

**A NEW MACHINE LEARNING-BASED COMPLEMENTARY
APPROACH FOR SCREENING OF NAFLD (HEPATIC
STEATOSIS)**

A Project Phase-II

Submitted in partial fulfillment of the requirements for the award of a
Bachelor of Engineering degree in Electronics and Communication Engineering

by

B N S Aditya(39130054)

A.ROHIT(39130037)



DEPARTMENT OF ELECTRONICS AND COMMUNICATION
ENGINEERING

SCHOOL OF ELECTRICAL AND ELECTRONICS

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with Grade "A" by NAAC

JEPPIAAR NAGAR, RAJIV GANDHISALAI, CHENNAI - 600119

APRIL - 2023



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with —All grade by NAAC

Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai – 600 119

www.sathyabama.ac.in

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Final Year Project Report is the bonafide work of **B ADITYA** (39130054), **A ROHIT**(39130037) and who carried out of the project title "**A NEW MACHINE LEARNING BASED COMPLEMENTARY APPROACH FOR SCREENING OF NAFLD(HEPATIC STEATOSIS)**" under my supervision from Jan 2023 to April 2023.

Internal Guide

Dr. POONGUZHALI S, M.E., Ph.D.

Head of the Department

Dr. T. RAVI, M.E., Ph.D.

Submitted for Viva voice Examination held on 24-04-2023

Internal Examiner

External Examiner

DECLARATION

We, B.ADITYA(39130054), A.ROHIT(39130142), hereby declare that the Project Report on "A NEW MACHINE LEARNING-BASED COMPLEMENTARY APPROACH FOR SCREENING OF NAFLD(HEPATIC STEATOSIS)" done by us under the guidance of Dr.POONGUZHALI S, M.E.,Ph.D., at Sathyabama Institute of Science and technology is submitted in partial fulfillment of the requirements of the award of Bachelor of Engineering degree in **Electronics and Communication Engineering**.

DATE: 24-04-23

PLACE: CHENNAI

1. A. Rohit
2. B.N.S. Aditya

SIGNATURE OF THE CANDIDATES

ACKNOWLEDGEMENT

We are pleased to acknowledge our sincere thanks to Board of Management of **SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY** for their kind encouragement in doing this project and for completing it successfully. We are grateful to them.

We convey our thanks to **Dr. N.M.NANDHITHA M.E., Ph.D., Dean**, School of Electronics, **Dr. T.RAVI M.E., Ph.D.**, Head of the Department of Electronics and Communication Engineering for providing us necessary support and details at the right time during the progressive reviews.

We would like to express our sincere and deep sense of gratitude to our Project Guide **Dr.POONGUZHALI,M.E.,Ph.D.**,for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of our project work.

We wish to express our thanks to all Teaching and Non-teaching staff members of the **Department of Electronics and Communication Engineering** who are helpful in many ways for the completion of the project.

ABSTRACT

Due to data science progress in biomedical and healthcare communities, accurate study of medical data benefits early disease recognition, patient care and community services. When the quality of medical data is incomplete the exactness of study is reduced. Moreover, different regions exhibit unique appearances of certain regional diseases, which may results in weakening the classification of disease outbreaks. In the proposed system, it provides data science techniques for effective classification of various disease occurrences in disease-frequent societies. It experiment the altered estimate models over real-life hospital data collected. To overcome the difficulty of incomplete data, it use a latent factor model to rebuild the missing data. It experiment on a regional chronic illness of cerebral infarction. Using structured and unstructured data from hospital it use data science techniques. It predicts probable diseases by mining data sets such as Malaria, Breast Cancer, Diabetes, Liver Disease, Kidney Disease and Heart Disease. To the best of our knowledge in the area of medical big data analytics none of the existing work focused on both data types. Compared to several typical estimate algorithms, the calculation exactness of our proposed algorithm reaches 94.8% with a convergence speed which is faster than that of the data science disease risk classification algorithm.

TABLE OF CONTENT

CHAPTER NO	TITLE	PAGE NO
	ABSTRACT	v
	LIST OF FIGURES	viii
1	INTRODUCTION	1
2	LITERATURE SURVEY	3
	2.1 Inference of literature suvery	3
	2.2 Existing system	6
3	REQUIREMENT ANALYSIS	11
	3.1 Feasiblity studies	11
	3.2 Software requirements	12
	3.3 Data Visualization	18
4	DESCRIPTION OF PROPOSED SYSTEM	21
	4.1 Proposed system	21
	4.2 Architecture design	22
	4.3 Input and output design	22
	4.4 System testing	25
	4.5 Hardware requirements	29
5	RESULTS	40
	DISCUSSION	44
6	CONCLUSION	46
	6.1 Conclusion	46
	6.2 Future Work	47
	REFERENCE	48

APPENDIX	
SOURCE CODE	51
RESEARCHPAPER	69

LIST OF FIGURES		
FIGURE NUMBER	FIGURE NAME	PAGE NUMBER
4.1	Architecture design	
4.2	Breast Cancer	32
4.3	Diabetes Complication	33
4.4	Heart Disease	34
4.5	Liver Disease	35
4.6	Kidney Disease	36
4.7	Malaria	37
4.8	Pneumonia	38
5.1	Login URL	40
5.2	Liver Disease Prediction	42
5.3	Liver Disease Prediction	42
5.4	Liver Disease Prediction	43
5.5	Liver Disease Prediction	43

CHAPTER -1

INTRODUCTION

The Earth is going through a purplish patch of technology where the demand of intelligence and accuracy is increasing behind it. Today's people are likely addicted to internet but they are not concerned about their physical health. People ignore the small problem and don't visit to visit hospital which turn into serious disease with time. Taking the advantage of this growing technology, our basis aim is to develop such a system that will predict the multiple diseases in accordance with symptoms put down by the patients without visiting the hospitals / physicians. Machine Learning is a subset of AI that is mainly deal with the study of algorithms which improve with the use of data and experience. Machine Learning has two phases i.e. Training and Testing. Machine Learning provides an efficient platform in medical field to solve various healthcare issues at a much faster rate. There are two kinds of Machine Learning — Supervised Learning and Unsupervised Learning. In supervised learning we frame a model with the help of data that is well labelled. On the other hand, unsupervised learning model learn from unlabeled data. The intent is to deduce a satisfactory Machine Learning algorithm which is efficient and accurate for the prediction of disease. In this paper, the supervised Machine Learning concept is used for predicting the diseases. The main feature will be Machine Learning in which we will be using machine learning algorithm which will help in early prediction of diseases accurately and better patient care. The analysis accuracy is reduced when the quality of medical data is incomplete. Moreover, different regions exhibit unique characteristics of certain regional diseases, which may weaken the prediction of disease outbreaks. However, those existing work mostly considered structured data. There is no proper methods to handle semi structured and unstructured. The proposed system will consider both structured and unstructured data. The analysis accuracy is increased by using Machine

Learning algorithm. In this work, our goal is to provide a tool to assist professionals and consumers in finding and choosing disease. To achieve this goal, we develop an approach that allows a user to query for disease that satisfy a set of conditions based on disease properties, such as disease indications and also takes into account patient profiles.

CHAPTER -2

2.1 LITERATURE SURVEY

Machine learning-based method for personalized and cost-effective detection of Alzheimer's disease

Diagnosis of Alzheimer's disease is often difficult, especially early in the disease process at the stage of mild cognitive impairment. Yet, it is at this stage that treatment is most likely to be effective, so there would be great advantages in improving the diagnosis process. We describe and test a machine learning approach for personalized and cost-effective diagnosis of AD. It uses locally weighted learning to tailor a classifier model to each patient and computes the sequence of biomarkers most informative or cost-effective to diagnose patients. Using ADNI data, we classified AD versus controls and MCI patients who progressed to AD within a year, against those who did not. The approach performed similarly to considering all data at once, while significantly reducing the number (and cost) of the biomarkers needed to achieve a confident diagnosis for each patient. Thus, it may contribute to a personalized and effective detection of AD, and may prove useful in clinical settings. Non-alcoholic fatty liver disease (NAFLD) is a growing health concern worldwide, and early detection is crucial for effective treatment. Machine learning has shown great promise in healthcare, including the early detection of various diseases.

Effect of Meteorological Conditions on Occurrence of Hand, Foot and Mouth Disease in Wuwei City, Northwestern China

The main objective of this paper is to supply scientific basics for preventing and forecasting the prevalence of hand, foot and mouth disease to explore the effect of different meteorological conditions on occurrence of hand, foot and mouth disease in Wuwei City, northwestern China. Here the data about the diseases and weather was collected from 2008-2010, and the correlation analysis, multiple linear regression and exponential curve fitting

methods were made. The results showed that 2688 cases of hand, foot and mouth disease were collected from 2008 to 2010, and the annual average incidence was 47.62/100,000. The average prevalence of hand, foot and mouth disease at Liangzhou District, Minqin County, Gulang County and Tianzhu Tibetan Autonomous County were 42.69, 38.52, 65.92 and 49.18 per 100,000 respectively. This disease occurred year-round in Wuwei City, but had a clear seasonal climax. Generally, the incidence increased from April and rose to the first peak in May, Jun, July respectively. The second peak was in September or October every year. Different meteorological factors had different impact on the epidemic of disease in four areas, such as average temperature, relative humidity, indicated that relative humidity and atmospheric pressure were the main influence factors in Liangzhou District, average temperature in Gulang County, atmospheric pressure in Tianzhu County. The incidence of the disease and average sunshine hours showed exponential function relationship in Minqin County. One potential approach for screening NAFLD using machine learning could involve analyzing a combination of clinical and imaging data. For example, a model could be trained on a dataset of patient records, including demographics, medical history, laboratory test results, and imaging scans such as ultrasound, CT or MRI. The model would then learn to identify patterns in the data that are associated with the presence of NAFLD. In conclusion, different weather conditions have different impact on the prevalence of hand, foot and mouth disease. A high correlation exists in four areas of Wuwei City between meteorological factors and hand, foot and mouth disease occurrence. And summer and autumn were the important seasons to prevent and control the disease.

Developing an Index for Detection and Identification of Disease Stages

Spectral data have been widely used to estimate the disease severity levels of different plants. However, such data have not been evaluated to estimate the disease stages of the plant. This study aimed at developing a

spectral disease index that is able to identify the stages of wheat leaf rust disease at various DS levels. To meet the aim of the study, the reflectance spectra of infected leaves with different symptom fractions and DS levels were measured with a spectroradiometer. To improve the accuracy of the model, additional features could be added such as genetic markers, dietary habits, or physical activity levels. Once the model has been trained, it could be used to screen new patients for the presence of NAFLD, potentially allowing for earlier detection and intervention. Then, pure spectra of the different disease symptoms at the leaf scale were analyzed, and a new function was developed to find the wavelengths most sensitive to disease symptom fraction. The reflectance spectra with highest sensitivity were found at 675 and 775 nm. Finally, the normalized difference of DS and the ratio p_{675}/p_{775} was used as a new SDI to discriminate three different levels of the disease stage at the canopy level. The suggested SDI showed a promising performance to improve the detection disease stages in precision plant protection.

Quantized Analysis for Heart Valve Disease based on Cardiac Sound Characteristic Waveform Method

In order to analyze heart valve disease accurately and effectively, a new quantized diagnosis method was proposed to analyze four clinical heart valve sounds, namely cardiac sound characteristic waveform. BIOPAC acquiring system was used to collect signal. The recorded data is transmitted to a computer by ethernet for storage ! analysis and display in real-time. Analytical model of single degree-of- freedom was established to extract characteristic waveform. Furthermore, diagnosis parameters were calculated to discriminate heart sound of normal and heart valve disease by easy-understanding graphical representation, so that, even for an inexperienced user is able to monitor his or her pathology progress easily. It is important to note that the development and implementation of any such machine learning-based screening approach for NAFLD would require careful validation and testing to ensure its safety and effectiveness. Additionally, any such approach should

always be used in conjunction with clinical judgment and follow-up testing as necessary. Finally, a case study on a heart valve disease patient before and after surgery is demonstrated to validate the usefulness and efficiency of the proposed method.

Non-Linear Analysis of Heart Rate Variability in Patients with Coronary Heart Disease

The article emphasizes clinical and prognostic significance of non-linear measures of the heartrate variability, applied on the group of patients with coronary heart disease and age-matched healthy control group. Three different methods were applied: Hurst exponent, Detrended Fluctuation Analysis and approximate entropy. Hurst exponent of the R-R series was determined by the range rescaled analysis technique. Non-alcoholic fatty liver disease (NAFLD) is a condition in which excess fat accumulates in the liver of individuals who consume little or no alcohol. NAFLD is a growing health concern, and there is a need for accurate and efficient screening methods to detect the disease early. Machine learning (ML) has shown promise in improving the accuracy of disease diagnosis and prediction, and could be applied to NAFLD screening. DFA was used to quantify fractal long-range-correlation properties of heart rate variability. Approximate entropy measures the unpredictability of fluctuations in a time series. It was found that the short-term fractal scaling exponent. The patients with CHD had lower Hurst exponent in each program of exercise test separately, as well as approximate entropy than healthy control group.

2.2 EXISTING SYSTEM

This study proposes a non-invasive method for diagnosing liver diseases using ultrasound images, by classifying liver tissue as normal, steatosis, or cirrhosis, using feature extraction, feature selection, and classification. First, the correlation, homogeneity, variance, entropy, contrast, energy, long run emphasis, run percentage, and standard deviation are determined. Second, the most efficient features are selected based on the Fisher discriminant and manual selection methods. Third, three voting-based subclassifiers are used,

namely, the normal/steatosis, normal/cirrhosis, and steatosis/cirrhosis classifiers. The final liver tissue classification is based on the majority function. There are several methods in the literature individually to diagnosis diabetes, kidney or heart disease. Parthiban *et al.* proposed Naïve Bayes based method to diagnose heart disease using diabetic dataset that contain no prior information related to heart disease. However, there is no machine learning based method to identify NAFLD risk from diabetic dataset with no prior information related to NAFLD risk; using risk factors based on ATP III clinical criteria proposed in 2005 for metabolic syndrome to our knowledge.

Early stage detection and diagnosis of NAFLD risk is needed for a variety of reasons. If detected at an early stage and contained promptly, it may be possible to check NAFLD from getting worse and decrease the quantity of fat in liver effectively. About 50% of individuals with compensated cirrhosis owing to NAFLD would either require liver transplant or pass away due other disorders triggered by liver associated diseases. NAFLD individuals demonstrate significantly higher premature mortality rate than the general population. Identification of novel treatments is bound on the early and reliable identification of NAFLD risk. Data mining has been of tremendous interest in healthcare community for some decades now, which identifies useful information by sifting through huge quantities of data using statistical as well as pattern recognition and mathematical techniques. In this setting, EMRs demonstrate a vital role through cognizing of repetitive clinical measurements related to a patient's condition over time along with vital signs, diagnosis, procedures, prescribed medications and demographics. In principle this comprehensive information from each medical encounter can be incorporated to build models that take the semantics of such data into account, use information and knowledge intelligently and effectively help disease prediction as well as progression. Hence it is needed to analyse the already available huge diabetic data sets to discover some incredible facts which may help in producing some prediction model.

To overcome the above-mentioned issues and provide for a rapid and detailed analysis of medical data the present study proposes a Decision Tree (DT)

based prediction model to investigate the risk of developing NAFLD in the Canadian population using risk factors proposed for MetS by ATP III. It may be noted that the risk factors used in our proposed method are those that are put forward in Adult Treatment Panel III (ATP III) clinical criteria proposed in 2005 to diagnose metabolic syndrome and are not direct indicators of NAFLD.

Methodology

HealthCare Data

The data used in this research is acquired from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) which is a pioneer multi-disease EMR-based surveillance system of Canada. Data from all participating networks, provided by family physicians and other primary care providers, are aggregated into a single national database (<http://cpcssn.ca/>). CPCSSN contains 667907 records for a period ranging from 2003 to Sept 30, 2013 and every record comprises of various attributes regarding vital signs, diagnosis and demographics. This dataset has previously been used by Mashayekhi *et al.* to assert the discriminability of the Framingham diabetes risk model in Canadian population. An abstract overview of CPCSSN dataset is given in Table. The consolidation of healthcare information from healthcare centers and hospitals in CPCSSN.

Predictors	Findings
Demographic (Gender, Age)	
Male, sample size (%)	287964 (43)
Female, sample size (%)	379561, (57)
Male age mean $\pm$ SD, Years	47.2 $\pm$ 25.1
Female age mean $\pm$ SD, Years	49.5 $\pm$ 24.8
Vital Signs/ clinical measures	
Diastolic BP mean $\pm$ SD, mm Hg	73.3 $\pm$ 12.4
Systolic BP, mean $\pm$ SD, mm Hg	121.9 $\pm$ 16.9
Unknown disease frequency (%)	393344 (59)
COPD frequency (%)	15926 (2.4)
Dementia frequency (%)	12007 (1.8)
Depression frequency (%)	62682 (10)
Diabetes Mellitus frequency (%)	40637 (6)
Epilepsy frequency (%)	5553 (0.8)
Hypertension frequency (%)	88615 (13)
Osteoarthritis frequency (%)	47606 (7)
Parkinson's Disease frequency (%)	1825 (0.2)
Lab Values	
FG, mean $\pm$ SD, mmol/L	5.54 $\pm$ 1.91
Triglycerides, mean $\pm$ SD, mmol/L	1.43 $\pm$ 1.21
HDL, sample size, mean $\pm$ SD, mmol/L	1.38 $\pm$ 0.41
BMI, mean $\pm$ SD, kg/m ²	26.54 $\pm$ 7.37

job; hence, not all the information related to risk factors considered for the NAFLD risk prediction are available for all individuals, thus restricting the size of data. At this stage the dataset on clinical measurements are partial, about 627,180 patients out of 667,907 do not bear information for all the factors that are considered in Characteristics of the population in the Canadian primary care sentinel surveillance network database. SD, standard deviation; BP, Blood Pressure, BMI, body mass index, FG, Fasting glucose, HDL, high-density lipoprotein, COPD, chronic obstructive pulmonary disease. *Some patients have more than 1 disease in the database. this research based on ATP III for the prediction of NAFLD. The records that contain information for all the related factors are considered in this research. Hence, the final research sample after preprocessing contains 40637 records that include approximately 59% women and 40% men. The liver is the largest organ of the body and it is essential for

digesting food and releasing the toxic element of the body. The viruses and alcohol use lead the liver towards liver damage and lead a human to a life-threatening condition. There are many types of liver diseases whereas hepatitis, cirrhosis, liver tumors, liver cancer, and many more. Among them liver diseases and cirrhosis as the main cause of death.

Therefore, liver disease is one of the major health problems in the world. Every year, around 2 million people died worldwide because of liver disease [2]. According to the Global Burden of Disease (GBD) project, published in BMC Medicine, one million peoples are died in 2010 because of cirrhosis and million are suffering from liver cancer [3]. Machine learning has made a significant impact on the biomedical field for liver disease prediction and diagnosis [4-6]. Machine learning offers a guarantee for improving the detection and prediction of disease that has been made an interest in the biomedical field and they also increase the objectivity of the decision-making process.

The attribute that divides the data efficiently is selected as a root node. Next, a child node is selected by calculating Information Gain or some other statistical measure. The branches coming from an internal node are labeled with values of the attribute that a particular node can assume and each branch from root to leaf node represent an if-then rule for the prediction of class for a newly seen instance. Decision trees are reasonable to build, easy to perceive and integrate with database system. The level of effectiveness of the classification methods can be distinguished with number of correct and incorrect classifications in each possible case of the variables being classified. Given a particular classification algorithm and a data instance, there exist four possibilities. To assess the discriminative capability of J48 classifier in both datasets as described above most frequently used performance measures such as Micro- and Macro-average of Precision,

CHAPTER-3

3.1 FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

- ◆ ECONOMICAL FEASIBILITY
- ◆ TECHNICAL FEASIBILITY
- ◆ SOCIAL FEASIBILITY

ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system. Classification is a procedure of building a model of class attributes from a dataset, to assign a class label to previously unseen record as accurately as possible. DT is a supervised classification model aimed at partitioning data into homogeneous

groups in term of variables to be predicted using entropy. If the partition of data is completely homogeneous, entropy will be zero. Entropy is a gauge to measure the level of disorder in data. Basically, it defines the quantity of information provided by an event. The lower the entropy of an event is (it is rare), the higher the information it provides. Information gain is based on decrease in entropy. DT is a tree like hierarchical structure that consists of branches (arcs) and three types of nodes, root, intermediate and leaf node respectively that correspond to the sequence of decision rules

SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

Non-alcoholic fatty liver disease (NAFLD) is a condition in which excess fat accumulates in the liver of individuals who consume little or no alcohol. NAFLD is a growing health concern, and there is a need for accurate and efficient screening methods to detect the disease early. Machine learning (ML) has shown promise in improving the accuracy of disease diagnosis and prediction, and could be applied to NAFLD screening.

3.2 SOFTWARE REQUIREMENTS

- Platform : Windows Xp/7/8/10
- Coding Language : Python

SOFTWARE ENVIRONMENT

Python:

Python is a high-level, interpreted, interactive and object-oriented scripting language. Python is designed to be highly readable. It uses English keywords

frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages.

- **Python is Interpreted** – Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is similar to PERL and PHP.
- **Python is Interactive** – You can actually sit at a Python prompt and interact with the interpreter directly to write your programs.
- **Python is Object-Oriented** – Python supports Object-Oriented style or technique of programming that encapsulates code within objects.
- **Python is a Beginner's Language** – Python is a great language for the beginner-level programmers and supports the development of a wide range of applications from simple text processing to WWW browsers to games.
- Our ultimate goal is to reduce mortality and morbidity from liver disease. We therefore aim to set up a targeted, easy-to-use, cost-beneficial screening program for detection of liver disease in high-risk groups within the general population, using transient elastography as screening tool.
- Validate the diagnostic accuracy of machine learning model in the general population and primary care, using Support Vector Classifier
- Assess the prevalence of significant liver disease as evidenced in the general population and stratified according to subjects at risk for liver disease
- Identify optimal screening intervals and define population subgroups at highest risk of progressive liver disease with the aim to design the ideal screening strategy
- Estimate the incremental cost-effectiveness ratios of the screening program

History of Python

Python was developed by Guido van Rossum in the late eighties and early nineties at the National Research Institute for Mathematics and Computer Science in the Netherlands.

Python is derived from many other languages, including ABC, Modula-3, C, C++, Algol-68, SmallTalk, and Unix shell and other scripting languages.

Python is copyrighted. Like Perl, Python source code is now available under the GNU GeneralPublic License (GPL).

Python is now maintained by a core development team at the institute, although Guido vanRossum still holds a vital role in directing its progress.

Python Features

Python's features include –

- **Easy-to-learn** – Python has few keywords, simple structure, and a clearly defined syntax. This allows the student to pick up the language quickly.
- **Easy-to-read** – Python code is more clearly defined and visible to the eyes.
- **Easy-to-maintain** – Python's source code is fairly easy-to-maintain.
- **A broad standard library** – Python's bulk of the library is very portable and cross-platformcompatible on UNIX, Windows, and Macintosh.
- **Interactive Mode** – Python has support for an interactive mode which allows interactivetesting and debugging of snippets of code.
- **Portable** – Python can run on a wide variety of hardware platforms and has the sameinterface on all platforms.
- **Extendable** – You can add low-level modules to the Python interpreter. These modulesenable programmers to add to or customize their tools to be more efficient.

- **Databases** – Python provides interfaces to all major commercial databases.
- **GUI Programming** – Python supports GUI applications that can be created and ported to many system calls, libraries and windows systems, such as Windows MFC, Macintosh, and the X Window system of Unix.
- **Scalable** – Python provides a better structure and support for large programs than shell scripting.

Apart from the above-mentioned features, Python has a big list of good features, few are listed below –

- It supports functional and structured programming methods as well as OOP.
- It can be used as a scripting language or can be compiled to byte-code for building large applications.
- It provides very high-level dynamic data types and supports dynamic type checking.
- It supports automatic garbage collection.
- It can be easily integrated with C, C++, COM, ActiveX, CORBA, and Java.
- Python is available on a wide variety of platforms including Linux and Mac OS X. Let's understand how to set up our Python environment.

Getting Python

The most up-to-date and current source code, binaries, documentation, news, etc., is available on the official website of Python <https://www.python.org>.

Windows Installation

Here are the steps to install Python on Windows machine.

- Open a Web browser and go to <https://www.python.org/downloads/>.
- Follow the link for the Windows installer python-XYZ.msifile where XYZ is the version you need to install.

- To use this installer python-XYZ.msi, the Windows system must support Microsoft Installer

2.0. Save the installer file to your local machine and then run it to find out if your machinesupports MSI.

- Run the downloaded file. This brings up the Python install wizard, which is really easy to use. Just accept the default settings, wait until the install is finished, and you are done.

The Python language has many similarities to Perl, C, and Java. However, there are some definite differences between the languages.

First Python Program

Let us execute programs in different modes of programming.

Interactive Mode Programming

Invoking the interpreter without passing a script file as a parameter brings up the following prompt

Project Structure:

- Flask files consist of template folder which has HTML pages, app.py file and .pkl files which are used for application building
- IBM folder has flask files and scoring endpoint.ipynb- model training code file.
- We need the model which is saved and the saved model in this content is Traffic volume. Pkl
- Templates folder which contains index.HTML file, chance.HTML file, noChance.HTML file.
- Scale.pkl for scaling, encoder.pkl file for encoding the categorical data, imputer.pkl file for filling out the missing values

Importing The Dataset

- You might have your data in .csv files, .excel files

- Let's load a .csv data file into pandas using `read_csv()` function. We will need to locate the directory of the CSV file at first (it's more efficient to keep the dataset in the same directory as your program).
- If your dataset is in some other location, Then
- **`Patients.data=pd.read_csv("Indian_liverpatient.csv")`**
- If the dataset is in the same directory of your program, you can directly read it, without giving raw as r.
- Our Dataset `indian_liverpatient.csv` contains the following Columns
- Age – of the person
- Total_Bilirubin
- Direct_Bilirubin
- Alkaline_Phosphate
- Alamine_Aminotransferase
- Aspartate_Aminotransferase
- Total_protiens
- Albumin
- Albumin_and_Globulin_Ratio
- Dataset

The output column to be predicted is liver disease on the input variables we predict the volume of disease..The predicted output gives them a fair idea whether a person have liver disease or not.

Analyse The Data

`head()` method is used to return top n (5 by default) rows of a DataFrame or series.

Handling the Missing values

1. The Most important step in data pre-processing is dealing with missing data, the presence of missing data in the dataset can lead to low accuracy.
2. Check whether any null values are there or not. if it is present then the following can be done.

There are missing values in the dataset, we will fill the missing values in the

columns.

3. We are using mean and mode methods for filling the missing values

- Columns such as temp, rain, and snow are the numeric columns, when there is a numeric column you should fill the missing values with the mean/median method. so here we are using the mean method to fill the missing values.
- Weather column has a categorical data type, in such case missing data needs to be filled with the most repeated/ frequent value. Clouds are the most repeated value in the column, so imputing with clouds value.

3.3 Data Visualization:

Data visualization is where a given data set is presented in a graphical format. It helps the detection of patterns, trends and correlations that might go undetected in text-based data.

Understanding your data and the relationship present within it is just as important as any algorithm used to train your machine learning model. In fact, even the most sophisticated machine learning models will perform poorly on data that wasn't visualized and understood properly.

- To visualize the dataset we need libraries called Matplotlib and Seaborn.
- The Matplotlib library is a Python 2D plotting library that allows you to generate plots, scatter plots, histograms, bar charts etc.

Let's visualize our data using Matplotlib and seaborn library.

Before diving into the code, let's look at some of the basic properties we will be using when plotting.

xlabel:Set the label for the x-axis.

ylabel:Set the label for the y-axis.

title:Set a title for the axes.

Legend:Place a legend on the axes.

- **data.corr()** gives the correlation between the columns

	Age	Gender	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Protiens	Albumin	Albumi
0	65	Female	0.7	0.1	187	16	18	6.8	3.3	
1	62	Male	10.9	5.5	699	64	100	7.5	3.2	
2	62	Male	7.3	4.1	490	60	68	7.0	3.3	
3	58	Male	1.0	0.4	182	14	20	6.8	3.4	
4	72	Male	3.9	2.0	195	27	59	7.3	2.4	

Table 3.1

Correlation of the Data

Correlation is a statistical term describing the degree to which two variables move in coordination with one another. If the two variables move in the same direction, then those variables are said to have a positive correlation. If they move in opposite directions, then they have a negative correlation.

- Correlation strength varies based on colour, lighter the colour between two variables, more the strength between the variables, darker the colour displays the weaker correlation

Data Header

	Age	Gender	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Protiens	Albumin	Albumi
0	65	Female	0.7	0.1	187	16	18	6.8	3.3	
1	62	Male	10.9	5.5	699	64	100	7.5	3.2	
2	62	Male	7.3	4.1	490	60	68	7.0	3.3	
3	58	Male	1.0	0.4	182	14	20	6.8	3.4	
4	72	Male	3.9	2.0	195	27	59	7.3	2.4	

Table 3.2

DATA HEAD AFTER SPLITTING IN BETWEEN MALE AND FEMALE

	Age	Gender	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphatase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Proteins	Albumin	Albumi
0	65	0	0.7	0.1	187	16	18	6.8	3.3	
1	62	1	10.9	5.5	699	64	100	7.5	3.2	
2	62	1	7.3	4.1	490	60	68	7.0	3.3	
3	58	1	1.0	0.4	182	14	20	6.8	3.4	
4	72	1	3.9	2.0	195	27	59	7.3	2.4	

Table 3.3

CHAPTER - 4

4.1 PROPOSED SYSTEM

The Proposed system of multiple disease classification using data science is that we have used algorithms and all other various tools to build a system which classification the disease of the patient using the symptoms and by taking those symptoms we are comparing with the system's dataset that is previously available. By taking those datasets and comparing with the patient's disease we will classify the accurate percentage disease of the patient. The dataset and symptoms go to the classification model of the system where the data is pre-processed for the future references and then the feature selection is done by the user where he will enter/select the various symptoms. Then the classification of those data is done with the help of data science techniques such as Logistic regression. Then the data goes in the recommendation model, there it shows the risk analysis that is involved in the system and it also provides the probability estimation of the system such that it shows the various probability like how the system behaves when there are n number of classifications are done and it also does the recommendations for the patients from their final result and also from their symptoms like it can show what to use and what not to use from the given datasets and the final results. It classifies probable diseases by mining data sets such as Malaria, Breast Cancer, Diabetes, Liver Disease, Kidney Disease and Heart Disease. To the best of our knowledge in the area of medical big data analytics none of the existing work focused on both data types.

ADVANTAGES OF PROPOSED SYSTEM

- Easily analyze the disease
- High Accuracy

4.2 ARCHITECTURE DIAGRAM

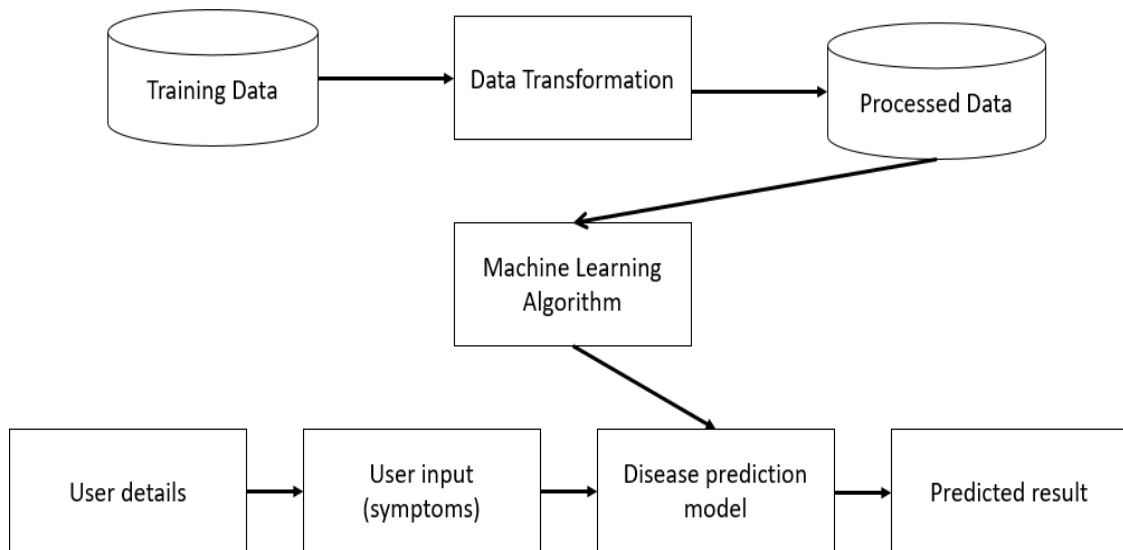


FIGURE 4.1

4.3 INPUT DESIGN AND OUTPUT DESIGN

INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. One potential ML-based approach for NAFLD screening is to use a combination of clinical and imaging data to predict the presence or absence of the disease. This could be done using a supervised learning algorithm, such as logistic regression or support vector machines, which is trained on a large dataset of patients with and without NAFLD. The algorithm would learn to identify the patterns and features in the data that are

most indicative of the disease, and use this information to make predictions on new patients. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

- What data should be given as input?
- How the data should be arranged or coded?
- The dialog to guide the operating personnel in providing input.
- Methods for preparing input validations and steps to follow when error occur.

OBJECTIVES

1.Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

2.It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

3.When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow

4.Another approach is to use unsupervised learning algorithms, such as clustering or anomaly detection, to identify subgroups of patients with similar clinical and imaging profiles, some of which may be indicative of NAFLD. This could be particularly useful in cases where NAFLD is not immediately obvious from imaging or clinical data alone.

OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. A third approach is to use deep learning techniques, such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs), to analyze medical images and extract features that are associated with NAFLD. This approach has shown promise in recent studies, which have demonstrated the ability of deep learning algorithms to accurately classify patients with NAFLD using ultrasound or MRI images. In summary, machine learning-based approaches have the potential to improve the accuracy and efficiency of NAFLD screening. However, further research and validation is needed to determine the best ML methods for NAFLD screening and to ensure that these approaches are safe and effective for use in clinical practice. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should identify the specific output that is needed to meet the requirements.

2. Select methods for presenting information.

3. Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

- ❖ Convey information about past activities, current status or projections of the

- ❖ Future.
- ❖ Signal important events, opportunities, problems, or warnings.
- ❖ Trigger an action.
- ❖ Confirm an action.

4.4 SYSTEM TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

When you are working on a model and you want to train it, you obviously have a dataset. But after training, we have to test the model on some test datasets. For this, you will have a dataset which is different from the training set you used earlier. But it might not always be possible to have so much data during the development phase. In such cases, the solution is to split the dataset into two sets, one for training and the other for testing.

- The train-test split is a technique for evaluating the performance of a machine learning algorithm.
- Train Dataset: Used to fit the machine learning model.
- Test Dataset: Used to evaluate the fit machine learning model.
- In general you can allocate 80% of the dataset to the training set and the remaining 20% to test.
- Now split our dataset into train set and test using `train_test_split` class from `sk learn` library.

TYPES OF TESTS

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce

valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application. It is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfactory, as shown by successful unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be

exercised. Systems/Procedures: interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

NAFLD is a type of fatty liver disease that is not related to heavy alcohol use. There are two kinds:

- Simple fatty liver, in which you have fat in your liver but little or no inflammation or liver cell damage. Simple fatty liver typically does not get bad enough to cause liver damage or complications.
- Nonalcoholic steatohepatitis (NASH), in which you have inflammation and liver cell damage, as well as fat in your liver. Inflammation and liver cell damage can cause fibrosis, or scarring, of the liver. NASH may lead to cirrhosis or liver cancer.

White Box Testing

White Box Testing is a testing in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is used to test areas that cannot be reached from a black box level.

Black Box Testing

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source

document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

Training And Testing

- Once after splitting the data into train and test, the data should be fed to an algorithm to build a model.

There are several Machine learning algorithms to be used depending on the data you are going to process such as images, sound, text, and numerical values. The algorithms that you can choose according to the objective that you might have it may be Classification algorithms are Regression algorithms.

Unit Testing:

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

Test objectives

- All field entries must work properly.
- Pages must be activated from the identified link.
- The entry screen, messages and responses must not be delayed.

Features to be tested

- Verify that the entries are of the correct format
- No duplicate entries should be allowed
- All links should take the user to the correct page.

Integration Testing

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce

failures caused by interface defects.

The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level –interact without error.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

4.5 HARDWARE REQUIREMENTS

- Processor : Core i3/i5/i7
- RAM : 2-4 GB

MODULES

- Data Collection Module
- Preparing the data Module
- Training a model
- Disease prediction Module

MODULE DESCRIPTIONS

Data Collection Module

Be it the raw data from excel, access, text files etc., this step (gathering past data) forms the foundation of the future learning. The better the variety, density and volume of relevant data, better the learning prospects for the machine becomes.

Preparing the data Module

Any analytical process thrives on the quality of the data used. One needs to spend time determining the quality of data and then taking steps for fixing issues such as missing data and treatment of outliers. Exploratory analysis is perhaps one method to study the nuances of the data in details thereby burgeoning the nutritional content.

Training a model

This step involves choosing the appropriate algorithm and representation of data in the form of the model. The cleaned data is split into two parts — train and test (proportion depending on the prerequisites); the first part (training data) is used for developing the model. The second part (test data), is used as a reference.

Disease prediction Module

Patient will specify the symptoms caused due to his illness. System will ask certain question regarding his illness and system predict the disease based on the symptoms specified by the patient.

Non-alcoholic fatty liver disease (NAFLD) is a complex disorder that involves the accumulation of fat in the liver in individuals who consume little or no alcohol. NAFLD is becoming increasingly common worldwide and is closely associated with obesity, type 2 diabetes, and metabolic syndrome. Screening for NAFLD is essential to diagnose the disease early and prevent its progression to more severe forms, such as non-alcoholic steatohepatitis (NASH) and cirrhosis.

Machine learning (ML) is a powerful tool that can be used to develop a complementary approach for screening of NAFLD. ML algorithms can analyze large datasets of patient information and identify patterns and trends that can be used to predict the risk of developing NAFLD. Here are some steps that can be taken to develop an ML-based screening approach for NAFLD:

Data collection: The first step in developing an ML-based screening approach for NAFLD is to collect large datasets of patient information. This information can include medical history, physical exam results, laboratory tests, and imaging studies.

Data preprocessing: Once the data has been collected, it needs to be preprocessed to remove any irrelevant or redundant information. This step can include data cleaning, normalization, and feature selection.

ML model development: ML models can be developed using a variety of algorithms, such as logistic regression, decision trees, random forests, and support vector machines. These models can be trained on the preprocessed data to identify patterns and trends that can be used to predict the risk of NAFLD.

Model validation: Once the ML model has been developed, it needs to be validated using an independent dataset. This step helps to ensure that the model is accurate and reliable.

Deployment: The final step in developing an ML-based screening approach for NAFLD is to deploy the model in a clinical setting. This can be done by integrating the model into electronic health records or developing a standalone software application.

An ML-based screening approach for NAFLD has the potential to be a powerful tool for identifying individuals at risk of developing the disease. However, it is important to note that ML models are only as good as the data they are trained on. Therefore, it is essential to collect high-quality data and ensure that the model is validated using independent datasets. Additionally, ML-based screening approaches should be used in conjunction with other diagnostic tests and clinical evaluations to ensure accurate and timely diagnosis of NAFLD.

Regenerate response

1. Breast Cancer

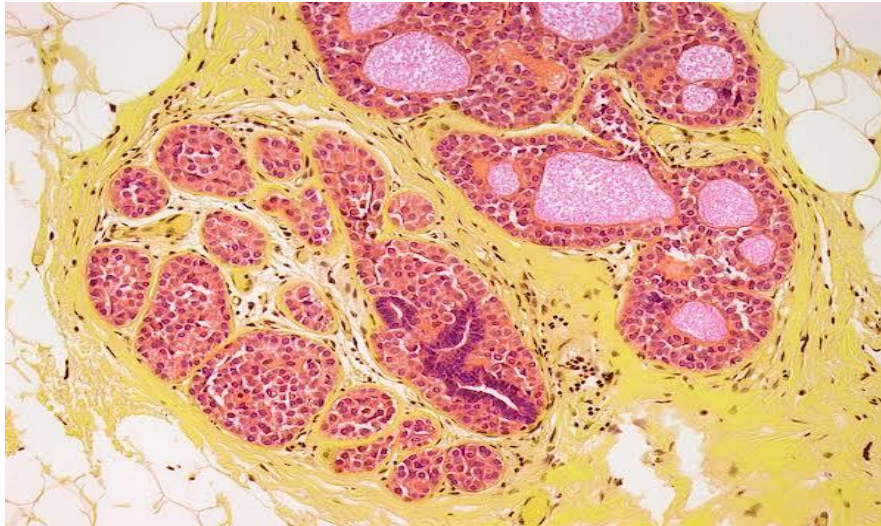


Fig 4.2 Breast Cancer

Breast cancer is cancer that develops from breast tissue. Signs of breast cancer may include a lump in the breast, a change in breast shape, dimpling of skin, fluid coming from the nipple, a newly inverted nipple, or a red or scaly patch of skin.

Symptoms

- Signs and symptoms of breast cancer may include.
- A breast lump or thickening that feels different from the surrounding tissue.
- Change in the size, shape or appearance of a breast.
- Changes to the skin over the breast, such as dimpling.
- A newly inverted nipple.

Peeling, scaling, crusting to flaking of the pigmented area of skin surrounding the nipple or breast skin • Redness or pitting of the skin over your breast like the skin of an orange

Causes

Doctors know the breast cancer occurs when some breast cells begin to grow abnormally. These cells divide more rapidly than healthy cells do and continue to accumulate, forming a lump or mass. Cells may spread (metastasize) through your breast to your lymph nodes or to other parts of your body.

Breast cancer most often begins with cells in the milk-producing ducts (invasive

ductal carcinoma), Breast cancer may also begin in the glandular tissue called

lobular carcinoma) or in other cells or tissue within the breast. Researchers have identified hormonal, lifestyle, and environmental factors that may increase your risk of breast cancer. But it's not clear why some people who have no risk factors develop cancer, yet other people with risk factors never do. It's likely that breast cancer is caused by a complex interaction of your genetic makeup and your environment.

2. Diabetes

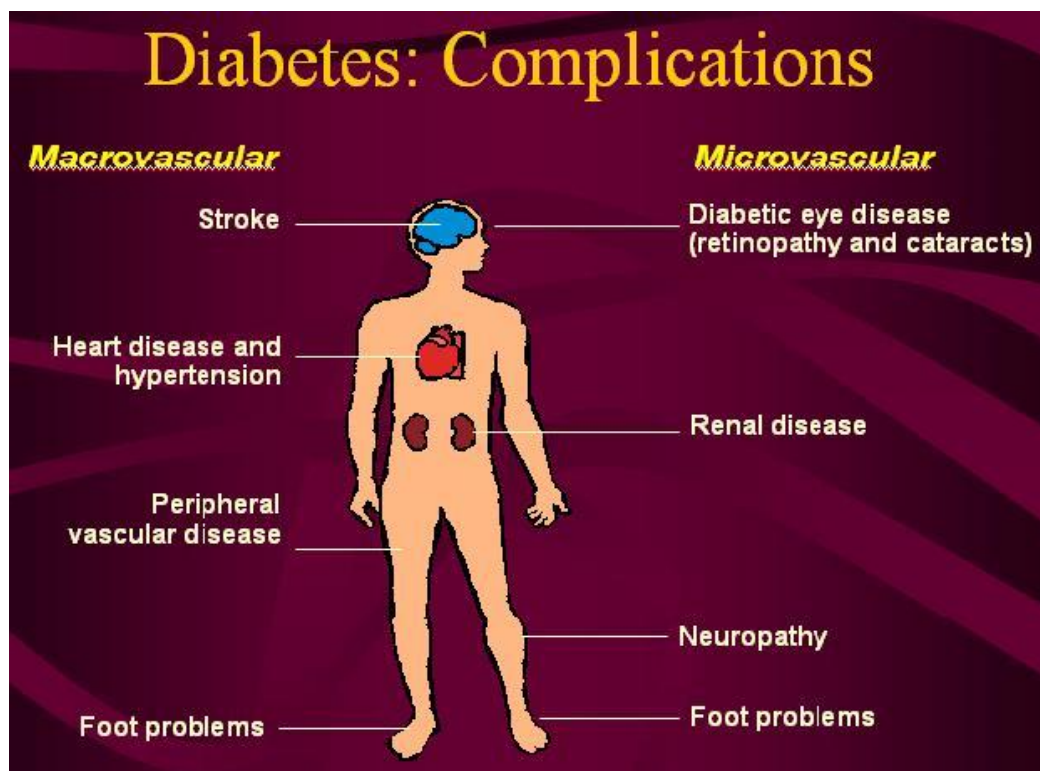


Fig 4.3 Diabetes Complications

Diabetes is a disease that occurs when your blood glucose, also called blood sugar, is too high. Blood glucose is your main source of energy and from the food you eat. Insulin, a hormone made by the pancreas, helps glucose from food get into your cells to be used for energy:

What health problems can people with diabetes develop? Over time, high blood glucose leads to problems such as

- Heart disease
- Stroke
- Kidney disease
- Eye problems
- Dental disease
- Nerve damage
- Foot problems

3. Heart Disease

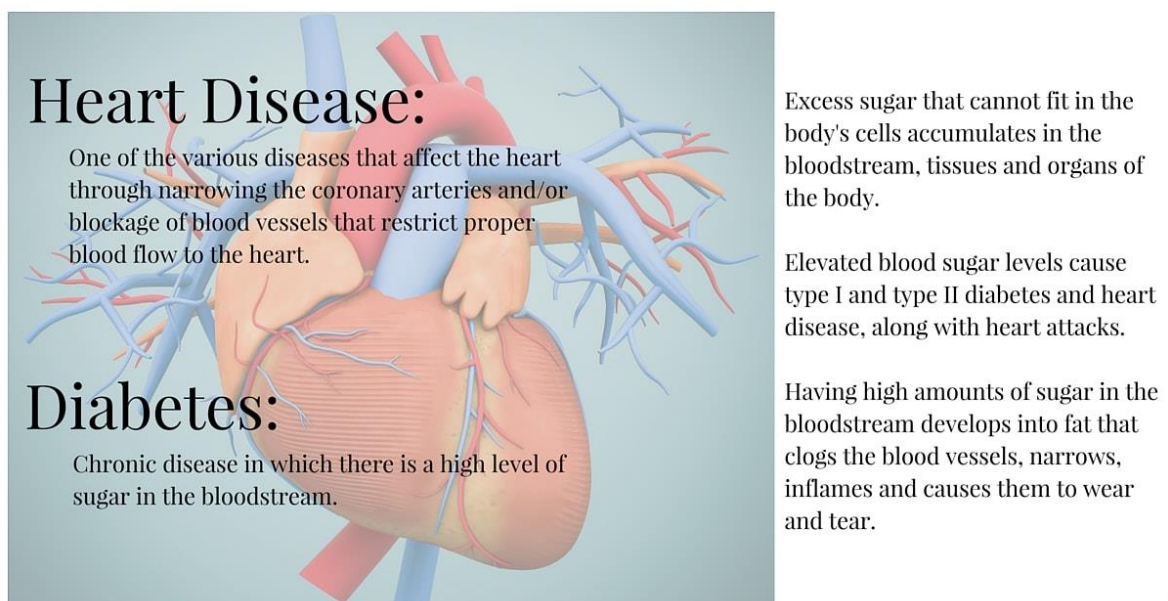


Fig 4.4 Heart Disease

Overview

Heart disease describes a range of conditions that affect your heart, Denies under the heart disease umbrella clude blood vesseldiseases such coronary arany disease heart rhythm problems anythies) and heart defects you're burn with (congenital heart defecta), among others The term "heart duas often used interchangeably with the term "cardiovascular disease. Cardiovascular disease generally raters to conditions that involve narrowed or blocked blood vessels that can lead to a heart attack, chest pain tanginal or stroka, Other heart conditions such as those that affect your heart's muscle valet or myth also are considered forms of heart disease

Many forms of heartse can be prevented or treated with healthy lifestyle choices

Symptoms

- Chest pain, chest tightness, chest pressure and chest discomfort (angina).
- Shortness of breath.
- Pain, numbness, weakness or coldness in your legs or arms it the blood vessels in those parts of your body are narrowed.
- Pain in the neck, a throat, upper abdomen or back.

4. Liver Disease

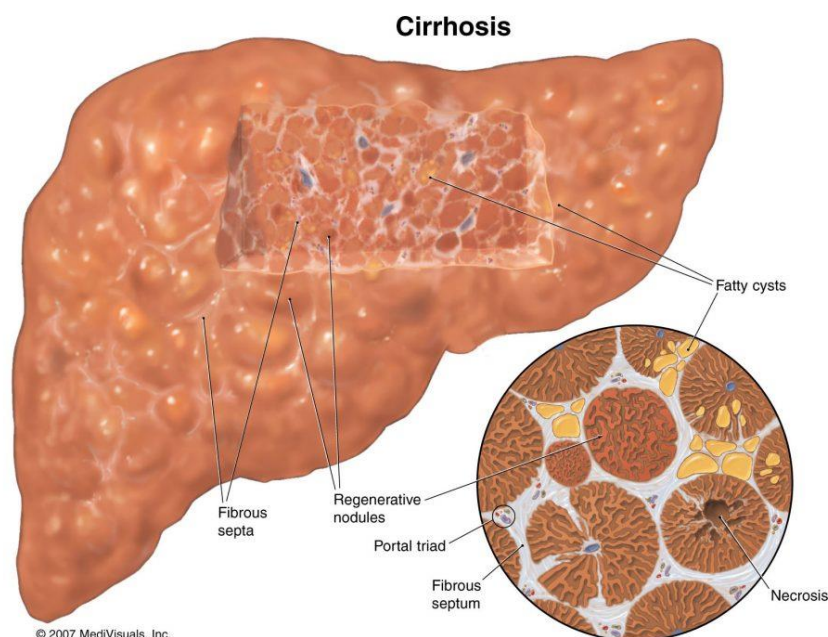


Fig 4.5 Liver Disease

Liver disease is any disturbance of liver function that causes illness. The thr is responsible for many citical functions within the body and should it become dinated or injured the loss of those functions can cause significant damage to the body. Liver disease also referred to as hepatic de Liveries abroad tem that covers all the potential problem that cause the low to fail to perform in designated functions thaly, mom than 75% three quarters of liver the needs to

be affected before a decrease in function occurs

Symptoms

classic symptoms of liver disease include

- Nausea
- Vomiting
- Right upper quadrant abdominal pain and
- Jaundice la yellow discoloration of the skin due to elevated bilirubin concentrations in the bloodstream

5. Kidney Disease



Fig 4.6 Kidney Disease

Overview

The kidneys play key roles in body function, not only by filtering the blood and getting rid of waste products, but also by balancing the electrolyte levels in the body, controlling blood pressure and stimulating the production of red blood cells. The kidneys are located in the abdomen toward the back, normally one on each side of the spine. They get their blood supply through the renal arteries directly from the aorta and send blood back to the heart via the renal veins to

the vena cava. (The term rent is derived from the Latin name for kidney)

- Lethargy Weakness
- Shortness of breath
- Generalized swelling (edema) Generalized weakness due to anemia
- Loss of appetite
- Lethargy
- Fatigue
- Congestive heart failure Metabolic acidosis
- High blood potassium (hyperkalemia)
- Fatal heart rhythm disturbances (arrhythmias) including ventricular tachycardia and ventricular fibrillation

6. Malaria



Fig 4.7 Malaria

Overview

Maleriertem musque-bure bloode, The Anopheles mosquitos traits it to humans
The parasites in mosquitos that spread malaria belong to the Podium genus.
Over 100 types of Plasmodium parashe can infect a variety of species

Different types replicate at different rates changing how quickly the symptoms escalate and the severity of the same

Symptoms

- Fever and chills
- Impaired consciousness
- Prostration, or adopting a prone position multiple convulsions
- Deep breathing and respiratory distress
- Abnormal bleeding and sign of anemia
- Cinical jaundice and guidance of vital organ dysfunction

7. Pneumonia

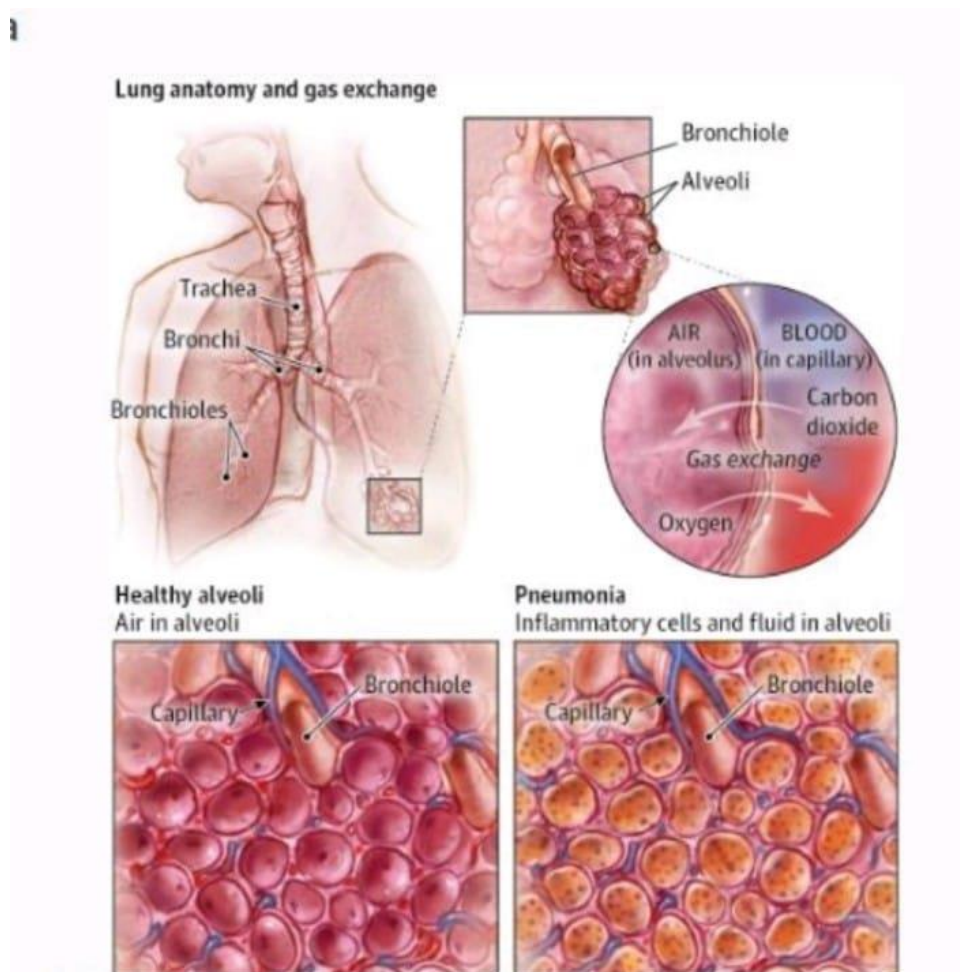


Fig 4.8 Pneumonia

Pneumonia is an infection in one or both lungs. It can be caused by bacteria, viruses or

fungi Bacterial pneumonia the most common type in adults

Pneumonias inflammation in the air sacs in your lungs, which are called alveoli.

Symptoms

- Common symptoms include:
- Cough
- Rusty or green phlegm, or sputum, coughed up from lungs
- Fast breathing and shortness of breath
- Shaking chills
- Chest pain that usually worsens when taking a deep breath, known as pleuritic pain
- Fast heartbeat
- Fatigue and weakness
- Nausea and vomiting
- Diarrhea

CHAPTER-5

RESULT AND DISSCUSION

We have used the logistic regression, Knn, SVC, random forest for the model evaluation

Random Forest -> 0.7542857142857143

Knn Classifier -> 0.7028571428571128

Logistic Regression -> 0.7657142857124857

Support Vector Classifier -> 0.7771428571428571

we have applied different type of algorithms for given data set to know which algorithm fits to our dataset

The above values are the respective accuracy values of Random Forest, Knn classifier, Logistic Regression, Support Vector Classifier models.

we have to take the best model having high accuracy in the above models.

Accuracy value for SVC is better, When compared with other models, so saving the SVC and deployed.

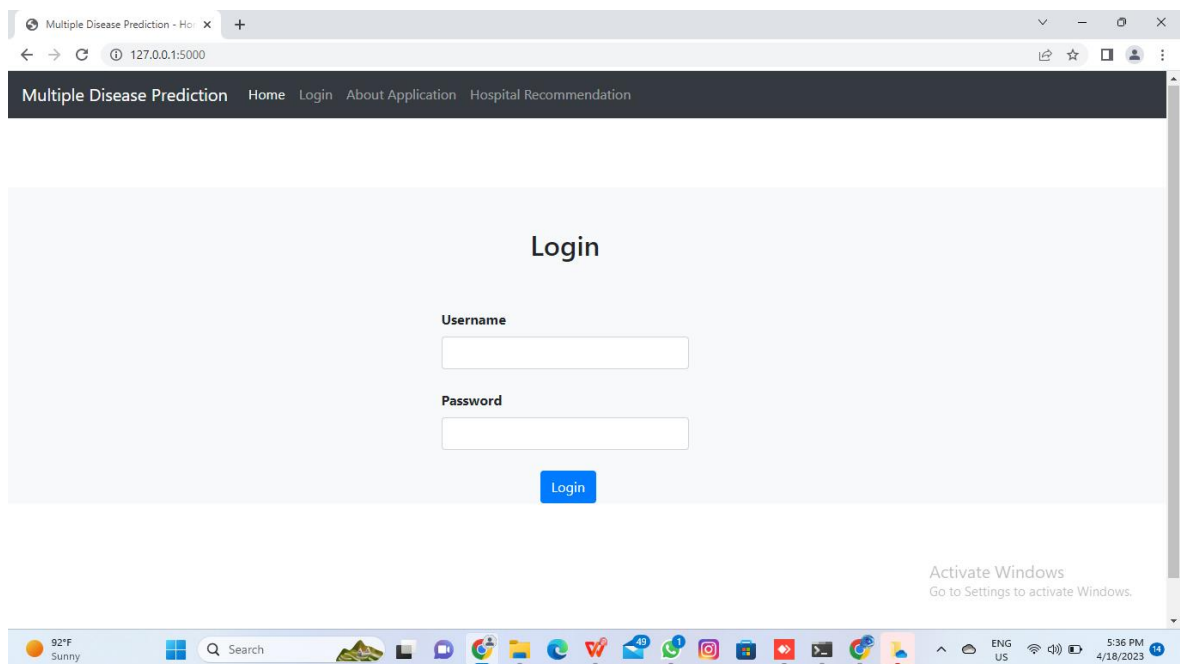


Fig 5.1 Login Url

Details

Welcome to the Home page of the application

This is mainly based on at the application of the machine learning meant to be employed in the remote areas

Overview

in todays time we see a lot of the shortage of the doctors in the world especially in Indie A lot of people are suffering a lot without the help of the proper medical checkup Also most of the cases many cases arise leading to death due to lack of timely medical checkup

So to cope up with all of those problems this app is designed which would prove its benefits upto much extent

Application

in todays time we see a lot of the shortage of the doctors in the world especially in Indie A lot of people are suffering a lot without the help of the proper medical checkup Also most of the cases many cases arise leading to death due to lack of timely medical checkup

So to cope up with all of those problems this app is designed which would prove its benefits upto much extent

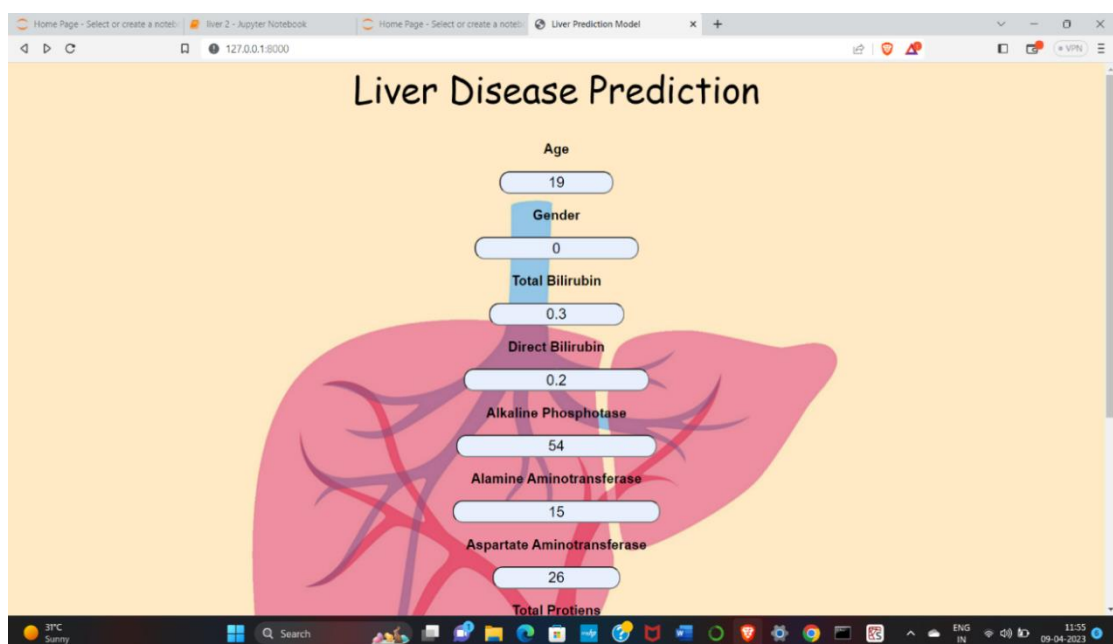


Fig 5.2

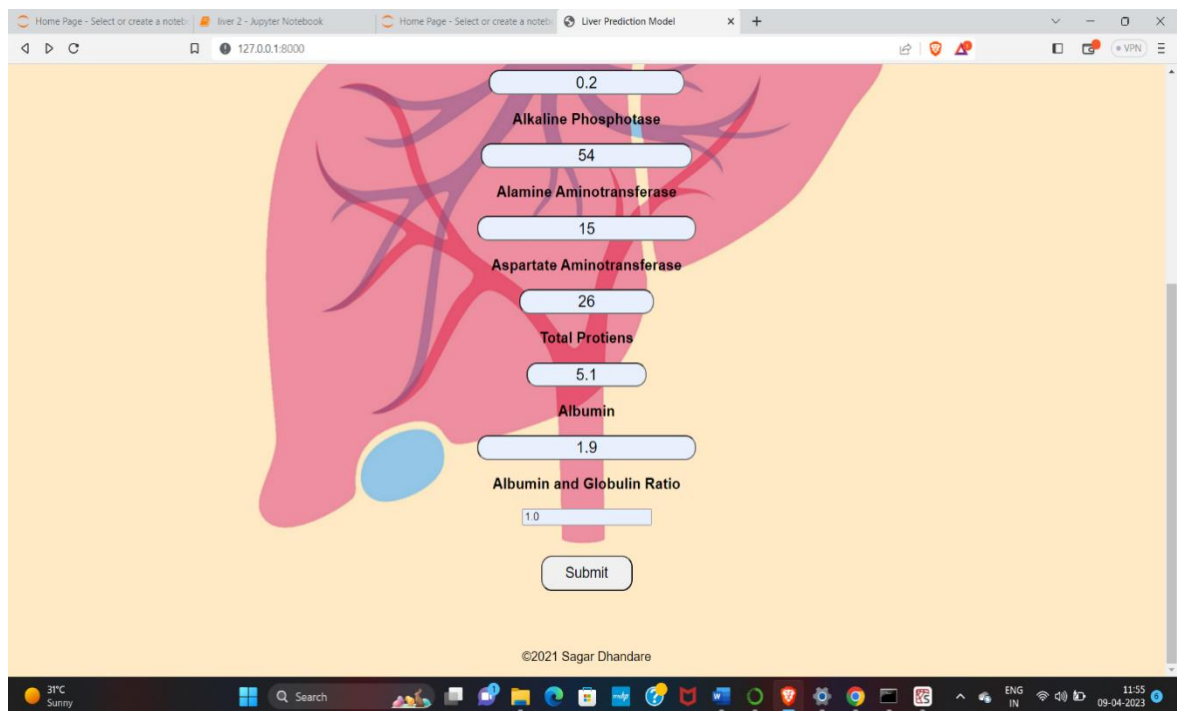


Fig 5.3



Fig 5.4

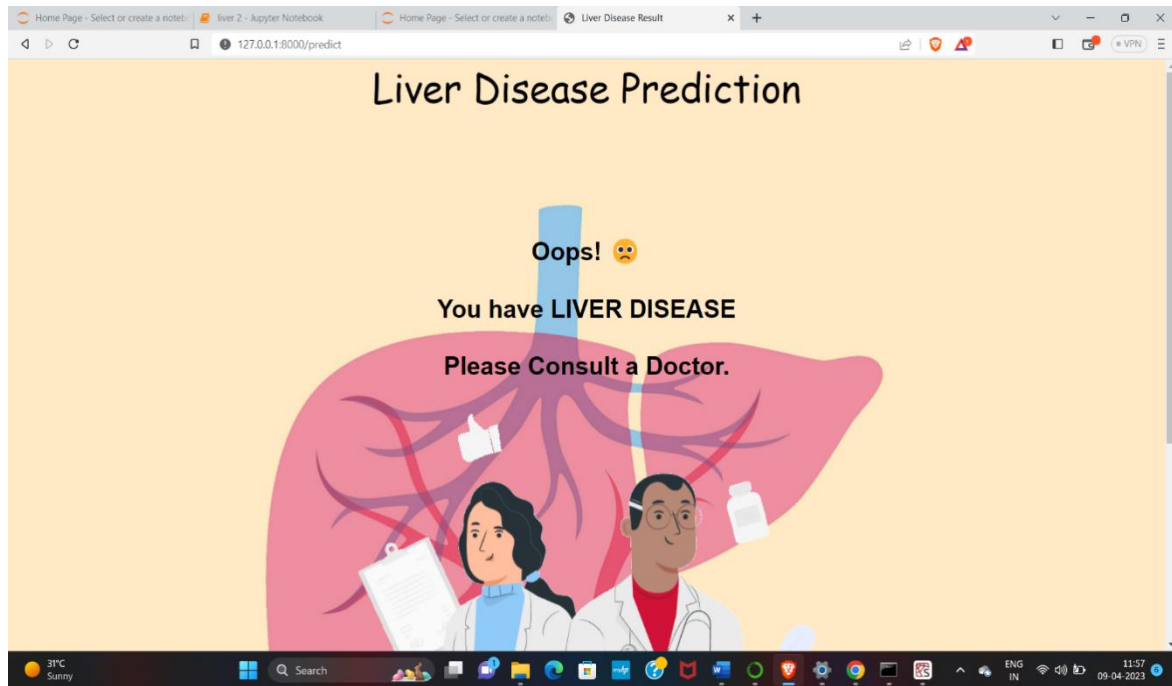


Fig 5.5

The proposed method was able to classify 76% of the input instances correctly without random under-sampling. To evaluate the overall discriminative capability of multivariate classifier in Canadian healthcare data without random under-sampling different performance measure are used a tool. It exhibited a precision μ of 66%, recall μ of 73%, F-measure μ of 67%, and AROC 73% on average, showing a fairly significant discriminative capability. The results for all the cases show that MCC range from 0.055 to 0.328.

Discussion

As mentioned earlier NAFLD is associated with metabolic disturbances and both are bi-directionally associated. It is a very complex clinical condition with different etiology involving a multitude of physiological mechanisms and symptoms. Selecting potentially relevant data is crucial for building an efficient model from EMRs. Therefore, the major clinical factors considered in the ATP III clinical criteria for MetS are incorporated in the context of NAFLD as a basis for early stage screening of individuals at risk for developing NAFLD. It is a very complex clinical condition with different etiology involving a multitude of physiological mechanisms and symptoms. Selecting potentially relevant data is

crucial for building an efficient model from EMRs. Therefore, the major clinical factors considered in the ATP III clinical criteria for MetS are incorporated in the context of NAFLD as a basis for early stage screening of individuals at risk for developing NAFLD. Diabetes mellitus, NAFLD and metabolic syndrome frequently co-exist as they potentially share common risk factors of, imbalanced triglycerides and insulin resistance. We have taken quantitative scores defined by the Italian Association for the Study of the Liver (IASF) depicted in Table along with a guideline explicitly defined for the Canadian population based on triglyceride (TRG) level to evaluate the impact of metabolic factors on NAFLD risk; defined in equations (), (), () and (). Tomizawa *et al.* performed multivariate regression analysis to evaluate the efficiency of various risk factors in the prediction of NAFLD. It is a very complex clinical condition with different etiology involving a multitude of physiological mechanisms and symptoms. Selecting potentially relevant data is crucial for building an efficient model from these factors include TRG, HDL, low-density lipoprotein cholesterol (LDL), blood glucose (BG) and hemoglobin A1c (HbA1c). Experimental results demonstrate that TRG was the parameter most significantly associated with NAFLD ($\chi^2=9.89$, $P=0.0017$) and also highlight that TRG is an elevated marker of NAFLD. A recent study also revealed that prevalence of NAFLD in individuals without metabolic syndrome was in this research we have taken quantitative scores defined by IASF along with a guideline explicitly defined for the Canadian population based on triglyceride (TRG) level. These defined levels are used as the reference value for determining of NAFLD risk. This was the first step in the development of NAFLD risk prediction model.

Early stage prediction of risk for developing NAFLD is not sufficient; a physician or decision maker may also want to know the causes for developing NAFLD risk. DT is one of the machine learning techniques possess ability to communicate the results in a simple self-explanatory symbolic and visual format with satisfactory accuracy levels in various domains. It incorporates multiple predictors in a simple step by step manner, whose semantics are intuitively clear and easy to interpret for experts, as they can see the structure of decisions in the classifying process.

Hence, we evaluated J48 decision tree algorithm to identify contributing factors in the onset of NAFLD as facts are crucial to comprehend the prognosis. The most promising attribute with maximum information gain in our case HDL is selected as root. The root node is evaluated first when assessing NAFLD risk in an individual.

CHAPTER - 6

CONCLUSION

6.1 CONCLUSION

This paper gives research of multiple researches done in this field. Our Proposed System aims at bridging gap between Doctors and Patients which will help both classes of users in achieving their goals. This system provides support for multiple disease prediction using different Machine Learning algorithms. The present approach of many systems focuses only on automating this process which lacks in building the user's trust in the system. By providing Doctor's recommendation in our system, we ensure user's trust side by side ensuring that the Doctor's will not feel that their Business is getting affected due to this System st, Kn Neighbor, Support vector classifier. Based on algorithm applied, it is observe. In this research we predicted the risk of developing NAFLD in an individual by incorporating noninvasive markers and gold standard machine learning method. Te rationale behind our approach is divided in two parts: firstly relevant risk factors selection using ATP III clinical criteria proposed in 2005 for MetS and allocation of class label with respect to triglycerides (TRG) level along with qualitative scoring criteria proposed by IASF for extracting knowledge from the input data and evaluating the NAFLD risk in an individual. Secondly rule based reasoning and visualization of predictive results that can be employed in better understanding of the phenomena involving a multitude of physiological mechanisms and symptoms. Te results demonstrate that the proposed technique is suitable with optimal discrimination for the assessment of NAFLD risk, understanding the contributing factors, producing accurate, specfic and decision oriented rules to facilitate physician and make informed choices about their management and improve health condition. Tis can be extended to predict other type of ailments which arise from metabolic syndrome.

The prediction of liver illness in patients has been examined and analysed in this project. By using various techniques the data has been cleaned by imputation of missing values with median, various classification algorithm have been applied such as Logistic regression, Random Forest, KNNeighbor, Support vector classifier. Based on algorithm applied, it is observed that models Support

vector classifier gives better accuracy than the other classification algorithm. Therefore concluding that SVC algorithm is appropriate for predicting liver disease.

6.2 FUTURE WORK

It is observed that some approaches do not show efficient outcomes when used on a large dataset and classifies a sample of datasets with greater accuracy than the remaining dataset. Different methods have a varied range of efficiency and sensitivity. It is also seen that specific approaches do not show the same level of accuracy when applied to real-time data. Collecting large-scale medical data for research and selecting the most significant features are the challenging aspects of using ML concepts for liver disease detection. These problems can be solved by carrying out large scale studies on multiple centres with emphasizing the collection, segmentation, and pre-processing of data..

The experiments were run with following settings: The confidence factor that represents a threshold value of allowed inherent error in data (whether an attribute is inside the confidence interval of the assigned class) while pruning the decision is set to 0.5 along with Subtree raising pruning. The minimum number of instances at a single leaf node for which confidence interval is computed was set to 20 in order to obtain simpler and smaller decision trees. Binary split is set to false basically this selection criteria control the visual outlook of the tree. The developed decision tree is shown. The level of effectiveness of the classification methods can be distinguished with number of correct and incorrect classifications in each possible case of the variables being classified. Given a particular classification algorithm and a data instance, there exist four possibilities. If the instance is positive and it is classified as positive, it is counted as a true positive (equation 8); if it is classified as negative, it is counted as a false negative. If the instance is negative and it is classified as negative, it is counted as a true negative; if it is classified as positive, it is counted as a false positive

REFERENCES

- [1] Singh et al., "Heart Disease Prediction Using Machine Learning Algorithms", *2020 International Conference on Electrical and Electronics Engineering (ICE3)*, pp. 452-457, February 2020.
- [2] A Narin, C Kaya and Z. Pamuk, "Automatic detection of coronavirus disease (COVID-19) using x-ray images and deep convolutional neural networks", Mar 2020.
- [3] Rajesh. Ranjan, "Predictions for COVID-19 outbreak in India using Epidemiological models", 2020.
- [4] Mohan Senthilkumar, ChandrasegarThirumalai and Gautam Srivastava, "Effective heart disease prediction using hybrid machine learning techniques", *IEEE Access*, vol. 7, pp. 81542-81554, 2019.
- [5] Mohan Senthilkumar, ChandrasegarThirumalai and Gautam Srivastava, "Effective heart disease prediction using hybrid machine learning techniques", *IEEE Access*, vol. 7, pp. 81542-81554, 2019.
- [6] M. Bayati, S. Bhaskar and A. Montanari, "Statistical analysis of a low cost method for multiple disease prediction", *Statistical Methods Med. Res.*, vol. 27, no. 8, pp. 2312-2328, 2018.
- [7] A. K. Shrivastava and S. Kumar Sahu, "Classification of Chronic Kidney Disease using Feature Selection Techniques", *IJCSE*, vol. 6, no. 5, pp. 649-653, 2018.
- [8] N Chaithra and B Madhu, "Classification Models on Cardiovascular Disease Prediction using Data Mining Techniques", *Journal of Cardiovascular Diseases & Diagnosis*, vol. 6, pp. 1-4, 2018.
- [9] M. Sameer and B. Gupta, "Beta Band as a Biomarker for Classification between Interictal and Ictal States of Epileptical Patients," in 2020 7th International Conference on Signal Processing and Integrated Networks

- (SPIN), 2020, pp. 567–570, doi: 10.1109/SPIN48934.2020.9071343.
- [10] S. K. B. Sangeetha, N. Afreen, and G. Ahmad, “A Combined Image Segmentation and Classification Approach for COVID-19 Infected Lungs,” J. homepage <http://iicta.org/journals/rcs>, vol. 8, no. 3, pp. 71–76, 2021.
- [11] M. Sameer, A. K. Gupta, C. Chakraborty, and B. Gupta, “Epileptical Seizure Detection: Performance analysis of gamma band in EEG signal Using Short-Time Fourier Transform,” in 2019 22nd International Symposium on Wireless Personal Multimedia Communications (WPMC), 2019, pp. 1–6, doi: 10.1109/WPMC48795.2019.9096119.
- [12] A. Mahajan, K. Somaraj, and M. Sameer, “Adopting Artificial Intelligence Powered ConvNet To Detect Epileptic Seizures,” in 2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES), 2021, pp. 427–432, doi:
- [13] N. Nasir, N. Afreen, R. Patel, S. Kaur, and M. Sameer, “A Transfer Learning Approach for Diabetic Retinopathy and Diabetic Macular Edema Severity Grading,” *Rev. d’Intelligence Artif.*, vol. 35, pp. 497–502, Dec. 2021, doi: 10.18280/ria.35060
- [14] Fang, R., Pouyanfar, S., Yang, Y., Chen, S. C. & Iyengar, S. S. Computational health informatics in the big data age: a survey. *ACM Comput. Surv. (CSUR)*, **49** (2016).
- [15] Agrawal, A. & Choudhary, A. Perspective: materials informatics and big data: realization of the “fourth paradigm” of science in materials science. *APL Mater.* **4**, 053208 (2016).
- [16] Miljkovic, D. *et al.* Machine Learning and Data Mining Methods for Managing Parkinson’s Disease. *Mach. Learn. Health. Inf.*, 209–220 (2016).
- [17] Zemmal, N., Azizi, N., Dey, N. & Sellami, M. Adaptive semi supervised support vector machine semi supervised learning with features cooperation for breast cancer classification. *J. of Med. Imaging Health. Inf.* **6**, 53–62 (2016).

- [18] Li, Y. *et al.* Classification of Parkinson's Disease by Decision Tree Based Instance Selection and Ensemble Learning Algorithms. *J. Med. Imaging Health. Inf.* **7**, 444–452 (2017).
- [19] Wang, Y., Coiera, E., Runciman, W. & Magrabi, F. Using multiclass classification to automate the identification of patient safety incident reports by type and severity. *BMC Med. Inf. Decis. Making*, **17** (2017).
- [20] Armañanzas, R., Iglesias, M., Morales, D. A. & Alonso-Nanclares, L. Voxel-Based Diagnosis of Alzheimer's Disease Using Classifier Ensembles. *IEEE j. of biomed. Health. inf.* **21**, 778–784 (2017).
- [21] Garg, R., Dong, S., Shah, S. & Jonnalagadda, S. R. A bootstrap machine learning approach to identify rare disease patients from electronic health records. *arXiv preprint arXiv:1609.01586*, (2016).

APPENDIX

SOURCE CODE

○ PYTHON CODE USED IN JUPYTER NOTEBOOK

```
#import libraries

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
import os

warnings.filterwarnings('ignore')


# Importing the Dataset
patients_data = pd.read_csv('indian_liver_patient.csv')


# Analysing the Data
Patients_data.head()

Patients_data.shape()

patients_data.info()


# Exploratory data analysis
patients_data['Age'].hist(bins=40, color='midnightblue')
plt.title("Distribution of Age among patients")
plt.xlabel("Age");

#distribution of gender in patients in dataset
patients_data['Gender'].value_counts().plot.bar(color='midnightblue')
plt.title("Distribution of Gender");
```

```

# changing the column name
patients_data.rename(columns={'Dataset': 'Diagnosis', 'Total_Protiens':
'Total_Proteins'}, inplace=True)

# changing labels
patients_data['Diagnosis'] = patients_data['Diagnosis'].apply(lambda x:1 if
x==1 else 0)

#let us check the countplot of our diagnosis column
patients_data['Diagnosis'].value_counts().plot.bar(color='midnightblue')
plt.title('Diagnosis of patients (1 - Liver disease | 0 - No Liver disease');

# Checking the null values
patients_data.isnull().any()
patients_data.isnull().sum()
# filling null values
mean_ratio = patients_data['Albumin_and_Globulin_Ratio'].mean()
mean_ratio
patients_data = patients_data.fillna(mean_ratio)
#we have now filled the null values
patients_data.isnull().sum()
data.isnull().sum()

# pair plot of patients data based on gender
plt.rcParams['figure.figsize']=(10,10)
sns.pairplot(patients_data,hue='Gender')

plt.rcParams['figure.figsize']=(10,10)
sns.pairplot(patients_data,hue='Diagnosis')

```

```

#GENDER BASED COMPARISON OF DIFFERENT METRICS
plt.figure(figsize=(8,6))
patients_data.groupby('Gender').sum()["Total_Proteins"].plot.bar(color='coral')
plt.title('Total proteins sum by gender');
plt.figure(figsize=(8,6))
patients_data.groupby('Gender').sum()['Albumin'].plot.bar(color='midnightblue')
plt.figure(figsize=(8,6))
patients_data.groupby('Gender').sum()['Albumin'].plot.bar(color='midnightblue')
#Bilirubin content for males and females as follows
plt.figure(figsize=(8,6))
patients_data.groupby('Gender').sum()['Total_Bilirubin'].plot.bar(color='midnightblue')
plt.title('Total bilirubin sum by gender');

#bilirubin content for males and females as follows
plt.figure(figsize=(8,6))
patients_data.groupby('Gender').sum()['Total_Bilirubin'].plot.bar(color='midnightblue')
plt.title('Total bilirubin sum by gender');
patients_data['Gender'] = patients_data['Gender'].apply(lambda x:1 if x=='Male' else 0)

patients_data.head()

# correlation between the features using a heatmap
corr=patients_data.corr()
plt.figure(figsize=(20,10))
sns.heatmap(corr,cmap="Greens",annot=True)

```

```

# Building the classification model
patients_data.columns
patients_data.drop("Diagnosis",axis=1).apply(lambda x:
x.corr(patients_data['Diagnosis']))
X = patients_data.drop('Diagnosis', axis=1)
y = patients_data['Diagnosis']
from sklearn.preprocessing import StandardScaler

# Scaling the feature columns
scaler = StandardScaler().fit(X)
X_scaled = scaler.transform(X)
#split data into train and test sets
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.3,random_state=21)
from sklearn.model_selection import train_test_split
# training a logistic regression classifier
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, confusion_matrix
logmodel = LogisticRegression()
logmodel.fit(X_train,y_train)
#calculating accuracy of model

from sklearn.metrics import accuracy_score
y_pred = logmodel.predict(X_test)
accuracy_score(y_test, y_pred)
from sklearn.metrics import classification_report
print(classification_report(y_test, y_pred))

#training a svc classifier
from sklearn.svm import SVC
svc = SVC(kernel='linear')
svc.fit(X_train, y_train)

```

```

y_pred = svc.predict(X_test)
accuracy_score(y_test, y_pred)
print(classification_report(y_test, y_pred))
#training a random forest classifier
from sklearn.ensemble import RandomForestClassifier
clf = RandomForestClassifier()
clf.fit(X_train, y_train)

y_pred = clf.predict(X_test)
accuracy_score(y_test, y_pred)
print(classification_report(y_test, y_pred))
#training a KNN CLASSIFIER
from sklearn.neighbors import KNeighborsClassifier
neigh = KNeighborsClassifier(n_neighbors=3)

neigh.fit(X_train, y_train)

y_pred = neigh.predict(X_test)
accuracy_score(y_test, y_pred)
print(classification_report(y_test, y_pred))
#INFERENCE
value = {
    "Age": 19,
    "Gender": 0,
    "Total_Bilirubin": 0.3,
    "Direct_Bilirubin": 0.2,
    "Alkaline_Phosphotase": 54,
    "Alanine_Aminotransferase": 15,
    "Aspartate_Aminotransferase": 26,
    "Total_Proteins": 5.1,
    "Albumin": 1.9,
    "Albumin_and_Globulin_Ratio": 1.0

```

```

    }
    data = pd.DataFrame([value])
    data
    svc.predict(data)[0]

#saving the model
pickle.dump(svc, open('liver _analysis.pkl','wb'))
pickle.dump(svc, open('liver _analysis.pkl','wb'))

```

PYTHON CODE USED FOR APP BUILDING

```

from flask import Flask, render_template, request
import numpy as np
import pickle
import pandas as pd

```

```

app = Flask(__name__)
model = pickle.load(open('liver _analysis.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'])
        Direct_Bilirubin = float(request.form['Direct_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'])

    Value =
np.array([[Age,Gender,Total_Bilirubin,Direct_Bilirubin,Alkaline_Phosphot
ase,Alamine_Aminotransferase,Aspartate_Aminotransferase,Total_Protie
ns,
Albumin,Albumin_and_Globulin_Ratio]])

    prediction = model.predict(value)[0]

    return render_template('result.html', prediction=prediction)

if __name__ == "__main__":
    app.run(threaded = True,debug=True,port="8000")

```

The screenshot shows the Spyder Python IDE interface. The main editor displays the code from the previous block. The console on the right shows the following output:

```

Python Console History
Type "copyright", "credits" or "license" for more information.
IPython 8.10.0 -- An enhanced Interactive Python.
In [1]: runfile('C:/Flask/app.py', wdir='C:/Flask')
* Serving Flask app 'app'
* Debug mode: on
C:\Users\gedel\anaconda\lib\site-packages\sklearn\base.py:299: UserWarning: Trying to unpickle
estimator SVC from version 1.0-2 when using version 1.2.1. This might lead to breaking code or invalid
results. Use at your own risk. For more info please refer to:
https://scikit-learn.org/stable/model_persistence.html#security-maintainability-limitations
warnings.warn(
Warning: This is a development server. Do not use it in a production deployment. Use a production WSGI
server instead.
* Running on http://127.0.0.1:8000
Press CTRL-C to quit
* Restarting with watchdog (windowsapi)

```

Let us build an app.py flask file which is a web framework written in python for server-side scripting. Let's see step by step procedure for building the backend application.

In order to develop web API with respect to our model, we basically use the Flask framework which is written in python.

We are importing necessary libraries like Flask to host our model request

Initialise the Flask application

Loading the model using pickle

Routes the API URL

Rendering the template. This helps to redirect to the home page. In this home page, we give our input and ask the model to predict

we are taking the inputs from the form

Feature Scaling the inputs

Predicting the values given by the user

if the output is false render no chance template If the output is True render chance template

The value of `__name__` is set to `__main__` when the module run as the main program otherwise it is set to the name of the module .

HTML CODES USED

- **Index.html**

```
<!DOCTYPE html>
```

```
<html lang="en">
```

```
<head>
```

```
    <meta charset="UTF-8">
```

```
    <title>Liver Prediction Model</title>
```

```

</head>
<body>
    <div class="container">
        <h2 class='container-heading'><span
class="heading_font">Liver Disease Prediction</span></h2>
    </div>

    <div class="ml-container">
        <form action="{{ url_for('predict') }}" method="POST">
            <br>
            <h3>Age</h3>
            <input id="first" name="Age" placeholder="in Year"
required="required">
            <br>
            <h3>Gender</h3>
            <input id="second" name="Gender" placeholder="Male =
1, Female=0" required="required">
            <br>
            <h3>Total Bilirubin</h3>
            <input id="third" name="Total_Bilirubin" placeholder="Total
Bilirubin" required="required">
            <br>
            <h3>Direct Bilirubin</h3>
            <input id="fourth" name="Direct_Bilirubin"
placeholder="Direct Bilirubin" required="required">
            <br>
            <h3>Alkaline Phosphotase</h3>
            <input id="fifth" name="Alkaline_Phosphotase"
placeholder="Alkaline Phosphotase" required="required">
            <br>
            <h3>Alamine Aminotransferase</h3>
            <input id="sixth" name="Alamine_Aminotransferase"

```

```

placeholder="Alamine Aminotransferase" required="required">
    <br>
    <h3>Aspartate Aminotransferase</h3>
    <input id="seventh" name="Aspartate_Aminotransferase"
placeholder="Aspartate Aminotransferase" required="required">
    <br>
    <h3>Total Protiens</h3>
    <input id="eight" name="Total_Protiens"
placeholder="Total Protiens" required="required">
    <br>
    <h3>Albumin</h3>
    <input id="ninth" name="Albumin" placeholder="Albumin"
required="required">
    <br>
    <h3>Albumin and Globulin Ratio</h3>
    <input id="tenth" name="Albumin_and_Globulin_Ratio"
placeholder="Albumin and Globulin Ratio" required="required">
    <br>
    <br>
    <br>
    <button id="sub" type="submit ">Submit</button>
    <br>
    <br>
    <br>
    <br>
    <p class='footer-description'>©2021 Sagar Dhandare</p>

</form>
</div>

```

```
<style>
```

```
/* Background Image */
```

```
body
```

```
{
```

```
background-
```

```
image:url("https://raw.githubusercontent.com/SagarDhandare/Liver-  
Disease-Prediction-Project/main/Images/Liver.jpg");
```

```
height: 100%;
```

```
/* Center and scale the image nicely */
```

```
background-position: center;
```

```
background-repeat: no-repeat;
```

```
background-size: 100% 100%;
```

```
}
```

```
/* Color */
```

```
body{
```

```
    font-family: Arial, Helvetica,sans-serif;
```

```
    text-align: center;
```

```
    margin: 0;
```

```
    padding: 0;
```

```
    width: 100%;
```

```
    height: 100%;
```

```
    display: flex;
```

```
    flex-direction: column;
```

```
}
```

```
/* Heading Font */
```

```
.container-heading{  
    margin: 0;  
}
```

```
.heading_font{  
    color: #black;  
    font-family: 'Pacifico', cursive;  
    font-size: 50px;  
    font-weight: normal;  
}
```

```
/* Box */  
    #first {  
        border-radius: 14px;  
        height: 25px;  
        width: 150px;  
        font-size: 20px;  
        text-align: center;  
    }  
  
    #second {  
        border-radius: 14px;  
        height: 25px;  
        width: 220px;  
        font-size: 20px;  
        text-align: center;  
    }  
  
    #third {  
        border-radius: 14px;
```

```
height: 25px;  
width: 180px;  
font-size: 20px;  
text-align: center;  
}
```

```
#fourth {  
border-radius: 14px;  
height: 25px;  
width: 250px;  
font-size: 20px;  
text-align: center;  
}
```

```
#fifth {  
border-radius: 14px;  
height: 25px;  
width: 270px;  
font-size: 20px;  
text-align: center;  
}
```

```
#sixth {  
border-radius: 14px;  
height: 25px;  
width: 280px;  
font-size: 20px;  
text-align: center;  
}
```

```
#seventh {  
border-radius: 14px;
```

```
        height: 25px;  
        width: 170px;  
        font-size: 20px;  
        text-align: center;  
    }
```

```
#eight {  
    border-radius: 14px;  
    height: 25px;  
    width: 150px;  
    font-size: 20px;  
    text-align: center;  
}
```

```
#ninth {  
    border-radius: 14px;  
    height: 25px;  
    width: 280px;  
    font-size: 20px;  
    text-align: center;  
}
```

```
#Tenth {  
    border-radius: 14px;  
    height: 25px;  
    width: 280px;  
    font-size: 20px;  
    text-align: center;  
}
```

/* Submit Button */

```
#sub {
```

```
        width: 120px;
        height: 43px;
        text-align: center;
        border-radius: 14px;
        font-size: 18px;
    }
```

```
<link rel="stylesheet" href="https://cdnjs.cloudflare.com/ajax/libs/font-
awesome/4.7.0/css/font-awesome.min.css">
```

```
</style>
```

```
</body>
```

```
</html>
```

. Output.html

```
<!DOCTYPE html>
```

```
<html lang="en">
```

```
<head>
```

```
    <meta charset="UTF-8">
```

```
    <meta name="viewport" content="width=device-width, initial-scale=1.0">
```

```

<title>Liver Disease Result</title>
</head>

<body>

    <div class="container">
        <form action="{{ url_for('predict')}}" method="post">
            <h2 class='container-heading'><span
class="heading_font">Liver Disease Prediction</span></h2>

            <br><br><br><br><br><br><br>

            <!-- Result -->
            <div class="results">
                {% if prediction==1 %}
                    <h1><span class='danger'>Oops! 😞<br><br>You
have LIVER DISEASE <br><br>Please Consult a Doctor.</span></h1>
                    
                {% elif prediction==0 %}
                    <h1><span class='safe'>🎉 Congratulation!
🎉<br><br>You DON'T have LIVER DISEASE.</span></h1>
                    
                {% endif %}
            </div>
        </form>

    </div>
</div>
    <br><br> <br><br><br><br><br><br>

```

<p class='footer-description'>©2023 NITEESH KUMAR</p>

</div>

<style>

/* Background Image */

body

{

background-

image:url("https://raw.githubusercontent.com/SagarDhandare/Liver-Disease-Prediction-Project/main/Images/Liver.jpg");

height: 100%;

/* Center and scale the image nicely */

background-position: center;

background-repeat: no-repeat;

background-size: 100% 100%;

}

/* Color */

body{

font-family: Arial, Helvetica,sans-serif;

text-align: center;

margin: 0;

padding: 0;

width: 100%;

height: 100%;

display: flex;

flex-direction: column;

```
}
```

```
/* Heading Font */
```

```
.container-heading{
```

```
    margin: 0;
```

```
}
```

```
.heading_font{
```

```
    color: #black;
```

```
    font-family: 'Pacifico', cursive;
```

```
    font-size: 50px;
```

```
    font-weight: normal;
```

```
}
```

A NEW MACHINE LEARNING-BASED COMPLEMENTARY APPROACH FOR SCREENING OF NAFLD (HEPATIC STEATOSIS)

BANDARU ADITYA

Department of electronics and
Communication Engineering
Sathyabama Institute of science and
Technology, Chennai-600119, India
adityabandaru2002@gmail.com

ARRA ROHITH

Department of electronics and
Communication Engineering
Sathyabama Institute of science and
Technology, Chennai-600119, India
Arrarohith304@gmail.com

ABSTRACT

With advances in information technology within the biomedical and health care groups, the correct take a look at of clinical facts drives early ailment detection, Community services and patient care. When medical records aren't full and of good quality, In addition, exclusive areas show particular manifestations of positive local illnesses, that may result in a weakening of the onion disease type. The proposed gadget gives facts processing techniques for the efficient type of various sickness cases in common sickness societies. Experiments with modified scoring models on actual sanatorium data amassed. To conquer the problems of incomplete statistics, the latent issue model is used to retrieve the missing facts. This is an experiment in continual nearby cerebral infarction disease. Using based and unstructured hospital data, he uses facts technology strategies. It predicts the maximum likely sicknesses via the analysis of the dataset inclusive of malaria, breast most cancers, diabetes, liver ailment, kidney ailment, and heart disorder. To our information, in the discipline of huge statistics analytics, no existing paintings focuses on each varieties of records. Compared with numerous ordinary scoring algorithms, the calculation accuracy of our proposed set of rules reaches ninety four.Eight%, and the convergence charge is higher than the expertise-primarily based disorder hazard type algorithm.

Keywords-sensitivity, Liver disease, Biological system modeling, sociology, Data processing, Physiology, Statistics

INTRODUCTION

The Earth is experiencing a technological crimson blip, after which the needs for intelligence and precision get up. Today, humans are probably hooked on the Internet, but they don't worry about their physical fitness.

People ignore a small hassle and do now not go to the health facility, which in the end leads to a serious contamination. Using this developingera, the

principle goal is to increase a machine that predicts more than one diseases based totally on patients' symptoms without visiting hospitals.

/ Doctors.

Machine getting to know is a subset of AI that is specially concerned with the observe of algorithms which might be stepped forward via information and revel in. The getting to know technique includes two tiers: training and checking out. Machine gaining knowledge of offers an effective platform in the medical field to solve diverse health problems much quicker.

There are styles of gadget mastering, predictive studying and unsupervised gaining knowledge of. In careful research we form a model with nicely-classified facts. On the other hand, an embedded studying version learns from disjointstatistics.

The goal is to derive an powerful machine gaining knowledge of set of rules, this is, greenand accurate for disease prediction. In this text, the idea of gadget studying control is used to expect diseases.

The primary characteristic may be machine learning, in which we will use a system learning

set of rules in an effort to be able to correctly expect diseases at an early level and improve affected person care.

IDENTIFYING THE PROBLEM

A clever healthcare prediction device is good at reducing healthcare prices. There are several functions that continue to be intact inside the health prediction gadget. So residing at the reducing fringe of technology, and but if we can't use it successfully and well, it's little need. To clear up this trouble, fitness studies is performed.

OBJECTIVES

Aim of the Project:

Accurate analysis is decreased whilst the fine of scientific statistics is incomplete.

Additionally, distinct places show distinct traits of some local diseases, which could make it harder to predict illness outbreaks.

However, the ones current works mainly handled information systems. There are not any right strategies for managing semi-structured and unstructured facts.

The proposed device will account for each based and unstructured statistics. The accuracy of the analysis is multiplied by way of using machine learning algorithms.

Scope of the Project

The objective of this effort is to provide a tool that will help consumers and experts locate and select a disorder.

To do this, we are creating a method that enables a person to search for a disease satisfying a set of criteria based on the characteristics of the disorder, including the symptoms of the illness, as well as paying attention to the affected individual's profile.

LITERATURE SURVEY

Personalized and economical Alzheimer's disease screening using machine learning

The analysis of Alzheimer's ailment is often difficult, particularly in the earliest stages of the condition when there is only a small cognitive deficit. However, in this case, treatment is most likely to be successful because a thorough diagnostic process may result in excellent benefits. A system learning technique for customised and cost-effective analysis has been developed and put on display. It determines the set of biotones that will be the most useful or informative for diagnosing patients using domestically weighted learning to customise a classifier model for each patient. We compare AD to controls and MCI patients who progressed forward to AD within a year to those who did not using ADNI information. The method is similar to taking into account all of the signs at once while also significantly reducing the number (and cost) of biomarkers required to produce a conclusive analysis for each patient. As a result, it can aid in the targeted and effective identification of AD and be helpful in a therapeutic setting.

Effect of Meteorological Conditions on Occurrence of Hand, Foot and Mouth Disease in Wuwei City, Northwestern China

This article's main goal is to provide a methodical framework for the prevention and forecasting of the superiority of cystic, foot, and mouth disorder while also taking into account the influence of various meteorological conditions on the occurrence of Wuwei cystic, foot, and mouth disorder. The dance of China is the city. Here, data on illnesses and climatic conditions for the years 2008 to 2010 were gathered, and correlation analysis, multiple regression analysis, and exponential curve fitting were carrier out

respectively 100,000 persons. This condition affected Wuwei City throughout the year, although it peaked significantly during the winter. In the world of fashion, the incidents started to rise in April and peaked in May, June, and July, respectively. Every year, in September or October, there is another top. The epidemic disease in the four places was impacted differently by many meteorological parameters, including the average temperature, relative humidity, atmospheric pressure, precipitation, and evaporative capacity. The outcome of multiple linear regressions indicates that relative humidity and atmospheric strain are the main factors in Liangzhou, whereas average temperature and atmospheric pressure are dominant in Gulang County and Tianzhu County, respectively. The frequency and typical wide range of daylight hours in Minqin County demonstrate an exponential growth. Climate factors thus have an impact on FMD's advantage in unique techniques. There is a close relationship between weather conditions and the prevalence of hand, foot, and mouth disease in Wuwei town's 4 districts. The seasons of summer and fall were both crucial for the prevention and management of illness.

Developing an Index for Detection and Identification of Disease Stages

It is frequently used to assess the severity of diseases in various flora using spectral recordings. Such data, however, stopped being utilised to predict the phases of plant disease. This investigation aims to raise a spectral illness index to identify the severity of wheat leaf rust sickness at several DS ranges. A spectroradiometer was used to measure the reflectance spectra of inflamed leaves with unusual fracture symptoms and levels of DS in order to achieve the study's objective. Following the evolution of pure spectra of diverse illness symptoms at the leaf scale, a new function was created to determine the equality of sensitive proportions of disease symptoms. It has been established that the reflection spectra with the best sensitivity are at 675 and 775nm. Finally, a new SDI was developed to distinguish between the three degrees of illness at the crown stage using the normalised difference in DS and the ratio of 675 to 775. Precision plant protection outcomes for detecting station diseases were to be improved by the SDI software.

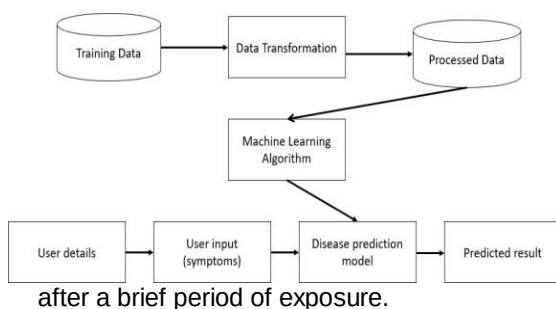
Quantized Analysis for Heart Valve Disease based on Cardiac Sound Characteristic Waveform Method

The four clinical heart valve sound analysis, specifically the sound of the coronary heart wave, is a novel quantitative diagnostic technique that is suggested for accurately and effectively examining coronary heart valve disease. The sign was recorded using a BIOPAC facts collecting device.

An ethernet network is used to send memory data to a computer network for storage, analysis, and real-time display. To extract the characteristic, a one-degree-of-freedom analytical model is built. Additionally, diagnostic criteria were computed to distinguish between common cardiac sounds and valves using an easy-to-understand visual depiction, allowing even a novice user to easily track the progression of their disease. Finally, a case study of a patient with heart valve disease before and after surgical therapy was validated, demonstrating the value and efficacy of the suggested strategy.

Non-Linear Analysis of Heart Rate Variability in Patients with Coronary Heart Disease

The non-linear coronary heart rate variability that was employed in a group of patients with coronary heart disease and a group of healthy controls of the same age was highlighted in the paper for its clinical and prognostic significance. Hurst exponent, wavelet analysis with DE trend, and approximative entropy were all used. The R-R series' Hurst exponent was determined through the rescaling evaluation technique. The fractal long-term correlation properties of coronary heart rate variability were quantified using DFA. The variation of a time series is quantified by approximate entropy. A fractal zoom was identified



EXISTING SYSTEM

This examine proposes a non-invasive approach for diagnosing liver illnesses using ultrasound images of liver tissue including ordinary, steatosis, or cirrhosis using feature extraction, feature choice, and class. First, correlations, homogeneity, tensions, entropy, contrast, electricity, lengthy-term importance, percentage crowning glory, trendy deviation are determined. Secondly, the simplest manufacturers are selected according to the standards and manual strategies of choosing the fisherman. Third, three subclassifiers were used, namely the ordinary/steatosis, normal/cirrhosis, and steatosis/cirrhosis classifiers. The very last class of liver tissue is primarily based on most important feature.

PROPOSED SYSTEM

The proposed gadget of classification of more than one illnesses using facts technology is that algorithms and all other exclusive tools to create a

machine that indicates the affected person's disease the usage of signs and symptoms and taking those signs and symptoms we compare with the gadget dataset that changed into formerly available. By taking those datasets and evaluating them to the affected person's disorder, we will inform the percentage of the patient's disorder. The dataset and records are exceeded to the type machine model, in which the information is pre- processed for future records, and then the consumer makes a desire of features wherein he's going to input/pick diverse symbols. The partitioning of this records is then executed the use of statistics technological know-how methods which includes logistic regression. The records then goes into the writer's version in which he shows the chance evaluation that is involved inside the system and additionally provides a possibility estimate to reveal the numerous probabilities, together with how the machine behaves with the variety n. Type of They also make suggestions to sufferers who're doing their quit, and also their indicators, to show what to apply and what now not to apply from the given schedules and final effects.

Probable illnesses thru the evaluation of datasets inclusive of malaria, breast cancer, diabetes, liver disorder, kidney ailment, and coronary heart disorder. To our knowledge, within the field of massive statistics analytics, no existing paintings specializes in each sorts of data.

ADVANTAGES OF PROPOSED SYSTEM

- Easily diagnose the sickness
- High accuracy

ARCHITECTURE DIAGRAM

SYSTEM REQUIREMENTS HARDWARE REQUIREMENTS

- Processor : Core i3/i5/i7
- RAM : 2-4GB
- HDD : 500 GB

SOFTWARE REQUIREMENTS

- Platform: Windows Xp/7/8/10
- Coding Language : Python

SOFTWARE ENVIRONMENT

Python:

Python is an extreme object-oriented, interpretive, interactive, and literal language. Python was intended to be easy to read. It has fewer syntactic structures than other languages and frequently employs English key phrases at the same time as other languages use punctuation.

Interpreted Python At runtime, Python is processed by an interpreter. This system doesn't require any configuration prior to use. The same may be said for PHP and PERL.

Python is interactive; you could sit at the Python prompt and immediately start writing your apps using the interpreter.

Python supports a prepared style of programming or a style of programming that wraps code in items.

Python is a beginner-friendly language. Python is a great language for beginning programmers since it enables the creation of many different kinds of programs, from simple phrase processing to web browsers and games.

Features of Python

Features of Python consist of –

Simple to study Python is a well-defined syntax language with a small number of important terms. This enables the pupil to quickly pick up the language.

Python code is more clearly defined and is therefore simpler to comprehend.

Python source code is easy to maintain and easy to secure.

Wide Standard Library - The vast majority of the Python library is cross-platform and highly portable, compatible with UNIX, Windows, and Macintosh.

Python offers an interactive mode that enables you to inspect and troubleshoot code snippets live.

Python that is portable can operate on a variety of hardware platforms and uses the same user interface across all of them.

The Python interpreter is extensible; you may add low-stage modules to it. These modules provide programmers the ability to upgrade or customise their hardware for improved performance.

Databases Python provides a user interface for all important commercial databases.

Python's support for various system calls, libraries, and Windows structures, such as Windows MFC, Macintosh, and the X Windows System on Unix, makes it easier to design and port GUI programmes.

Scalability - Compared to shell scripts, Python provides larger packages with better structure and direction.

MODULES

Data collection module preparing the data module
Training a model

Disease prediction module

MODULE DESCRIPTIONS

Data Collection Module

Whether it is raw facts from Excel, Access, textual content files, and many others., this step (gathering beyond statistics) paperwork the foundation for destiny studyingThe expectations for device learning increase with increasing range, density, and breadth

of pertinent data.

Preparing the data Module

Every analytical method is dependent on the accuracy of the data employed. Spend some time figuring out how the numbers stand out, and then move to fix concerns like handling difficulties and missing statistics. Perhaps one of the best ways to specifically uncover the nuances of facts and so raise the nutritional content is through exploratory evaluation.

Training a model

In this stage, the optimum set of rules is chosen, and the information version is represented. The cleaned data is split into two sections: training and testing (the ratio depends on the circumstances); the first section, consisting of school records, is used to raise the version. For reference, utilise the second element.

Disease prediction Module

The patient names disease-related symptoms. The computer will ask a specific question about the patient's problem and predict the disease based on the information provided by the patient.

CONCLUSION

This article provides an overview of numerous research conducted on this region. The proposed system ambitions to be an middlemanamong doctors and sufferers, so one can assist both forms of customers reap their goals. This machine provides help in predicting multiple illnesses the usage of various machine gaining knowledge of algorithms. The current technique of maximum structures focuses handiest on automating this system, which does not anything to build consider in the person device. By imparting doctor pointers in our gadget, we ensure that user self belief is cozy so that physicians do no longer experience that their business is harmed via this machine.

We have used the logistic regression, Knn, SVC,random forest for the model evaluation

| | |
|---------------------------|-----------------------|
| Random Forest | -> |
| 0.7542857142857143 | |
| Knn Classifier | -> 0.7028571428571128 |
| Logistic Regression | -> 0.7657142857124857 |
| Support Vector Classifier | -> |
| 0.7771428571428571 | |

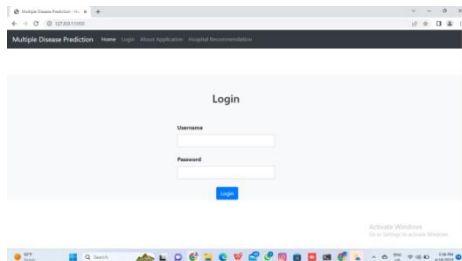
we have applied different type of algorithms for given data set to know which algorithm fits to our dataset

The above values are the respective accuracy values of Random Forest, Knn classifier, Logistic Regression, Support Vector Classifier models.

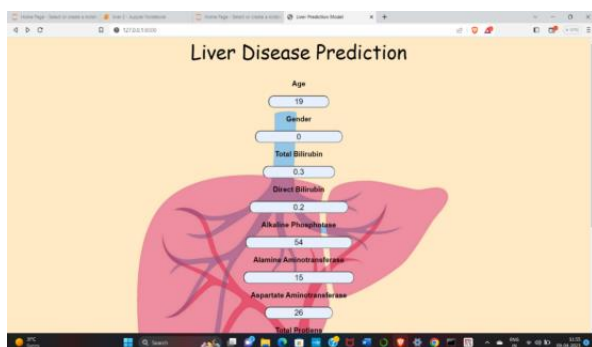
we have to take the best model having high accuracy

in the above models.

Accuracy value for SVC is better, When compared with other models, so saving the SVC and deployed.

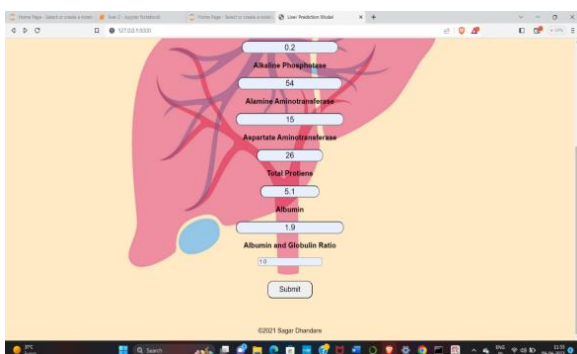


Login URL



Details

Welcome to the Home page of the application. This is mainly based on at the application of the machine learning meant to be employed in the remote areas



Overview

In todays time we see a lot of the shortage of the doctors in the world especially in Indie A lot of people are suffering a lot without the help of the proper medical checkup Also most of the cases many cases arise leading to death due to lack of timely medical checkup. So to scope up with all of those problems this app is designed which would prove its benefits upto much extent



Application

in todays time we see a lot of the shortage of the doctors in the world especially in Indie A lot of people are suffering a lot without the help of the proper medical checkup Also most of the cases many cases arise leading to death due to lack of timely medical checkup. So to cope up with all of those problems this app is designed which would prove its benefits upto much extent



REFERENCES

- Singh et al., "Heart Disease Prediction Using Machine Learning Algorithms", *2020 International Conference on Electrical and Electronics Engineering (ICE3)*, pp. 452-457, February 2020.
- A Narin, C Kaya and Z. Pamuk, "Automatic detection of coronavirus disease (COVID-19) using x-ray images and deep convolutional neural networks", Mar 2020.
- Rajesh. Ranjan, "Predictions for COVID-19 outbreak in India using Epidemiological models", 2020.

Mohan Senthilkumar, Chandrasegar Thirumalai and Gautam Srivastava, "Effective heart disease prediction using hybrid machine learning techniques", *IEEE Access*, vol. 7, pp. 81542-81554, 2019.

Mohan Senthilkumar, Chandrasegar Thirumalai and Gautam Srivastava, "Effective heart disease prediction using hybrid machine learning techniques", *IEEE Access*, vol. 7, pp. 81542-81554, 2019.

M. Bayati, S. Bhaskar and A. Montanari, "Statistical

analysis of a low cost method for multiple disease prediction", *Statistical Methods Med. Res.*, vol. 27, no. 8, pp.2312-2328, 2018.

A. K. Shrivastava and S. Kumar Sahu, "Classification of Chronic Kidney Disease using Feature Selection Techniques", *IJCSE*, vol. 6, no. 5, pp. 649-653, 2018.

N Chaithra and B Madhu, "Classification Models on Cardiovascular Disease Prediction using Data Mining Techniques", *Journal of Cardiovascular Diseases & Diagnosis*, vol. 6, pp. 1-4, 2018.

