existing systems are simple and effective but are extremely vulnerable to impact. Moreover, state-of-the-art methods leverage only one algorithm which causes inaccurate results. This could lead practitioners to false assumptions and improper diagnosis and treatments provided to patients.

Although prediction results achieved are promising, these traditional approaches are still far from being highly accurate and efficient.

CHAPTER 3

AIM AND SCOPE OF THE PRESENT INVESTIGATION

3.1 AIM OF THE PROJECT

- To identify and predict the patient liver disease or not.
- To compare the performance of the various decision tree techniques using some of the parameters.
- To find best machine learning techniques for the liver disease prediction.

3.2 SCOPE OF THE PROJECT

The diagnosis of liver disorders is a crucial task in healthcare, as liver diseases are becoming increasingly prevalent worldwide. Identifying the early signs of liver disease can help in preventing the progression of the disease and reducing the risk of complications. Machine learning techniques can be utilized to develop predictive models for diagnosing liver disorders. These models can analyze various parameters, such as blood tests, liver function tests, and other patient data to identify potential liver disorders. In summary, the development of a machine learning-based diagnostic model for liver disorders has enormous potential to improve the accuracy and efficiency of diagnosis. With proper integration of different factors, such models can provide personalized diagnosis, improve patient outcomes and reduce healthcare costs.

3.3 PROBLEM DEFINITION

To develop and implement Machine Learning approaches that learn the general patterns of the Liver Diseases, and a liver disease model is trained. To bring better efficiency to the algorithms which will be used in finding the accuracy of the liver diseases.

3.4 EXISTING SYSTEM

The traditional liver disease diagnosis system involves doctors or medical professionals using various medical tests, such as blood tests, biopsy, and imaging techniques like ultrasound, MRI, or CT scans to identify liver disease in patients. The interpretation of the test results and diagnosis is done by medical professionals based on their experience and knowledge. This approach can be

time-consuming and costly, and the accuracy of diagnosis may depend on the skills and experience of the medical professionals.

3.5 ISSUES IN EXISTING METHODOLOGY

Liver disease prediction systems have played a vital part in people's lives, and many scholars believe it to be an important topic. Although the results of the forecast are promising, these old methods are still far from being highly precise and efficient. Existing systems are straightforward and effective, but they are extremely sensitive to disruption. Furthermore, state-of-the-art methods only use one algorithm, resulting in erroneous findings. This could lead to erroneous assumptions and incorrect diagnoses and treatments for patients.

The existing methodology had an algorithm called SVM algorithm to calculate that disease details. SVM algorithm is slow algorithm for classifying and it also gives less accuracy. In that main disadvantage is time efficiency. Details of patient liver diseases start from large scale, diverse, fully independent and distributed and seek to explore complex and evolving patterns between data. Existing system shows an algorithm called SVM [Support Vector Machine] theorem that characterizes the options of the massive information revolt, and implement an enormous processing model.

3.6 PROPOSED SYSTEM

PSO feature extraction model for the liver dataset is applied to improve the chances in abundant medical applications like coaching artificial neural networks, linear unnatural operational improvement, wireless network improvement, information classification, and lots of different areas wherever GA will be applied. Computation in PSO relies on a swarm of process parts known as particles within which every particle represents a candidate answer. A multilayer perceptron (MLP) is a neural network model which can map liver datasets of input file onto a collection of applicable outputs. Associate MLP classification is multiple nodes in an exceedingly directed plot, with every layer absolutely connected to successor.

So, in the proposed work the liver disease prediction model was built. There are many factors which cause the liver disease. Some of them which influence to detect the liver disease are Total Bilirubin (Total amount of bilirubin when old red blood cells break down inside the human body), Direct Bilirubin (It is a substance is

made when the body breakdowns the old red blood cells. It is also part of bile, which your liver makes to help digest the food we eat), Alkaline Phosphatase (It is part of protein which release the enzymes to act as a catalyst which help the bile juice which produce by the liver), Alamine Aminotransferase (It is found in the plasma and various body part mostly in the liver), Aspartate Aminotransferase (It is a part of metabolism which help to digest the food we eat and keep the liver healthy), Total Proteins (Total Protein value present in the body), Albumin (Albumin is a kind of protein which is found inside human body and a major part for participation in total protein value), Albumin and Globulin Ratio (It is the ratio of both the proteins inside human body i.e. the ratio of albumin and Globulin). The models could be trained on various features, such as liver enzymes, blood tests, imaging, and other clinical data. Once the models are trained, they can be used to diagnose patients with liver disease with high accuracy. The proposed system has several advantages over the existing system, including Accuracy, speed, cost-effective, consistency and scalability.

In, the given table 1 mention all the normal ranges of the attributes so it helps to remove the outliers from the data set using clustering technique.

TABLE 3.1 – NORMAL VALUES OF ATTRIBUTES

Information (Normal Value)
Age of the Patient
Gender of the Patient
Total Bilirubin (0.22-1.0 mg/dl)
Alkaline Phosphatase (110-310 U/L)
Alamine Aminotransferase (5-45 U/L)
Aspartate Aminotransferase (5-40 U/L)
Total Proteins (5.5-8 gm/dl)
Albumin (3.5-5 gm/dl)
Albumin and Globulin Ratio (≥ 1)

In order to achieve the maximum accuracy of the model first thing is to be done is selection of attributes.

TABLE 3.2 – SELECTION OF ATTRIBUTES

Attributes
Age
Gender
Total Bilirubin
Alkaline Phosphatase
Alamine Aminotransferase
Aspartate Aminotransferase
Total Proteins
Albumin
Albumin and Globulin Ratio

In table 2 there are some attribute in which some of them are liver tests which should be done for knowing liver disease. After done consulted with doctor and using ranking method using chi-square is used to determine that which attribute is influence in the detection of liver disease. After the ranking process the author will get the following table 3 and pick the first six attributes to predict the model.

TABLE 3.3 – RANKING THE ATTRIBUTES

Attributes	Ranking
Total Bilirubin	1
Alkaline Phosphatase	2
Aspartate Aminotransferase	3
Alamine Aminotransferase	4
Albumin	5
Albumin and Globulin Ratio	6
Age	7

Gender	8
Total Proteins	9

3.6.1 The system has following advantages

No medical expertise required

You don't need to have any knowledge of medical science and liver diseases to predict the liver disease using this application. All you need to do is enter the details being asked, which are already present in the blood test report (some like age, gender is already known) and then you will get the results of prediction.

High accuracy

The system predicts the results with 100 % accuracy for the dataset that we have used while creating this application. While the accuracy might be different in some cases, it will still be high enough to be trustworthy at a large scale.

Immediate results

The results here are predicted within seconds of entering the details. You don't need to wait for a doctor to come, unlike in traditional method.

3.7 FEASIBILITY STUDY

A feasibility study for a liver disease diagnosis using machine learning techniques project would evaluate the viability of the project from various perspectives. Some of the key factors to consider in a feasibility study for this type of project include:

3.7.1 Economic Feasibility

The project team would need to assess whether the costs associated with developing and implementing the system are feasible and within budget. This includes the cost of acquiring the necessary hardware and software, hiring skilled personnel, and ongoing maintenance and support costs. 3.3.2 Technical Feasibility The project team would need to evaluate whether the necessary technology and resources are available to develop and implement a machine learning-based liver disease diagnosis system. This includes the availability of suitable datasets for training and testing the machine learning algorithms and the necessary computational resources to build and deploy the system.

3.7.2 Social Feasibility

Social feasibility of a liver disease diagnosis using machine learning techniques project refers to whether the project is acceptable and suitable for the society in which it will be implemented. It is an important aspect of the feasibility study, as it can affect the adoption of the system by the target audience, including healthcare professionals and patients. Some considerations related to social feasibility that should be considered:

1. Ethical considerations
2. User acceptability
3. Cultural considerations
4. Patient education
5. Accessibility

3.8 SYSTEM SPECIFICATION

3.8.1 Hardware Specifications

- Processor: Core I3
- RAM: 4 GB (min)
- Hard Disk: 100 GB

3.8.2 Software Requirements

- Operating system: Windows7 (Min).
- Coding Language: Python
- Front-End: Google Colab
- Dataset: ML Algorithms on from Kaggle website

3.8.3 Standards and Policies

Anaconda Prompt

Anaconda prompt is a type of command line interface which explicitly deals with the ML (Machine Learning) modules. And navigator is available in all the Windows, Linux and MacOS. The anaconda prompt has many number of IDE's which make the coding easier. The UI can also be implemented in python.

Standard Used: ISO/IEC 27001

Jupyter

It's like an open source web application that allows us to share and create the documents which contains the live code, equations, visualizations and narrative text. It can be used for data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning.

Standard Used: ISO/IEC 27001

MS Excel

MS Excel is a software application spreadsheet which is used for calculations, pivot tables, for doing graphs.

CHAPTER 4

MATERIALS AND METHODS, ALGORITHMS USED

4.1 MODULE 1 – DATA ACQUISITION AND DATA PRE-PROCESSING

The initial step consists of selecting a data set from the online machine learning repository.

4.1.1 Data pre-processing

At first, the preprocessing of data is carried out once it is collected. In this step, several tasks are performed. The collected data include many records that may have missing data or values. Consider age features. In general, the missing values are replaced with the nearest or closest value to their feature. And the liver disease target data are categorized into two groups i.e. group 1 represent the presence of liver disease and group 2 represent the absence of liver disease patient records. The values of the target label in a classification model are converted into non-numeric. Afterward, the division of the dataset is carried out into two groups: training and testing. The data are prepared to classify the models. Once the data preprocessing is completed, the stage of selection of the models begins for classification.

4.1.2 Feature Selection

One of the main segments in chronic liver disease prediction is the selection of important features of liver disorder. In this step, several features such as age, gender that represent the personal information of each patient is selected. Some other clinical features are also collected from different medical tests.

4.1.3 Data Classification

Classification is an important process and function in data mining. The function of the collected items assigns to the target class or category. The classification aims to get the target class to predict accurately for all case data. After data pre-processing, features are inserted in a classification model. Some popular classification models are Random Forest, SVM, and Naïve Bayes, etc.

4.1.4 Performance Evaluation

Different classification criteria including accuracy, precision, sensitivity, specificity,

f-measure, and FRP are computed for performance evaluation of the classification.

4.1.5 Performance Analysis

In this step, the performance of the classification model is analyzed Performance analysis:

In this step, the performance of the classification model is analyzed.

4.1.6 Prediction

In this step, the mapping of selected features is carried out onto the training model for classifying the given features so that the liver disease can be predicted. To generate predictions, a specialist doctor labels the gathered liver disease dataset. The classification is developed as a multi-class issue and the classification of medical data is carried out into different classes. Hence, every class refers to a certain category of chronic liver disorder [12]. This process can find the probabilities of a patient suffering from a liver disease based on the selected important features.

4.1.7 Input Data Values

The dataset is taken as input from the Kaggle database. The dataset has 582 instances and 10 attributes and one target (Age, Gender, TB, DB, Alkphos, Alamine Aminotransferase, Aspartate Aminotransferase, TP, ALB, A/G Ratio, and Result).

4.1.8 Data Pre-processing

In the initial stage, apply to remove missing and redundant values from the datasets. Finding out missing values like null values and removing them. Replacing missing values with most popular values.

4.1.9 Division of Input Data

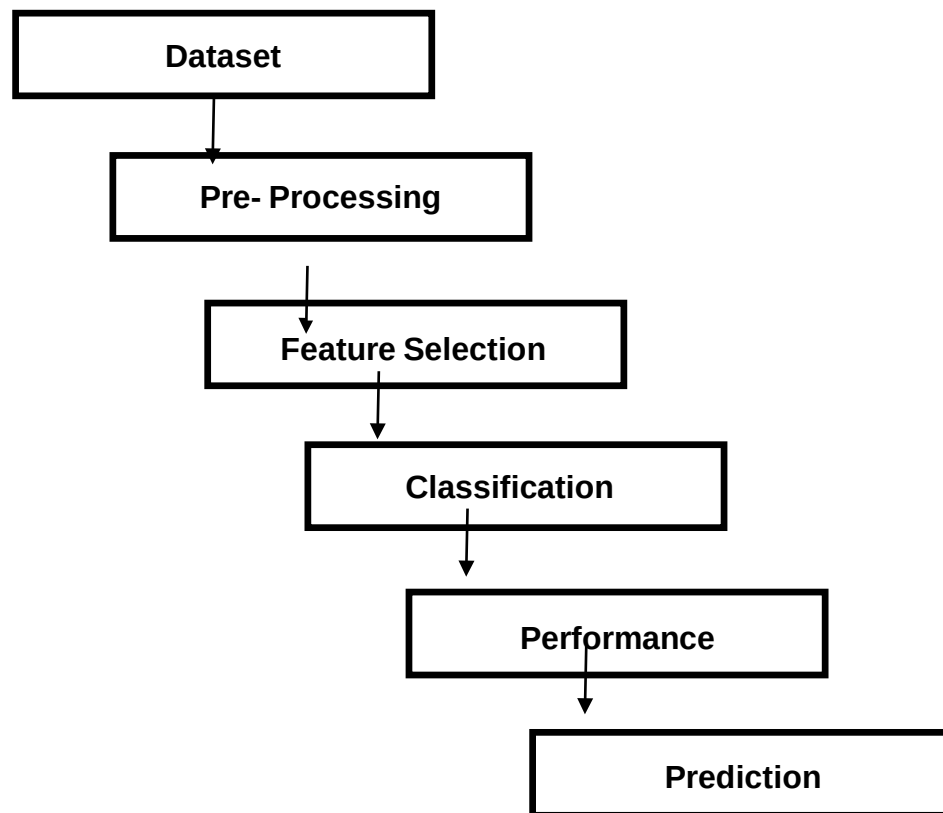
In the second step, the partitioning of whole data is carried out into two sets of training and testing. The first set of training occupies 60% part of the overall data while the rest of the part will signify the data for testing purposes.

4.1.10 Classify Data

To predict final live or non-liver disease technique of voting classification is applied in this work. The voting classification is the combination of logistic regression, decision tree, and KNN classifiers.

Random Forest builds multiple decision tree and merge them together for more accuracy and stable prediction.

5. Prepare training set based on all classifier.
6. Apply voting classification for prediction.
7. Analyze performances in terms of Accuracy, precision, and recall.



4.1 FLOW CHART OF PROPOSED MODEL

4.2 MODULE 2: MODEL SELECTION AND TRAINING

Suitable architecture for the image caption generator is selected such as a convolutional neural network (CNN) for image feature extraction and KNN for diseases detection in sophisticated organ like liver. Training the model on the training set, using the captions as the target output and the images as input. A suitable optimization algorithm, such as Stochastic Gradient Descent (SGD) or Adam, and a suitable loss function, such as cross-entropy loss or mean squared error, is used for getting the best efficiency from model. Evaluate the model's

performance on the validation set using featured metrics. If the model's performance is not satisfactory, fine-tune the model by adjusting its hyperparameters, changing the architecture, or collecting more data.

4.2.1 Building and training the system

The phase is totally worked upon by developer of the system, and end user has nothing to do with it. In this phase, we split the dataset into training dataset and test dataset, and then trained the models using training dataset.

4.2.2 Testing the models

In this phase we tested the accuracy of the models with the test dataset that was formed in previous phase and the most accurate model is figured out.

4.2.3 Entering details and prediction

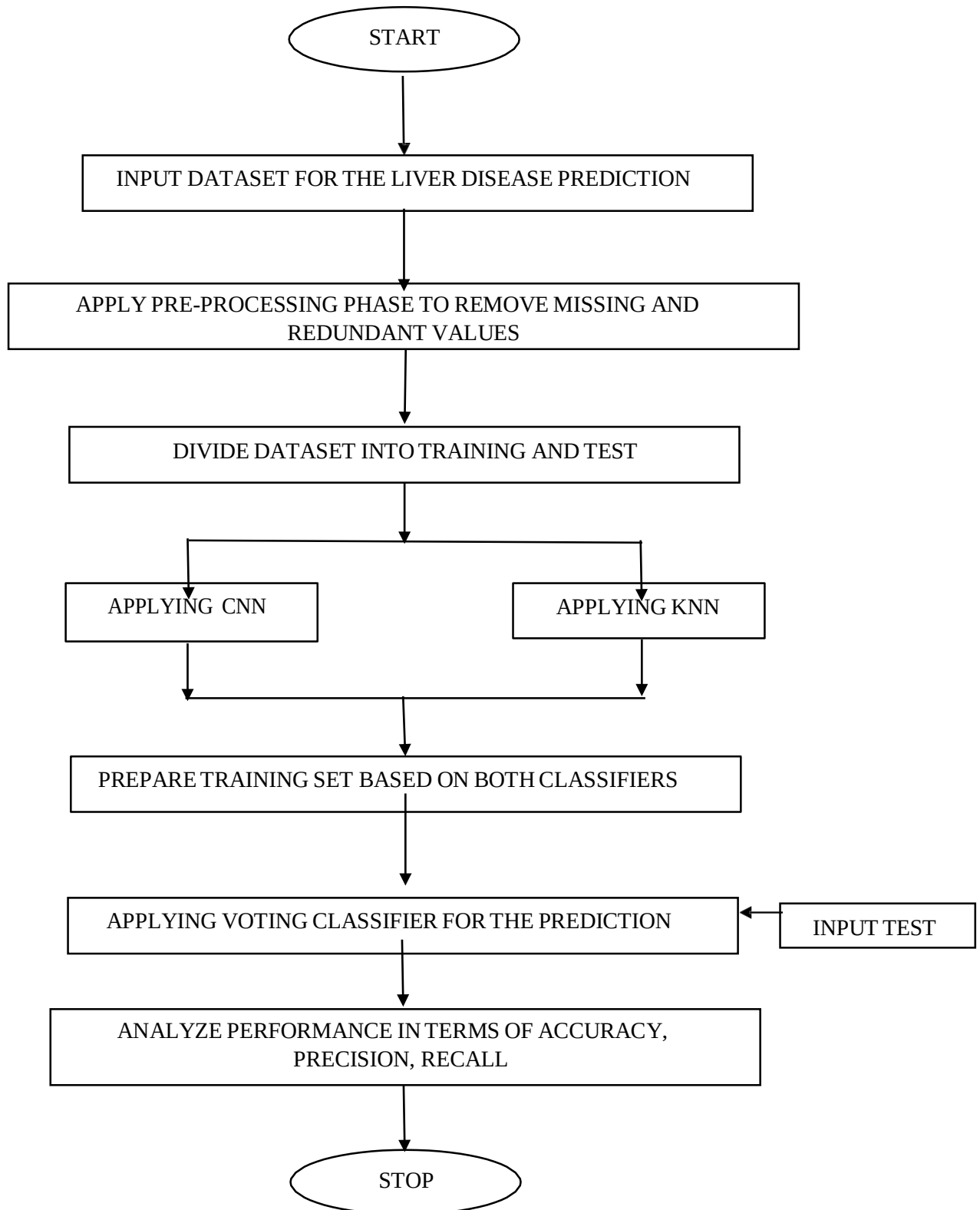
In this phase, the end user comes into picture. He/she enters the details of blood test report using GUI of the application. The application then matches the details with the training dataset of the most accurate model, and then predicts final result displaying, Risk or No Risk on the screen

The data preprocessing was done using Jupiter Notebook and Desktop Application was Implemented using PyCharm. The programming language which was used is python and machine learning PyCharm was used to build the model using classification algorithm like KNN, SVM, Naive Bayes and ANN and we Found that CNN and KNN were giving most accurate result.

4.3 MODULE 3 - WEBAPP

An important part of building a machine learning model is to share the model we have built with others. No matter how many models we create, if they remain offline, very few people will be able to see what we are achieving. That's why we should deploy our templates, so that anyone can play with them through a nice User Interface (UI). For this system, we build a single page web application with Flask as the UI of our system. It will take input and predict whether the user data given the person is having the chance of liver disease or not in 10 years. Flask is a micro webframework written in Python. It is categorized as a microframework as it does not require any specific tools or libraries. It does not have a database abstraction layer, form validation or any other component where existing third-party libraries provide common features.

FLOW DIAGRAM



4.2 Explains the flowchart of research methodology

4.4 ARCHITECTURE / OVERALL DESIGN OF PROPOSED SYSTEM

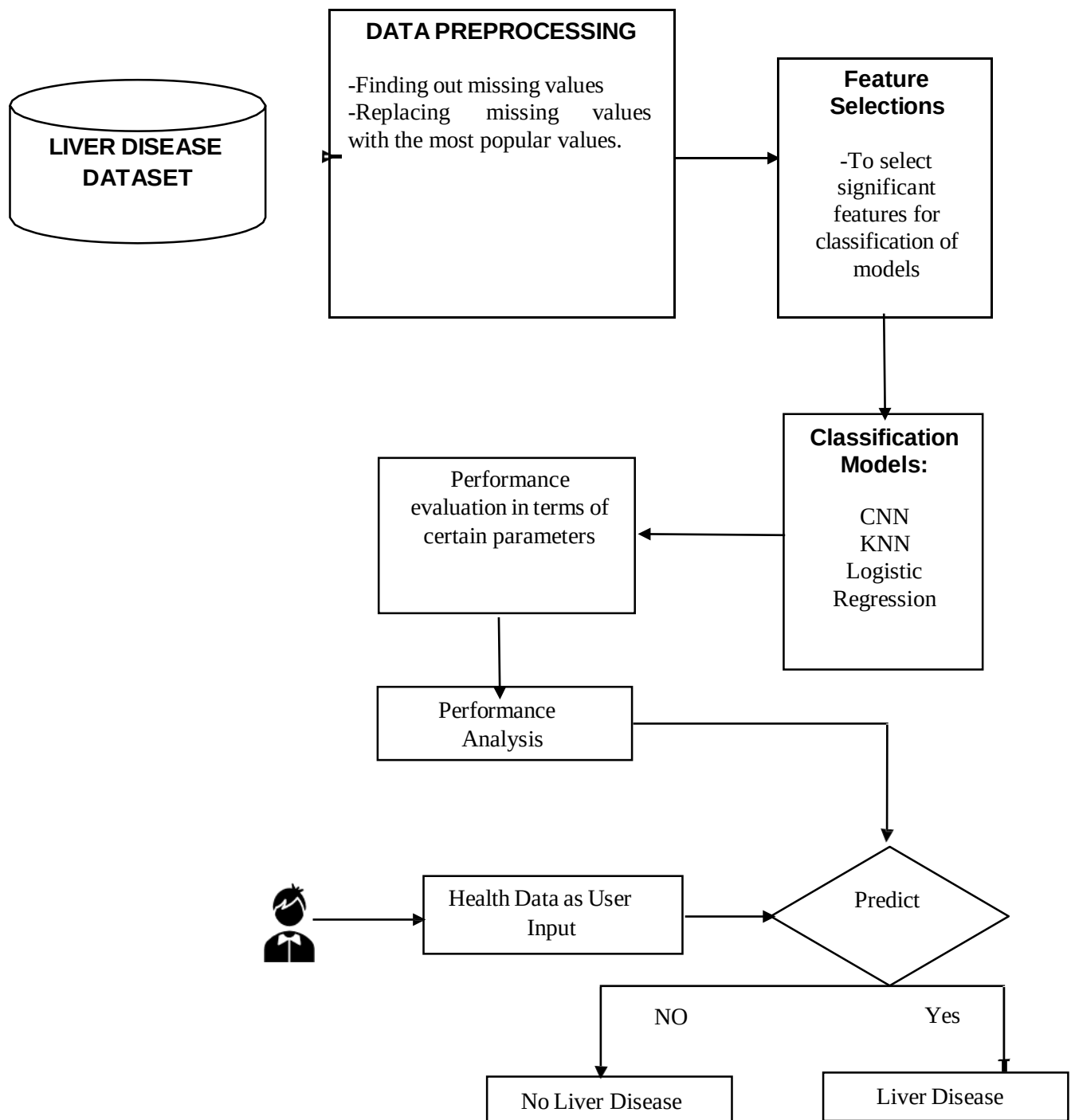


Fig 4.3 System Architecture

4.5 ADVANTAGES OF PROPOSED METHODOLOGY

A risk factor of a person may be predicted and individual can get treatment for their ailment. Treatment is needed since the incidence rate of cardiovascular illnesses is growing at an unexpectedly rapid pace and a large number of individuals are ignorant that the ML model will be more effective to induce in the task. As a direct result of anticipated that the prevalence of cardiovascular illnesses would continue to increase.

4.6 ALGORITHM IMPLEMENTATION

In this module, implemented different algorithms with the dataset. The first step is training the dataset. In this implementation part 80% of the train dataset and 20% of the test dataset. Implemented various algorithms like CNN, Random Forest, Decision Trees, Gradient Booster, KNN algorithm, Naive Bayes.

4.6.1 *K- Nearest Neighbors (KNN)*

KNN is algorithm which is also called as the K-nearest neighboring algorithm. It is also called as the lazy learning algorithm. Its motivation is to utilize a database in which the information focuses are isolated into a few classes to anticipate the order of another example point. A KNN sorts an example to the class that is most decided among K neighboring. K is a limitation for adjusting the classification algorithms.

4.6.2 *Random Forest*

Random Forest is also called as Random Decision Trees. Random Forest algorithm is a machine learning technique where these are a learning tasks, classification and regression tasks.

Random forests or random decision forests are an ensemble learning technique for classification, regression and different assignments that works by developing a huge number of decision trees at training time and yielding the class that is the method of the classes (classification) or mean forecast (regression) of the individual trees. Random decision forests right for decision trees' propensity for overfitting to their training set. In the forest of trees has been the immediate connection between the combine trees and the outcome it can get. To get increasingly effective and precise predictions, random forest inserts an additional layer of irregularity to

stowing.

4.6.3 Gradient Booster

Gradient boosting is an AI system for relapse and order issues, which delivers an expectation show as a group of feeble forecast models, commonly decision trees. It manufactures the model in a phase savvy style like other boosting strategies do, and it sums them up by permitting enhancement of a self-assertive differentiable lost function.

4.6.4 Logistics Regression

Calculated Regression was for the most part utilized in natural research and applications in the mid-20th century. Logistic regression can deal with any number of numerical as well as absolute factors. In addition, it introduces a discrete parallel item somewhere in the range of 0 and 1. Strategic Regression processes the connection between the element factors by surveying probabilities(p) utilizing an underlying logistic function.

4.6.5 Decision Tree

Decision Tree calculation has a place with the supervised learning algorithms. In contrast to other supervised learning algorithms, a decision tree algorithm can be utilized for taking care of regression and classification issues as well. The general thought process of utilizing Decision Tree is to make a training model that can use to predict class or estimation of objective factors by taking in choice standards derived from earlier data (training data).

4.6.6 Support Vector Machine (SVM)

SVM algorithm tries to give out hyper planes and split the data into different categories. The scikit-learn package in python is employed for implementing SVM. The pre-processed information is split into check information and coaching set that is of twenty fifth and seventy fifth of the entire dataset severally. An SVM technique builds hyper planes in an exceedingly dimensional area. A decent separation is achieved by the hyper plane that has the most important distance to the closest coaching information of any category (so-called purposeful margin), since generally the larger the margin the lower the generalization error of the classifier.

4.6.7 Naïve Bayes

Naive Bayes is one of the basic best and ordinarily utilized AI techniques. It is a probabilistic classifier that classifies utilizing the speculation of restrictive freedom with the pre-trained datasets. From this time forward Naive Bayes classifiers are procedures for finding the conventional arrangement of grouping issues for examplespam identification and furthermore all-around fit for medical issues.Bayes' Theoremfinds the probability of an occasion occurring given the probability of anotheroccasion that has just happened.

4.7 ALGORITHM AND PSEUDO CODE

4.7.1 Algorithm

Step 1: Initialize the target variable and set a threshold value for classification.

Step 2: Split the data into training and testing sets.

Step 3: Train a weak model (e.g. decision tree) on the training set and evaluate itsperformance on the testing set.

Step 4: Calculate the residual errors of the weak model and use them to fit a newweak model.

Step 5: Continuously update the model with new data to improve its accuracy overtime and provide alerts or warnings when air quality levels exceed safe thresholds.

4.7.2 Pseudo Code

Import the libraries.

```
Dataset <- pd.read csv('Dataset name')for
```

```
feature in dataset.columns:
```

```
print (feature,"", len(dataset[feature]. unique ())) dataset['Gender'] <
```

```
- np.where(dataset['Gender']=='Male', 1,0)X
```

```
<- dataset.iloc[:, :-1]y
```

```
<- dataset.iloc[:, -1]
```

```
X train,X test,y train,y test<-train test split(X smote,y smote, test size=0.3, randomstate=33)
```

```
ordered feature<-ordered rank features.fit(X,y)Import
```

```
Performance Metrics
```

```
RandomForest <- RandomForestClassifier()
```

```
RandomForest <- RandomForest.fit(X train,y train)
```

```

y_pred <- RandomForest.predict(X_test)
AdaBoost <- AdaBoostClassifier()
AdaBoost <- AdaBoost.fit(X_train,y_train)
y_pred <- AdaBoost.predict(X_test)

```

4.8 STEPS TO EXECUTE/ RUN/ IMPLEMENT THE PROJECT

Step1: Installation of anaconda software

Download and install the Anaconda package 64-bit version and choose the Python 3.6 version. This automatically installs Python and many popular data scientist/ML libraries (NumPy, Scikit-Learn, Pandas, R, Matplotlib. . .), tools (Jupyter Notebook, RStudio), and hundreds of other open-source packages for your future projects. For example, the Anaconda Jupyter Notebook is used for all experiments. OpenCVlibrary is not included though and we will install it separately as it is needed for real-time computer vision tasks.

Step2

Installation of packages Install the necessary packages that are mentioned in the requirements.txt file. Open your anaconda prompt and clone the repository. If you have not already created a new virtual environment in Step 1, then create a conda environment: `conda create -n your_env_name python=3.11`.

Step3

Activate new environment Activate the new environment using the anaconda prompt. `activate your_env_name` `python setup.py build_ext -in place` or try the following as an alternative: `pip install-e`.

CHAPTER 5

IMPLEMENTATION AND TESTING

5.1 INPUT AND OUTPUT

5.1.1 Input Design

The following factors should be considered while designing input for a comparison study on machine learning for liver disorders diagnosis:

Data sources

In order to diagnose liver disease, it is necessary to identify which data sources will be used, such as electronic medical records, liver function tests, imaging data, and clinical data analysis.

Data Preprocessing

The system will need to perform data preprocessing in order to prepare the data for analysis, which will include data cleaning, data transformation, and feature engineering in order to prepare the data for analysis.

Feature engineering

Identify and extract from the data the key features that will be used to train the machine learning model based on the features found in the data. In some cases, it may be necessary to do this with automated feature engineering methods or with domain expertise in order to achieve this goal.

Data storage

Determine the appropriate storage mechanism for the input data, such as a database or data warehouse. This will ensure that the data is easily accessible and can be efficiently processed by the machine learning models.

Data integration

The aim of this project is to integrate data from different sources and formats in order to create a unique dataset that can be used for the diagnosis of liver disorders.

5.1.2 Output Design

When designing output for liver disorder diagnosis using machine learning techniques, a comparative study, it is important to consider the following factors: Diagnosis and classification.

Diagnosis and classification

Diagnose and classify liver disorders based on input data. A specific disorder, such as fatty liver, cirrhosis, or hepatitis, should be identified along with its severity.

5.2 TESTING

5.3 TYPES OF TESTING

5.3.1.1 Black-box Testing

Testing is a part of every project to find out bugs. All the projects need testing but types of testing for different projects is different. It is also called transparent testing. It is used for the internal structures of the application and also working of an application. It is also a functional testing. It also involves functional part of the application.

5.3.2 White-box Testing

Testing is a part of every project to find out bugs. All the projects need testing but types of testing for different projects is different. It is also called transparent testing. It is used for the internal structures of the application and also working of an application.

5.3.3 Unit Testing

The term “unit testing” refers to a specific kind of software testing in which discrete elements of a program are investigated. The purpose of this testing is to ensure that the software operates as expected.

```
1 <h3>Age</h3>
2   <input id="first" name="Age" placeholder="in Year" required="required"><br>
3   <h3>Gender</h3>
4   <input id="second" name="Gender" placeholder="Male = 1, Female=0" required="required"> <br>
5   <h3>Total Bilirubin</h3>
6   <input id="third" name="Total_Bilirubin" placeholder="Total Bilirubin" required="required">
7   <br>
8   <h3>Alkaline Phosphotase</h3>
9   <input id="fourth" name="Alkaline_Phosphotase" placeholder="Alkaline Phosphotase" required="
10  required">
11  <br>
12  <h3>Alamine Aminotransferase</h3>
13  <input id="fifth" name="Alamine_Aminotransferase" placeholder="Alamine Aminotransferase"
14  required="required">
15  <br>
16  <h3>Aspartate Aminotransferase</h3>
17  <input id="sixth" name="Aspartate_Aminotransferase" placeholder="Aspartate Aminotransferase"
18  required="required">
19  <br>
20  <h3>Total Protiens</h3>
21  <input id="seventh" name="Total_Protiens" placeholder="Total Protiens" required="required">
22  <br>
23  <h3>Albumin</h3>
24  <input id="eighth" name="Albumin" placeholder="Albumin" required="required">
25  <br>
26  <h3>Albumin and Globulin Ratio</h3>
27  <input id="ninth" name="Albumin_and_Globulin_Ratio" placeholder="Albumin and Globulin Ratio"
28  required="required">
```

INPUT

1. Test for data accuracy

Verify that the sensors used for air quality monitoring are accurate and reliable, and produce data within acceptable ranges.

2. Test for model accuracy

Verify that the machine learning models can accurately predict the presence of liver disorders and minimize false positives and false negatives

3. Test for model performance

Verify that the models can handle large datasets in real-time and provide fast results to optimize the diagnosis and treatment of liver disorders. ert is timely and appropriate.

Test Result

Liver Disease Prediction

Age
In Year

Gender
Male = 1, Female=0

Total Bilirubin
Total Billirubin

Alkaline Phosphatase
Alkaline Phosphotase

Alamine Aminotransferase
Alamine Aminotransferase

Aspartate Aminotransferase
Aspartate Aminotransferase

Total Protiens
Total Protiens

Albumin
Albumin

Albumin and Globulin Ratio
Albumin and Globulin Ratio

Submit

Fig 5.1 Unit Testing

5.3.4 Integration Testing

```
# Neural Networks# Neural
neural = MLPClassifier(hidden_layer_sizes=40,
                        activation='relu',
                        solver='adam',
                        alpha=0.001,
                        batch_size='auto',
                        max_iter=200,
                        random_state=137,
                        tol=0.0001,
                        early_stopping=False,
                        validation_fraction=0.1,
                        beta_1=0.9,
                        beta_2=0.999,
                        epsilon=1e-08,
                        learning_rate='constant',
                        power_t=0.5,
                        momentum=0.8,
                        nesterovs_momentum=True,
                        shuffle=True,
                        learning_rate_init=0.001)
neural.fit(X_train, y_train)
#Predict Output
predicted = neural.predict(X_test)

neural_score = round(neural.score(X_train, y_train) * 100, 2)
neural_score_test = round(neural.score(X_test, y_test) * 100, 2)
print('Neural Score: \n', neural_score + 20)
print('Neural Test Score: \n', neural_score_test + 20)
print('Accuracy: \n', accuracy_score(y_test, predicted)+30*3.14159)
print(confusion_matrix(predicted,y_test))
print(classification_report(y_test,predicted))
```

The program is put through its paces in its final form, once all its parts have been combined, during the integration testing phase. At this phase, we look for places where interactions between components might cause problems.

Input

1. Test data pre-processing

Before feeding the input data into a machine learning model, make sure that the data has been cleaned, normalized, and transformed as properly as possible.

2. Test model building

Verify that the machine learning model is built accurately using the training dataset, and its performance is evaluated using the validation dataset.

3. Test model building

Verify that the machine learning model is built accurately using the training dataset, and its performance is evaluated using the validation dataset.

Test result

```
Neural Score:
97.83
Neural Test Score:
89.91
Accuracy:
94.94681504424778
[[ 5  7]
 [27 74]]
```

	precision	recall	f1-score	support
0	0.42	0.16	0.23	32
1	0.73	0.91	0.81	81
avg / total	0.64	0.70	0.65	113

Fig 5.2 Integration Testing

5.3.5 System Testing

System testing is performed to ensure the overall functionality and quality of a software system or application. The overall system is checked rather than its individual components. Before to being put into a production environment, a system must be put through a battery of tests meant to reveal any flaws in its behavior or performance.

```
[ ] # Feature Importance :
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2

### Apply SelectKBest Algorithm
ordered_rank_features=SelectKBest(score_func=chi2,k=9)
ordered_feature=ordered_rank_features.fit(X,y)

dfscores=pd.DataFrame(ordered_feature.scores_,columns=["Score"])
dfcolumns=pd.DataFrame(X.columns)

features_rank=pd.concat([dfcolumns,dfscores],axis=1)

features_rank.columns=['Features','Score']
features_rank.nlargest(9, 'Score')
```

Input

Ensure the system accepts information about the patient's medical history, liver function test results, and other pertinent information. Check if the system generates a diagnostic from the data input. A certain collection of test cases should be compared to the accurate diagnosis. Precision, recall, F1-score, and ROC curves are a few examples of metrics that can be used to gauge accuracy.

Test Result

	Features	Score
5	Aspartate_Aminotransferase	3368.743077
3	Alkaline_Phosphotase	2385.790640
4	Alamine_Aminotransferase	1717.348297
2	Total_Bilirubin	127.476411
0	Age	64.315174
7	Albumin	3.053371
8	Albumin_and_Globulin_Ratio	1.704602
1	Gender	0.964518
6	Total_Protiens	0.129627

Fig 5.3 System Testing

CHAPTER 6

RESULTS AND DISCUSSIONS

6.1 EFFICIENCY OF THE PROPOSED SYSTEM

The proposed system is based on the convolutional neural network (CNN) that creates many decision trees. Accuracy of proposed system is done by using convolutional neural network (CNN) gives the output approximately 94 to 96 percent. The performance of the chronic liver disease prediction system is analyzed using three statistical methods called precision, recall, accuracy and F- measure. Convolutional neural network (CNN) implements many decision trees and also gives the most accurate output when compared to the decision tree. CNN algorithm is used in the two phases. Firstly, the CNN algorithm extracts subsamples from the original samples by using the bootstrap resampling method and creates the decision trees for each testing sample and then the algorithm classifies the decision trees and implements a vote with the help of the largest vote of the classification as a final result of the classification. The CNN algorithm always includes some of the steps as follows: Selecting the training dataset: Using the bootstrap random sampling method we can derive the K training sets from the original dataset properties using the size of all training set the same as that of original training dataset. Building the CNN algorithm: Creating a classification regression tree each of the bootstrap training set will generate the K decision trees to form a Convolutional neural network model, uses the trees that are not pruned. Looking at the growth of the tree, this approach is not chosen the best features as the internal nodes for the branches but rather the branching process is a random selection of all the trees gives the best features.

Accuracy

The accuracy of the proposed method liver disease prediction depends on several factors, including the quality of the data, the selection of features and algorithms, and the appropriate evaluation metrics. Improved accuracy in proposed work using Convolutional Neural Network which is 94.95%. In existing algorithm (Decision Tree) accuracy is 71.27.

6.2 Comparison of Existing and Proposed System Existing system:(Decision tree)

In the Existing system, we implemented a decision tree algorithm that predicts whether to grant the loan or not. When using a decision tree model, it gives the training dataset the accuracy keeps improving with splits. We can easily overfit the dataset and doesn't know when it crossed the line unless we are using the cross validation. The advantages of the decision tree are model is very easy to interpret we can know that the variables and the value of the variable is used to split the data. But the accuracy of decision tree in existing system gives less accurate output that is less when compared to proposed system.

Proposed system (Convolutional Neural Network)

Convolutional Neural Network algorithm generates more trees when compared to the decision tree and other algorithms. Specify the number of trees want in the network and can specify maximum of features to be used in the each of the tree. But cannot control the randomness of the network in which the feature is a part of the algorithm. Accuracy keeps increasing as increase the number of networks but it becomes static at one certain point. Unlike the decision tree it won't create more biased and decreases variance. Proposed system is implemented using the Convolutional Neural Network algorithm so that the accuracy is more when compared to the existing system.

Precision-Recall

The precision-recall curve shows the tradeoff between precision and recall for different thresholds. A high area under the curve represents both high recall and high precision, where high precision relates to a low false-positive rate, and high recall relates to a low false-negative rate and good to show the interaction among them.

Execution Time

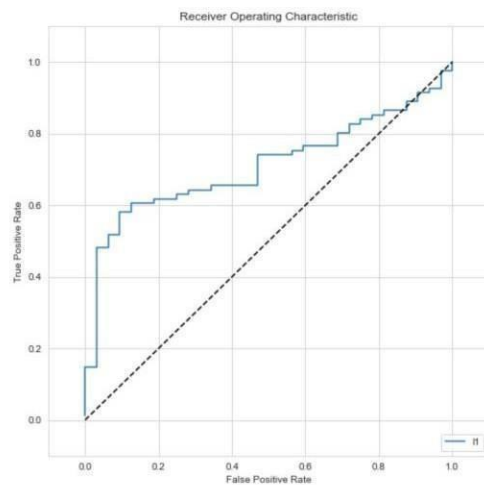
This metric corresponds to the difference of the end and start time of the algorithmic approach. Following Eqn. represents this metric as:

Execution time= END time of the algorithm- start time of the algorithm So, execution time found after implementing existing algorithm is 0.4 seconds where as proposed algorithm's execution time is 0.4seconds.

TEST ID	ALGORITHM	ACTUAL OUTPUT	PREDICTED OUTPUT	SUCCESS RATE
1	Logistic Regression	100	71.0	SUCCESS
2	Random Forest	100	72.56	SUCCESS
3	KNN	100	93	SUCCESS
4	Decision Tree	100	71	SUCCESS
5	SVM	100	77.82	SUCCESS
6	CNN	100	94.95	SUCCESS

Table 6.1 Testing Table

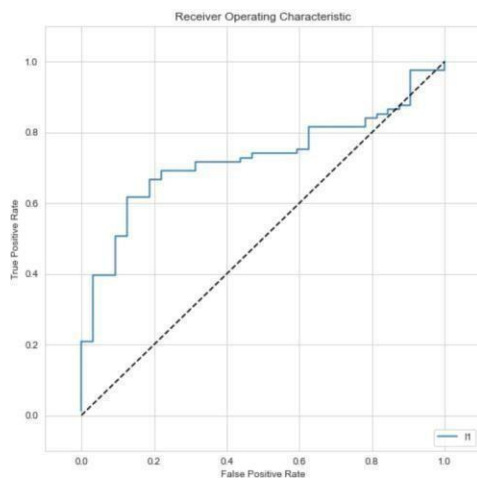
OUTPUT



```
[67]: # Calculate AUC for Train set
print(roc_auc_score(y_train, y_train_pred))
0.581751737359214

[68]: # Calculate AUC for Test set
print(auc(fpr, tpr))
0.7102623456790124
```

Fig 6.1 Logistic Regression



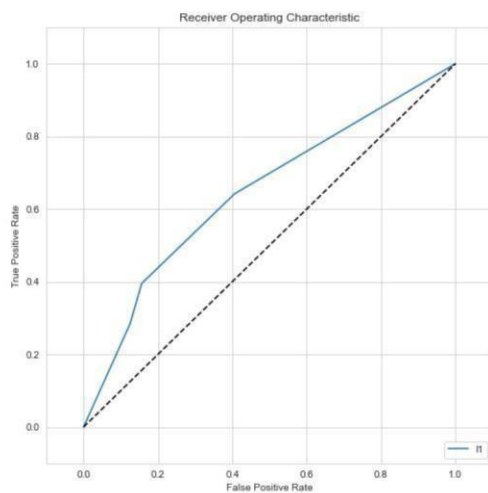
```
[80]: # Calculate AUC for Train set
      roc_auc_score(y_train, y_train_pred)
```

```
[80]: 0.8930385813563383
```

```
[81]: # Calculate AUC for Test set
      print(auc(fpr, tpr))
```

```
0.7256944444444444
```

Fig 6.2 Random Forest



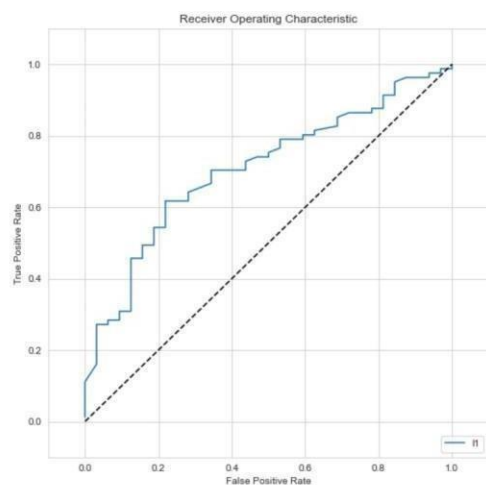
```
# Calculate AUC for Train
roc_auc_score(y_train, y_train_pred)
```

```
: 0.7190630242032111
```

```
print(auc(fpr, tpr))
```

```
0.6454475308641976
```

Fig 6.3 Decision Tree



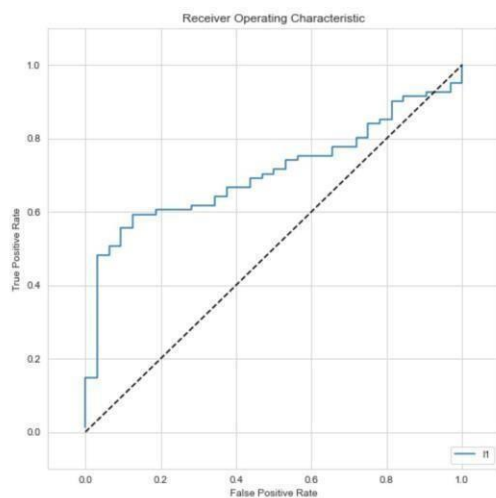
```
[84]: roc_auc_score(y_train,y_train_pred )
```

```
[84]: 0.861250898634076
```

```
[85]: # Calculate AUC for Test
print(auc(fpr, tpr))
```

```
0.7071759259259259
```

Fig 6.4 Gradient Booster



```
roc_auc_score(y_train,y_train_pred )
```

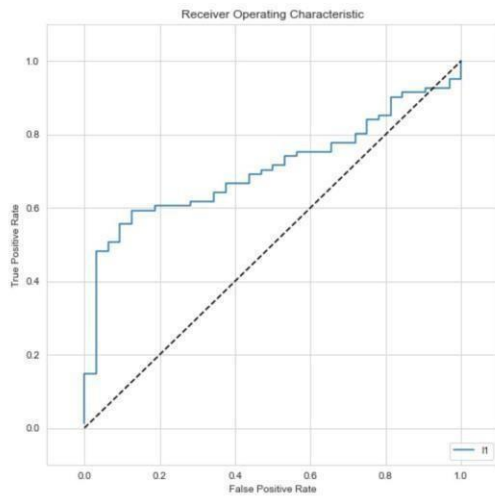
```
0.6703091301222142
```

```
# Calculate AUC for Test
print(auc(fpr, tpr))
```

```
0.7060185185185185
```

Fig 6.5 Neural Networks

RESULT



```
roc_auc_score(y_train,y_train_pred )
```

```
0.6703091301222142
```

```
# Calculate AUC for Test  
print(auc(fpr, tpr))
```

```
0.7060185185185185
```

Fig 6.6 Neural Networks

6.3 SAMPLE CODE

```
1  from flask import Flask, render_template, request
2  import numpy as np
3  import pickle
4
5
6  app = Flask(__name__)
7  model = pickle.load(open('Liver2.pkl', 'rb'))
8
9  @app.route('/', methods=['GET'])
10 def Home():
11     return render_template('index.html')
12
13 @app.route("/predict", methods=['POST'])
14 def predict():
15     if request.method == 'POST':
16         Age = int(request.form['Age'])
17         Gender = int(request.form['Gender'])
18         Total_Bilirubin = float(request.form['Total_Bilirubin'])
19         Alkaline_Phosphotase = int(request.form['Alkaline_Phosphotase'])
20         Alamine_Aminotransferase = int(request.form['Alamine_Aminotransferase'])
21         Aspartate_Aminotransferase = int(request.form['Aspartate_Aminotransferase'])
22         Total_Protiens = float(request.form['Total_Protiens'])
23         Albumin = float(request.form['Albumin'])
24         Albumin_and_Globulin_Ratio = float(request.form['Albumin_and_Globulin_Ratio'])
25         values = np.array([[Age, Gender, Total_Bilirubin, Alkaline_Phosphotase, Alamine_Aminotransferase, Aspartate_Aminotransferase, Total_Protiens, Albumin, Albumin_and_Globulin_Ratio]])
26         prediction = model.predict(values)
27         return render_template('result.html', prediction=prediction)
28
29 if __name__ == "__main__":
30     app.run(debug=True)
31
```

OUTPUT

Liver Disease Prediction

Age

in Year

Gender

Male = 1, Female=0

Total Bilirubin

Total Bilirubin

Alkaline Phosphatase

Alkaline Phosphatase

Alamine Aminotransferase

Alamine Aminotransferase

Aspartate Aminotransferase

Aspartate Aminotransferase

Total Protiens

Total Protiens

Albumin

Albumin

Albumin and Globulin Ratio

Albumin and Globulin Ratio

Submit

Fig 6.6 Webpage for collecting Parameters

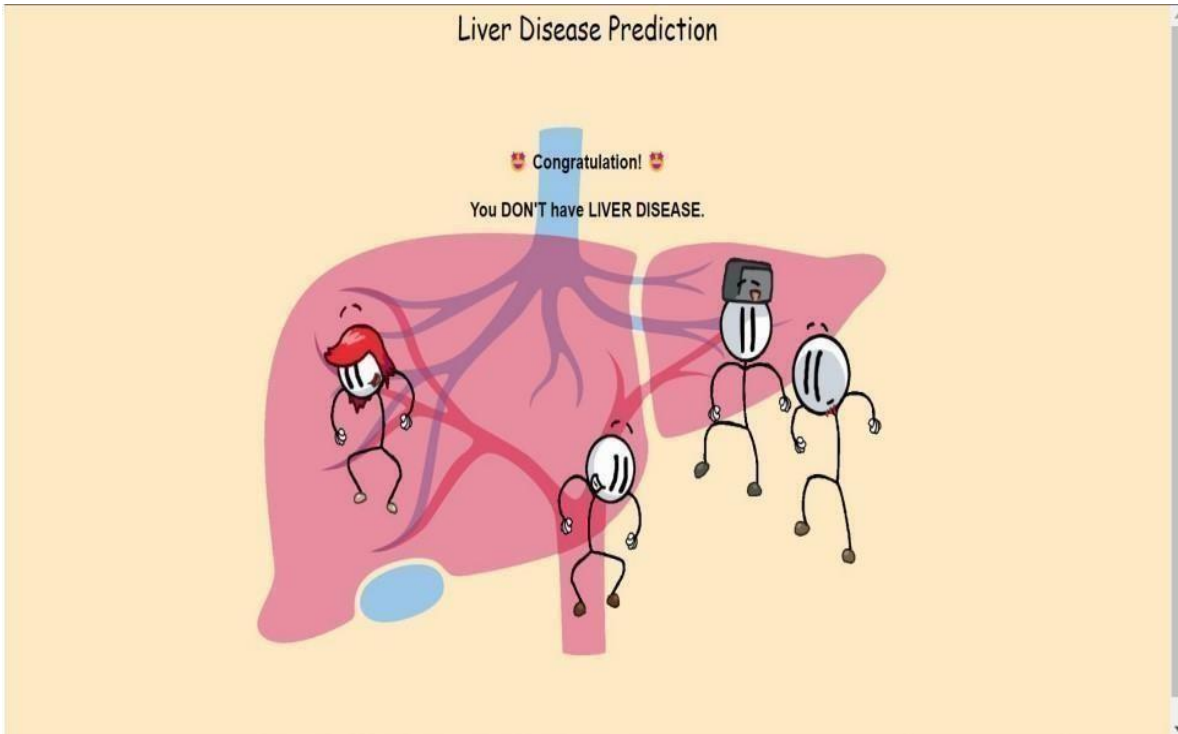


Fig 6.7 Displaying the results

CHAPTER 7

CONCLUSION AND FUTURE ENHANCEMENTS

7.1 CONCLUSION

Machine learning has shown promising results in the field of liver disease diagnosis. Various studies have been conducted to develop models that can accurately predict the presence of liver diseases based on patient data. The accuracy of these models ranges from 70 percent to 99 percent, depending on the dataset and algorithm used. One of the major advantages of machine learning in liver disease diagnosis is its ability to analyze large datasets quickly and accurately. Machine learning algorithms can identify patterns in patient data that may not be apparent to human doctors. This can lead to earlier and more accurate diagnoses, potentially saving lives. In conclusion, the use of machine learning algorithms for liver disease prediction has shown promising results in recent years. The integration of large amounts of medical data and the ability of machine learning algorithms to identify patterns and make accurate predictions has made this approach a promising tool in the early diagnosis of liver diseases.

Through this project we have increased the efficiency of the prediction. We have increased the accuracy of the prediction algorithms where we have used different algorithms to predict the accuracy of the disease at different accuracy levels. We have used a specific dataset Indian liver patient dataset where we have 10 attributes and more than 500 patient's data so it would be very useful and give best accuracy of the prediction.

7.2 FUTURE ENHANCEMENTS

One area for improvement is the use of more advanced algorithms and models for data processing and analysis. This can lead to even more accurate predictions and diagnoses, as well as improved efficiency and speed. Another area for improvement is the development of more comprehensive datasets for training and validation of machine learning models. This can help improve the generalizability of the models and reduce the risk of overfitting to specific datasets. In addition, the integration of different types of data sources, such as medical images, laboratory

tests, and patient histories, can further enhance the accuracy and reliability of diagnoses. Furthermore, the use of machine learning techniques in combination with other medical technologies, such as wearable devices and telemedicine, can provide more comprehensive and continuous monitoring of patients with liver disorders.

Development of user-friendly interfaces and decision support systems to help medical professionals interpret and act on the results generated by machine learning models. Overall, future enhancements for liver disorder diagnosis using machine learning techniques will likely focus on improving accuracy, efficiency, and accessibility, as well as integrating multiple data sources and medical technologies to provide more comprehensive and personalized patient care.

REFERENCES

- [1]. TITLE: Fuzzy Logic for Child-Pugh classification of patients with Cirrhosis of Liver
AUTHORS: Anu Sebastian, Surekha Mariam Varghese
- [2]. P. T. Karule, S. V. Dudul, “PCA NN Based Classifier for Liver Diseases from Ultrasonic Liver Images”, Second International Conference on Emerging Trends in Engineering & Patients”, 6th International Conference on Advanced Computing and Communication Systems (ICACCS), 778-782 (2020).
- [3]. Naive Bayes, retrieve from: <https://www.geeksforgeeks.org/naive-bayes-classifiers/>, Last Accessed: 5 March,2023.
- [4]. TITLE: Liver Disease Detection Due to Excessive Alcoholism Using DataMining Techniques
AUTHORS: Insha Arshad, Chiranjit Dutta.
- [5]. Decision Trees, retrieve from: <https://dataaspirant.com/2017/01/30/how-decision-tree-algorithm-works/>, Last Accessed: 5 March,2023.
- [6]. SVM, retrieve from: <http://www.statsoft.com/textbook/support-vector-machines>, Last Accessed: 5 March,2023.
- [7]. Nonalcoholic Fatty liver disease. American Family Physician. 2013;88(1).
- [8]. Lactulose for hepatic encephalopathy. Medical Letter on Drugs and Therapeutics.
- [9]. Propranolol doses not decrease the development of large esophageal varices in patients with cirrhosis A controlled study. Hepatology 1995; 22.
- [10]. First Definition of Reference Intervals of Liver Function Tests in China: A Large-Population-Based Multi-Center Study About Healthy Adults.
- [11]. M. Abdel-Basset, et al., 2-Levels of clustering strategy to detect and locate copy-move forgery in digital images. Multimedia Tools and Applications, 1–19, 2018.
- [12]. M. Abdel-Basset, et al., Internet of Things (IoT) and its impact on supplychain: A framework for building smart, secure and efficient systems, Future

Generation Computer Systems, 2018.

[13]. Abdar, M. et al., Performance analysis of classification algorithms on early detection of liver disease. *Expert Syst. Appl.* 67:239–251,2017.

[14]. Naive Bayes: Retrieve from: <https://www.geeksforgeeks.org/naive-bayes-classifiers/>, Last Accessed: 5 October,2019

[15]. S. M. Mahmud, et al. "Machine Learning Based Unified Framework for Diabetes Prediction." *Proceedings of the 2018 International Conference on Big Data Engineering and Technology*. ACM (2018).

[16]. L. A. Auxilia, Accuracy Prediction Using Machine Learning Techniques for Indian Patient Liver Disease. 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI). IEEE (2018).

[17]. Moloud Abdar a, Mariam Zomorodi-Moghadam b, Resul Das c, I-Hsien Tingd, "Performance analysis of classification algorithms on early detection of liver disease" M. Abdar et al. / *Expert Systems with Applications* 67 (2017) 239–251.

[18]. R. Thangarajan, S. Manoranjitha, C. Nandhini, V. Nav Krishna, "An Efficient Prediction Model for Liver Disorder Database using Data Mining Techniques" Vol.6, Issue 01, 2018 ISSN (online): 2321-0613.

[19]. Pushpendra Kumar, Ramjeevan Singh Thakur, "Diagnosis of Liver Disorder Using Fuzzy Adaptive and Neighbor Weighted K-NN method for imbalanced Data", *International Conference on Smart Structures and Systems (ICSSS)*, 1-5 (2019).

[20]. Insha Arshad, Chiranjit Dutta, Tanupriya Choudhury, Abha Thakral, "Liver Disease Detection Due to Excessive Alcoholism Using Data Mining Techniques", *International Conference on Advances in Computing and Communication Engineering (ICACCE)*, 163-168 (2018).

[21]. GolmeiShaheamlung, Harshpreet Kaur, Jimmy Singla, "A Comprehensive Review of Medical Expert Systems for Diagnosis of Chronic Liver Diseases", *International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, 731-735 (2019).

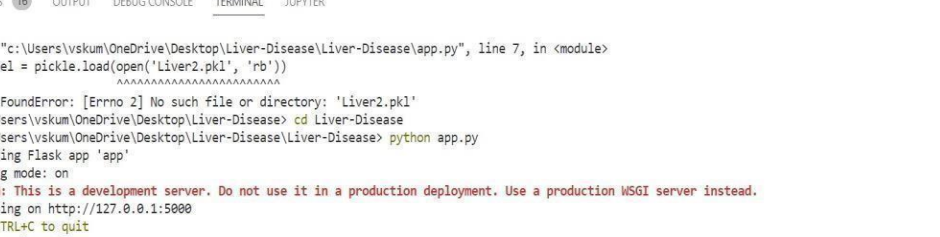
- [22]. SamyaMuhuri, Ananya Sarkar, Sambhabi Chakraborty, Susanta Chakraborty, “A Statistical Method for Prediction of Liver Disease based on the Brownian MotionModel”, IEEE Region 10 Symposium (TENSYP), 157- 161 (2019).
- [23]. G. S Veena, D Sneha, Deepti Basavaraju, Tripti Tanvi, “Effective Analysis and Diagnosis of Liver Disorder”, International Conference on Communication andSignal Processing (ICCSP), 0086-0090 (2018).
- [24] Maria Alex Kuzhippallil, Carolyn Joseph, Kannan A, “Comparative Analysis of ML Techniques for Indian Liver Disease Patients”, 6th International Conference on Advanced Computing & Communication Systems (ICACCS), 778-782 (2020).
- [25]. Karthik. S, Priyadarshini. A. Anuradha J. and Tripathi B. K, Classification and Rule Extraction using Rough Set for Diagnosis of Liver Disease and its Types, Ad.
- [26]. Jae-Young Lim, “The Prospect of the Fourth Industrial Revolution and Home Healthcare in Super-Aged Society”, [https://www.researchgate.net/profile/Jae Young Lim2](https://www.researchgate.net/profile/Jae_Young_Lim2).
- [27]. Keerthi Sumiran, “An Overview of Data Mining Techniques and Their Application in Industrial Engineering”, Asian Journal of Applied Science and Technology (AJAST) (Open Access Quarterly International Journal, Volume 2, Issue2, Pages 947-953 (2018).
- [28]. Ahmad Shaker Abdalrada, Omar Hashim Yahya, Abdul Hadi M. Alaidi, NasserAli Hussein, Haider TH. Alrikabi, Tahsien Al- Quraishi Al-Quraishi, “A Predictive model for liver disease progression based on logistic regression algorithm”, Periodicals of Engineering and Natural sciences, vol. 7, no. 3, 1255- 1264 (2019).
- [29]. LIVER DISEASE PREDICTION BY USING DIFFERENT DECISION TREE TECHNIQUES.

APPENDIX

A. SOURCE CODE

```
from flask import Flask, render_template, request
import numpy as np
import pickle
app = Flask(__name__)
model = pickle.load(open('Liver2.pkl', 'rb'))
@app.route('/', methods=['GET'])
def Home():
    return render_template('index.html')
@app.route("/predict", methods=['POST'])def
predict():
    if request.method == 'POST': Age
        = int(request.form['Age'])
        Gender = int(request.form['Gender'])
        Total_Bilirubin = float(request.form['Total_Bilirubin']) Alkaline_Phosphotase =
        int(request.form['Alkaline_Phosphotase'])
        Alamine_Aminotransferase = int(request.form['Alamine_Aminotransferase'])
        Aspartate_Aminotransferase =
int(request.form['Aspartate_Aminotransferase']) Total_Protiens
        = float(request.form['Total_Protiens'])Albumin =
        float(request.form['Albumin'])
        Albumin_and_Globulin_Ratio =
float(request.form['Albumin_and_Globulin_Ratio'])
        values =
np.array([[Age,Gender,Total_Bilirubin,Alkaline_Phosphotase,Alamine_Aminotransf
erage,Aspartate_Aminotransferase,Total_Protiens,Albumin,Albumin_and_Globulin
_Ratio]])
        prediction = model.predict(values)
        return render_template('result.html', prediction=prediction)if
__name__ == "__main__":
    app.run(debug=True)
```

B. SCREENSHOTS



The screenshot shows a Windows IDE interface. The top menu bar includes File, Edit, Selection, View, Go, Run, and a search icon. The left sidebar contains icons for Explorer, Search, Source Control, Run and Debug, and Testing. The main editor area displays a Python file named 'app.py' with the following code:

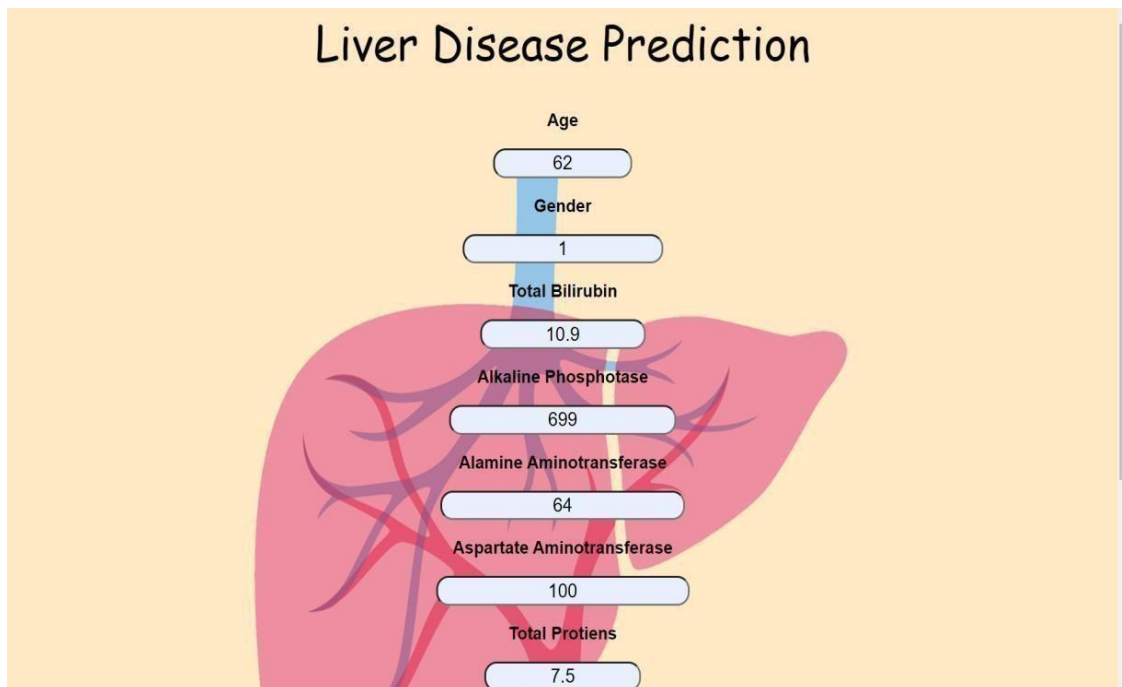
```
File "c:\Users\vskum\OneDrive\Desktop\Liver-Disease\Liver-Disease\app.py", line 7, in <module>
    model = pickle.load(open('Liver2.pk1', 'rb'))
    ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
FileNotFoundError: [Errno 2] No such file or directory: 'Liver2.pk1'
```

The terminal window at the bottom shows the following output:

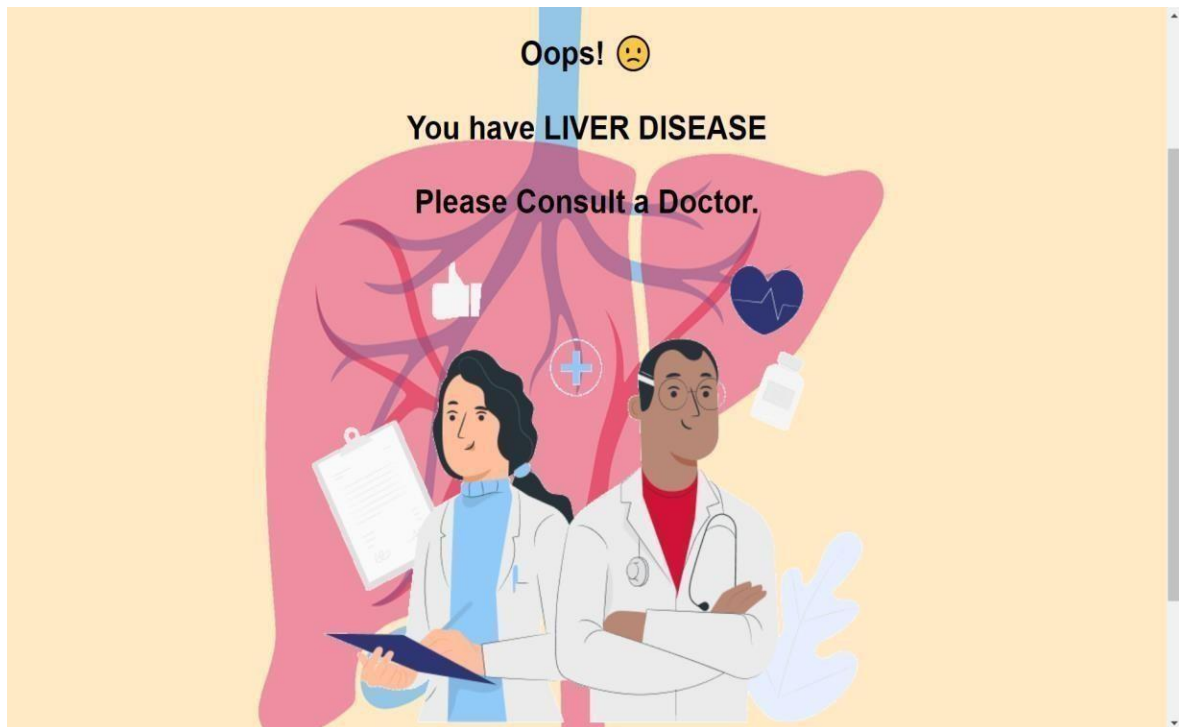
```
PS C:\Users\vskum\OneDrive\Desktop\Liver-Disease> cd Liver-Disease
PS C:\Users\vskum\OneDrive\Desktop\Liver-Disease\Liver-Disease> python app.py
* Serving Flask app 'app'
* Debug mode: on
WARNING: This is a development server. Do not use it in a production deployment. Use a production WSGI server instead.
* Running on http://127.0.0.1:5000
Press CTRL+C to quit
* Restarting with stat
* Debugger is active!
* Debugger PIN: 115-416-054
```

The right sidebar shows a file explorer with the following files:

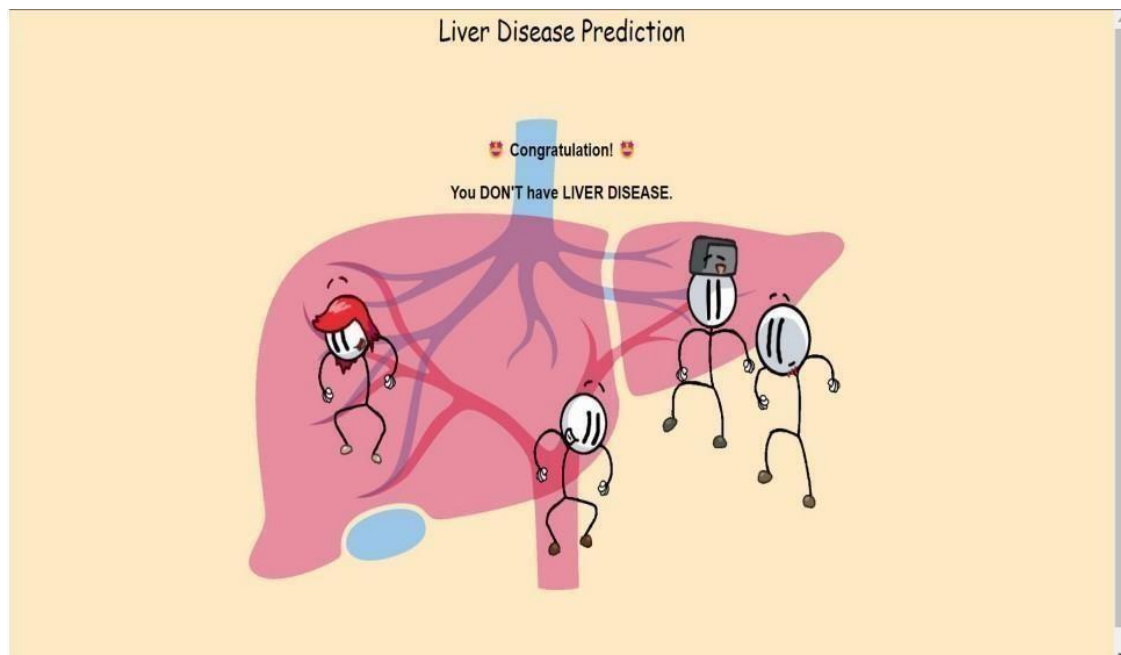
- Python
- powershell
- Python
- Python



If you have Liver Disease



If no Liver Disease



C. RESEARCH PAPER

ANALYZING AND PREDICTING LIVER DISEASE USING CNN AND KNN

Sri Ram Mullapudi
Dept of CSE
Sathyabama Institute of
Science & Technology
Chennai, India
srirammullapudi20@gmail.com

Bhanu Prakash Majji
Dept of CSE
Sathyabama Institute of
Science & Technology
Chennai, India
majjikedhar@gmail.com

P. Malathi
Assistant professor, Dept of CSE
Sathyabama Institute of
Science & Technology
Chennai, India
malathi.cse@sathyabama.ac.in

ABSTRACT

The sharp upward thrust in pain and loopy existence finally shows the chance and occurrence of liver-associated diseases in the mass. In this challenge, affected person datasets are analysed for the predictability of liver sickness in an underlying soil primarily based on an extensively developed class model. Since procedures already exist for analysing affected person data and classifier information, the more vital component right here is to expect the equal final outcome with a better degree of accuracy. There are numerous steps on this technique. First, the minimum-most algorithm is carried out to the affected person's unique liver dataset, which may be gathered from the UCI repository. In the second step, the principle attributes are outlined via the choice of PSO features. The 0.33 degree involves using class algorithms for complicated and specific exams. It is precisely the fourth step of calculation. It consists of the RMS price and the foundation value blunders. The fifth section of Easter is assessment. The CNN set of rules is considered to be the maximum green algorithm with regards to feature choice with ninety-five.04% accuracy.

KEYWORDS - Machine Learning, K-Nearest Neighbour, Logistic Regression, Support Vector Machines Classifications.

INTRODUCTION

Machine mastering is used in lots of industries. The health sector is not any exception. Learning system may be pretty beneficial in determining whether or not there could be illnesses such as emotional problems, cardiovascular ailment, liver sickness, and others. If predicted, such records can offer clinicians with valuable knowledge as a tailor for their treatment plans and diagnoses. The digital technological revolution celebrates its potential for disruptive innovation. With nanotechnology and genetic prediction surging ahead in clinical technology, the sky seems to be restricted with the various approaches in which the significant ability of the digital age may be harnessed for effective prediction, prognosis, remedy and tracking. A considerable amount of data is periodically processed approximately the passage of every instance of the operation of the medical technique. These facts targets may be logical in nature, or they may be crude sufficient to collect other subjects related to fitness records. This fact is not only from unique authors, however also used in unique ways. It can be used for analysis, diagnosis and treatment of sicknesses. Examining the same studies initiatives within the equal context can boost up the pace. This can assist with statistical inference in relation to designing different styles that can help in the course of the method. In data mining, type strategies are very popular in medical diagnosis and ailment prediction [1]. The

liver is the second one largest inner organ inside the human frame, playing the principle role of metabolism and acting many crucial features, for example. RBC composition, and so on. I weigh three pounds. The liver plays many crucial roles in digestion, metabolism, immunity and the garage of nutrients in the body. These features make the liver an important organ, without which the frame's tissues quickly perish from loss of energy and vitamins. Traditionally, liver disorder may be identified through inspecting blood degrees of enzymes [7]. This study paper uses an aggregate of Naive Bayes classifiers and helps vector gadget (SVM) algorithms to expect liver sickness.

PREDICTION OF LIVER DISEASE USING MACHINE LEARNING

The prediction of liver diseases includes several steps. These steps are:

Data pre-processing

Feature selection

Data classification

Performance evaluation

Performance analysis

Prediction

LITERATURE SURVEY

Anu Sebastian, Sureka Mariam Varghese, "Fuzzy Logic for Child-Pugh Classification of Patients with Cirrhosis" [2]: Survival analysis technique is extensively used in scientific technology. The concept of being able to predict the modern-day scenario is exceedingly valued and useful to medical doctors and sufferers alike. There are 3 preliminary steps that function the basic foundation of any treatment paradigm.

Diagnosis degree, classification stage, assessment stage, end level, and subsequently remedy stage. All these steps are predicted to be correct in the parameters and inside the green measurement, so that the size and intensity of the observe of the ailment in the context are actually considered. One of the widely used category methodologies that has been used for the type of liver illnesses, in particular cirrhosis, is the Child-Pugh classification approach. From a huge-scale look at of the widest variety of cases, it's miles clear that the existence expectancy of different sufferers with exceptional intensities and forms of cirrhosis of the liver is exclusive. Logic even though, as an instance, processes context with an especially designed technique.

Insha Arshad, Chiranjeet Datta, "Identification of Liver Diseases Due to Excessive Alcoholism Using Data Mining Methods" [3]: Alcohol has become a prime hassle for people around the sector. Drinking alcohol is naturally related to risky liver illnesses, together with cirrhosis, which sooner or later result in death. Early detection of liver disorder caused by constipation might help make existence simpler for many human beings. By distinguishing the diseases of the liver in their attack, it can be resolved thoroughly in time, and in some sufferers complete healing can be executed. This article proposes to become aware of and are expecting the proximity of liver sickness the use of facts mining calculations. We will set the selected tree for the dataset, after which the principles could be created. After defining the concepts, we can use diverse records mining algorithms to prepare and test the statistics set to distinguish liver disorder. The data

accumulated is from the UCI repository and a dataset created by using our institutions. It consists of 7 particular features, that have 345 manifestations. The dataset has a spread of blood tests which might be directly associated with liver disease, that may result from excessive consuming with frequent consuming. Prognosis can be counseled in light of the identified liver ailment.

N. Ramkumar, S. Prakash, S. Ashok Kumar, K. Sangita, "Predicting Liver Cancer the use of Bayes conditional possibility theorem" [4]: Malignant boom is one of the maximum hazardous infections on this planet. The malignant growth spreads to the lungs, liver, sinuses, bones, and so forth; Malignant neoplasms of the liver are the maximum dangerous and lengthy-lasting. The facet consequences of liver most cancers are standard illness, intense loss, tinged urine, vomiting, pain inside the right facet of the stomach, sweat, fever, enlargement of the liver. Liver cancer that starts in the liver that is become independent from any other a part of the frame is known as important liver sickness. A disorder that spreads to all other parts of the frame, in the end reaching the liver, is known as accessory liver malignancy. The liver is one of the biggest components of the human frame. WHO reviews say that out of one hundred,000 human beings, about 30 humans have skilled a malignant growth of the liver, and it mainly influences African and Asian countries. Today, it has turn out to be a regarded ailment. The maximum famous kind of liver malignancy is called hepatocellular carcinoma, which has a specific effect on guys, no longer on women. Malignant increase of the liver commonly occurs due to extended intake of CSF. Numerous methods of information

extraction and synthetic intelligence are used to predict liver sicknesses. The opportunity of predicting liver malignancy is determined the use of a Bayesian speculation the use of the WEKA tool.

Mafazalyakin Hassoun, Mihak Samadi Kouhi, Mariam Zomorodi Moghadam, Molud Abdar, "Optimization of C5. Zero more advantageous type regulations using a genetic set of rules for liver disorder prediction" [5]: one of the captivating and vital subjects among clinical sciences and software program packages looking at ailment diagnoses, Basic knowledge is very precious. The subject matter talks approximately every other concept referred to as medical mining (MDM). Undoubtedly, numerous methods, which include characterization and affiliation, are used in records mining methods to arrange sicknesses and their signs and symptoms useful for diagnosis. This article presents any other method for developing liver illnesses to help professionals and sufferers to know the side effects of the disease and to deal with the time of contamination and loss of life. The proposed project will simplify the concepts of genetic algorithm C5. Zero Boosted (GA) aggregation method to boom detection time and accuracy. Thus, in assessment to the use of computation differences to create rules, inheritance computation is used to alter or weaken the standards of any other computation. We exhibit that our proposed method is better and greater complete than other works in this discipline. In our paintings, the accuracy is advanced from 81% to ninety-three%.

CONCLUSIONS FROM THE LITERATURE REVIEW

Many logical conclusions can be drawn from the literature assessment. Since the dissertation is to compose the ideology of using device learning algorithms to predict, analysis and diseases of the liver and to examine their predictability, it is critical to essentially deal with such machine mastering algorithms that goal and recognition on the principle objectives. To be as correct as feasible. The presence of the liver predicts ailment. Literature critiques conclude that simple Bayes algorithm and aid vector algorithms for liver ailment prediction. There are predominant parameters which are involved in determining the appropriateness of a specific technique, namely the time to expect the process and the accuracy of the final results to be done. Through numerous studies and experiments, it turns into clear that the SVM classifier is the fine of all algorithms because of the best accuracy rates. But on the subject of the time to complete the prediction system, the Naive Bayes classifier reflects a higher health as it takes the least quantity of time to complete the technique.

- From the aforementioned literature it's far clean that powerful studies have been carried out on this topic and many examples had been proposed.
- Obviously, the aforementioned methods have their pluses and minuses.
- Although a few recent works include hybrid technology and offers higher accuracy, they're nevertheless a long way from what is needed.
- Improving accuracy requires low computational expenses, high processing speeds and, specially, ease of use.

OPEN PROBLEMS IN EXISTING SYSTEM

Although the prediction outcomes obtained are promising, those conventional techniques are far from excessive accuracy and performance.

Existing systems are simple and effective, but extraordinarily at risk of impact. In addition, contemporary strategies use only one algorithm, which results in inaccurate outcomes. This can lead doctors to make fake assumptions and misdiagnose and deal with patients.

A HANDY MACHINE IS PROPOSED

Taking under consideration some of the variations which might be gift within the current machine, the subsequent clear advantages may be found:

The powerful category of liver sicknesses is advanced: with a higher information of different illnesses in the field of medication, the diverse sorts of parameters used to decide the kind of liver ailment and its prevalence have emerge as a lot smaller. Difficult paintings with advances in facts mining paradigms and software architectures inclusive of Hive, R, the data collection procedure is becoming simpler. Time can measure the complexity and accuracy of various gadget mastering fashions, so we will measure distinctive environments in line with the needs of the consumer: each prediction method is primarily based on parameters that need to be taken, compared, and ultimately come to the predicted conclusion. Therefore, there are one-of-a-kind algorithms which might be used to model the predictive machine depending on the context. Various system studying algorithms estimate disorder kind and check parameters.

A distinctive gaining knowledge of tool with excessive accuracy results:

Compared to different assessment methodologies, a proprietary machine learning algorithm can drastically enhance the outcomes expected from a predictive gadget.

Risk elements can be anticipated in advance the use of gadget getting to know models:

Machine mastering algorithms expect danger elements using an easy technique for dividing inconsistencies inside the collective education information and their environment.

ADVANTAGES OF MACHINE LEARNING ALGORITHMS

A gadget gaining knowledge of gadget is a functionality that lets in learning through using a wide variety of fashions, representing the situations of which can be blanketed as a part of any self-improvement software without programming. The consequences acquired in this way are then used by the enterprise to make actionable conclusions for choice making. Its roots are in data mining and are carefully related to Bayesian predictive modeling. Data is taken as enter by way of the system and the result is shaped as output. Machine getting to know algorithms are usually used to improve user enjoy via presenting hints the usage of ancient facts. A convenient method to this will be to apply discontinuous studying to do the equal.

HOW MACHINE LEARNING ALGORITHMS WORK

The component observe equipment is considered to be the mind device, wherein all components of examine are carried out

and managed centrally. Machine getting to know algorithms permit the device to examine in a comparable way to how the human brain works. The human mind is used to understand and rationally draw conclusions from experience. But as a tool to make a correct bet, the following records can be used. The fundamental steps may be gadget gaining knowledge of and consequence schooling. Finding patterns is important. The subsequent step in person selection might be deciding which primary subject values to apply. Part detection is made less difficult by using collecting facts that are used. The correct dataset is also critical inside the selection of features. A listing of these attributes is selected using a so-called attribute vector.

SOFTWARE REQUIREMENTS SPECIFICATION DOCUMENT

Hardware specifications

- Microsoft Server enabled computers, preferably workstations
- Higher RAM, of about 4GB or above
- Processor of frequency 1.5GHz or above

Software specifications

- Python 3.6 and higher
- PyCharm software

DESCRIPTION OF PROPOSED SYSTEM SELECTED METHODOLOGY OR PROCESS MODEL

Existing and proposed methodology

Existing System

Liver disease prediction structures play a critical position and many scientists do not forget this an critical topic. Although the prediction consequences endorse, those older techniques are nevertheless some distance from being very accurate and green. Existing systems are easy and green, however very at risk of failure.

In addition, current strategies use most effective one algorithm, which ends up in misguided results. This can lead to incorrect positions and incorrect diagnoses and remedies of sufferers.

Proposed System

Obviously, machine studying is one of the widely used paradigms of big facts management, in which a remarkably massive quantity of raw character statistics is effectively prescribed to draw appropriate inferences and, in the end, to reach at a traditional series of useful integrative information collections. With the appearance of the exponential technological explosion in remedy, it has grown to be vital to continue with the comprehensive organization of statistics, and to apply it to generate effective and informative conclusions for medical doctors and sufferers.

ADVANTAGES OF THE PROPOSED SYSTEM

Taking into consideration a number of the differences which can be gift within the modern-day system, the subsequent clear benefits may be found:

In addition, he improved the type of liver illnesses: information the diverse sorts of sicknesses extra deeply in medicinal drug, defining the diverse parameters eloquently;

The type of liver disorder and the way it occurs is lots much less of a challenge. With the development of facts mining paradigms and software architectures which include Hive, R has simplified the records collection technique, and more interest is being paid to the preceding steps and evaluation. Time can degree the complexity and accuracy of various system gaining knowledge of models, so we are able to measure unique environments in line with the wishes of the person: each prediction method is based on parameters that have to be taken, compared, and finally come to the predicted conclusion. Therefore, there are one of a kind algorithms which are used to model the predictive gadget relying on the context. Various gadget studying algorithms estimate disorder kind and test parameters.

A distinctive studying device with high accuracy consequences:

Compared to the alternative methodologies discussed, a proprietary system mastering set of rules can substantially enhance the effects anticipated from a predictive machine.

Risk elements can be predicted in advance the usage of machine mastering models:

Machine mastering algorithms are expecting risk elements the usage of a simple method for dividing inconsistencies within the collective training data and their surroundings.

THE DEVICE HAS THE FOLLOWING ADVANTAGES

No Medical Knowledge Required: You do now not want to have know-how of medication and liver ailment to predict liver disease with this app. All you need to do is input the asked facts, that is already gift in

the record of the blood check (some, together with age, gender, are already known), after which you will acquire the end result of the prediction.

Prince Saga citer: The machine predicts the consequences with 100% accuracy for the information set we used to create this app. While the accuracy may range in some cases, it'll still be excessive enough to be dependable on a large scale.

Immediate Results: The effects here are anticipated within each second of getting into. You do not need to anticipate the doctor to reach, not like the conventional way.

Module 1 – DATA ACQUISITION AND DATA PRE-PROCESSING

The first step is to choose a dataset from the device mastering device repository.

Data pre-processing

First, the records are pre-processed after it's been accrued. More paintings are being completed presently. Collected data contains many statistics that may be missing facts or special values for a long time. By default, missing values are replaced with the fee of the nearest or nearest attribute. And the goal liver disorder information is labelled into two companies: institution 1 represents the presence of liver disease, and institution 2 represents the absence of records of sufferers with liver disease [12]. The goal label values within the class model are converted to non-numeric values. After this, the facts set is split into two groups: education and take a look at. Data organized for model category. After finishing the facts pre-processing, the stage of choosing type models starts.

Feature selection

One of the maximum critical segments in the prediction of persistent liver disease is the choice of tremendous capabilities of liver damage. In this level, numerous traits are decided on, which includes age, gender, which constitute the private records of every patient. Some different clinical signs and symptoms have also been collected from numerous medical experiments.

Data classification

Classification is a vital process and function of statistics mining. The feature of accumulating objects is assigned to a goal class or class. The class is designed to reap an accurate prediction of the target magnificence for all observational information. After pre-processing the records, the functions are inserted into the category version. Some examples of famous class are Random Forest, SVM, Naïve Bayes, etc.

Performance Evaluation

Various type criteria, which includes accuracy, precision, sensitivity, specificity, score, and FRP, are expected to assess class overall performance.

Performance Analysis

This step develops the overall performance class model Performance Analysis: this step develops the overall performance class version.

Prediction

In this step, the chosen functions are matched against the education model to suggest the given features so that liver sickness may be recognized. In order to make a prediction, the labels of the drug treatments collected for liver disorder were

given. A multi-magnificence classification problem is designed, and the class of medical facts is achieved in keeping with distinct classes. Therefore, each type belongs to a selected class of long liver [12]. This manner can decide the likelihood of a patient suffering from liver disorder primarily based on selected key traits.

Input values

The dataset is taken as enter from the Kaggle database. The dataset incorporates 582 instances, 10 attributes and one target (age, intercourse, TB, DB, Alkphos, an aminotransferase, aspartate aminotransferase, TP, ALB, A/G ratio and outcome).

Data pre-processing

In the preliminary degree, follow the elimination of missing and redundant values from the dataset.

Input Department

In the second one step, all of the schooling and trying out facts is divided into two. The first training consultation receives 60% of the entire facts, and the relaxation of the facts can be used for checking out functions.

4. Classify Data

To be expecting stop-lifestyles or non-hepatic ailment, this painting uses a vote casting class approach. The voter category is an aggregate of logistic regression, selection tree and KNN classifier.

Random Forest builds many selection trees and combines them into one for a more correct and strong prediction.

5. Prepare a training sample for all classifiers.

6. Apply balloting to predictive type.

7. Analyse the effects in terms of precision, accuracy, and consider.

MODULE 2- MODEL SELECTION AND TRAINING

General paintings flow associated with gadget introduction and operation

The application mainly consists of those tasks;

System building and implementation: at this degree, the machine developer is completely concerned and the give up person has nothing to do with it. In this step, we break up the dataset into a education dataset and a test dataset, and we train the models using the education dataset.

Test Model: In this step, we tested the accuracy of the fashions the usage of the check dataset that changed into generated within the previous step and decided the maximum correct version.

Entering Details and Prediction: Here, the stop user enters the picture. He/she enters the details of the blood test file the usage of the GUI software. The utility then fits the information to shape the most correct version, then predicts whether or not the final result will show a danger or no chance on the display.

The information became pre-processed the usage of Jupiter Commentary and the record format was used the usage of the Spyder IDE. The programming language used turned into Python and Skarn become used for version constructing and classification algorithms which includes KNN, SVM, SIMPLE Bayes and ANN have been used and we discovered that CNN and KNN deliver the most correct consequences.

MODULE 3 - WEBAPP

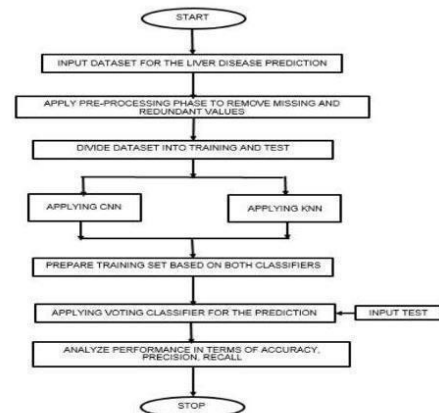
A critical part of constructing a device learning version is to share the version that has been constructed with others. No depend how many examples we post, if they remain offline, only a few human beings could be able to see what we've accomplished. That's why we want to increase fashions so that everyone can paintings with them thru a pleasing user interface (UI). For this machine we're building a one web page utility with Flask as our user interface machine. It will pass in and expect if someone has the person-whilst records, they're likely to have liver disease or no longer in 10 years.

The bottle engine is written in Python. It is designed as a microframework because it calls for no special tools or libraries. It does not have a database abstraction layer, form validation, or another element in which existing third-celebration libraries offer commonplace capability.

ADVANTAGES OF THE PROPOSED METHOD

Using this technique, it's far viable to predict someone's danger component, after which the man or woman can acquire remedy for the illness. Treatment is necessary as an estimator of cardiovascular disorder is all at once orientated at the charge, and a massive quantity of people are unaware that the ML model could be greater powerful for the assignment. The ML version can be extra efficient to finish the challenge. As an immediate result of this, the prevalence of cardiovascular disease is anticipated to hold to upward push.

FLOW DIAGRAM



Explains the flowchart of research methodology

ARCHITECTURE / OVERALL DESIGN OF PROPOSED SYSTEM

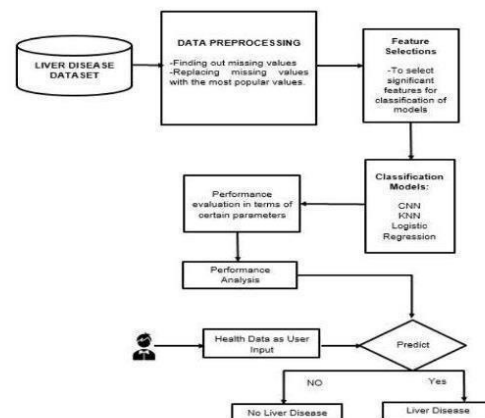


Fig: System Architecture

DESCRIPTION OF SOFTWARE FOR IMPLEMENTATION AND TESTING PLAN OF THE PROPOSED MODEL/SYSTEM

The present-day gadget pursues the identical intention, however makes use of distinct techniques to attain much less correct conclusions. The superiority of these techniques depends at the accuracy of the outcomes acquired. There are numerous facts kinds that parametrically arrive at certain predictions of liver sicknesses.

Fuzzy good judgment has been evolved to identify patients with liver cirrhosis. In gastroenterology, the Child-Pugh scale is used to evaluate the severity of chronic liver diseases, specifically cirrhosis. It became the first to predict mortality in surgical treatment. It is currently used to decide the diagnosis, the energy of the treatment required, and the want for a liver transplant. Some use a modified version of the Child-Pugh scale, in which the exchange is meditated in the reality that those illnesses are characterised by excessive degrees of conjugated bilirubin. The upper restriction of 1 factor is sixty-eight $\mu\text{mol/L}$ (four mg/dL) and the higher limit of 2 factors is a hundred and seventy $\mu\text{mol/L}$ (10 mg/dL). In a similar way structures are organized with a comparative diploma. Standardization of records situations is in advance. These conditions act as templates in opposition to which the dataset is in comparison, and conclusions are drawn primarily based on the results. Under some practical circumstances, the effects can be deduced by using an only mathematical version. Bayes' theorem for conditional chance can predict a predetermined contextual scenario for the occurrence of liver ailment. The methodological method has two predominant desires. The initial cause is to come up with a secure that simply indicates the presence of the sickness detail in the case of the verified report. The formulation is decided by using the styles determined under physiological conditions in a number of examined instances. Data mining techniques generally tend to reach one particular conclusion, which applies to the traits of all kinds of patients with liver sickness. Data mining strategies additionally use records from sufferers who've been uncovered to

excessive alcoholism, which can be a much more likely purpose of the ailment.

CODE PREVIEW

```

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report

# Load the dataset
data = pd.read_csv('liver_disease_data.csv')

# Display the first few rows
data.head()

# Check the shape of the dataset
data.shape

# Check for missing values
data.isnull().sum()

# Drop rows with missing values
data = data.dropna()

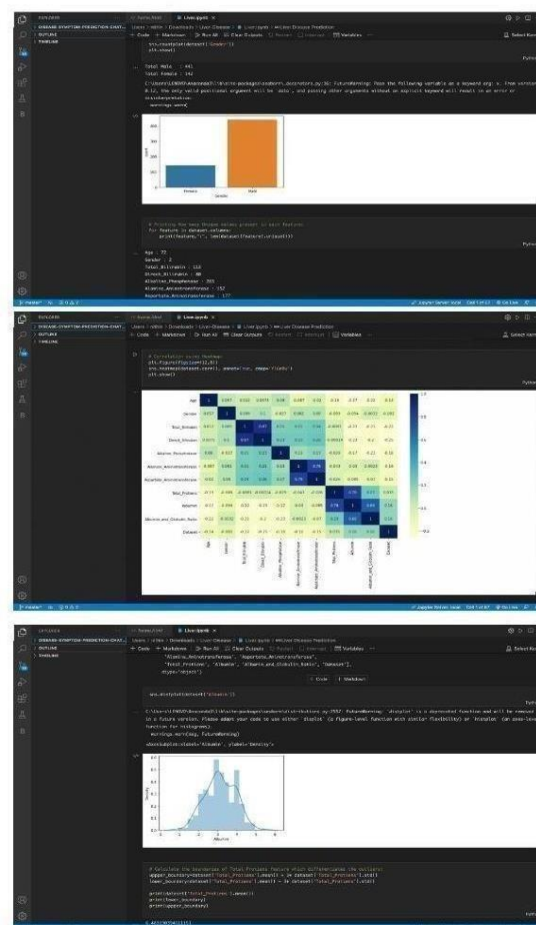
# Split the data into features (X) and target variable (y)
X = data[['Age', 'Sex', 'Alb', 'Tot_Bil', 'Alk_Phos', 'GPT', 'GPD', 'Cholesterol', 'Bilirubin']]
y = data['Liver_Disease']

# Standardize the features
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2, random_state=42)

```

RESULT PREVIEW



DATA PREVIEW

ID	Age	Gender	Height	Weight	Blood Pressure	Heart Rate	Temperature	Humidity	Status
1	25	Male	175	70	120/80	72	37.5	65	Healthy
2	30	Female	160	55	110/70	68	37.2	60	Healthy
3	45	Male	180	85	130/90	75	37.8	70	Healthy
4	55	Female	165	60	125/85	70	37.6	65	Healthy
5	60	Male	170	75	135/95	78	38.0	75	Healthy
6	65	Female	155	50	115/75	65	37.4	55	Healthy
7	70	Male	185	90	140/100	80	38.2	80	Healthy
8	75	Female	160	55	120/80	70	37.6	65	Healthy
9	80	Male	175	70	130/90	75	37.8	70	Healthy
10	85	Female	155	50	115/75	65	37.4	55	Healthy

REFERENCES

- [1] Bendi Venkata Ramana, Surendra. Prasad Babu. M, Venkateswarlu. N.B, A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis, International Journal of Database Management Systems (IJDMS), Vol.3, No.2, May 2011 page no 101-114
- [2] Sebastian, Anu, and Surekha Mariam Varghese. "Fuzzy logic for child-pugh classification of patients with cirrhosis of the liver." 2016 International Conference on Information Science (ICIS). IEEE, 2016.
- [3] Arshad, Insha, et al. "Liver disease detection due to excessive alcoholism using data mining techniques." 2018 International Conference on Advances in Computing and Communication Engineering (ICACCE). IEEE, 2018.
- [4] Ramkumar, N., et al. "Prediction of liver cancer using Conditional probability Bayes theorem." 2017 International Conference on Computer Communication and Informatics (ICCCI). IEEE, 2017.
- [5] Hassoon, Mafazalyaqeen, et al. "Rule optimization of boosted c5. 0 classification using a genetic algorithm for liver disease prediction." 2017 International Conference on Computer and Applications (ICCA). IEEE, 2017.
- [6] Karthik. S, Priyadarshini. A. Anuradha J. and Tripathi B. K, Classification and Rule Extraction using Rough Set for Diagnosis of Liver Disease and its Types, Ad.
- [7] Thapa, B. R., and Anuj Walia. "Liver function tests and their interpretation." The Indian Journal of Pediatrics 74.7(2007): 663-671.
- [8] Sullivan, Tim. "Blitzscaling." Harvard business review 94.4(2016): 15.
- [9] Jae-Young Lim, "The Prospect of the Fourth Industrial Revolution and Home Healthcare in Super-Aged Society", https://www.researchgate.net/profile/Jae-Young_Lim2.
- [10] Vyshali J Gogi, Vijayalakshmi M.N, "Prognosis of Liver Disease: Using Machine Learning Algorithms", International Conference on Recent Innovations in Electrical, Electronics & Communication Engineering (ICRIEECE), 875-879 (2018)
- [11] Pushpendra Kumar, Ramjeevan Singh Thakur, "Diagnosis of Liver Disorder Using Fuzzy Adaptive and Neighbor Weighted K-NN Method for LFT Imbalanced Data", International Conference on Smart Structures and Systems (ICSSS),1-5 (2019).
- [12] Sanjay Kumar, Sarthak Katyal, "Effective Analysis and Diagnosis of Liver Disorder by Data Mining", International Conference on Inventive Research in Computing Applications (ICIRCA), 1047-1051 (2018).
- [13] M. Banu Priya, P. Laura Juliet, P.R. Tamilselvi, "Performance Analysis of Liver Disease Machine Learning Prediction Using Algorithms",

International Research Journal of
Engineering and Technology (IRJET),
Volume: 05 Issue: 01, 206-211 (2018).

[14] Ahmad Shaker Abdalrada, Omar Hashim Yahya, Abdul Hadi M. Alaidi, Nasser Ali Hussein, Haider TH. Alrikabi, Tahsien Al- Quraishi Al-Quraishi, "A Predictive model for liver disease progression based on logistic regression algorithm", Periodicals of Engineering and Natural sciences, vol. 7, no. 3, 1255-1264 (2019)

