Liver Disease Classification by Pruning Data Dependency utilizing Ensemble Learning based Feature Selection

Md. Asif Bin Khaled [1] Md. Mahin Rahman [2] Md. Golam Quaiyum[3] Sumiya Akter [4]

Department of Computer Science and Engineering
Independent University, Bangladesh
Dhaka, Bangladesh.
{[1]mdasifbinkhaled,[2]mahin.arrean,[3]mgquaiyum,[4]aazadkonok}@gmail.com

Abstract

Liver disease is responsible for over 2 million additional deaths globally each year. Therefore, early detection and treatment may lower the likelihood of liver disease-related death. Many researchers have been using artificial intelligence to detect liver disease. Inaccurate and disorganized data, however, make it difficult for them to choose an approach for determining the condition. Additionally, disproportionate data worsens dataset biases, reducing the validity of the research. As a result, it becomes necessary to develop techniques for dealing with this sort of challenge. This study suggests a methodology that integrates approaches for classifying liver disease by reduction of data dependency, which gives the advantage of getting more accurate predictions even with less data. Two imputation strategies were employed to tackle missing value and were contrasted with each other. Despite showing slight differences, no statistically significant distinctions between them were found. Machine Learning (ML) methods such as Random Forest, Extra Trees, Support Vector Machine, and K-Nearest Neighbor and neural network such as Multilayer Perceptron were employed to categorize liver diseases. The Extra Trees classifier outperformed other approaches in both of the imputed datasets, achieving **accuracy of 98.37% and 99.18%, F1-Score of 98.37% and 99.17% while achieving 99.3% and 99.4% area under the ROC curve (AUC)** respectively. This unorthodox method delivers cutting-edge accuracy with few feature dependencies. Hence, the suggested technique will make it easier for medical practitioners to identify liver diseases more quickly, resulting in a classification with lower data reliance that is less susceptible to error.

Introduction

Liver diseases are the major causes of morbidity and mortality in many parts of the world. According to the World Health Organization (WHO), it is one of the top ten principal causes of death in developing and underdeveloped countries. The liver is in charge of metabolism and protein absorption. It is affected due to a variety of conditions, such as cirrhosis, fibrosis, and hepatitis. It is imperative to have precise methods for the diagnosis and detection of liver diseases. The application of machine learning and data mining approaches can be effective in addressing this issue. In various inferential and decision-making applications, researchers and lab workers have used a variety of methodologies, including statistical techniques and machine learning approaches. However, collecting medical records in large quantities is often too hard to complete. Since numerous tests are necessary, they could also be costly and time-consuming.

Therefore, this study aims to reduce data dependency by removing features that are less contributing to model training through selective feature selection processes. Additionally, to study their impact, two different imputation techniques were compared. Moreover, many algorithms were employed on two differently imputed datasets to gauge models and propose a better-performing model for the detection of liver disease.

Problem Statement

Early detection of liver disease might be challenging. Numerous people don't exhibit many symptoms until their illness has advanced, at which point treatment may be ineffective. Liver function tests are frequently requested as part of a primary care blood test panel. But on their own, these diagnostic techniques are insufficient. Since numerous tests are necessary, they could also be costly and time-consuming. Thus, a precise and affordable diagnosis of liver disease is required.

Proposed Methodology

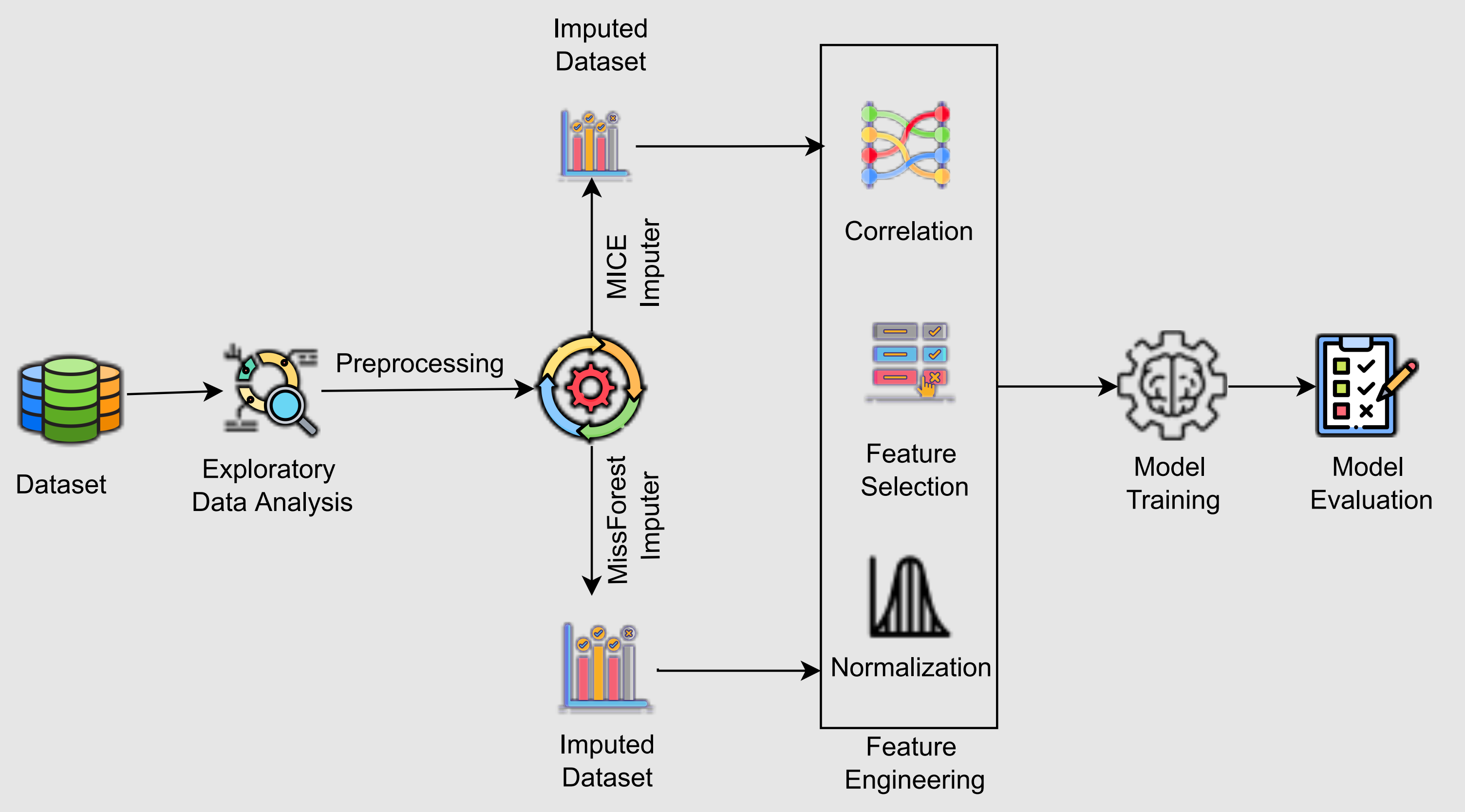


Figure 1. Methodology Employed in this Study.

Preprocessing

There are 10 haematological features present in the data, Aspartate Aminotransferase (AST), Alkaline Phosphatase (ALP), Albumin (ALB), Alanine Aminotransferase (ALT), Creatinine (CREA), Bilirubin (BIL), Cholesterol (CHOL), Choline Esterase (CHE), Gamma-glutamyl Transferase (GGT), and Total Protein (PROT). Missing Values are imputed using two different way.

- **MICE**
- **MissForest**

Feature Selection

A random forest feature importance test was conducted which is calculated using Gini importance. It was discovered that CHE and BIL were engaging in the process practically identically. From Pearson's correlation, BIL had a positive correlation whereas CHE had a negative correlation with the target feature. Therefore only one of the similarly important features was decided to be discarded picking only 5 features. The Recursive Feature Elimination approach was used to confirm the reduction and selection of features, where Random Forest was choosen as internal model and it was discovered that ALP, ALT, AST, BIL, and GGT were chosen by the algorithm, bearing almost 73.11 % and 73.07 % importance for MICE and MissForest imputed data, respectively.

Balancing Imbalance Dataset

The dataset appears to be slightly unbalanced, so Synthetic Minority Over-sampling Technique (SMOTE) technique approach have been applied to oversample the minor category.

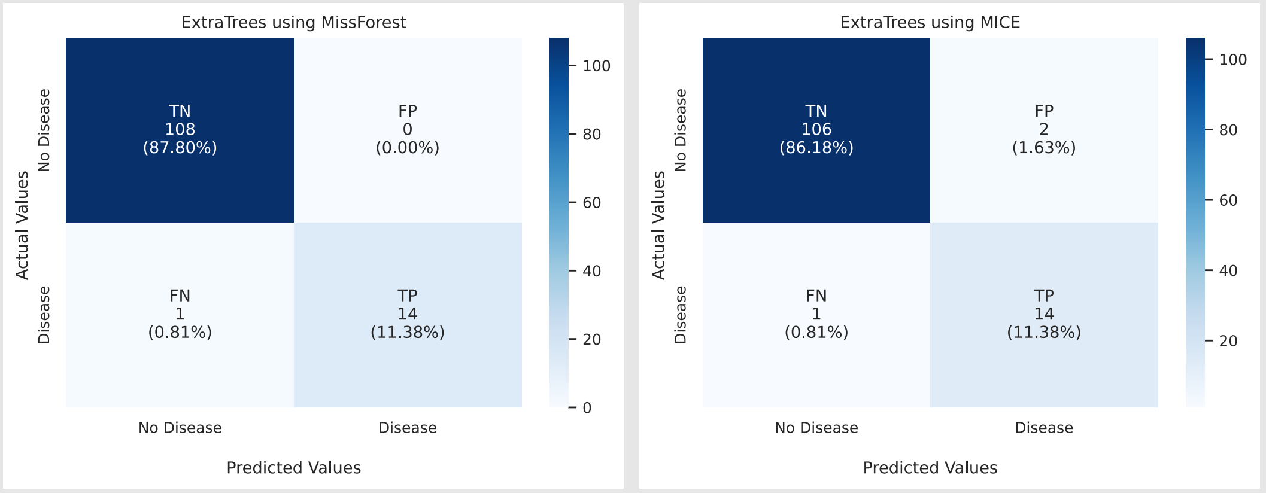
Results and Discussion

The target feature was divided into binary category as 'Liver Disease' and 'No Liver Disease'. Then the result of the classification task has been analyzed in different segments. Among them the confusion matrix, ROC curves and the accuracy of different matrices are shown bellow.

Model Data	Random Forest Classifier	Extra Trees Classifier	Multilayer Perceptron	SVM Classifier	KNN Classifier
MICE	95.12	98.37	97.56	95.12	93.49
MissForest	98.37	99.18	96.74	95.93	95.12

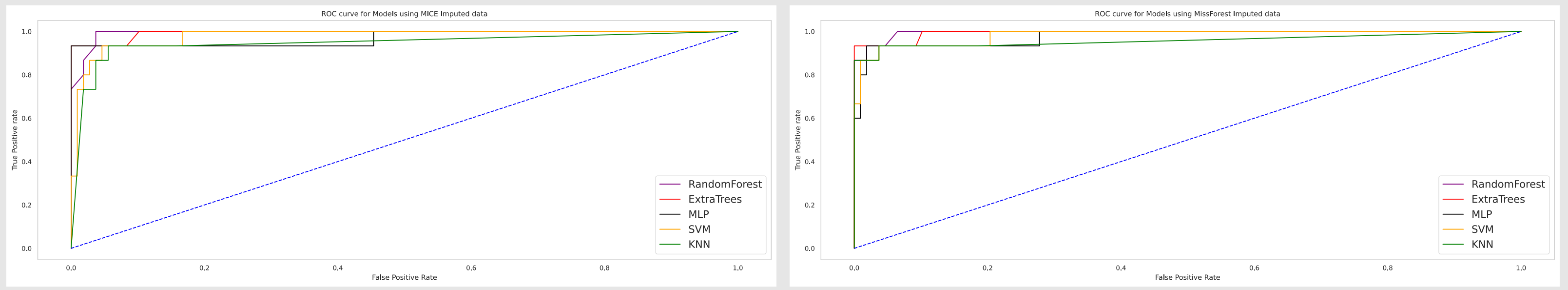
Confusion Matrix

Here is the confusion matrix of Extra Trees in two differently imputed data.



ROC Curve

Here is the ROC curve of models in two differently imputed data.



Precision, Recall, F1-Scores and AUC-ROC

Metric	Imputed Dataset	Extra Trees Classifier(%)	Random Forest Classifier(%)	Multilayer Perceptron (%)	Support Vector Machine (%)	K - Nearest Neighbor (%)
Precision	MICE	98.374	94.931	97.655	95.946	95.074
	MissForest	99.194	98.404	96.748	96.454	95.946
Recall	MICE	98.374	95.122	97.561	95.122	93.496
	MissForest	99.187	98.374	96.748	95.935	95.122
F1-Score	MICE	98.374	94.97	97.595	95.363	93.945
	MissForest	99.175	98.323	96.748	96.091	95.363
AUC-ROC	MICE	99.414	99.444	96.975	97.901	94.599
	MissForest	99.383	99.29	97.716	98.21	95.802

Conclusion

To predict Liver Disease, this study made use of methods like imputation, feature selection, and over sampling. A similar approach to this study can be taken and used to assist to some extent in the work of healthcare professionals. Also, this approach can classify liver disease patients with a higher probability while using much fewer data, where choosing only 5 features and achieving good success rate.