Hepatitis Disease Detection by Hybrid Algorithm Machine Learning Model

CSCI 3430 Machine Learning Assignment 2

Group Members

- Mohamed Moubarak Mohamed Misbahou Mkouboi 1820705
- Karim Fazlul 1822049
- Mohammad Afif Muhajir 1816203

Introduction

Hepatitis Disease

Hepatitis is usually the result of a viral infection or liver damage caused by alcohol, and symptoms include yellowing of the eyes and skin (jaundice), loss of appetite, a high temperature, dark urine, and feeling unusually tired all the time

Machine Learning in Medical Field

Machine Learning in medical field provides methods, tools, and technique that can help solving diagnostic and prognostic problems in a variety of medical domains.

Hybrid Algorithm/Model

Hybrid Algorithm/Model refers to a technique under Machine Learning and Deep Learning that is to combine more than 1 base algorithm to solve the same problem.

Technical Report Sections

Computation Methods

Dataset

Algorithm Explanation

Use of Hybrid Algorithm

In this project we are tasked to implement hybrid model for our program. The chosen technique is called “Blending”.

Hepatitis Dataset

This project will be conducted using an Hepatitis dataset prepared by a third-party.

Blending Algorithm

Blending is a colloquial term for ensemble learning with a stacking-type architecture model

Technical Report Sections

Data Preprocessing

Analysis of Result

Flowchart

Dataset Preparation for ML Model

The explanation of steps we are taking to preprocess our dataset will be explained here

Result Comparison

The result of our hybrid algorithm model will be compared with another model for accuracy value comparison to determine how good is our model.

Flowchart Explanation

This section will show our project executions stages in a flowchart

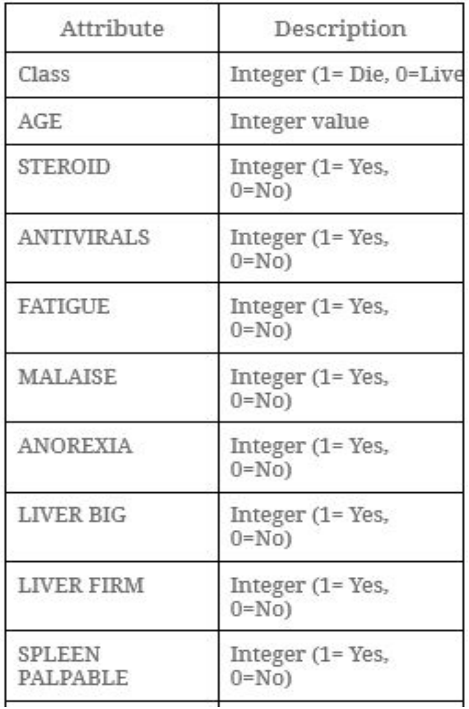
Technical Report

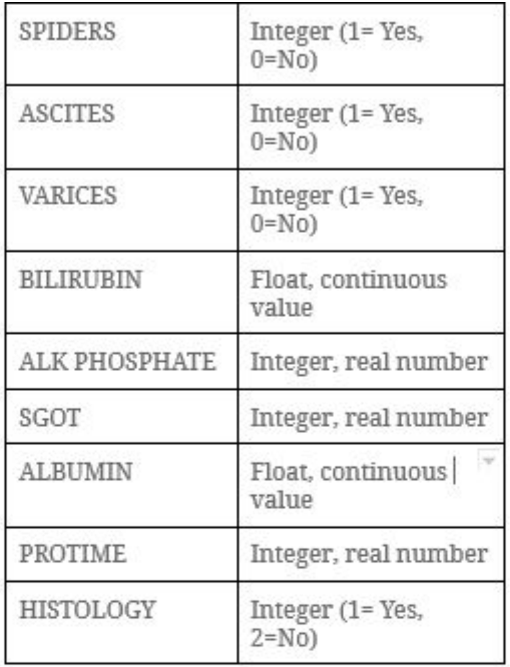
Computation Method and Dataset

- For this project, we are tasked to experiment with a branch of Machine Learning model which is the “Hybrid Model/Algorithm”.

- Our Dataset is an Hepatitis electronic health record that has been prepared by a third-party on an online repository. The dataset consists of 20 attributes and 142 instances of Data.

Dataset Attributes Table





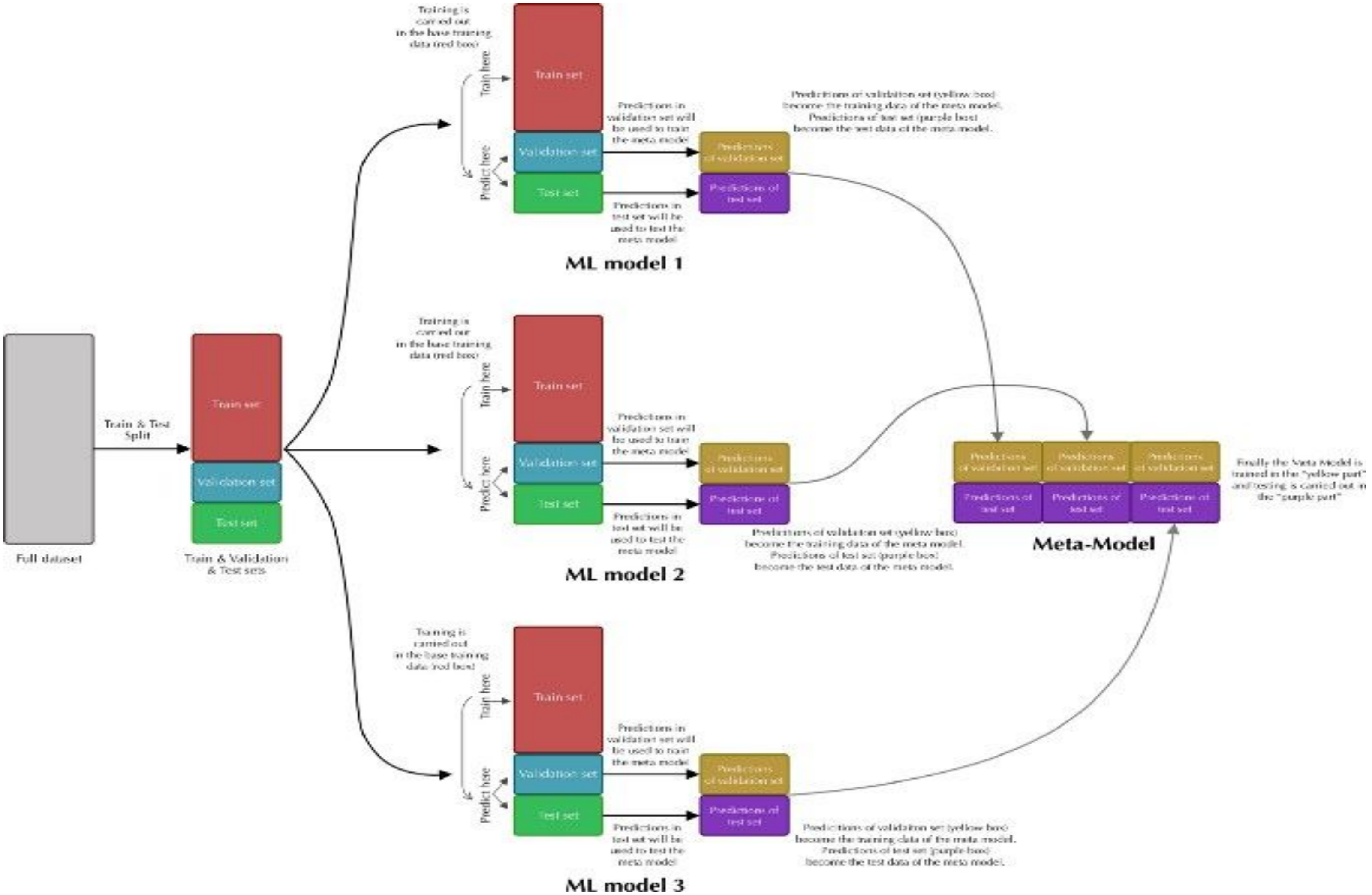
Attribute	Description
Class	Integer (1= Die, 0=Live
AGE	Integer value
STEROID	Integer (1= Yes, 0=No)
ANTIVIRALS	Integer (1= Yes, 0=No)
FATIGUE	Integer (1= Yes, 0=No)
MALAISE	Integer (1= Yes, 0=No)
ANOREXIA	Integer (1= Yes, 0=No)
LIVER BIG	Integer (1= Yes, 0=No)
LIVER FIRM	Integer (1= Yes, 0=No)
SPLEEN PALPABLE	Integer (1= Yes, 0=No)
SPIDERS	Integer (1= Yes, 0=No)
ASCITES	Integer (1= Yes, 0=No)
VARICES	Integer (1= Yes, 0=No)
BILIRUBIN	Float, continuous value
ALK PHOSPHATE	Integer, real number
SGOT	Integer, real number
ALBUMIN	Float, continuous value
PROTIME	Integer, real number
HISTOLOGY	Integer (1= Yes, 2=No)

Algorithm Explanation

Our project uses a form of hybrid model named “Blending”. The way that the algorithm works is similar to another technique called “stacked generalization”.

It can be understood that these two models works by involving two or more base models, often referred to as level-0 models, and a meta-model that combines the predictions of the base models referred to as a level-1 model. The meta-model is trained on the predictions made by base models on out-of-sample data.

The only difference is blending doesn’t accommodate a validation technique that is essential in stacked generalization.

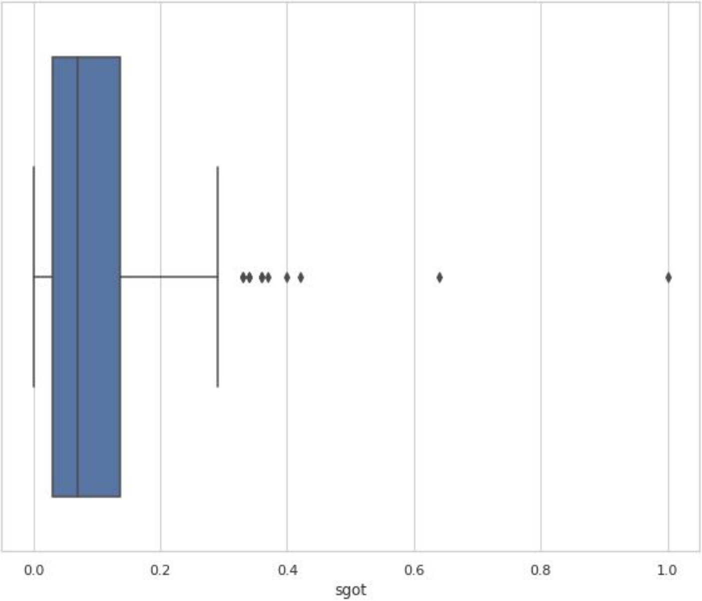
Full dataset
Train & Test Split
Train set
Validation set
Test set
Train & Validation & Test sets
Training is carried out in the base training data (red box)
Train here
Predict here
Train set
Validation set
Test set
Predictions in validation set will be used to train the meta model
Predictions in test set will be used to test the meta model
Predictions of validation set
Predictions of test set
Predictions of validation set (yellow box) become the training data of the meta model. Predictions of test set (purple box) become the test data of the meta model.
ML model 1
Training is carried out in the base training data (red box)
Train here
Predict here
Train set
Validation set
Test set
Predictions in validation set will be used to train the meta model
Predictions in test set will be used to test the meta model
Predictions of validation set
Predictions of test set
Predictions of validation set (yellow box) become the training data of the meta model. Predictions of test set (purple box) become the test data of the meta model.
ML model 2
Training is carried out in the base training data (red box)
Train here
Predict here
Train set
Validation set
Test set
Predictions in validation set will be used to train the meta model
Predictions in test set will be used to test the meta model
Predictions of validation set
Predictions of test set
Predictions of validaiton set (yellow box) become the training data of the meta model. Predictions of test set (purple box) become the test data of the meta model.
ML model 3
Predictions of validation set
Predictions of validation set
Predictions of validation set
Predictions of test set
Predictions of test set
Predictions of test set
Meta-Model
Finally the Meta Model is trained in the "yellow part" and testing is carried out in the "purple part"

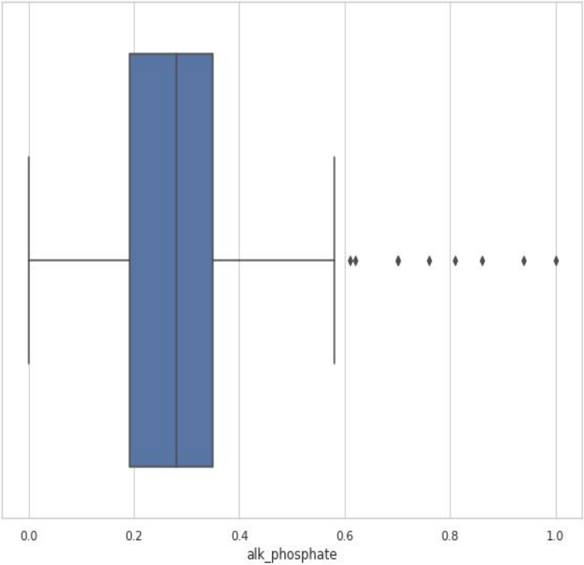
Implementation and Process

In this portion of the study, we will explain how our programme model manages data. This will comprise building the model, coding it, training it, and testing it before assessing the results.

Analysis of Results

Model Training and Testing Results





Explanation of the process

Data Acquired

Data Normalization

- Distribution of data has a significant effect on the ML model as most of the machine learning model assumes that data are in a normal distribution, data without normal distribution can cause the model to learn the wrong feature because of the variance difference of such feature.
- We have used standard normalization to normalize our data and it is mathematically expressed as

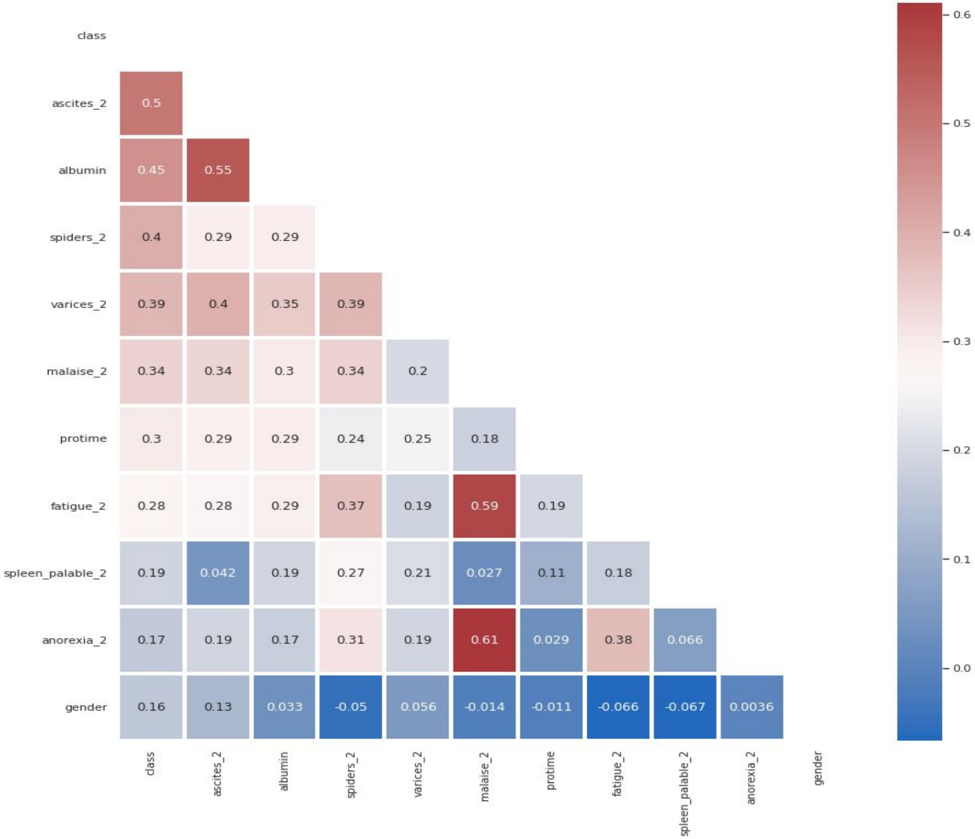
$$Z = (x - \mu) \ / \ \sigma$$

Outlier Removal

- Outliers are data points that are 3 standard deviations apart from the mean. These values can have a significant influence on our model performance. They are not common in natural events, and most of the time they are caused by mistakes while data is being prepared for analysis.
- We used a numpy built-in function to construct a mask of data points that are three standard deviations away from the mean, then filtered them out using the thai mask.

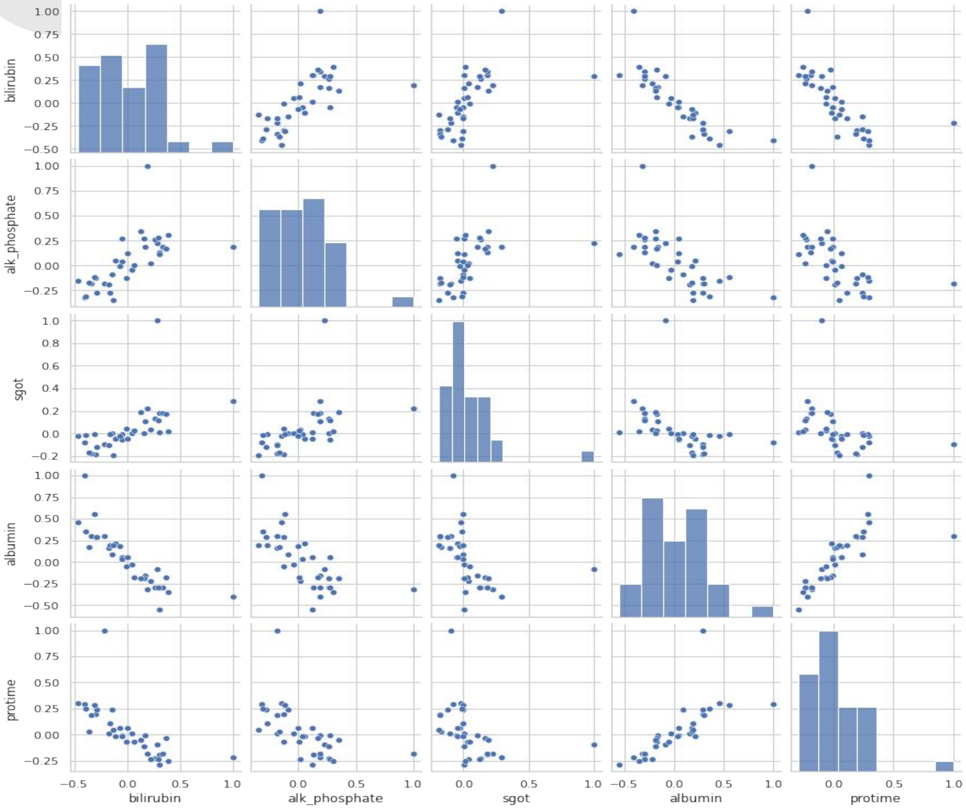
Explanation of results

Figure- Correlation Plot



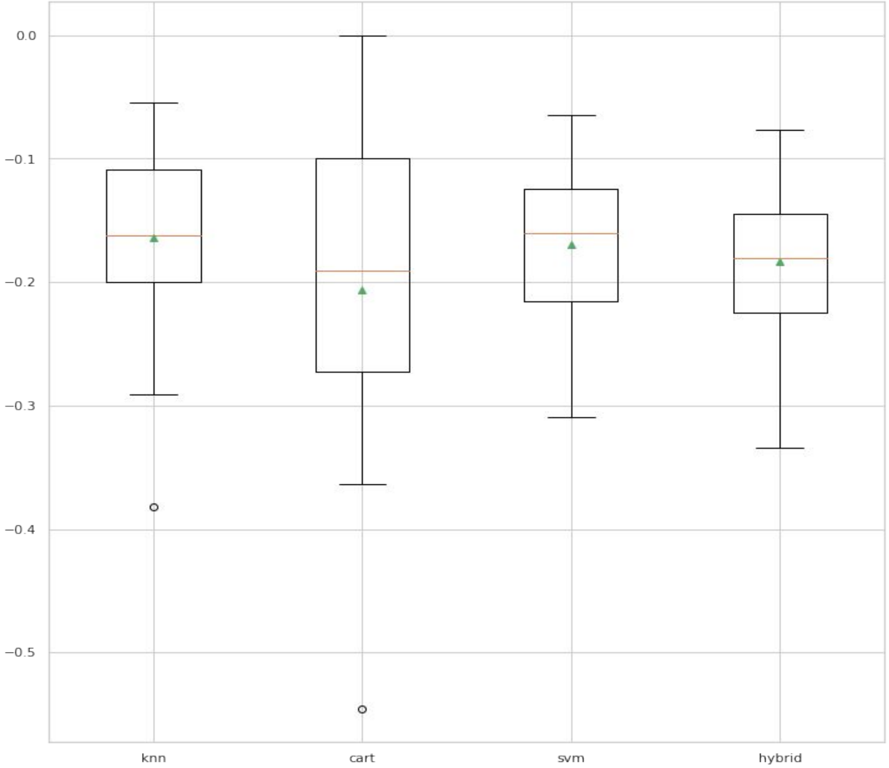
Ascites and albumin are highly correlated with each other so we can eliminate one of them for feature reduction. Ascites have the highest correlation while gender has the lowest correlation with the predicted label, as shown in the figure below. Figure shows the distribution of correlation of attributes with another attribute in a pair.

Figure- Linear Relationship between top five features in the dataset



Graph represents the correlation of the top 5 features (based on statistical analysis) of the data set. The diagonal components of the figure represent the relation with the feature itself, that is why the shape of the graph appears to be in natural form. Protamine has a negative correlation with bilirubin while a strong correlation with albumin.

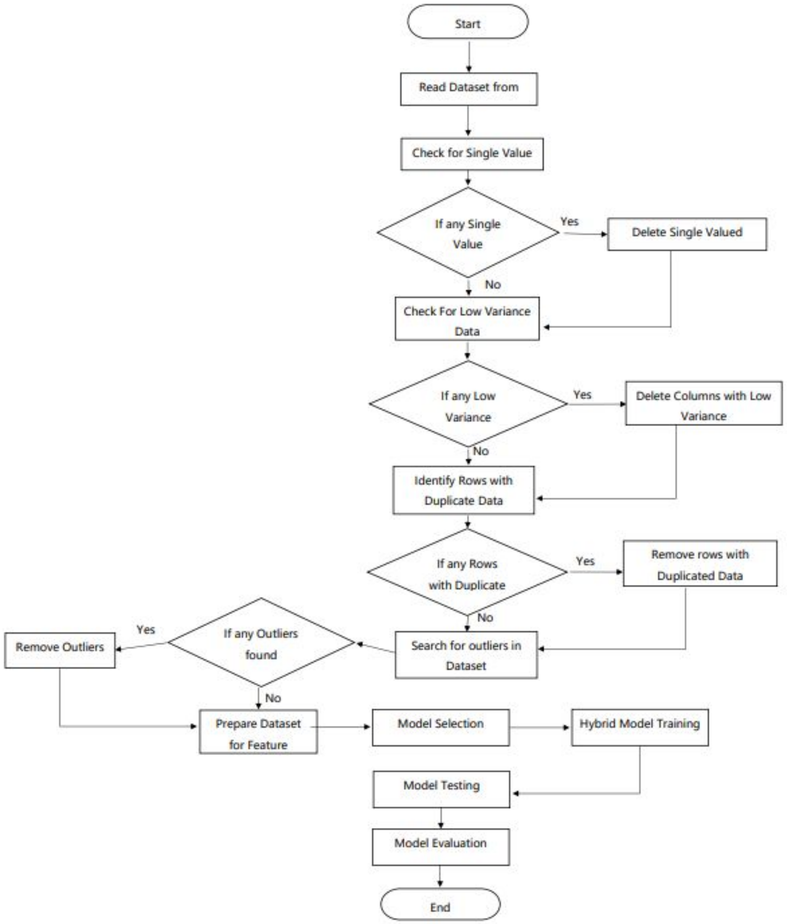
Figure - Analysis of the result



Cross-validation is a popular approach for evaluating a machine learning model's generalization performance. In this graph, the average of the support vector machines(SVM) has been compared to our hybrid model. The last one has almost the same result as our hybrid and is underfitting.

Flowchart

FLOWCHART FOR Hepatitis Detection Deep Learning


Start
Read Dataset from
Check for Single Value
If any Single Value
Yes
Delete Single Valued
No
Check For Low Variance Data
If any Low Variance
Yes
Delete Columns with Low Variance
No
Identify Rows with Duplicate Data
If any Rows with Duplicate
Yes
Remove rows with Duplicated Data
No
Search for outliers in Dataset
If any Outliers found
Yes
Remove Outliers
No
Prepare Dataset for Feature
Model Selection
Hybrid Model Training
Model Testing
Model Evaluation
End