Final Report

Supervised Machine Learning Algorithms for Liver Disease Prediction

Lee Bing Qian BI20110252[1,*], Chan Khai Khee BI20110118[1], Chang Zi Yin BI20110125[1], Lee Yi Feng BI20110003[1], and Tee Geok Huan BI20110096[1]

1 Faculty of Computing and Informatics, Universiti Malaysia sabah, Kota Kinabalu 88400, Sabah, Malaysia; lee_bing_bi20@iluv.ums.edu.my; chan_kai_bi20@iluv.ums.edu.my; chang_zi_bi20@iluv.ums.edu.my; lee_yi_bi20@iluv.ums.edu.my; tee_geok_bi20@iluv.ums.edu.my;
* Correspondence: lee_bing_bi20@iluv.ums.edu.my

Abstract: Liver disease is a global killer which took away 70 % of people's lives every year. If a liver disease patient is diagnosed in its early stage before cirrhosis, it is still possible to cure. Thus, a good predictor must be developed by applying machine learning (ML) in identifying the feature of a patient with liver disease. Supervised ML such as classification is especially popular in the health care sector. In this study, an experiment was carried out on Support Vector Machine (SVM) Classifier, Random Forest Classifier, and Multi-Layer Perceptron Neural Network (MLP) to find the best-performed model and identify the important features in predicting liver disease. Two different sets of data were obtained to represent synthetic datasets and real-world datasets. The synthetic dataset was built by using Multiple Imputations By Chained Equations (MICEs) to replace the missing data and using the Synthetic Minority Oversampling Technique (SMOTE) to balance the training data. In contrast, the real-world dataset was obtained without any help of artificial intelligence-generated value. These two sets of data were used to train the classification models. Confusion matrices were used to evaluate the models' performances. The RF gives the best result with 98.37% of accuracy for synthetic data, and 80.50% of accuracy for real-world data. The original dataset was looped up to 100 times and performed train-test split with different random state numbers. The RF obtained 86.22% of mean accuracy, 3.95% of accuracy standard deviation, and a 95% confidence interval of accuracy between 0.8544 and 0.8699.

Keywords: supervised machine learning, classification, SMOTE algorithm, MICE algorithm, synthetic data, liver disease, Hepatitis C, Support Vector Machine, Random Forest, Neural Network

1.0 Introduction

The liver is a huge, meaty, and reddish-brown colored organ in the human body. It is one of the crucial organs of humans as it is responsible in blood detoxification in the human body. It filters the blood directed from the digestive system and eliminates toxic metabolized wastes. Furthermore, it also performs glucose synthesis, digestive enzyme production, erythrocyte regulation, protein synthesis, and other metabolic functions. Although the liver does humans a great favor, liver disease is one of the most "death-dealing" diseases globally. It claims the lives of up to 2 million people worldwide every year [1]. In 2010, based on the Global Burden of Disease

Project, the statistics claimed one million people died due to cirrhosis, and millions of people have liver cancer [2].

Chronic liver diseases can be classified into hepatitis, fibrosis, and cirrhosis. Hepatitis is usually caused by a viral infection or an auto-immune reaction. In contra, fibrosis and cirrhosis are subordinated by alcoholism and non-alcoholic fatty liver disease [3]. Fibrosis occurs when liver tissues become moderately damaged and scarred, making the liver performs poorly. It then causes the accumulation of toxins in humans' brains, affecting people's ability to focus, memorize, and fall asleep, and other mental functions [4]. However, suppose fibrosis is detected and treated early on before cirrhosis. Then, there is still a chance for the liver to heal itself and avoid liver failure. Cirrhosis is a severe liver disease caused by many other liver diseases, including inflammation, hepatitis, and fibrosis. It is considered a late stage of liver scarring—most healthy liver tissue is replaced with damaged liver tissue. Therefore, the liver diagnosed with cirrhosis will fail at this point and lose all its function gradually.

The comprehensive metabolic panel (CMP) is usually conducted to detect liver disease. Albumin (ALB), alkaline phosphatase (ALP), alanine aminotransferase (ALT), aspartate aminotransferase (AST), gamma glutamyl-transferase (GGT), creatine (CREA), total protein (PROT), and bilirubin (BIL) can all be detected by a CMP test with a liver function panel. The analysis of liver disease is made by studying the disposal of liver-associated molecules detected in the CMP test with the values of the normalization of the patient's age, gender, and BMI. When AST levels are higher than ALT levels, the patient has alcoholic liver disease. [3] In contrast, when AST and ALT levels are equal, the patient has fatty liver or non-alcoholic liver disease [5,6]. ALB increases in the intermediate range can indicate hepatitis and cirrhosis, but these findings are less distinct except they are confirmed by liver-specific enzymes like GGT [7]. In addition, the concentration of ALB synthesized in the liver decreases with chronic liver disease [8].

Society should treat liver disease issues seriously as it accounts for 70% of deaths worldwide [9]. Therefore, a more reliable and efficient way is required to solve the issue. Machine learning is usually applied to describe modification to a scheme made by conducting recognition, diagnosis, planning, robot control, and prediction associated with Artificial Intelligence (AI) tasks to enhance the system's achievement [10]. More recently, machine learning is the novel technique that has established the ability to decode large medical datasets into meaningful pieces of information. Machine learning helps the humanities solve many healthcare issues by applying machine learning and brings multiple benefits. For instance, the machine learning model helps improve patient safety [11], enhances the quality of health care [12], and decreases the costs of medication [13]. On top of that, machine learning assists clinicians' work by processing the billions of patient information and related data recorded in Electronic Healthcare Record(EHR). It has already been widely used in many therapeutic applications, including diagnosing lung disease from X_rays photos of the human chest [14], detecting lung cancer in the early stage [15], as well as predicting liver-related disease [16]. With the aid of machine learning, the liver condition can be identified and analyzed early to prevent it from deteriorating. If liver disease is detected early, it can avoid liver failure, and the possibility of death from liver disease will decrease. Hence, liver disease prediction using machine learning skills is essential to society.

The main purposes of conducting this study are to determine which machine learning models perform better in liver inflammation diagnosis and determine important attributes in liver disease prediction among all the variables such as patients' age, sex, and other variables. Classification technique is much significant in medical diagnosis and forecasting diseases [17]. In this study, the

problem statement predicts the relationship between liver disease with other variables such as patient's age, sex, and others using a classification algorithm. The targeted group of the prediction contains five categories, which are 0=Blood Donor and 0s=suspect Blood Donor indicates the condition of blood donor is healthy while 1=Hepatitis, 2=Fibrosis, and 3=Cirrhosis indicate blood donor is diagnosed with liver disease. This research applies three classification algorithms, such as Support Vector Machine classifier, Random Forest classifier, and Artificial Neural Network (ANN), to predict liver disease. These binary classification models will study the training data fitted to them and learn the patterns of the data. Once the models have been trained, they might accurately predict the labels such as "healthy" or "liver disease" [4].

The Support Vector Machine (SVM) is a discriminative classifier with a splitting hyperplane. This hyperplane is a two-dimensional plane line that divides a plane into two parts, each class on either side. It does an excellent job of distinguishing between the two categories. Random forests, also known as random decision forests, are an ensemble learning technique for classification, regression, and other assignments that works by training a large number of decision trees and yielding the class that is the method of the individual trees' classes (classification). Random decision forests are appropriate for decision trees' proclivity to overfit their training set. There is an immediate association between the combination of trees and the output obtained in a forest of trees. Random forest adds an extra layer of irregularity to stowing to produce increasingly accurate and precise predictions. Furthermore, the third model that is appropriate for classification problems is the Artificial Neural Network (MLP). Using a set of weights, a Multi-Layer Perceptron predicts the class label of tuples. It has several hidden layers as well as an input and output layer. MLP is a directed graph with nodes connected without the input node, and each node is a neuron with a non-linear simulation function. MLP can separate data that cannot be separated linearly [18].

An experiment is conducted on three selected classification models, including SVM, Random Forest, and MLP, to identify the best-performing model in forecasting liver disease patients and studying the most crucial features of liver disease patients. Besides, two sets of data, synthetic and original datasets, are also applied in training the models. The synthetic dataset is generated by applying the MICE algorithm in filling null values and SMOTE algorithm in balancing data sample, while the original dataset is generated by dropping the rows with null values and using the NearMiss algorithm in balancing the sample of data. Finally, the performance of models fitted with these two datasets is compared.

2.0 State of The Art

Machine learning approaches were able to determine which blood donors were healthy and which had liver disease with high accuracy by using the correlation of each attribute with the risk of liver disease to train the model.

A machine learning approach by Mostafa et al. [3] incorporates the significant predictors for liver disease while predicting the liver disease to improve better inference-based diagnosis of patients. Result of comparing binary classifier machine learning algorithms such as support vector machine, random forest (RF), and artificial neural network showed that Random forest contributed to a higher accuracy score with sensitivity, specificity, and accuracy of 0.9904, 0.9729, and 0.9814 compared to other methods. In addition, AST, ALT, GGT, BIL, and ALP were ranked using the Gini index as the five most important risk factors for liver disease diagnosis.

2.1 Support Vector Machine

Vijayarani et al.[19] in their proposed work used Support Vector Machine (SVM) and Naïve Bayes to predict liver disease. Patients liver function test results predict acute hepatitis, cirrhosis, liver cancers, and chronic hepatitis. The dataset was acquired from the UCI repository and includes variables such as Gender, TB, DB, ALP, Sgpt, Sgot, TP, ALB, and A/G Ratio. In terms of accuracy, SVM performs better with the accuracy of 79.66% than Naive Bayes of 61.28% accuracy in predicting acute hepatitis, cirrhosis, liver cancers, and chronic hepatitis from patient liver function test results.

2.2 Random Forest

Ghosh et al.[20] proposed to determine the accuracy of several classification algorithms such as logistic regression, XGBoost, random forest, support vector machine (SVM), AdaBoost, K-NN, and decision tree to predict liver disease. The random forest algorithm outperformed the other algorithms in predicting liver illness with an accuracy, sensitivity, and precision of 83.70%, 93.5%, 87%. And

2.3 Deep learning approach on Artificial Neural Network (MLP)

S. Sontakke et al.[21] proposed a paper to diagnose liver disease using three algorithms Logistic regression, K-NN, SVM, and ANN. The models used Indian Liver Patient Dataset comprised of 10 variables and showed that the backpropagation algorithm developed by Rumelhart and McClelland which are also known as the classic multi-layered neural network algorithm has sensitivity, specificity, and accuracy of 73.3%, 87.7%, and 73.2% as compared to SVM which has sensitivity, specificity, and accuracy of 71.5%, 88.3%, 71%. A convolutional neural network (CNN) model was proposed by Phan and Chan et al.[22] in predicting liver cancer among hepatitis patients in Taiwan. The model shows CNN yields a high accuracy of 0.980. Besides, ANN was also used in predicting liver cancer risk factors for patients with type II diabetes, developed by Rau et al.[23]. The result showed that this model could be an effective predictor with a sensitivity of 0.757. With a specificity of 0.755, 75.70% of diabetic patients can be correctly predicted to obtain a future liver cancer diagnosis.

3.0 Motivation

By referring to the related works conducted by Vijayarani et al., Ghosh et al., and S. Sontakke et al., the machine learning algorithms that outperformed in liver disease prediction were Random Forest, Support Vector Machine (SVM) and Artificial Neural Network (MLP). The question that arises is which algorithm amongst these three algorithms gives the best result in predicting liver disease? Besides, in the study proposed by Mostafa et al., there are several missing values have been found in the liver disease dataset. The method used by researchers in vanishing the null values is Multiple Imputation by Chained Equation (MICE). Furthermore, to solve the problem of unbalanced data, researchers have also implemented the Synthetic Minority Oversampling Technique (SMOTE) function. In both of the MICE and SMOTE methods, synthetic data is generated. Synthetic data is artificial information-created data to serve as replacement data for research purposes. It is very conducive in healthcare machine learning projects as it protects the patients' privacy [24]. However, the synthetic is not exactly the same as the real original data. Thus, the problem is will the original real-world data make a difference in the liver disease prediction? Last but not least, what is the feature that a model relies on the most?

As a further investigation from the previous research, Random Forest, Support Vector Machine (SVM) and Artificial Neural Network (MLP) are utilised in this study. An original real-world

dataset is produced by dropping the missing values are instead of using the mice algorithm, and using undersampling instead of oversampling. The original data is compared with the synthetic data to investigate the difference. Also, the feature importance test is carried out to determine the most important features for the best model in forecasting the presence of liver disease. Then, the performance of each model is evaluated by calculating the accuracy, specificity, sensitivity, precision, and f1 score based on the confusion matrix. To gain more confident and trustworthy results, the model evaluations are looped 100 times with different numbers of random state values in train-test split operation. The mean, standard deviation, and confidence interval of the accuracy, specificity, sensitivity, precision, and f1 score are then calculated to identify the most suitable model in predicting liver disease.

4.0 Methodology

4.1 Dataset Acquisition

The dataset used in this study was obtained from the UCI Machine Learning Repository. The dataset contains 615 observations and 14 variables of blood donors and non-blood donors diagnosed with Hepatitis and some personal information such as sex and age. The variables included in the dataset are shown in the table below. The targeted variable in this dataset is the Category variable with five specific values, 0=Blood Donor, 0s=suspect Blood Donor, 1=Hepatitis, 2=Fibrosis, and 3=Cirrhosis. These values can be classified into two main categories, 0=Blood Donor and 0s=suspect Blood Donor are classified into the group of healthy blood donors. On the other hand, 1=Hepatitis, 2=Fibrosis, and 3=Cirrhosis are a group of non-blood donors diagnosed with liver disease. In addition, the dataset also consists of independent variables such as Age, Sex, ALB, ALP, ALT, AST, BIL, choline esterase (CHE), CREA, cholesterol (CHOL), GGT, and PROT.

Table 1: Variables in the dataset with their definition and data type

Variables	Definition	Data type
Unnamed: 0	Index	Integer
Category	Category of respondent	Object
Age	Age	Integer
Sex	Sex	Object
ALB	Albumin	Float
ALP	Alkaline Phosphatase	Float

ALT	Alanine Aminotransferase	Float
AST	Aspartate Aminotransferase	Float
BIL	Total Bilirubin	Float
CHE	Choline Esterase	Float
CHOL	Cholesterol	Float
CREA	Creatine	Float
GGT	Gamma Glutamyl-transferase	Float
PROT	Total Protein	Float

4.2 Data Pre-processing

A crucial stage in machine learning. The quality of data and the reliability of its information directly impact our model's ability to learn. Hence, pre-processing data before feeding it into our model is necessary to improve the efficiency and accuracy of a machine learning model. The dataset obtained from the UCI Machine Learning Repository was first loaded into a data frame using the read_csv(). Then, the head() function was used to print the first five rows of the data frame. The purpose of printing figure 1 was to show an overview of the dataset. Figure 1 shows that the dataset includes 14 attributes. It helps the researchers have a quick check about the data type's suitability.

Out[3]:

	Unnamed: 0	Category	Age	Sex	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT
0	1	0=Blood Donor	32	m	38.5	52.5	7.7	22.1	7.5	6.93	3.23	106.0	12.1	69.0
1	2	0=Blood Donor	32	m	38.5	70.3	18.0	24.7	3.9	11.17	4.80	74.0	15.6	76.5
2	3	0=Blood Donor	32	m	46.9	74.7	36.2	52.6	6.1	8.84	5.20	86.0	33.2	79.3
3	4	0=Blood Donor	32	m	43.2	52.0	30.6	22.6	18.9	7.33	4.74	80.0	33.8	75.7
4	5	0=Blood Donor	32	m	39.2	74.1	32.6	24.8	9.6	9.15	4.32	76.0	29.9	68.7

Figure 1:First five rows in data frame

To have a more statistical understanding of the dataset, the describe() function was applied to the data frame to shows the mean, standard deviation, quartiles, the smallest value, and the largest value of the numerical variables. As shown in figure 2, some simple statistical information could be retrieved by the researcher. For example, the age of the respondents generally fell between 19

and 77 years old. Besides, the presence of missing data was discovered as a result of some of the variables with different total counts of value. For categorical variables, the summary statistics show the count of values, unique values, top and frequency of occurrence. Figure 3 depicts the majority of the observations were classified as blood donors, while the majority of the respondents were females.

Out[4]:

	Unnamed: 0	Age	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT
count	615.000000	615.000000	614.000000	597.000000	614.000000	615.000000	615.000000	615.000000	605.000000	615.000000	615.000000	614.000000
mean	308.000000	47.408130	41.620195	68.283920	28.450814	34.786341	11.396748	8.196634	5.368099	81.287805	39.533171	72.044137
std	177.679487	10.055105	5.780629	26.028315	25.469689	33.090690	19.673150	2.205657	1.132728	49.756166	54.661071	5.402636
min	1.000000	19.000000	14.900000	11.300000	0.900000	10.600000	0.800000	1.420000	1.430000	8.000000	4.500000	44.800000
25%	154.500000	39.000000	38.800000	52.500000	16.400000	21.600000	5.300000	6.935000	4.610000	67.000000	15.700000	69.300000
50%	308.000000	47.000000	41.950000	66.200000	23.000000	25.900000	7.300000	8.260000	5.300000	77.000000	23.300000	72.200000
75%	461.500000	54.000000	45.200000	80.100000	33.075000	32.900000	11.200000	9.590000	6.060000	88.000000	40.200000	75.400000
max	615.000000	77.000000	82.200000	416.600000	325.300000	324.000000	254.000000	16.410000	9.670000	1079.100000	650.900000	90.000000

Figure 2: Summary statistics of the numerical variables

	Category	Sex
count	615	615
unique	5	2
top	0=Blood Donor	m
freq	533	377

Figure 3: Summary statistics of the categorical variables

After overlooking the data, some changes have been made to the data frame. The column of Unnamed: 0 was removed since it did not have any specific meaning. Apart from that, the values of the Category variable were recategorized into 0 and 1 where 0 indicates healthy blood donors, and 1 indicates non-blood donors diagnosed with liver disease. This variable Category was used as the target variable in prediction. Besides, an encoding algorithm was also applied to the Sex variable to convert objects into integers for machine learning purposes.

4.3 Visualization and Analysis

The first analysis conducted was categorical variable analysis. A pie chart of the percentage of category types was drawn in this analysis. Figure 4 shows the number of healthy blood donors stood at 87.8%, while the liver disease case stood at 12.2%.

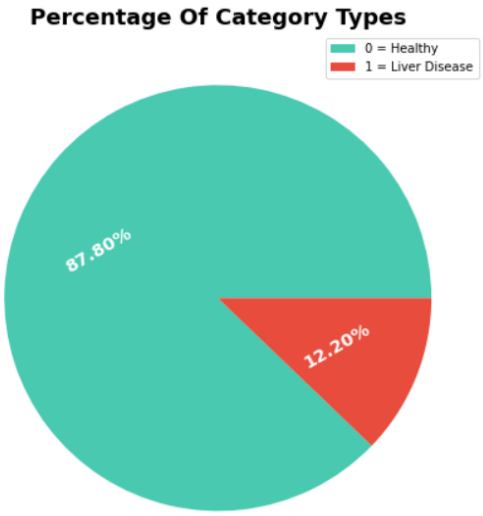


Figure 4: Pie chart of the percentage of category types

Next, another pie chart was sketched to show the percentage of sex types. Figure 5 represents the groups of gender involved in the data, 61.3% of males and 38.7% of females.

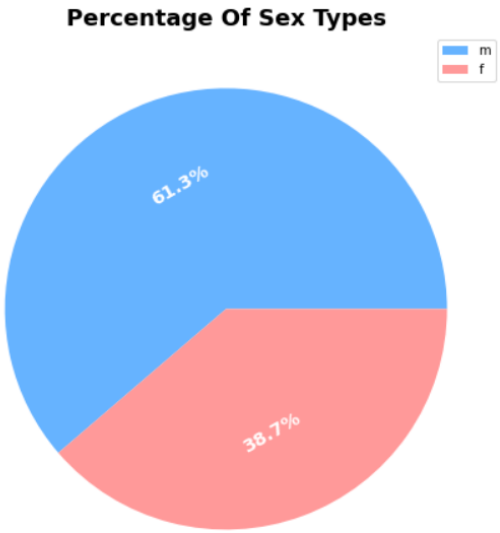


Figure 5: Pie chart of the percentage of sex types

After having some clues about the proportion of variable Category and Sex, the count of healthy blood donors and liver disease patients in each gender group was compared by plotting the categorical plot the figure 6. It shows that male liver disease patients were higher than females.

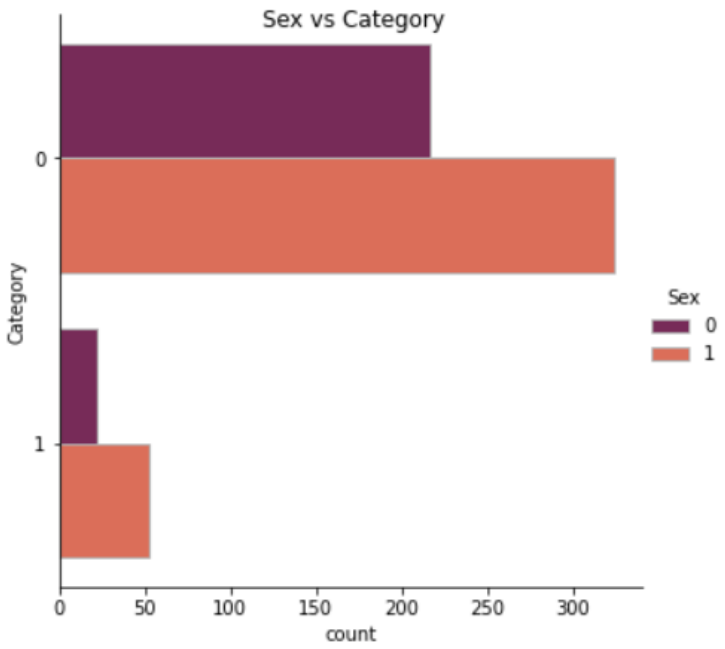


Figure 6: Categorical plot of variable Sex and Category

4.4 Exploratory Data Analysis(EDA)

EDA is a commonly used analysis technique by data scientists in analyzing and investigating different kinds of datasets. It outlines the essential characteristics and delineates the visualizations of the data. The main purpose of using this method is to help the scientists find the anomalies, observe the correlation between the variables, conduct meaningful hypotheses, and uncover the data patterns. It has been proved that the EDA technique performs greatly in analyzing medical field data. [25]

For numerical variables, a boxplot grouped by category was plotted for each variable as depicted in figure 7. The boxplot showed that each variable did not show an apparent relationship with liver disease independently except the AST, BIL, and GGT. The percentage of liver disease was higher if the person contained AST, BIL, or GGT in their body. The apparent difference is shown in the figure showing that the person will get liver disease if a high amount of the element has been found in their body, excluding the outliers. It showed that three of them were the significant variables that will affect the percentage of getting liver disease. However, all the other variables were also assumed to be important in determining liver disease as each variable was correlated with liver disease.

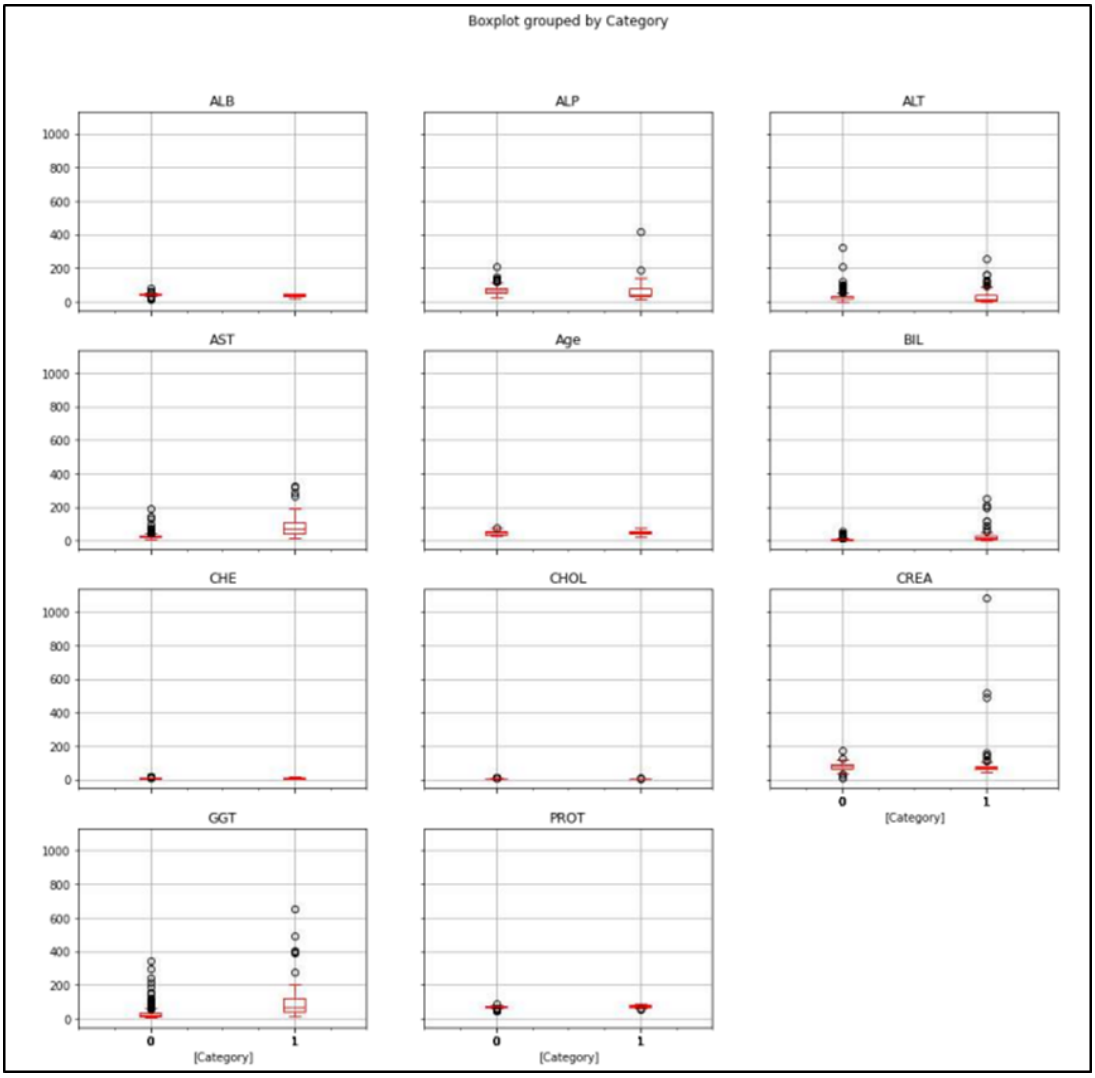


Figure 7: Each variable was visualized using a boxplot grouped by category.

A heatmap was constructed to define the correlation of the variables. In figure 8, the heatmap showed that AST, BIL, and GGT had the highest correlation value among the variables. The close to 1 the correlation is, the more positively correlated. Thus, this showed that three of them were the significant variables that will affect the percentage of getting liver disease. These three variables had a stronger relationship with liver disease.

Figure 8: Each variable was visualized using a heatmap

4.5 Multi Imputation Chained Equation (MICE) and dropna() Function for Missing Values

Firstly, the isna() function was used to check the missing values in the dataset. It has shown that there was 1 missing value in ALB, 18 in ALP, 1 in ALT, 10 in CHOL, and 1 in PROT, as shown in figure 9. To better understand the missing values, the matrix() function from the missingno library was used to visualize the missing data in the dataset in figure 10.

```
Out[15]: Category     0
         Age          0
         Sex          0
         ALB          1
         ALP         18
         ALT          1
         AST          0
         BIL          0
         CHE          0
         CHOL        10
         CREA         0
         GGT          0
         PROT         1
         dtype: int64
```

Figure 9: Missing values of each variable

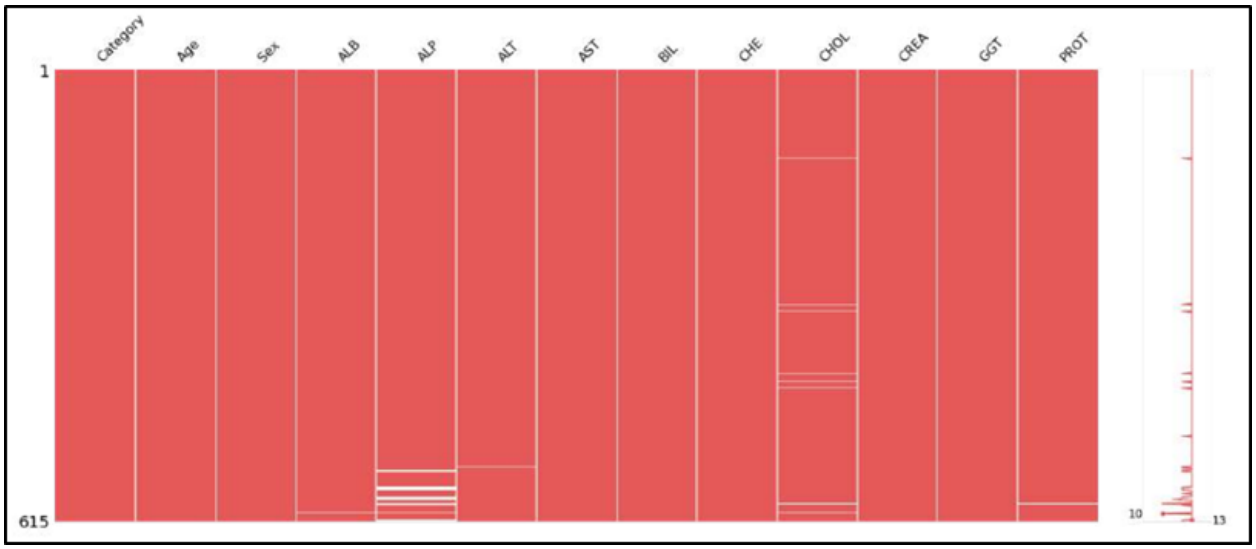


Figure 10: Missing values were visualized by matrix() function

To generate two different sets of data —-synthetic dataset and original dataset, different methods were used to deal with the missing values in each set of data. Firstly, the MICE imputation algorithm-generated synthetic data to replace the null values [26]. The null values were simply dropped away from the dataset for original data.

After solving the problems of the null value, check the data frame for the existence of missing data.

```
Out[49]: Category    0
         Age         0
         Sex         0
         ALB         0
         ALP         0
         ALT         0
         AST         0
         BIL         0
         CHE         0
         CHOL        0
         CREA        0
         GGT         0
         PROT        0
         dtype: int64
```

Figure 11: Missing values of each variable after solving missing values.

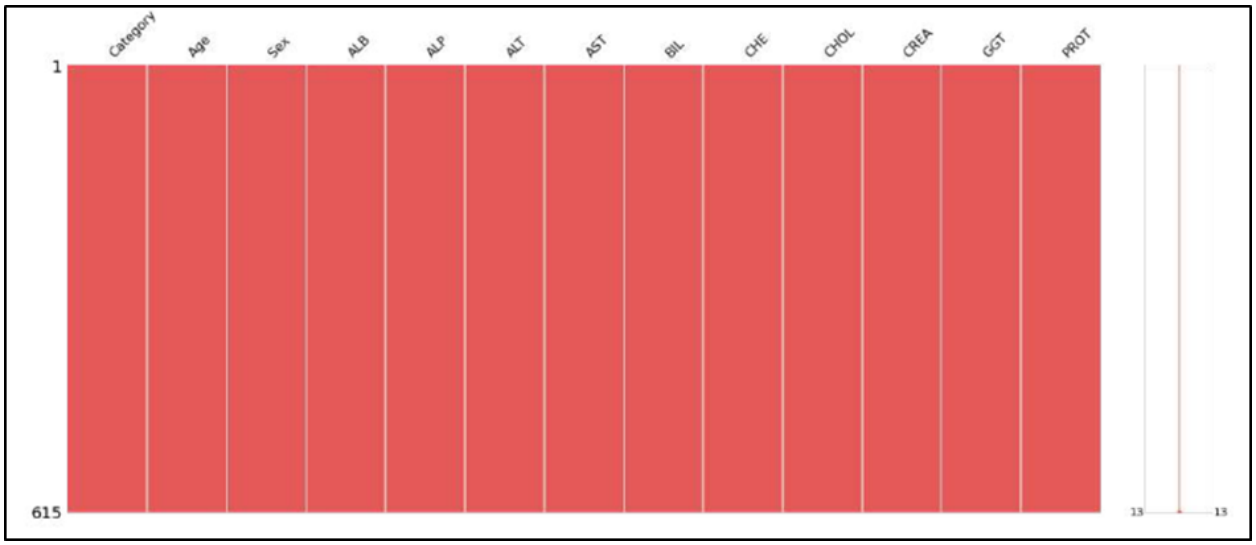


Figure 12: Missing values were visualized after replacing null values by MICE technique

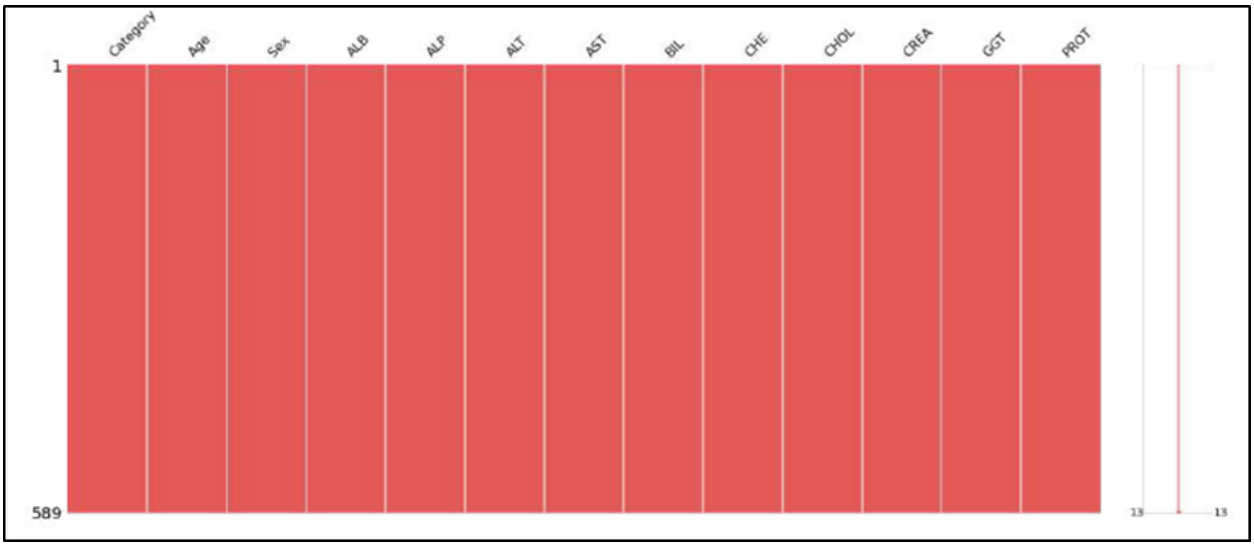


Figure 13: Missing values were visualized by matrix() function after applying dropna() function

4.6 Training and Testing Data

In supervised machine learning, the samples of training data must be selected in order to feed the machine learning models. According to Huang, C. et al. [27], the size of the training is a critical factor in machine learning as it affects the performance of the classification models. The study proposed by Huang, C. et al. conclude that a large training sample is a key to getting good accuracy in classification. 80 percent of the training sample was obtained from the dataset in this research. The other 20 percent was used as testing data.

A sub-set of the training sample was used to validate the model. Since the binary classification methods were applied in this research, the training data was used to train the machine learning classifier in predicting the liver disease patient, while the testing data was used to evaluate the performance of the classifier.

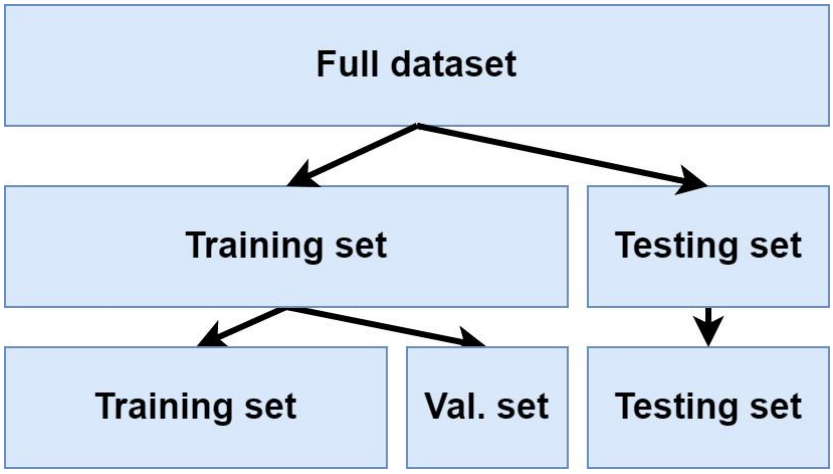


Figure 14: Train-test split of data sampling

4.7 Balancing Data Using Oversampling and Undersampling

The histograms have been constructed to visualize the unbalanced data in the dataset. The data imbalance problem will lead to inaccurate model performance and overfitting problems. Figure 15 shows the data imbalance problem in both synthetic (processed by MICE) and original data (drop NaN). The number of data labeled with liver disease patients was smaller. Thus, the poor performance of the classifier prediction will occur. To avoid this problem, oversampling and undersampling techniques must be applied to balance the data.

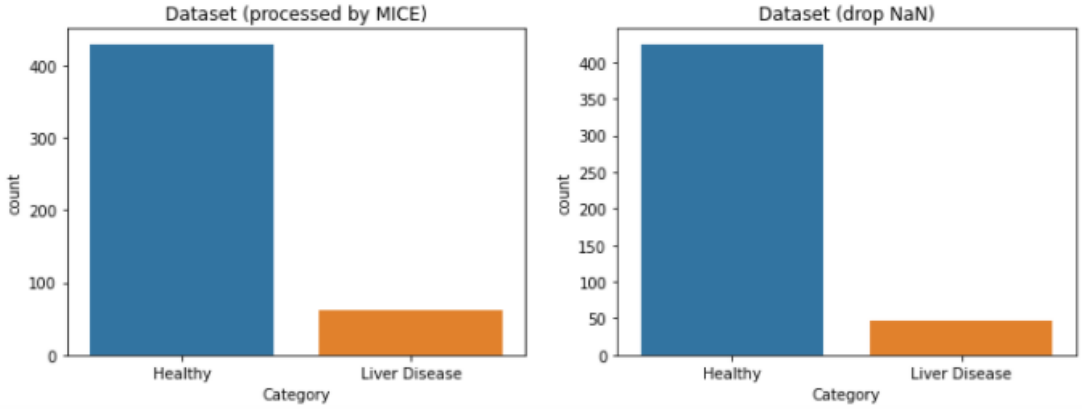


Figure 15: Histograms show balanced data problems in the synthetic dataset (processed by MICE) and original dataset (drop NaN).

As mentioned in the previous subsection, synthetic data generated by the AI technique serves as replacement data for research purposes [24]. The Synthetic Minority Oversampling Technique (SMOTE) was applied to the synthetic data to generate balanced data. SMOTE uses the nearest neighbor algorithm to produce a bunch of new data used for the training model. After applying SMOTE, 368 synthetic data was produced to fill the minority class. Then, both of the healthy and liver disease classes were balanced as shown in Figure 16. In contrast, the oversampling method was not suitable for original data. Since the original data must not contain any auto-generated values, the NearMiss undersampling method would be better. The NearMiss undersampling was used to cut off the 379 surplus data of the majority class to balance the data.

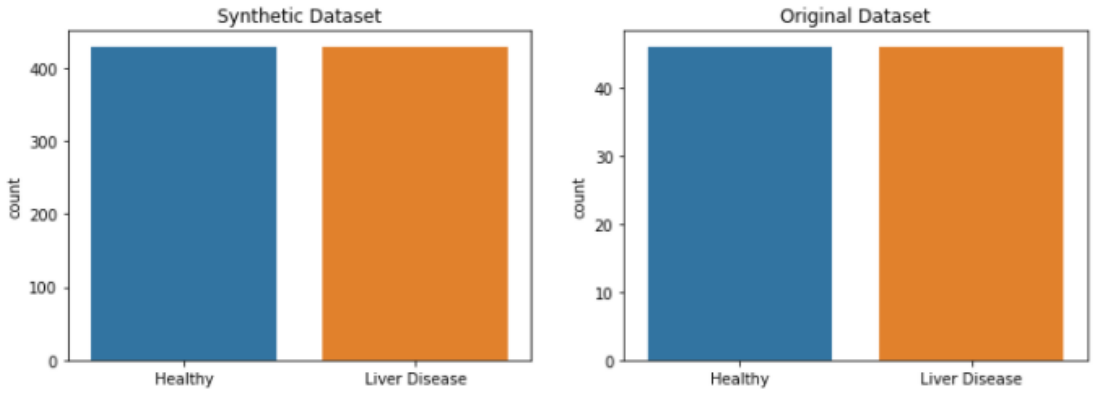


Figure 16: Histograms show synthetic dataset and original dataset.

4.8 Support Vector Machine Classification

SVM is a relatively new machine learning algorithm that was oriented in 2000. Its identical solid performance when dealing with messy and noisy data is why many data scientists prefer it [28]. A Support Vector Classifier will divide the data into two groups to classify and build a hyper-plane that best divides the class.

SVM will choose the best hyper-plane to separate the vectors of two classes. A vector is a collection of attributes that characterize a single case. The algorithm of SVM will recursively generate many different hyperplanes. The best hyperplane would be chosen to classify the vectors [29]. In figure 17, the dots located on both sides of the line of the best hyperplane are vectors. In the study of liver disease, circle dots indicate healthy blood donors, while triangular dots indicate liver disease patients.

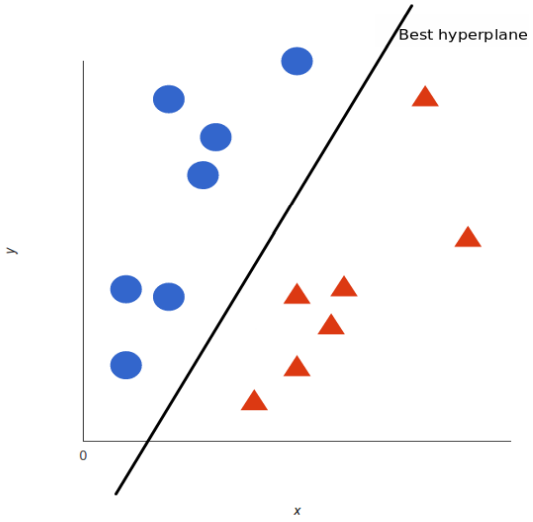


Figure 17: Illustration of SVM

4.9 Random Forest Classification

Random Forest is a supervised machine learning algorithm proposed by Leo Breiman in 2001. [30] The random forest classification was widely used in medical research [20]. The algorithm constructs multiple decision trees at training time to get accurate predictions. It is developed with random feature selection and bagging for avoiding the trees constructed from very close to each other. Since the random forest generates many trees when training the model, it is suitable for handling large sizes of data.

In the case of liver disease prediction, the training data were fed into different decision trees. The trees will split their nodes randomly by selecting those attributes and observations in the training sample. The best tree was voted by the algorithm from the forest of trees. The prediction result was produced by following the majority-voting tree.

4.10 Artificial Neural Network Classification (Multi-Layer Perceptron)

Multi-Layer Perceptron(MLP) is the simplest type of artificial neural network. It is a combination of multiple perceptron models[31]. Perceptrons in MLP are extensively linked and parallel. This parallelization aids in the speeding up of computations.

Perceptron was introduced by Frank Rosenblatt in 1950[31]. It is like the human brain, thus it is capable of learning complex things. Sensory Unit (Input Unit), Associator Unit (Hidden Unit), and Response Unit (Output Unit) will make up a perceptron network as shown in figure 18. The input layer in figure 18 might represent the variables of the dataset and the output layer might represent whether the person has liver disease diagnosed. The Multi-Layer Perceptron iteratively trains the model. The parameters were updated using partial derivatives of the loss function in each iteration cycle. This iteration occurs in the hidden layers.

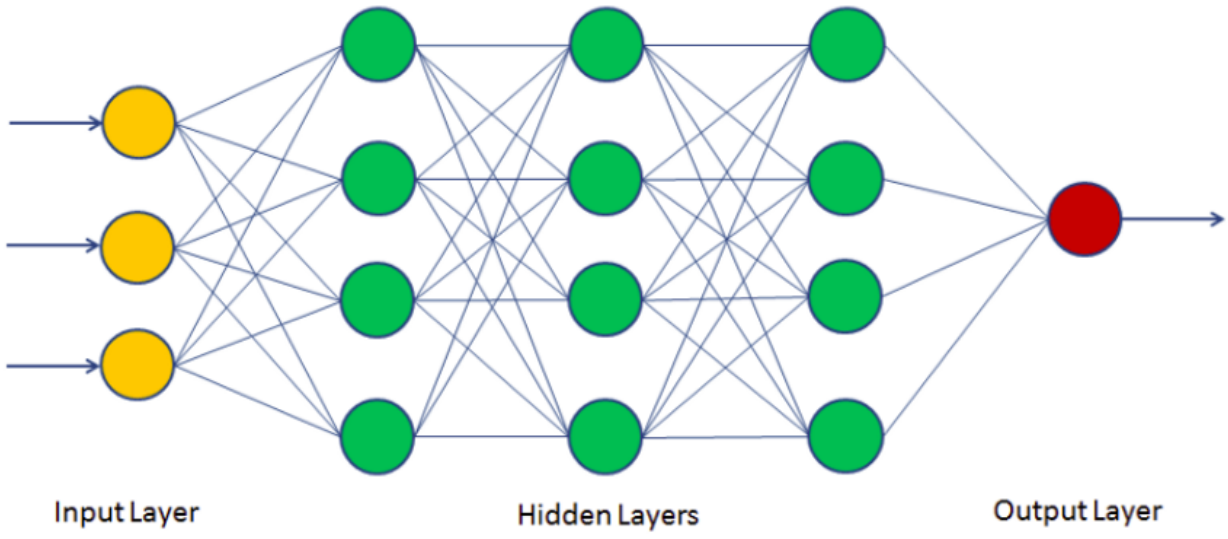


Figure 18: Illustration of perceptron network

4.11 Model Evaluation

After performing data pre-processing, solving data imbalance problems, and solving missing value problems, the models of SVM, RF, and MLP were fed with both the synthetic dataset and original dataset. In this step, each model will perform supervised learning. Then, the performances of these models were evaluated using a confusion matrix. Figure 19 shows the confusion matrix, in which the TP indicates correctly predicted healthy observation, TN indicates accurately predicted liver disease cases. The FP represents falsely classified healthy cases, and FN represents falsely classified liver disease cases.

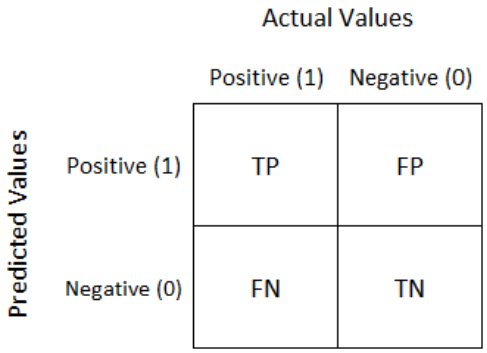


Figure 19: Confusion Matrix

Firstly, the accuracy was calculated to measure the ability of the models to accurately identify the observations of healthy blood donors and liver disease patients.

$$Accuracy = (TP + TN)/(TP + TN + FP + FN)$$

Secondly, the precision was also calculated to measure the ability of the models to discriminate between positive and negative cases.

$$Precision = TP/(TP + FP)$$

Moreover, the sensitivity and specificity were also calculated to measure the ability of the model to classify patients with liver disease and without liver disease.

$$Sensitivity = TP/(TP + FN)$$

$$Specificity = TN//(TN + FP)$$

$$F1 = (2 \times Precision \times Recall)/(Precision + Recall)$$

After evaluating the models, the model fed with the original dataset was looped up to 100 times and performed train-test split with selecting the random state from 0 to 100. The model training and model testing were conducted in each iteration, and the accuracy, precision, sensitivity, specificity, and F1 score were calculated. Then, the mean, standard deviation, and confidence interval of the accuracy, precision, sensitivity, specificity, and F1 score were also calculated.

5.0 Result

The performance of the 3 machine learning models in this paper was evaluated by using the receiver operating characteristic curve (ROC curve) and the confusion matrix. From the confusion matrix, 5 performance results were obtained which are accuracy, sensitivity, specificity, precision, and f1-score. Synthetic data and original data were generated from the Liver Disease Dataset to train and test the models.

5.1 Synthetic Data Performance

Support Vector Classifier:

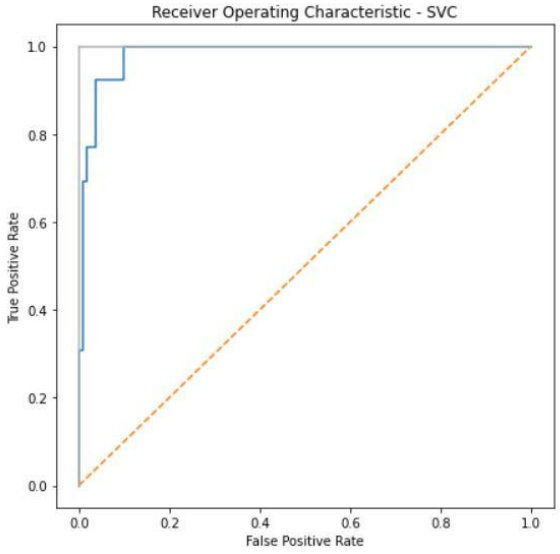


Figure 20: ROC curve for SVC

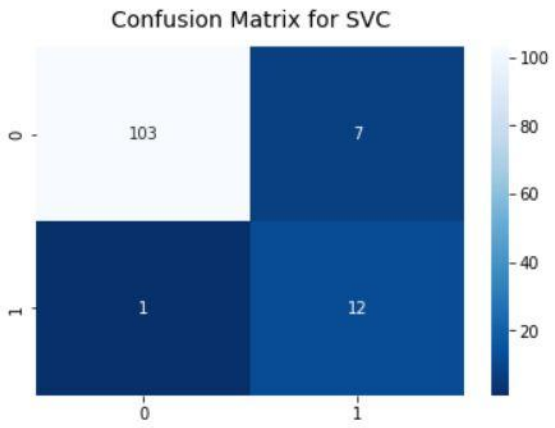


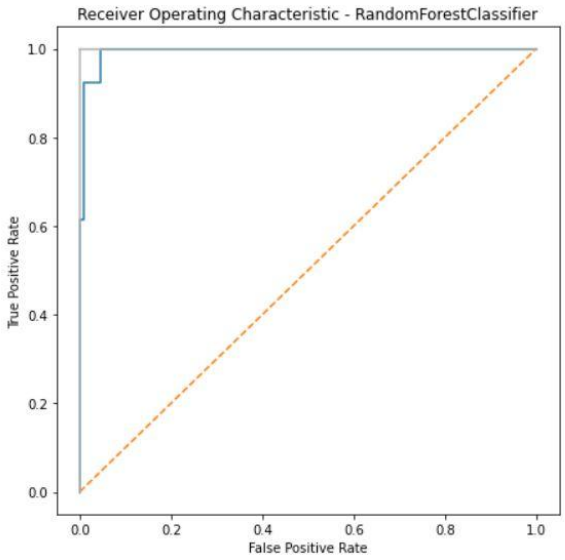
Figure 21: Confusion matrix for SVC

According to Figure 20, the ROC area under curve score of the support vector classifier model is 0.9818. From Figure 21, the accuracy of the training data is 93.95% and the accuracy of the testing data is 93.49%. The performances of SVC prediction are shown in table 2.

Table 2: Performance of SVC prediction

Performance of SVC prediction (%)	
Accuracy	93.49
Sensitivity	99.03
Specificity	63.15
Precision	93.63
F1 Score	96.26

Random Forest Classifier:



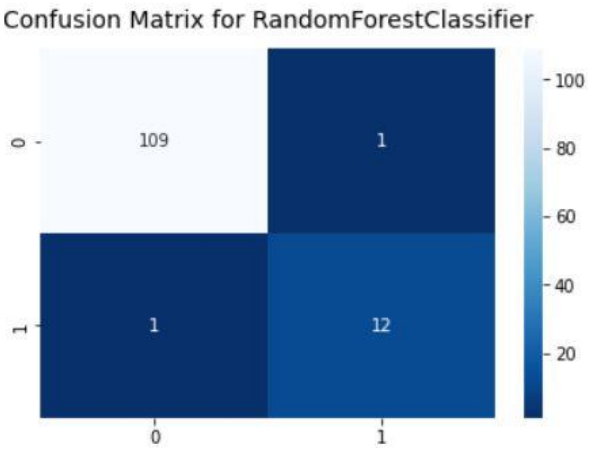


Figure 22: ROC curve for RF

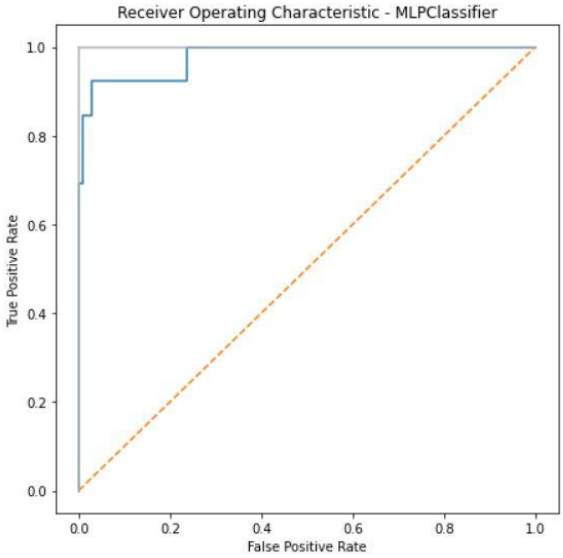
Figure 23: Confusion matrix for RF

According to Figure 22, the ROC area under curve score of the random forest classifier model is 0.9937. From Figure 23, the accuracy of the training data is 100% and the accuracy of the testing data is 98.37%. The performances of SVC prediction are shown in table 3.

Table 3: Performance of RF prediction

Performance of RF prediction (%)	
Accuracy	98.37
Sensitivity	99.09
Specificity	92.30
Precision	99.09
F1 Score	99.09

Artificial Neural Network (Multi-Layer Perceptron):



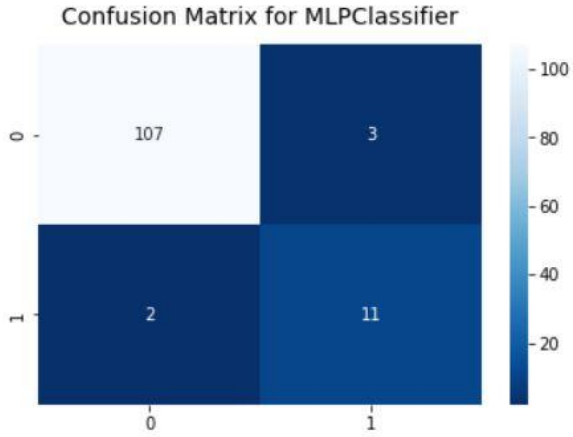


Figure 24: ROC curve for MLP

Figure 25: Confusion matrix for MLP

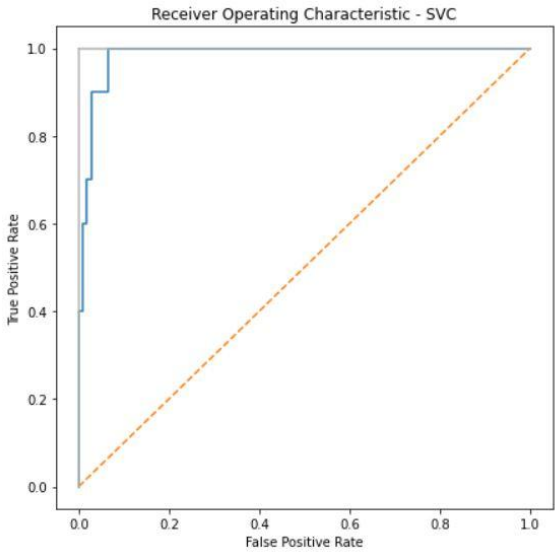
According to Figure 24, the ROC area under curve score of the multi-layer perceptron classifier model is 0.9783. From Figure 25, the accuracy of the training data is 98.60% and the accuracy of the testing data is 95.93%. The performances of MLP prediction are shown in table 4.

Table 4: Performance of MLP prediction

Performance of MLP prediction (%)	
Accuracy	95.93
Sensitivity	98.16
Specificity	78.57
Precision	97.27
F1 Score	97.71

5.2 Original Data Performance

Support Vector Classifier:



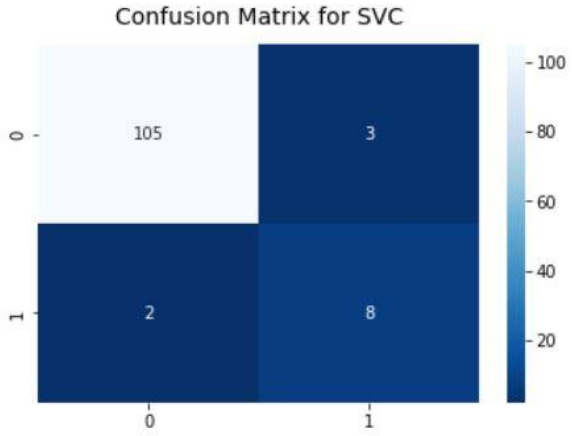


Figure 26: ROC curve for SVC

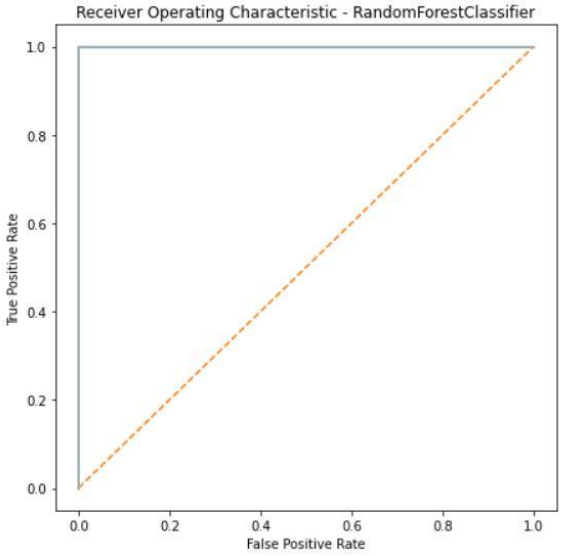
Figure 27: Confusion matrix for SVC

According to Figure 26, the ROC area under curve score of the support vector classifier model is 0.9842. From Figure 27, the accuracy of the training data is 90.21% and the accuracy of the testing data is 95.76%. The performances of SVC prediction are shown in table 5.

Table 5: Performance of SVC prediction

Performance of SVC prediction (%)	
Accuracy	95.76
Sensitivity	98.13
Specificity	72.72
Precision	97.22
F1 Score	97.67

Random Forest Classifier:



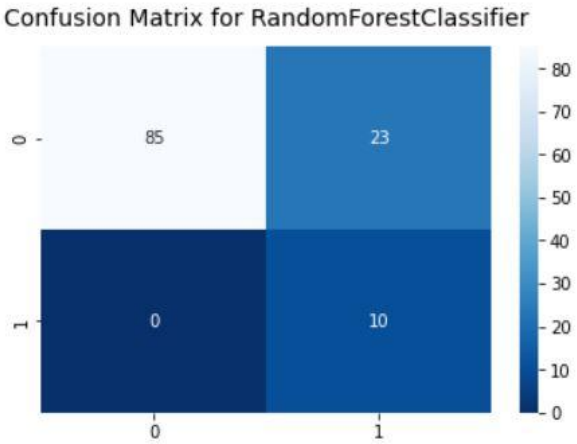


Figure 28: ROC curve for RF

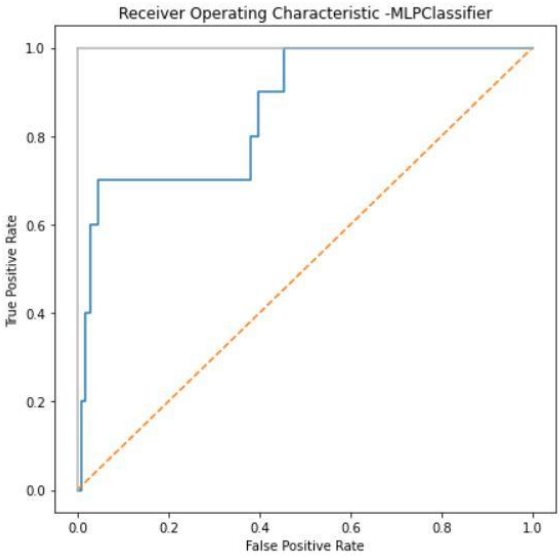
Figure 29: Confusion matrix for RF

According to Figure 28, the ROC area under curve score of the random forest classifier model is 1.0. From Figure 29, the accuracy of the training data is 100% and the accuracy of the testing data is 80.50%. The performances of SVC prediction are shown in table 6.

Table 6: Performance of RF prediction

Performance of RF prediction (%)	
Accuracy	80.50
Sensitivity	100.0 0
Specificity	30.30
Precision	78.70
F1 Score	88.08

Neural Network(Multi-Layer Perceptron):



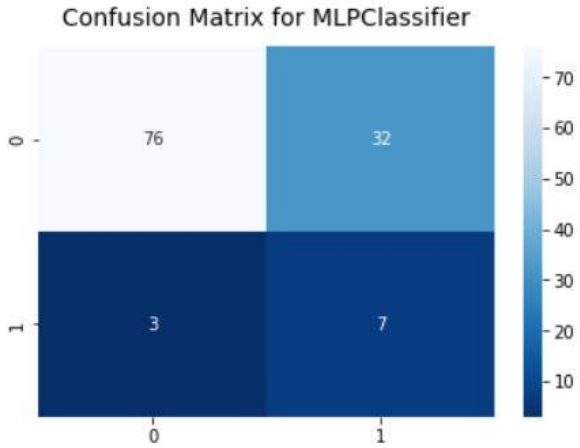


Figure 30: ROC curve for MLP

Figure 31: Confusion matrix for MLP

According to Figure 30, the ROC area under curve score of the multi-layer perceptron classifier model is 0.8611. From Figure 31, the accuracy of the training data is 84.78% and the accuracy of the testing data is 70.33%. The performances of MLP prediction are shown in table 7.

Table 7: Performance of MLP prediction

Performance of MLP prediction (%)	
Accuracy	70.33
Sensitivity	96.20
Specificity	17.94
Precision	70.37
F1 Score	81.28

5.3 Most Important Variable

There are some important variables in the liver disease dataset that can be used to determine the diagnosis of liver disease. The graphs were plotted to compare the most important variable in this dataset.

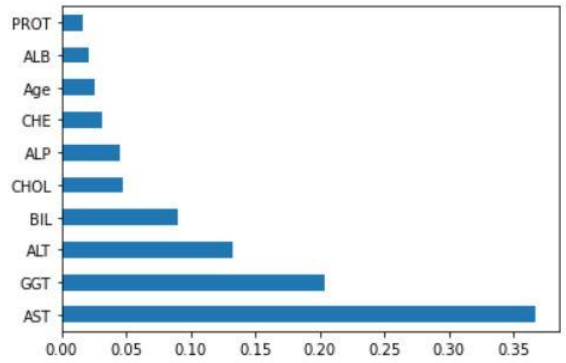


Figure 32: Feature Importance Graph

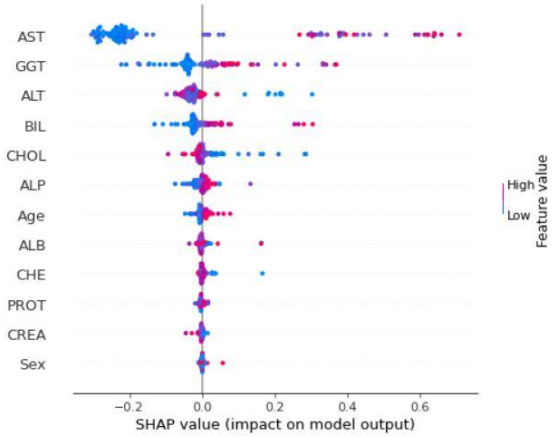


Figure 33: SHAP Feature Importance

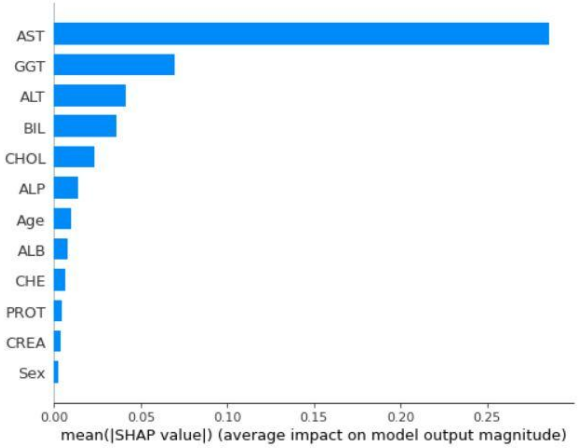


Figure 34: Mean SHAP Value Graph

According to Figure 32, 33 and 34, the graph shows that the most important variable is Aspartate amino-transferase(AST).

5.4 Mean and Standard Deviation of Model's Performance

To study the impact of random state(0-100) on the performance of the model, mean and standard deviation of model's performance were calculated and shown in table 8.

Table 8: The Mean and Standard Deviation of Model's Performance

	Mean					Standard Deviation				
Model	Accuracy	Sensitivity	Specificity	Precision	F1 Score	Accuracy	Sensitivity	Specificity	Precision	F1 Score
SVC	94.79 %	97.81%	68.98%	96.39%	79.90%	0.0207	0.0150	0.1177	0.0154	0.0876
RF	86.22 %	99.66%	40.37%	85.04%	54.34%	0.0395	0.0055	0.0835	0.0454	0.0827
MLP	73.66 %	95.59%	21.16%	74.43%	32.52%	0.0401	0.0216	0.0631	0.0481	0.0789

5.5 Confidence Interval of Model's Performance

To determine the best performance between 3 models, the size of 95% confidence intervals of 3 models were calculated and shown in table 9.

Table 9: The 95% Confidence Interval of Model's Performance

	Model								
	SVC			RF			MLP		
Performance	Confidence Interval		Difference	Confidence Interval		Difference	Confidence Interval		Difference
Accuracy	0.9439	0.9520	0.0081	0.8544	0.8699	0.0155	0.7288	0.7445	0.0157
Sensitivity	0.9752	0.9811	0.0059	0.9955	0.9977	0.0022	0.9517	0.9602	0.0085
Specificity	0.6667	0.7128	0.0461	0.3874	0.4201	0.0327	0.1992	0.2240	0.0248
Precision	0.9609	0.9670	0.0061	0.8415	0.8593	0.0178	0.7349	0.7537	0.0188
F1 Score	0.7818	0.8162	0.0344	0.5272	0.5596	0.0324	0.3097	0.3407	0.0310

6.0 Discussion of Result

Table 10: The AUC Score of the Model.

	Model					
	Synthetic Data			Real World Data		
	SVC	RF	MLP	SVC	RF	MLP
AUC	0.9818	0.9937	0.9783	0.9842	1.0	0.8611

In this paper, each model was plotted the Receiver Operator Characteristic (ROC) curve to determine the better the performance of the model at distinguishing between the positive and negative classes [32]. The ROC curve is a binary classification problem evaluation metric. It's a probability curve that plots the TPR against the FPR at different threshold values, effectively separating the 'signal' from the 'noise.' The Area Under the Curve (AUC) is a summary of the ROC curve that measures a classifier's ability to distinguish between classes [33]. From table 10, the AUC score of all models is higher than 0.5. These show that there's a good chance the classifier will be able to tell the difference between positive and negative class values. Because the classifier can detect more True positives and True negatives than False negatives and False positives.

Table 11: The Accuracy of the Model.

	Model					
	Synthetic Data			Real World Data		
	SVC	RF	MLP	SVC	RF	MLP
Accuracy(%)	93.49	98.37	95.93	95.76	80.50	70.33
Sensitivity(%)	99.03	99.09	98.16	98.13	100.00	96.20
Specificity(%)	63.15	92.30	78.57	72.72	30.30	17.94
Precision(%)	93.63	99.09	97.27	97.22	78.70	70.37
F1 Score(%)	96.26	99.09	97.71	97.67	88.08	81.28

By comparing the model performance when using synthetic data, the Random Forest Classifier has the highest accuracy of 98.37%. While using the original data, the highest accuracy model is Support Vector Classifier with accuracy 95.76%. However the SVC model has the highest accuracy when using original data, but the SVC model has an overfitting problem. This problem can be observed from the training accuracy and the testing accuracy of the SVC model. The training accuracy of the SVC model is 90.21% which was lower than the testing accuracy of the model, 95.76%. When a model learns the detail and noise in the training data to the point where it degrades the model's performance on new data, this is known as overfitting [34]. This means that the model picks up on noise or random fluctuations in the training data and learns them as

concepts. Each model in this paper was trained by using the synthetic data and original data (real-world data). From table 11, the models' accuracy for the synthetic data is higher than the original data. This is because synthetic data is data that can be created at any scale, at any time, and in any location. Synthetic data, on the other hand, closely resembles the balance and composition of real data, making it ideal for training machine learning models.

The metric that measures a model's ability to predict true positives in each available category is called sensitivity. The Random Forest model has the highest sensitivity(100%) while using real world data. Therefore, this model has the highest sensitivity of a test to correctly identify patients with a disease. The metric that measures a model's ability to predict true negatives in each available category is called specificity. The Random Forest model has the highest specificity(99.09%) while using the synthetic data. Therefore, this model has the highest specificity of a test to correctly identify people without the disease. Precision is one measure of a machine learning model's performance – the accuracy of a model's positive prediction. The number of true positives divided by the total number of positive predictions is known as precision. The Random Forest model has the highest precision (99.09%) in the synthetic data. Therefore, this model has the highest precision when predicting the patient has liver disease. F1 score is a metric for how accurate a model is on a given dataset. F1 score is used to assess binary classification systems that divide examples into 'positive' and 'negative' categories. The Random Forest model has the highest F1 score (99.09%) while using synthetic data [35]. By comparing all the model performances, the Random Forest model (in synthetic data) has the best overall performance. Therefore, this model is the best model to predict liver disease.

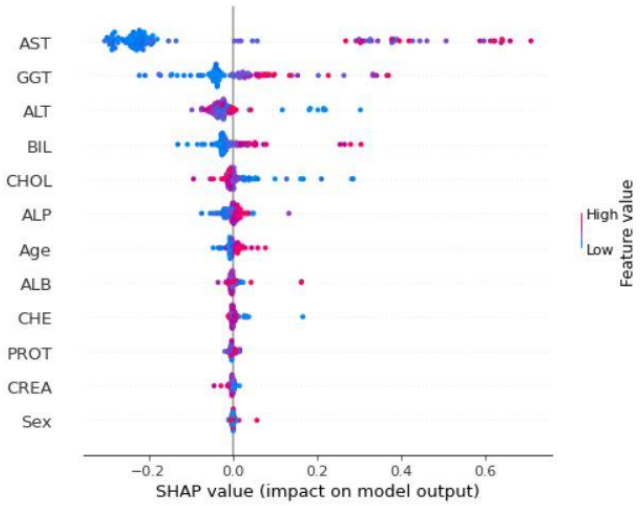


Figure 35: SHAP Feature Importance

The importance of SHAP features is an alternative to the importance of permutation features [36]. Both importance measures have a significant difference: The importance of the permutation feature is determined by the decrease in model performance. The SHAP algorithm is based on the size of feature attributions. The summary plot combines the importance of features with the effects of features. A Shapley value for a feature and an instance is represented by each point on the summary plot. The feature determines the position on the y-axis, while the Shapley value determines the position on the x-axis. The color represents the feature's value, which ranges from low to high. Overlapping points are jittered in the y-axis direction to give us a sense of the Shapley

value distribution per feature. The features are ranked in order of importance. From figure 35, Aspartate amino-transferase(AST) is the most important variable in the liver disease dataset.

Table 12: The Mean and Standard Deviation of Model's Performance

	Performance									
	Mean					Standard Deviation				
Model	Accuracy	Sensitivity	Specificity	Precision	F1 Score	Accuracy	Sensitivity	Specificity	Precision	F1 Score
SVC	94.79 %	97.81%	68.98%	96.39%	79.90%	0.0207	0.0150	0.1177	0.0154	0.0876
RF	86.22 %	99.66%	40.37%	85.04%	54.34%	0.0395	0.0055	0.0835	0.0454	0.0827
MLP	73.66 %	95.59%	21.16%	74.43%	32.52%	0.0401	0.0216	0.0631	0.0481	0.0789

After training and testing the model with synthetic data and real-world data, Random Forest is the highest accuracy model. To study the impact of random state from 0 to 100 on the model performance, the mean and standard deviation of the model's performance were calculated. From table 12, the mean of the sensitivity of random forest model has the highest percentage, 99.66%. The standard deviation of the sensitivity of the random forest model has the lowest figure, 0.0055. These show that the random forest model can predict the disease more accurately with the high sensitivity of the model.

Table 13: The Confidence Interval of Model Performance

	Model								
	SVC			RF			MLP		
Performance	Confidence Interval		Difference	Confidence Interval		Difference	Confidence Interval		Difference
Accuracy	0.9439	0.9520	0.0081	0.8544	0.8699	0.0155	0.7288	0.7445	0.0157
Sensitivity	0.9752	0.9811	0.0059	0.9955	0.9977	0.0022	0.9517	0.9602	0.0085
Specificity	0.6667	0.7128	0.0461	0.3874	0.4201	0.0327	0.1992	0.2240	0.0248
Precision	0.9609	0.9670	0.0061	0.8415	0.8593	0.0178	0.7349	0.7537	0.0188
F1 Score	0.7818	0.8162	0.0344	0.5272	0.5596	0.0324	0.3097	0.3407	0.0310

A confidence interval is a set of bounds on a population variable's estimation [37]. It's an interval statistic that's used to quantify the degree of uncertainty in a prediction. In applied machine learning, confidence intervals can be used in the presentation of the skill of a predictive model. The smaller the confidence interval, the more precise estimate. The larger the confidence interval, the less precise estimate.

By comparing the confidence interval of the model performance, the sensitivity of Random Forest has the narrowest confidence interval which is 0.0022 indifference. The sensitivity of the model has the ability of a test to identify patients with a disease correctly.

7.0 Conclusion

In this paper, the correlation of each variable is used to accurately predict blood donors who were healthy and those who had liver disease. The dataset is distributed into two parts: training and testing, with a weighted average of 80% to 20%. The dependent variable is "Category" which includes 0=Blood Donor, 0s=suspect blood donor, 1=Hepatitis, 2=Fibrosis, and 3=Cirrhosis, while the remaining variables are "Age", "Sex", "ALB", "ALP", "ALT", "AST", "BIL", "CHE", "CHOL", "CREA", "GGT", and "PROT". To solve with unbalanced data, an oversampling method by processing the MICE algorithm is used to generate synthetic data while the undersampling method by dropping the column with null values is performed to generate original data. The outputs of synthetic data and original data are compared. Furthermore, classification problem performance metrics such as the ROC curve, confusion matrix, precision, recall, and F1 score are used to compare which model is more effective in predicting liver disease. For three of the models, a random state of 0 to 100 is used to calculate the average mean and standard deviation of the model's accuracy. The project's problem statements of determining which machine learning models perform better accuracy in liver disease diagnosis by comparing the confidence interval and standard deviation, as well as determining the most important attribute in liver disease prediction among all variables, have been met.

To summarize, Random Forest is the best-performing model in this Liver Disease Prediction. As for the synthetic data, it has the highest accuracy of 98.37% in predicting test data and the highest sensitivity of 99.09%. This prediction has the highest specificity of 92.30% and the highest ROC area under curve score of 0.9937. The Area Under the Curve (AUC) is a measure of a classifier's ability to distinguish between classes and is used to summarize the ROC curve. The greater the AUC, the better the model's performance in distinguishing between positive and negative classes. The Random Forest model has the highest precision score and F1 score of 99.09%. In contrast, for the case of the original dataset which includes dropping the column with null values, Random Forest is still considered as the best model although the Support Vector Classifier (SVC) has the highest accuracy of 95.76% when compared to the Random Forest model and Artificial Neural Network model. This is due to the reason that the accuracy of SVC evaluated on training data is 90.22% while the accuracy of SVC evaluated on testing data is 95.76%. Higher accuracy of testing data as compared to the accuracy of training data will lead to an overfitting problem. The overfitting problem occurs because an undersampling method is used to balance the original data in the process of liver disease prediction. As a result, the shape of the training dataset becomes very small and leads to insufficient data in the training model. A lower accuracy of training data is gained. Thus, the Random Forest is considered the best model in the case of the original dataset since it has the second-highest of accuracy which is 80.50%. It has a sensitivity of 100%, specificity of 30.30%, a precision of 78.70%, and an F1 score of 88.08%. From this project, the most important variable such as Aspartate amino-transferase (AST) is identified. By using the heatmap technique, the variable of AST has the highest correlation with the category of patients which is the dependent variable in this liver disease prediction.

In addition, the test on the random state of 0 to 100 for Support Vector Machine (SVM), Random Forest (RF), and Artificial Neural Network (MLP) is carried out to study the impact of the random state on the accuracy of model prediction and to calculate the mean and standard deviation of models' accuracy, specificity, sensitivity, precision, and F1 score. The mean accuracy of the Random Forest Classifier is 86.22% while the standard deviation of accuracy is 3.95%. By using a 95% confidence interval, the confidence interval of accuracy for Random Forest is between 0.8544 and 0.8699. A model with a narrower confidence interval is usually more precise in its estimation.

In the future, more machine learning models which include K-Nearest Neighbor and Logistic Regression models can be used to predict liver disease even in real-world situations. As a result, more algorithms can be selected to build an increasingly precise model of liver disease

prediction, and the model's accuracy and performance can be incrementally improved. In addition, instead of binary classification, multinomial classification of separating the types of liver disease can be carried out in liver disease prediction [3]. By employing this method, it is possible to create a statistical record of the types of liver disease that are most prevalent in society, and the diseases can be treated as soon as possible in order to reduce the death rate caused by liver disease. Besides, a web application that allows medical professionals to input various liver functioning data that can determine whether a person is suffering from liver disease or not based on the prediction algorithm can be created [38]. This can shorten the time needed for a medical health report to be created.

In a nutshell, advances in data science and machine learning have provided a powerful tool for forecasting the future of the liver disease. Data science and machine learning have been applied to real-world data such as medical images, blood test results, and genetic data. This has resulted in the development of better and more accurate models for forecasting the future of liver disease, which will have a significant impact on society in the future. These models have been used to forecast the likelihood of developing cirrhosis, liver cancer, and liver failure. This means that a patient can be diagnosed earlier and has a better chance of survival in the long run.

8.0 Reference

[1] A.A. Mokdad, A.D. Lopez, S. Shahraz, R. Lozano, A.H. Mokdad, J. Stanaway, et al, Liver cirrhosis mortality in 187 countries between 1980 and 2010: a systematic analysis. BMC Med 2014; 12:145.

[2] Byass, Peter, The global burden of liver disease: a challenge for methods and for public health. BMC medicine 12.1 (2014); 159.

[3] Mostafa, F.; Hasan, E.;Williamson, M.; Khan, H. Statistical Machine Learning Approaches to Liver Disease Prediction. Livers 2021,1, 294–312. https://doi.org/10.3390/livers1040023

[4] Performance Evolution of Different Machine Learning Algorithms for Prediction of Liver Disease. (2019). International Journal of Innovative Technology and Exploring Engineering, 9(2), 1115–1122. https://doi.org/10.35940/ijitee.l3619.129219

[5] Torkadi, P. P., Apte, I. C., & Bhute, A. K. (2013). Biochemical Evaluation of Patients of Alcoholic Liver Disease and Non-alcoholic Liver Disease. Indian Journal of Clinical Biochemistry, 29(1), 79–83. https://doi.org/10.1007/s12291-013-0310-7

[6] Ceriotti, F., Henny, J., Queraltó, J., Ziyu, S., Özarda, Y., Chen, B., … Panteghini, M. (2010). Common reference intervals for aspartate aminotransferase (AST), alanine aminotransferase (ALT) and γ-glutamyl transferase (GGT) in serum: results from an IFCC multicenter study. Clinical Chemistry and Laboratory Medicine, 48(11). https://doi.org/10.1515/cclm.2010.315

[7] Chalasani, N., Younossi, Z., Lavine, J. E., Charlton, M., Cusi, K., Rinella, M., … Sanyal, A. J. (2017). The diagnosis and management of nonalcoholic fatty liver disease: Practice guidance from the American Association for the Study of Liver Diseases. Hepatology, 67(1), 328–357. https://doi.org/10.1002/hep.29367

[8] BORRONI, G., CERIANI, R., CAZZANIGA, M., TOMMASINI, M., RONCALLI, M., MALTEMPO, C., … SALERNO, F. (2006). Comparison of simple tests for the non-invasive diagnosis of clinically silent cirrhosis in

chronic hepatitis C. Alimentary Pharmacology and Therapeutics, 24(5), 797–804. https://doi.org/10.1111/j.1365-2036.2006.03034.x

[9] Asrani, S. K., Devarbhavi, H., Eaton, J., & Kamath, P. S. (2019). Burden of liver diseases in the world. Journal of hepatology, 70(1), 151-171, doi:https://doi.org/10.1016/j.jhep.2018.09.014

[10] Nilsson, N. J. (1996). Introduction to machine learning. An early draft of a proposed textbook.

[11] Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., ... & Lee, S. I. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nature biomedical engineering, 2(10), 749-760.

[12] Kuo, C. C., Chang, C. M., Liu, K. T., Lin, W. K., Chiang, H. Y., Chung, C. W., ... & Chen, K. T. (2019). Automation of the kidney function prediction and classification through ultrasound-based kidney imaging using deep learning. NPJ digital medicine, 2(1), 1-9.

[13] Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health affairs, 33(7), 1123-1131.

[14] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., ... & Ng, A. Y. (2017). Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225.

[15] Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, J. J., Peng, L., ... & Shetty, S. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nature medicine, 25(6), 954-961.

[16] Wu, C. C., Yeh, W. C., Hsu, W. D., Islam, M. M., Nguyen, P. A. A., Poly, T. N., ... & Li, Y. C. J. (2019). Prediction of fatty liver disease using machine learning algorithms. Computer methods and programs in biomedicine, 170, 23-29.

[17] Venkata Ramana, B., Babu, M. Surendra. P., & Venkateswarlu, N. B. (2011). A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis. International Journal of Database Management Systems, 3(2), 101–114. https://doi.org/10.5121/ijdms.2011.3207

[18] Ain Najwa Arbain, & Pillay, Y. (2019). A Comparison of Data Mining Algorithms for Liver Disease Prediction on Imbalanced Data. International Journal of Data Science and Advanced Analytics (ISSN 2563-4429), 1(1), 1–11. http://ijdsaa.com/index.php/welcome/article/view/2

[19] Vijayarani, S., & Dhayanand, S. (2015). Liver disease prediction using SVM and Naïve Bayes algorithms. International Journal of Science, Engineering and Technology Research (IJSETR), 4(4), 816-820.

[20] Ghosh, M., Mohsin, M., Raihan, M., Akter, L., Bairagi, A. K. et al. (2021). A Comparative Analysis of Machine Learning Algorithms to Predict Liver Disease. Intelligent Automation & Soft Computing, 30(3), 917–928.

[21] S. Sontakke, J. Lohokare and R. Dani, "Diagnosis of liver diseases using machine learning," *2017 International Conference on Emerging Trends & Innovation in ICT (ICEI)*, 2017, pp. 129-133, doi: 10.1109/ETIICT.2017.7977023.

[22] Phan, D.V.; Chan, C.L.; Li, A.A.; Chien, T.Y.; Nguyen, V.C. Liver cancer prediction in a viral hepatitis cohort: A deep learning approach. Int. J. Cancer 2020, 147, 2871–2878.

[23] Rau, H.-H., Hsu, C.-Y., Lin, Y.-A., Atique, S., Fuad, A., Wei, L.-M., & Hsu, M.-H. (2016). Development of a web-based liver cancer prediction model for type II diabetes patients by using an artificial neural network. Computer Methods and Programs in Biomedicine, 125, 58–65. https://doi.org/10.1016/j.cmpb.2015.11.009

[24] Bhanot, K., Qi, M., Erickson, J. S., Guyon, I., & Bennett, K. P. (2021). The problem of fairness in synthetic healthcare data. *Entropy*, *23*(9), 1165.

[25] Kawaguchi, T., Suzuki, F., Imamura, M., Murashima, N., Yanase, M., Mine, T., ... & Suzuki, K. (2019). Rifaximin‐altered gut microbiota components associated with liver/neuropsychological functions in patients with hepatic encephalopathy: An exploratory data analysis of phase II/III clinical trials. Hepatology Research, 49(4), 404-418.

[26] Van Buuren, S., & Oudshoorn, K. (1999). Flexible multivariate imputation by MICE (pp. 1-20). Leiden: TNO.

[27] Huang, C., Davis, L. S., & Townshend, J. R. G. (2002). An assessment of support vector machines for land cover classification. International Journal of remote sensing, 23(4), 725-749.

[28] Cristianini, N., & Shawe-Taylor, J. (2000). An introduction to support vector machines and other kernel-based learning methods. Cambridge university press.

[29] Sorich, M. J., Miners, J. O., McKinnon, R. A., Winkler, D. A., Burden, F. R., & Smith, P. A. (2003). Comparison of linear and nonlinear classification algorithms for the prediction of drug and chemical metabolism by human UDP-glucuronosyltransferase isoforms. *Journal of chemical information and computer sciences*, *43*(6), 2019-2024.

[30] Breiman, L. (2001), the prediction result was produced. Random forests. Machine learning, 45(1), 5-32.

[31] Avinash Navlani. (April23,2021). *Multi-Layer Perceptron Neural Network using Python.*

[32] Mandrekar, J. N. (2010). Receiver Operating Characteristic Curve in Diagnostic Test Assessment. Journal of Thoracic Oncology, 5(9), 1315–1316. https://doi.org/10.1097/jto.0b013e3181ec173d

[33] Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition, 30(7), 1145–1159. https://doi.org/10.1016/s0031-3203(96)00142-2

[34] Jason Brownlee. (2016, March 20). Overfitting and Underfitting With Machine Learning Algorithms. Machine Learning Mastery. https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/#:~:text=Overfitting%20happens%20when%20a%20model,as%20concepts%20by%20the%20model.

[35] Salma Ghoneim. (2019, April 2). Accuracy, Recall, Precision, F-Score & Specificity, which to optimize on? Medium; Towards Data Science. https://towardsdatascience.com/accuracy-recall-precision-f-score-specificity-which-to-optimize-on-867d3f11124

[36] Gopi Krishnan Rajbahadur, Wang, S., Gustavo Ansaldi Oliva, & Hassan, A. E. (2021, January 31). The impact of feature importance methods on the interpretation of defect classifiers. ResearchGate; Institute of Electrical and Electronics Engineers. https://www.researchgate.net/publication/348936721_The_impact_of_feature_importance_methods_on_the_interpretation_of_defect_classifiers

[37] Jason Brownlee. (2018, May 27). Confidence Intervals for Machine Learning. Machine Learning Mastery. https://machinelearningmastery.com/confidence-intervals-for-machine-learning/

[38] Mahaboob, M. (2021). Prediction of Liver Disease Using Machine Learning Algorithm and Genetic Algorithm. *Annals of the Romanian Society for Cell Biology*, 2347–23572347–2357. Retrieved from https://www.annalsofrscb.ro/index.php/journal/article/view/2768