PROJECT REPORT

STATISTICAL MACHINE LEARNING APPROACHES TO LIVER DISEASE PREDICTION

Done by :

- SNEGA S
- SURUTHI LAKSHMI E
- KEERTHANA L
- SHARUMATHI C V

:

ABSTRACT

The improvement of patient care, research, and policy is significantly impacted by medical diagnoses. Medical practitioners employ a variety of pathological techniques to make diagnoses based on medical records and the conditions of the patients. Disease identification has been significantly enhanced by the application of artificial intelligence and machine learning in conjunction with clinical data. Datadriven, machine learning (ML) techniques can be used to test current approaches and support researchers in potentially innovative judgments

1.INTRODUCTION

1.1.Project Overview

Liver diseases avert the normal function of the liver. Mainly due to the large amount of alcohol consumption liver disease arises. Early prediction of liver disease using classification algorithms is an efficacious task that can help the doctors to diagnose the disease within a short duration of time. Discovering the existence of liver

disease at an early stage is a complex task for the doctors. The main objective of this project is to analyze the parameters of various classification algorithms and compare their predictive accuracies so as to find out the best classifier for determining the liver disease.

1.2.Purpose

This Project examines data from liver patients concentrating on relationships between a key list of liver enzymes, proteins, age and gender using them to try and predict the likeliness of liver disease. Here we are building a model by applying various machine learning algorithms find the best accurate model. And integrate to flask based web application. User can predict the disease by entering parameters in the web application.

2.LITERATURE REVIEW

Bendi et al. [1] authors used two different input dataset and evaluate that the AP datasets has better than UCLA dataset for all the different selected algorithms. Based on performance on their classification KNN, Backward propagation and SVM are giving better results. The AP data set is better than UCLA for the entire selected algorithm. And found out Naïve Bayes, C4.5, KNN, Backward propagation and SVM has 95.07, 96.27, 96.93, 97.47, & 97.07% accuracy respectively.

• Bendi et al. [2] proposed a paper based on Modified Rotation Forest, used two dataset as an input UCI liver dataset and Indian liver dataset. And results show that MLP algorithm with random subset gives better accuracy of 94.78% for UCI dataset than CFS achieved accuracy of 73.07% for Indian liver dataset. • Yugal Kuma & G. Sahoo [3] proposed a paper based on different classification technique and used north east area of Andhra Pradesh (India) liver dataset. And the results shows that Decision tree(DT) algorithm has better than other algorithm and provide accuracy of 98.46%.

• S.Dhamodharan [4] proposed a paper based on two classification technique naïve Bayes and FT tree and used WEKA (Waikato Environment for Knowledge and Analysis) dataset. Naïve Bayes is 75.54% accuracy and FT Tree is 72.6624% accuracy and concluded Naïve Bayes gas better algorithm compare to other algorithms.

• Han Ma et al. [9] in this paper 11 different classification are evaluated and Demonstrated in China Zhejiang University, College of medicine and concluded Bayesian network accuracy of 83%, specificity 83%, sensitivity of 0.878 and F-measure of 0.655.

• Heba Ayeldeen et al. [5] propose a paper for prediction of liver fibrosis stages using decision tree technique and used Cario university data set and result shows that decision tree classifier accuracy is 93.7%.

• D.Sindhuja& R. JeminaPriyadarsini [6] survey a paper for classification of liver disease. In this survey different classification techniques of data mining are study and used dataset of dataset of AP liver has better than Dataset of UCLA, and concluded C4.5 achieved better results than other algorithms.

Somaya Hashem et al. [8] presented a paper for diagnosis of liver disease. In this paper they used two algorithms, SVM & Backpropagation and used UCI machine repository dataset. And concluded SVM has accuracy 71% better result than Backpropagation accuracy 73.2%.

• Joel Jacob et al. [10] proposed a paper to diagnosis of liver disease by using three different algorithms, Logistic regression, K-NN, SVM, and ANN and used Indian Liver Patient Dataset comprised of 10 different attributes of 583 patients. And concluded Logistic regression, KNN, SVM,& ANN has 73.23, 72.05, 75.04 & 92.8% accuracy respectively.

• Sivakumar D et al. [11] proposed a paper for prediction of chronic liver disease by using two different techniques K-means and C4.5. UCI repository.

• Mehtaj Banu H [12] in this paper authors study different machine learning technique, Supervised, unsupervised & reinforcement and also analysis UCI dataset database and concluded that KNN and SVM improved better performance and exactness of liver disease prediction.

• VasanDurai et al. [13] proposed a paper based on liver disease prediction by using three different techniques, SVM, NB & J48 using UCI repository dataset and concluded that J48 algorithm has better performance in terms of Feature selection and has accuracy of 95.04%.

Sl no	Authors	Year	Disease	Machine learning algorithm	Datase tinput	Remarks	Conclusion

1	BendiVe nkataRa mana et al.[1]	2011	Liverdisease	Naïve Bayes,C4.5, Backwardpropagation, KNNand SVM	AP liver dataset and UCLAliverdataset	Naïve Bayes, C4.5KNN, Backwardpropagation and SVMhas95.07, 96.27, 96.93,97.47, & 97.07% accuracyrespectively	KNN, Backwardpropagation and SVMare giving more betterresults. AP data set arebetterthanUCLAfor alltheselectedalgorithm
2	BendiVe nkataRa mana andM.SurendraPrasad Babu[2]	2012	Liverdisease	Modified Rotation Forest	UCI liver dataset andIndiandataset	MLP algorithm withrandom subset givesbetter accuracy 74.78%than NN with CFS ofaccuracy73.07%	MLP algorithm withUCIliver dataset hasbetter accuracy thanNN with Indian liverdataset

3	YugalK UMA&G . Sahoo[3]	2013	Liverdisease	DT, SVM, NB andANN	north east area of AndhraPradesh (India) liverdataset	Decision tree(DT) hasbetter accuracy of98.46%	Rule basedclassification with DTalgorithm has betteraccuracy
4	S.Dhamodharan [4]	2014	Liver cancer,Cirrhosis andHepatitis	Naïve-Bayes, FTTree	WEKA (WaikatoEnvironment forKnowledgeand Analysis) dataset	Naïve Bayes is 75.54%accuracy and FT Treeis72.6624% accuracy	Naïve Bayes algorithmhas better compare tootheralgorithms
5	HebaAyeldeenet al. [5]	2015	Liverfibrosis	Decisiontree	department of MedicalBiochemistry andMolecular Biology, FacultyofMedicine, CairoUniversity.		decision tree classifieraccuracyis 93.7%
6	D Sindhuja &RjeminaPriyadarsini[6]	2016	Liver disease disorder	C4.5,NaïveBayes,SVM,BPNN ,Regression andDT Data	APhas better datasetresultthanUCLA	Survey paper suggestC4.5has better resultsthanothers	C4.5 has betteraccuracy result thanotheralgorithms
7	Somaya Hashemetal [8]	2016	Liverfibrosis	PSO, GA, MReg& ADT	Egyptian nationalcommitteeforcontrolof viralhepatitisdatabase	PSO, GA, MReg&ADTare66.4, 69.6.69.1,&84.4%	ADT has moreaccuracy resultthan otheralgorithms

8	Sumedh Sontakke etal	2017	Liverdisease	SVM & Backpropagation	(UCI)Machine LearningRepository	SVM (accuracy71%))& Backpropagation(accuracy73.2%)	More accuracy result inBackpropagation
9	Hanma etal	2018	Nonalcoholicfatty liverdisease	Using 11classificationalgorithms	First Affiliated Hospital,Zhejiang University China,Collegeofmedicine FirstAffiliated	Bayesian networkaccuracy83%	Concluded Bayesiannetwork has bestperformance than otheralgorithms
10	Joel Jacob etal[10]	2018	Liverdisease	Logisticregression, K-NN,SVM,& ANN	IndianLiverPatientDataset comprised of 10different attributes of 583patients.	Logistic regression, K-NN, SVM,& ANN has73.23,72.05, 75.04& 92.8% accuracyrespectively	ANN has higheraccuracy than others
11	Sivakumar Det al[11]	2019	Liverdisease	K-means & C4.5algorithms	UCIRepository	C4.5 algorithm has94.36%precision.	C4.5 has betteraccuracythanK-means algorithms
12	MehtajBanuH[12]	2019	Liverdisease	Supervised,unsupervised&reinforcement	UCIrepositorydatabases.	Note: Only explainingnot implementingpractically	KNN and AVM hasimprovedprediction performanceaccur

							acy
13	VasanDu raiet al[13]	2019	Liv erdis ease	SVM,NB&J4 8	UCIrepository	J48 algorithm hasbetter feature selectionwith95. 04%accuracy	J48 algorithm isaccuracy rate of95.04%.

3. IDEATION AND PROPOSED SOLUTION

3.1 Empathy Map Canvas

3.2 Ideation and Brainstorming

3.3 Proposed Solution

1) **Problem Statement (Problem to be solved)**: Discovering the existence of liver disease at an early stage is a complex task for the doctors. The main objective of the project is to examine data from liver patients concentrating on relationships between a key list of liver enzymes, proteins, age and gender using them to try and predict the likeliness of liver disease

2) **Idea / Solution description**: Our solution is to build a model by applying various machine learning algorithms and find the best accurate model to predict whether as a liver disease or not. We plan to perform data pre-processing and data visualization methods to increase the accuracy of the model. And integrate the chosen model into Flask based web application where the User can predict the disease by entering parameters in the web application.

3) **Novelty / Uniqueness:** Data pre-processing which includes Data Cleaning, Data transformation, and Data Reduction is performed to increase the accuracy of the model.

   Various Machine model is implemented and the highest accurate model is chosen.

   Model output is evaluated using MSE, confusion matrix and other various metrics.

   ROC-AUC considers the rank of the output probabilities and intuitively measures the likelihood that model can distinguish between a positive point and a negative point. We will use AUC to select the best model among the various machine learning models.

4) **Social Impact / Customer Satisfaction:** Since the likeliness of the liver disease

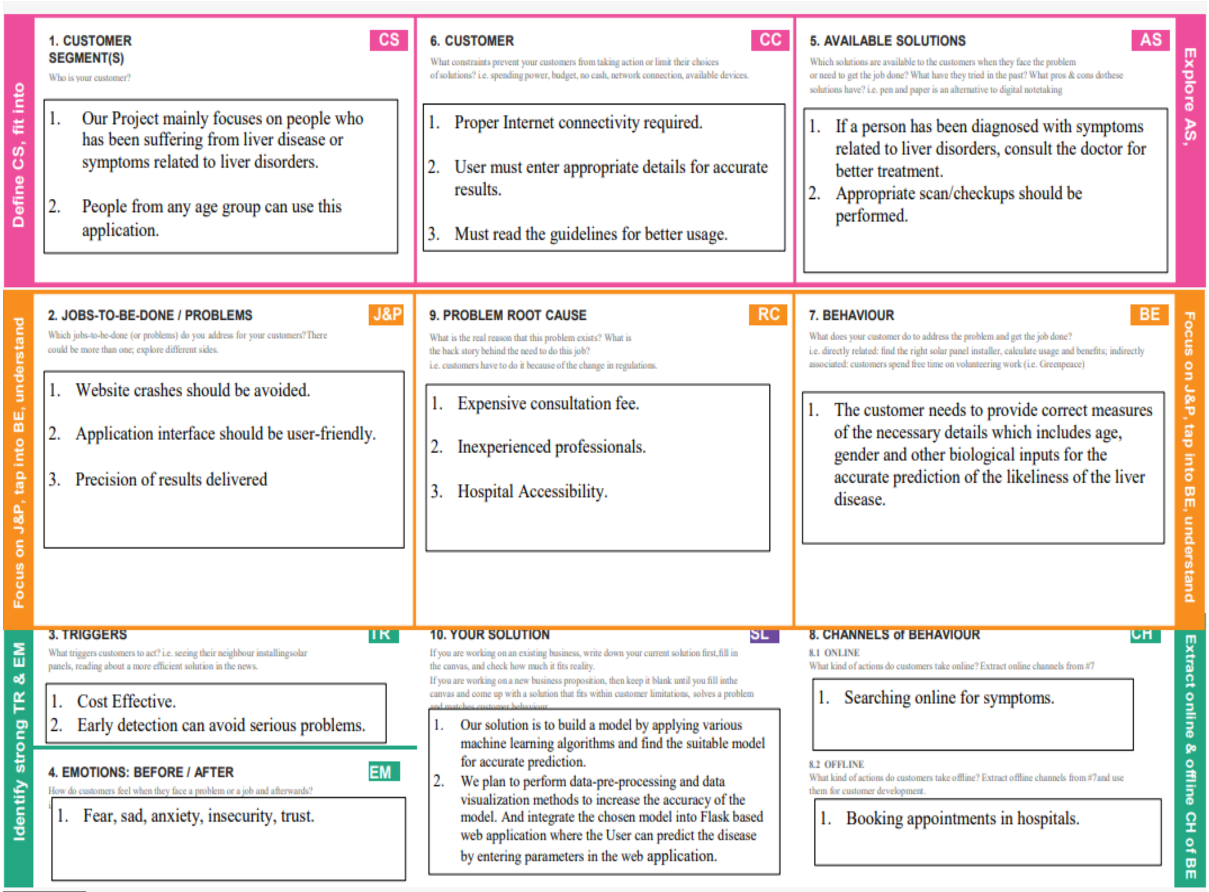
   is predicted with high accuracy, user can able to take remedial measures.

5) **Business Model (Revenue Model):** Revenue can be made by collaborating with Hospitals and other health related companies and Integrating subscription services to the application

6) **Scalability of the Solution:** Accuracy of the model can be increased by training with large data. The model can be made to learn from the user

input. Model is deployed in the web where the public from across the world can use to predict the likeliness of liver disease.

3.4 Proposed Solution Fit



4. REQUIREMENT ANALYSIS

Followingarethefunctionalrequirementsoftheproposedsolution.

FR.NO	FUNCTIONALREQUIREMENT(EPIC)	SUBREQUIREMENT(STORY /SUB-TASK)
FR.1	UserRegistration	Registration through FormRegistration through GmailRegistrationthroughLinkedIN
FR.2	UserConfirmation	Confirmation via EmailConfirmationviaOTP

FR.3	UserInput	Getnecessarydetailsforprediction
FR.4	DataProcessing	Data cleaning,Data scaling,Featureselec tion
FR.5	Prediction	Predictingwhethertheuserhaslive rdiseaseornot

Non-functionalRequirements:

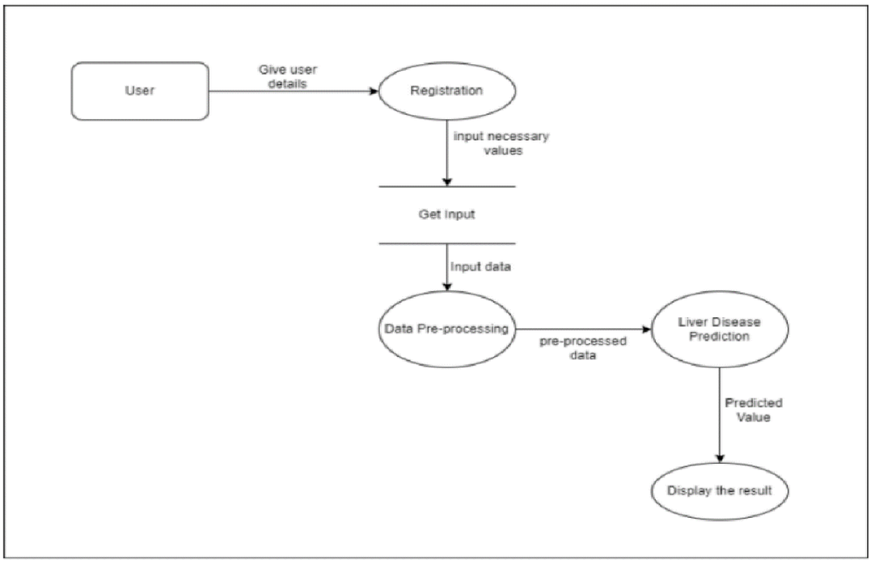
Followingarethenon-functionalrequirementsoftheproposedsolution.

FR.NO	**Non-FunctionalRequi rement**	**Description**
NFR-1	Usability	Tocheckwhetherthepatienthasliver diseaseornot
NFR-2	Security	Implement necessarytechniquestoprovidesecu ritytotheuserdata
NFR-3	Reliability	Makeensurethatthemodelisreliabl e
NFR-4	Performance	UseefficientMLtechniquesforbett eraccuracy
NFR-5	Scalability	Predictsvarioustypesofliverdisease

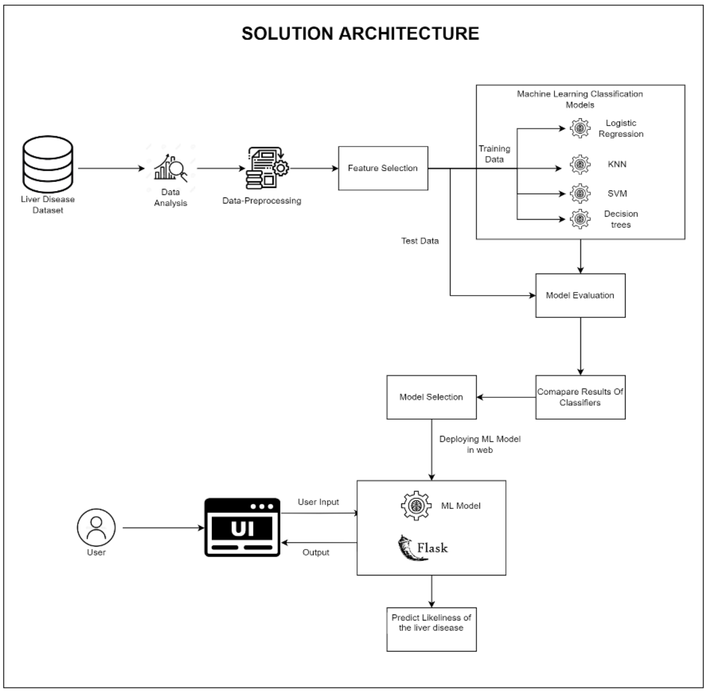
5. PROJECT DESIGN

5.1 Data Flow Diagram

Data Flow Diagram

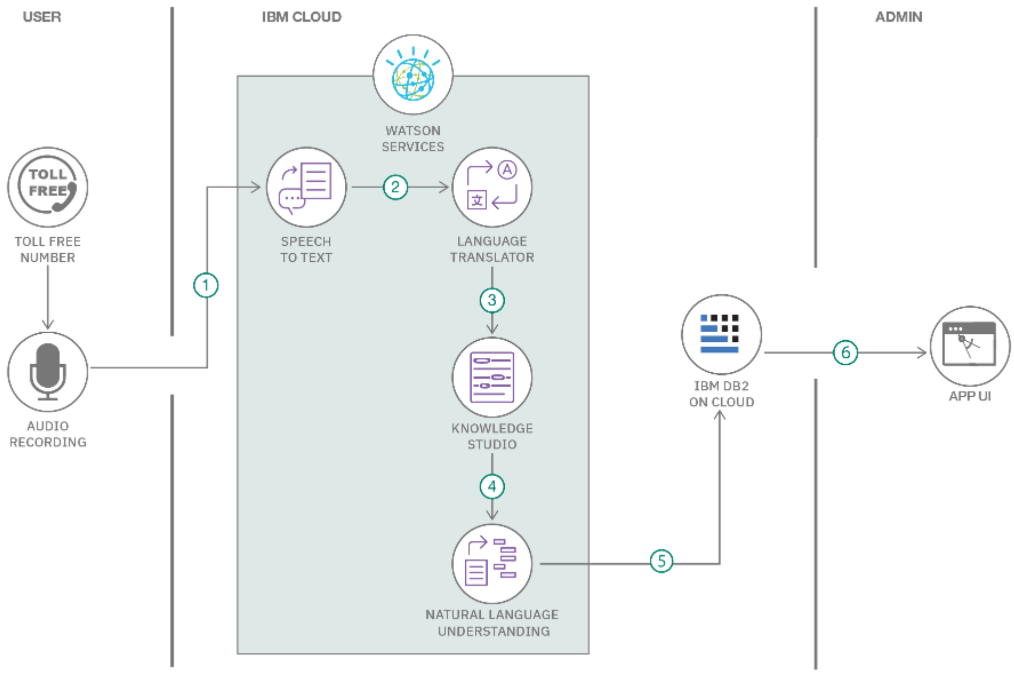


5.2 Solution and Technology Architecture



Technical Architecture:

The Deliverable shall include the architectural diagram as below and the information as per the table1 & table 2



Components & Technologies:

S.No	Component	Description	Technology
1.	User Interface	How user interacts with application e.g. Web UI, Mobile App, Chatbot etc.	HTML, CSS, JavaScript / Angular Js / React Js etc.
2.	Application Logic-1	Logic for a process in the application	Java / Python
3.	Application Logic-2	Logic for a process in the application	IBM Watson STT service
4.	Application Logic-3	Logic for a process in the application	IBM Watson Assistant
5.	Database	Data Type, Configurations etc.	MySQL, NoSQL, etc.
6.	Cloud Database	Database Service on Cloud	IBM DB2, IBM Cloudant etc.

7.	File Storage	File storage requirements	IBM Block Storage or Other Storage Service or Local Filesystem
8.	External API-1	Purpose of External API used in the application	IBM Weather API, etc.
9.	External API-2	Purpose of External API used in the application	Aadhar API, etc.
10.	Machine Learning Model	Purpose of Machine Learning Model	Object Recognition Model, etc.
11.	Infrastructure (Server / Cloud)	Application Deployment on Local System / Cloud Local Server Configuration: Cloud Server Configuration :	Local, Cloud Foundry, Kubernetes, etc.

Application Characteristics:

S.No	Characteristics	Description	Technology
1.	Open-Source Frameworks	List the open-source frameworks used	Technology of Opensource framework
2.	Security Implementations	List all the security / access controls implemented, use of firewalls etc.	e.g. SHA-256, Encryptions, IAM Controls, OWASP etc.
3.	Scalable Architecture	Justify the scalability of architecture (3 – tier, Micro-services)	Technology used
4.	Availability	Justify the availability of application (e.g. use of load balancers, distributed servers etc.)	Technology used
5.	Performance	Design consideration for the performance of the application (number of requests per sec, use of Cache, use of CDN’s) etc.	Technology used

6.PROJECT PLANNING AND SCHEDULING

6.1 Sprint Planning and Estimation

Project Planning Phase
Milestone and Activity List

TITLE	DESCRIPTION	DATE
Literature Survey & Information Gathering	Literature survey on the selected project & gathering information by referring the, technical papers, research publications etc.	28 SEPTEMBER 2022
Prepare Empathy Map	Prepare Empathy Map Canvas to capture the user Pains & Gains, Prepare list of problem statements	24 SEPTEMBER 2022
Ideation	List the by organizing the brainstorming session and prioritize the top 3 ideas based on the feasibility & importance.	25 SEPTEMBER 2022
Proposed Solution	Prepare the proposed solution document, which includes the novelty, feasibility of idea, business model, social impact, scalability of solution, etc.	23 SEPTEMBER 2022
Problem Solution Fit	Prepare problem - solution fit document.	30 SEPTEMBER 2022
Solution Architecture	Prepare solution architecture document.	28 SEPTEMBER 2022

Customer Journey	Prepare the customer journey maps to understand the user interactions & experiences with the application (entry to exit).	20 OCTOBER 2022
Functional Requirement	Prepare the functional requirement document.	8 OCTOBER 2022
Data Flow Diagrams	Draw the data flow diagrams and submit for review.	9 OCTOBER 2022
Technology Architecture	Prepare the technology architecture diagram.	3 NOVEMBER 2022
Prepare Milestone & Activity List	Prepare the milestones & activity list of the project.	4 NOVEMBER 2022
Project Development - Delivery of Sprint-1, 2, 3 & 4	Develop & submit the developed code by testing it.	IN PROGRESS...

6.2 Sprint Delivery Plan

SPRINT DELIVERY PLAN

Product Backlog, Sprint Schedule, and Estimation

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-1	Registration	USN-1	As a user, I can register for the application by entering my email, password, and confirming my password.	5	High	Suruthi Lakshmi
Sprint-1		USN-2	As a user, I will receive confirmation email once I have registered for the application	5	High	Snega
Sprint-1	Login	USN-3	As a user, I can log into the application by entering email & password	10	High	Keerthana
Sprint-2	Input Necessary Details	USN-4	As a user, I can give Input Details to Predict Likeliness of Liver Disease.	15	High	Sharumathi
Sprint-2	Data Pre-Processing	USN-5	Transform raw data into suitable format for prediction.	5	High	Keerthana
Sprint-3	Prediction of Liver Disease	USN-6	As a user, I can predict Liver Disease using machine learning model.	15	High	Sharumathi
Sprint-3		USN-7	As a user, I can get accurate prediction of liver disease.	5	Medium	Snega
Sprint-4	Deployment	USN-8	Deploy ML model into flask	5	High	Keerthana
Sprint-4	Deployment	USN-9	Deploy Website into real world	10	High	Suruthi Lakshmi
Sprint-4	Deployment	USN-8	As a user, I can give feedback of the application.	5	High	Snega

Project Tracker and Velocity:

Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date (Planned)
Sprint-1	20	6 Days	24 Oct 2022	9 Nov 2022
Sprint-2	20	6 Days	9 Nov 2022	11 Nov 2022
Sprint-3	20	6 Days	11 Nov 2022	16 Nov 2022
Sprint-4	20	6 Days	16 Nov 2022	19 Nov 2022

Velocity:

Imagine we have a 10-day sprint duration, and the velocity of the team is 20 (points per sprint). Let's calculate the team's average velocity (AV) per iteration unit (story points per day)

$$AV = \frac{sprint\ duration}{velocity} = \frac{20}{10} = 2$$

7.CODING AND SOLUTIONING

7.1 Model Building

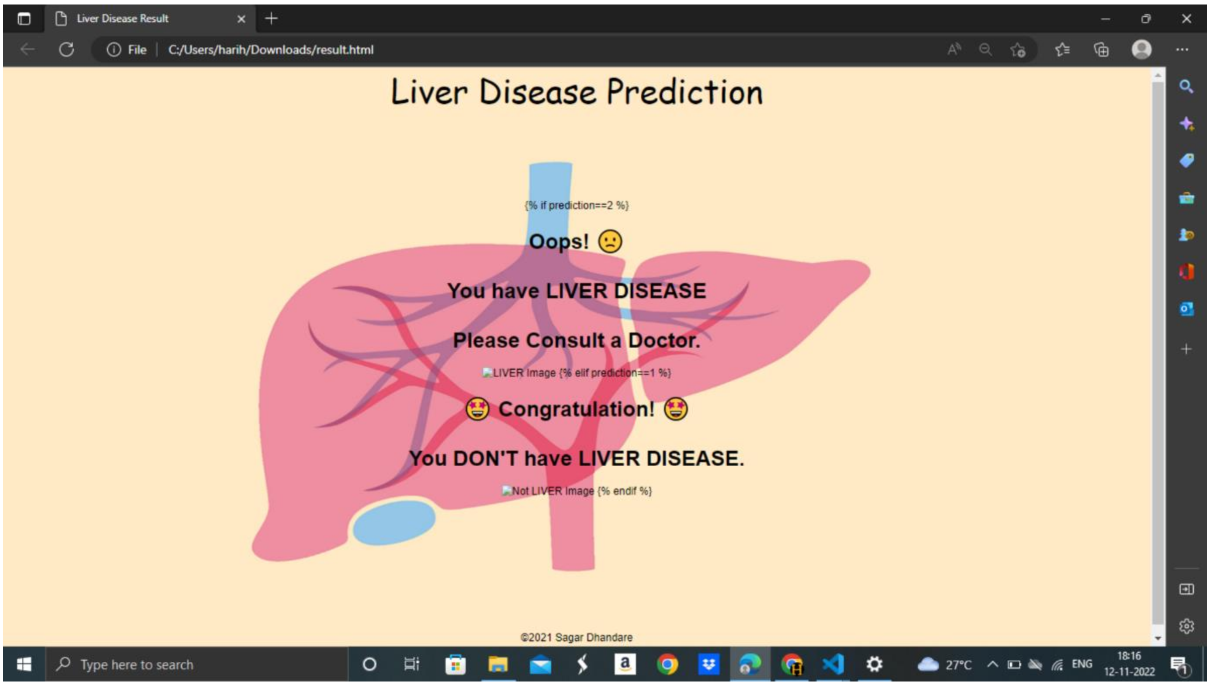
Model_Building.ipynb

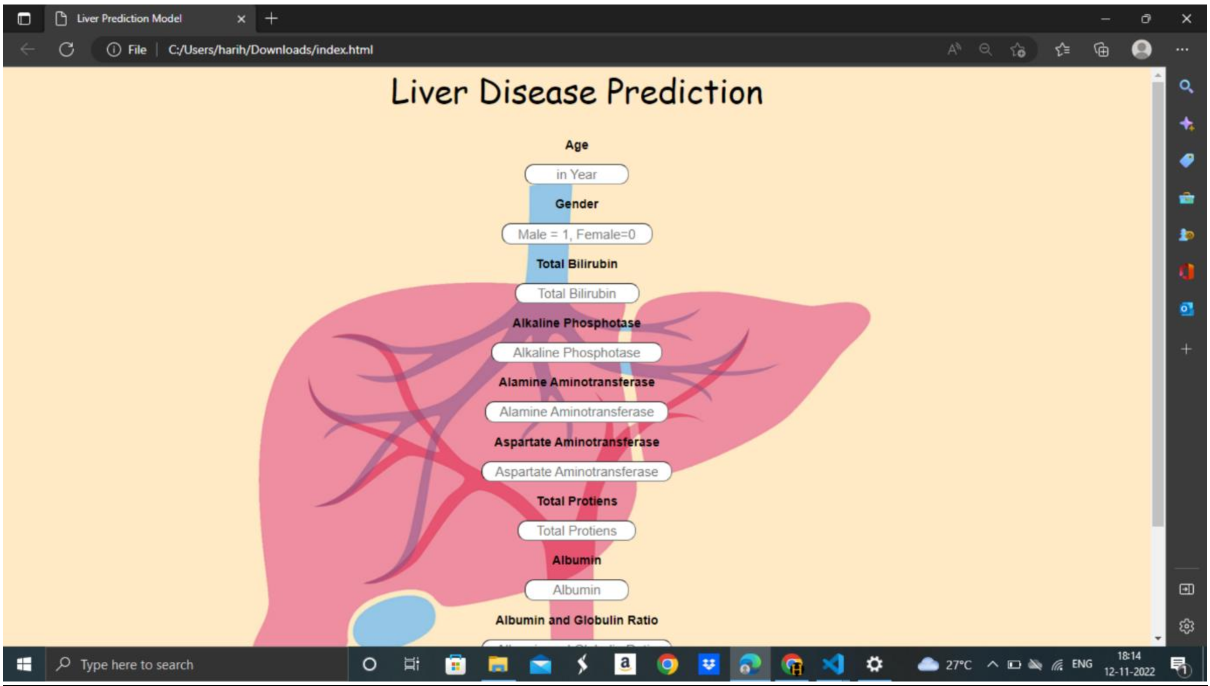
7.2 Application Building

app.py

```
from flask import Flask, render_template, request

import numpy as np

import pickle

app = Flask(__name__)

model = pickle.load(open('Liver2.pkl', 'rb'))

@app.route('/',methods=['GET'])

def Home():

    return render_template('index.html')

@app.route("/predict", methods=['POST'])

def predict():

    if request.method == 'POST':

        Age = int(request.form['Age'])

        Gender = int(request.form['Gender'])

Total_Bilirubin = float(request.form['Total_Bilirubin'])

Alkaline_Phosphotase = int(request.form['Alkaline_Phosphotase'])

Alamine_Aminotransferase = int(request.form['Alamine_Aminotransferase'])

Aspartate_Aminotransferase = int(request.form['Aspartate_Aminotransferase'])

Total_Protiens = float(request.form['Total_Protiens'])

        Albumin = float(request.form['Albumin'])

Albumin_and_Globulin_Ratio =
float(request.form['Albumin_and_Globulin_Ratio'])
```

```
values = np.array([[Age,Gender,Total_Bilirubin,Alkaline_Phosphotase,Alamine_Aminotransferase,Aspartate_Aminotransferase,Total_Protiens,Albumin,Albumin_and_Globulin_Ratio]])
    prediction = model.predict(values)

    return render_template('result.html', prediction=prediction)
if __name__ == "__main__":
app.run(debug=True)
```





8.TESTING

1. **Unit Testing:** A level of the software testing process where individual units/components of a software/system are tested. The purpose is to validate that each unit of the software performs as designed.

2. **Integration Testing:** A level of the software testing process where individual units are combined and tested as a group. The purpose of this level of testing is to expose faults in the interaction between integrated units.

3. **System Testing:** A level of the software testing process where a complete, integrated system/software is tested. The purpose of this test is to evaluate the system's compliance with the specified requirements.

4. **Black Box Testing:** The technique of testing in which the tester doesn't have access to the source code of the software and is conducted at the software interface without any concern with the internal logical structure of the software is known as black-box testing.

5. **White-Box Testing:** The technique of testing in which the tester is aware of the internal workings of the product, has access to its source code, and is conducted by making sure that all internal operations are performed according to the specifications is known as white box testing.

9. MODEL PERFORMANCE TESTING

ALGORITHM	ACCURACY	F1 SCORE	PRECISION	RECALL
RANDOM FOREST	0.80%	0.80	0.82	0.80
ADA BOOST	0.78	0.78	0.80	0.79
GRADIENT BOOSTING	0.76	0.76	0.76	0.76

10. ADVANTAGES AND DISADVANTAGES

ADVANTAGES

- Accurate Results
- Faster Prediction
- User Friendly Website

DISADVANTAGES

- Requires continuous support of Internet since this is a website
- Can be trained with more images to improve accuracy

11.CONCLUSION

Clinicians who are skilled at identifying noteworthy observations and categorising them as normal or abnormal using background knowledge and other context clues can detect chronic liver disease. Similar to how ML algorithms may help medical professionals, these algorithms can be trained to recognise the potential for liver illness. ML approaches were able to distinguish between blood donors with and without liver disease with high accuracy by using the correlation of each variable with the risk of liver disease to train the model. By increasing awareness of risk factors and diagnostic variables, the application of ML approaches can aid in lowering the overall burden of liver disease on public health globally. More importantly, for chronic liver illness, ML could reduce liver-related mortality, transplants, and/or hospitalizations by identifying liver disease in its early stages or in concealed cases.

12. FUTURE SCOPE

In future work, the use of fast datasets technique like Apache Hadoop or Spark can be incorporated with this technique. In addition to this, we can use distributed refined algorithms like Forest Tree implemented in Apache Hadoop to increase scalability and efficiency.

13. REFERENCES

[1] Asrani, S.K.; Devarbhavi, H.; Eaton, J.; Kamath, P.S. "Burden of liver diseases in the world". J. Hepatol. 2019.

[2] Chalasani, N.; Younossi, Z.; Lavine, J.E.; Charlton, M.; Cusi, K.; Rinella, M.; Harrison, S.A.; Brunt, E.M.; Sanyal, A.J. "The diagnosis and management of nonalcoholic fatty liver disease: Practice guidance from the American Association for the Study of Liver Diseases". Hepatology 2018.

[3] Wang, Y.; Li, Y.; Wang, X.; Gacesa, R.; Zhang, J.; Zhou, L.; Wang, B. "Predicting Liver Disease Risk Using a Combination of Common Clinical Markers: A Screening Model from Routine Health Check-Up". Dis. Markers 2020

[4] . Hughes, R.A.; Heron, J.; Sterne, J.A.; Tilling, K. Accounting for missing data in statistical analyses: Multiple imputation is not always the answer. Int. J. Epidemiol. 2019, 48, 1294–1304. [CrossRef] 20.

[5] Hoffmann, G.; Bietenbeck, A.; Lichtinghagen, R.; Klawonn, F. Using machine learning techniques to generate laboratory diagnostic pathways—A case study. J. Lab. Precis. Med. 2018, 3, 58. [CrossRef] 28.

[6] Rizwan, A.; Iqbal, N.; Ahmad, R.; Kim, D.H. WR-SVM Model Based on the Margin Radius Approach for Solving the Minimum Enclosing Ball Problem in Support Vector Machine Classification. Appl. Sci. 2021, 11, 4657. [CrossRef]

[7] Pianykh, O.S.; Guitron, S.; Parke, D.; Zhang, C.; Pandharipande, P.; Brink, J.; Rosenthal, D. Improving healthcare operations management with machine learning. Nat. Mach. Intell. 2020, 2, 266–273. [CrossRef]

[8] Jaganathan, K.; Tayara, H.; Chong, K.T. Prediction of Drug-Induced Liver Toxicity Using SVM and Optimal Descriptor Sets. Int. J. Mol. Sci. 2021, 22, 8073. [CrossRef]

[9] Munish, G.; Kaplan, H.C. Measurement for quality improvement: Using data to drive change. J. Perinatol. 2020, 40, 962–971

[10] Couronné, R.; Philipp, P.; Anne-Laure, B. Random forest versus logistic regression: A large-scale benchmark experiment. BMC Bioinform. 2018, 19, 1–14. [CrossRef]