Prediction of Heart Disease and Liver Disease

Abstract— **The coexistence of heart and liver diseases poses significant clinical challenges, requiring effective predictive models for early detection and intervention. In this study, we employed decision tree and logistic regression algorithms to predict the likelihood of liver disease in individuals diagnosed with heart disease. Two distinct datasets were utilized, one for heart disease prediction and the other for liver disease prediction, each containing relevant clinical attributes. Through rigorous experimentation and evaluation, our models demonstrated promising performance in identifying the presence of liver disease in individuals with heart disease.**

Keywords—machine learning, Logistic Regression, Decision Tree,

I. INTRODUCTION

In the ever-evolving field of healthcare, predicting and preventing heart and liver diseases have become paramount to ensuring the well-being of individuals and communities. Today, we will delve into two powerful machine learning techniques, Logistic Regression and Decision Tree, which have shown significant potential in predicting the likelihood of these diseases. Logistic Regression is a statistical method that allows us to model the relationship between predictor variables and a binary outcome, such as the presence or absence of heart and liver diseases. This technique is particularly useful when we want to understand the effect of various factors on the probability of a specific disease. Decision Trees, on the other hand, are a non-parametric method used for both classification and regression tasks. They work by recursively splitting the data into subsets based on the most significant predictor variables, thus creating a tree-like model that can be easily interpreted and understood. In the context of predicting heart and liver diseases, decision trees can help identify the most important risk factors and provide a visual representation of the decision-making process. Combining these two techniques can lead to more accurate and robust predictions, as well as a deeper understanding of the complex interplay between various risk factors and the likelihood of developing heart and liver diseases. By employing these machine learning algorithms, researchers and healthcare professionals can develop personalized preventive measures, early detection strategies, and more effective treatments to improve overall patient outcomes. In conclusion, the integration of Logistic Regression and Decision Tree in predicting heart and liver diseases holds great promise for advancing healthcare and saving lives. As we explore these techniques further, we can expect to gain valuable insights into disease risk factors and contribute to the development of more effective, personalized healthcare strategies.

II. LITERATURE SURVEY

From [1] it is observed that heart disease is known to strike men more frequently than it does women. Ageing, daily cigarette smoking, and fluctuating blood pressure all these factors raise the chance of acquiring heart disease.

In [2] the results of the proposed work depict that Logistic Regression is better than the other supervised classifiers in terms of the discussed performance metrics – accuracy, precision, sensitivity (or recall), specificity and F1 score. The model gives the results with the highest accuracy of 92.30%

In [4] out of all the classifier used, logistic regression gives the most elevated order exactness 75% dependent on F1 measure to predict the liver disease.

III. UNIQUENESS OF THE PROPOSED WORK

The uniqueness of the proposed work lies in its integration of logistic regression and decision tree classifiers to predict both heart and liver diseases simultaneously, leveraging a common attribute present in two separate datasets. Here is a breakdown of its uniqueness:

Simultaneous Prediction of Multiple Diseases: Most studies focus on predicting a single disease outcome. However, the proposed work aims to predict both heart and liver diseases concurrently. This approach provides a more holistic understanding of patients' health status, as these diseases often coexist and share common risk factors.

Integration of Disparate Data Sources: By combining data from two distinct datasets, each pertaining to different disease domains, the proposed work creates a comprehensive dataset. This integration allows for a more nuanced analysis of shared risk factors, comorbidities, and interactions between heart and liver diseases that may not be apparent when studying each disease in isolation.

Common Attribute Utilization: The unique aspect of using a common attribute across both datasets facilitates the integration process and enables more robust predictive modeling. Whether it is demographic information, clinical measurements, or biomarkers, this shared attribute serves as a bridge between the datasets, enriching the feature space and potentially capturing latent relationships between heart and liver diseases.

Synergistic Modeling Techniques: Integrating logistic regression and decision tree classifiers offer complementary advantages. Logistic regression provides a probabilistic interpretation of the relationships between predictors and disease outcomes, while decision trees offer intuitive, interpretable rules for classification.

Research Contribution: While predictive modeling in healthcare is not new, the proposed work contributes to the advancement of the field by addressing the challenge of simultaneous prediction of multiple diseases using disparate datasets. This research fills a gap in the literature and paves the way for future studies exploring similar cross-domain predictive modeling tasks in healthcare.

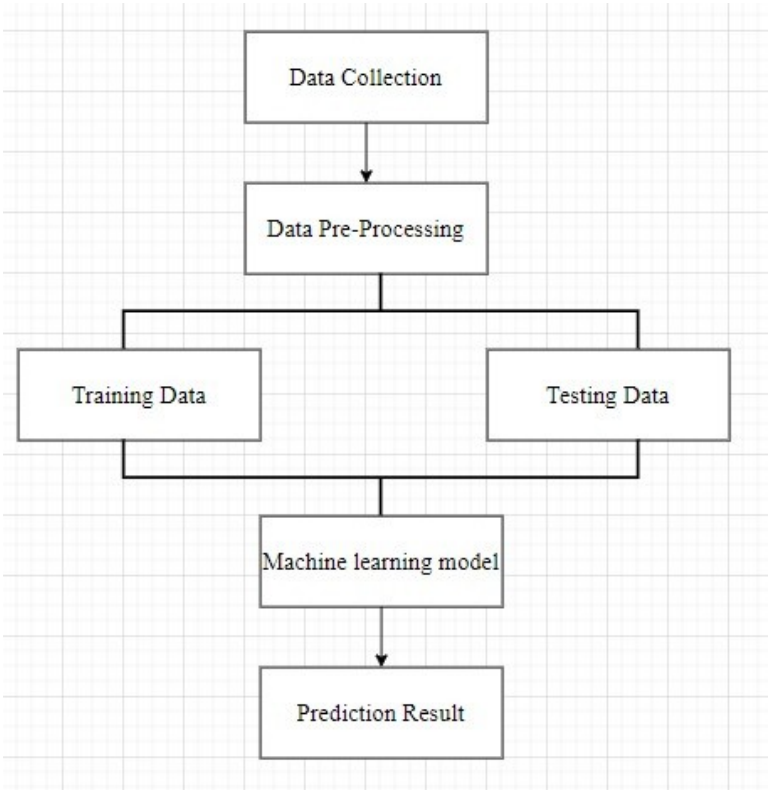
In essence, the uniqueness of the proposed work lies in its holistic approach to disease prediction, leveraging the integration of disparate datasets, the utilization of a common attribute, and synergistic modeling techniques to simultaneously predict heart and liver diseases, thereby offering novel insights and practical applications in healthcare.

IV. DATASET USED

Liver dataset contains 416 liver patient records and 167 non liver patient records collected from North East of Andhra Pradesh, India. Number of instances in dataset is 583 and total of attributes is 12. The "Dataset" column is a class label used to divide groups into liver patient (liver disease) or not (no disease). This data set contains 441 male patient records and 142 female patient records. Any patient whose age exceeded 89 is listed as being of age "90".

Heart dataset contains number of instances in dataset is 1025 and 14 attributes. We have 165 people with heart disease and 138 people without heart disease. The “target” column is a class label used to divide groups into heart patient or not.

V. PROPOSED ARCHITECTURE



Data Collection

The liver dataset was collected from the northeast of Andhra Pradesh, India. This dataset consists of 583 liver patient’s data whereas 75.64% male patients and 24.36% are female patients. This dataset has contained 12 parameters where we choose 11 parameters for our further analysis and 1 parameter as a target class. Such as,

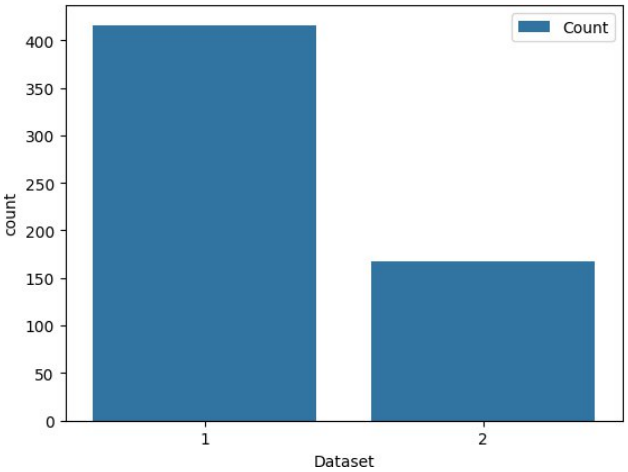
1. Age of the patient
2. Gender of the patient
3. Total Bilirubin
4. Direct Bilirubin
5. Alkaline Phosphatase
6. Alanine Aminotransferase
7. Aspartate Aminotransferase
8. Total Proteins
9. Albumin
10. Albumin and Globulin Ratio
11. Cholesterol
12. Dataset: field used to split the data into two sets (patient with liver disease, or no disease)

The heart dataset is multivariate type of dataset which means providing or involving a variety of separate mathematical or statistical variables, multivariate numerical data analysis. The dataset contained 14 parameters where we choose 13 parameters for our further analysis and 1 parameter as a dataset class. Such as,

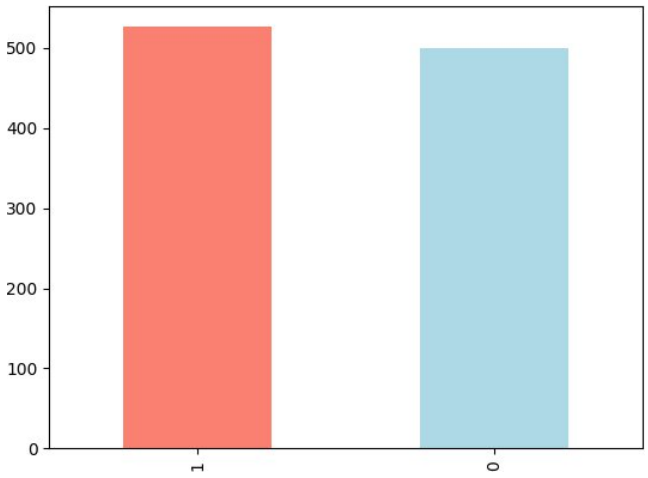
1. Age
2. Sex
3. Chest Pain
4. Person's resting blood pressure
5. Cholesterol
6. Person's fasting blood sugar
7. resting electrocardiographic results
8. The person's maximum heart rate achieved
9. Exercise induced angina
10. ST depression induced by exercise relative to rest
11. The slope of the peak exercise ST segment
12. number of major vessels
13. Thalassemia Value

Data Preprocessing

Dataset is first loaded and then data cleaning and finding missing values was performed on all records. We analyzed 583 liver patient's data where 416 samples are liver patient and 167 samples are non-liver patients.



We analyzed 303 heart patient's data where 165 samples are heart patient and 138 samples are non-liver patients.



The two dataframes (heart dataset and liver dataset are merged together based on a commom column "cholesterol")

Tool and Language

In this study we used the google colaboratory as a tool and python 3.10.14 as programming language.

VI. IMPLEMENTATION AND RESULTS

The implemented system is a heart and liver disease prediction system. The liver dataset and heart dataset are combined and both the dataset possess a common attribute which is "cholesterol." The dataset is divided into training and testing sets which is then trained using logistic regression and decision tree. Description of the classification algorithms

1. Logistic Regression

   One of the simplest and best ML classification algorithms is logistic regression. LR is a supervised ML binary classification algorithm widely used in most applications. It operates on a categorical dependent variable, the result can be a discrete or binary categorical variable 0 or 1.

   Logistic sigmoid function:

$$prob(Y=1)=\frac{e^{z}}{1+e^{z}}$$

2. Decision Tree

   Decision trees are one of the most powerful tools in supervised learning algorithms used for both classification and regression tasks. This algorithm has several advantages such as interpretability, ability to handle imbalanced data, variable selection, handling of missing values, and its non-parametric nature.

VII. OUTCOMES OF THIS RESEARCH

The outcomes of predicting heart and liver disease using Logistic Regression and Decision Tree classifiers can be categorized into two main areas:

1. Model Performance Metrics:

Accuracy: This measure indicates the overall percentage of the model's predictions that were accurate. A high accuracy (over 80%) signifies that the model functions effectively with the provided dataset. The logistic regression model has an accuracy of 0.858. Following feature engineering, the accuracy was 1.0.

The decision tree model's accuracy of 1.0 suggests overfitting; to address this, we used minimal cost-complexity pruning (CCP), a regularization technique. Currently, the accuracy is 0.86.

The training and testing score for both the model is shown below.

	Model	Training Score	Test Score
0	Logistic Regression	85.67	86.95
1	Decision Tree	85.67	86.95

The model's performance metrics are based on the specific dataset used for training and testing. Performance might vary on different datasets. It is crucial to evaluate both accuracy and other metrics like precision and recall to understand the model's strengths and weaknesses. Performance metric for logistic regression is given below

```
Accuracy: 0.8580750407830342
Confusion Matrix:
 [[227  63]
 [ 24 299]]
Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.78      0.84       290
           1       0.83      0.93      0.87       323

    accuracy                           0.86       613
   macro avg       0.87      0.85      0.86       613
weighted avg       0.86      0.86      0.86       613
```

Performance metric for decision tree is given below

```
Accuracy: 0.8694942903752039
Confusion Matrix:
 [[186  80]
 [  0 347]]
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.70      0.82       266
           1       0.81      1.00      0.90       347

    accuracy                           0.87       613
   macro avg       0.91      0.85      0.86       613
weighted avg       0.89      0.87      0.86       613
```

Feature selection can improve model performance by focusing on the most relevant features for each disease. Hyperparameter tuning can optimize the performance of both Logistic Regression and Decision Tree models. By analyzing the model performance metrics and the predicted disease status for new data points, you can gain insights into the effectiveness of these classifiers in predicting heart and liver disease.

VIII. CONCLUSION AND FUTURE WORK

The conclusion and future work for predicting heart and liver diseases using logistic regression and decision tree classifier can be summarized as follows: Using logistic regression and decision tree classifier models, it is possible to predict the likelihood of heart and liver diseases based on various risk factors and patient data. These machine learning algorithms can help identify high-risk individuals and assist healthcare professionals in making informed decisions for early diagnosis and intervention.

From the study it is observed that as the training set ratio increases, the model's performance on the training data generally improves, as expected. This is indicated by the increasing trend in training accuracy. However, the performance on the testing data may not necessarily follow the same pattern. It may peak at a certain point and then start to decrease due to overfitting.

A higher value of ccp_alpha increases the regularization strength, leading to simpler trees. In this case, it is set to 0.04, indicating a moderate level of regularization.
Managing the tree's depth aids in avoiding overfitting. While a deeper tree may be able to identify more complex patterns in the training set, overfitting could result from the tree learning to remember noise. The tree more effectively generalizes to unknown data by restricting the depth.

Feature Selection and Model Optimization: Future work should focus on selecting the most relevant features that contribute significantly to the prediction of heart and liver diseases. This can be achieved through feature selection techniques and model optimization. By reducing the number of input features, the models can be made more efficient and accurate.

The future work includes:
Ensemble Methods and Hybrid Models:
Incorporating ensemble methods and hybrid models can further improve the predictive power of the logistic regression and decision tree classifier models. By combining multiple models or algorithms, it is possible to achieve better accuracy and robustness in disease prediction.

Incorporating Advanced Techniques:
Exploring advanced machine learning techniques such as deep learning, random forests, and gradient boosting can lead to better disease prediction models. These techniques can capture complex patterns and relationships in the data, which may not be evident in logistic regression and decision tree classifier models.

Handling Imbalanced Datasets:
As heart and liver diseases are relatively rare compared to other health conditions, datasets may be imbalanced. Future research should address techniques to handle imbalanced datasets, such as oversampling, undersampling, and synthetic minority oversampling technique (SMOTE), to improve the models' performance in predicting rare disease cases.

Multi-class Prediction and Interpretability:
Extending the current binary classification models to multi-class prediction can help identify various heart and liver disease types. Additionally, ensuring the models' interpretability will allow healthcare professionals to understand the factors contributing to the disease prediction, leading to better decision-making and patient care.

In conclusion, the prediction of heart and liver diseases using logistic regression and decision tree classifier models has shown promising results. Further research and development in feature selection, model optimization, advanced techniques, real-world implementation, handling imbalanced datasets, and multi-class prediction will contribute to more accurate and reliable disease prediction models in the future.

IX. REFERENCE

[1] R., Shijitha & M, Kavya & C, PrathanyaSree & M, Deepasindhu & B, Nowshika. (2023).Heart Disease Prediction Using Logistic Regression. 11. 573-579.

[2] Mehdi, Mrs., Khundmir Iliyas, Mr. Imran and Sadekh Shaikh. “Prediction of Heart Disease Using Decision Tree.” (2019).

[3] Redrowthu Ph.D, Vijaya & Gajavelly, Kovid & Nikhath, A. & Vasavi, R. & Anumasula, Rakshith Reddy. (2022). Heart Disease Prediction Using Decision Tree and SVM. 10.1007/978-981-16-7389-4_7.

[4] Karthick K, Aruna SK, Samikannu R, Kuppusamy R, Teekaraman Y, Thelkar AR. Implementation of a Heart Disease Risk Prediction Model Using Machine Learning. Comput Math Methods Med. 2022 May 2;2022:6517716. doi: 10.1155/2022/6517716. Retraction in: Comput Math Methods Med. 2023 Jul 19;2023:9764021. PMID: 35547562; PMCID: PMC9085310.

[5] Gupta, Chiradeep & Saha, Athina & Reddy, N V Subba & Acharya, Dinesh. (2022). Cardiac Disease Prediction using Supervised Machine Learning Techniques.. Journal of Physics: Conference Series. 2161. 012013. 10.1088/1742-6596/2161/1/012013.

[6] Gupta, Ketan & Jiwani, Nasmin & Afreen, Neda & D, Divyarani. (2022). Liver Disease Prediction using Machine learning Classification Techniques. 221-226. 10.1109/CSNT54456.2022.9787574.

[7] A. Dhyani et al., "Comparative Analysis of Supervised Machine Learning Algorithms for Liver Disease Prediction with SMOTE Enhancement," 2023 3rd Asian Conference on Innovation in Technology (ASIANCON), Ravet IN, India, 2023, pp. 1-6, doi: 10.1109/ASIANCON58793.2023.10270381

[8] Rahman, A. K. M. & Shamrat, F M & Tasnim, Zarrin & Roy, Joy & Hossain, Syed. (2019). A Comparative Study On Liver Disease Prediction Using Supervised Machine Learning Algorithms. 8. 419-422.

[9] Nahar, Nazmun & Ara, Ferdous. (2018). Liver Disease Prediction by Using Different Decision Tree Techniques. International Journal of Data Mining & Knowledge Management Process. 8. 01-09. 10.5121/ijdkp.2018.8201.

[10] Zhang, Yingjie & Diao, Lijuan & Ma, Linlin. (2021). Logistic Regression Models in Predicting Heart Disease. Journal of Physics: Conference Series. 1769. 012024. 10.1088/1742-6596/1769/1/012024. H. Poor, *An Introduction to Signal Detection and Estimation*. New York: Springer-Verlag, 1985, ch. 4.