IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Early Prediction Of Eye Diseases Using Machine Learning Models

[1]Madhab Paul Choudhury
[1]Research Scholar, PhD(Engg.)
[1]Computer Science Engineering IIT ISM Dhanbad
[1]Dhanbad, Jharkhand

[2]Jagannibas Paul Choudhury
[2]Professor(Commputer Science Engg.)
[2]Narula Institute of Technology
[2] Kolkata, West Bengal

Abstract: Artificial Intelligence (AI) has changed something so that it becomes much better in various aspects of our lives, offering solutions to numerous problems and bridging gaps between reality and business. Within the domain of AI, emerging technologies such as machine learning and deep learning models have played an important role in transforming the way so that we can analyse data, make decisions, and can take action or give attention to difficult situations or problems, with an objective to understand them to find solutions.

Liver disease encompasses a wide range of conditions that impair the liver's ability to function properly, leading to serious health issues. Risk factors for liver disease include viral infections (e.g., hepatitis B and C), excessive alcohol consumption, obesity, metabolic disorders, and genetic predisposition. Symptoms vary depending on the disease's severity but commonly include fatigue, jaundice, abdominal pain, and swelling. If the lever disease can be detected in early life, certain preventive measures can be taken so that the patient can recover from lever disease and can enjoy a real and happy life.

Here machine learning models and ensemble methods of machine learning models have been applied for selecting a proper model for prediction of lever disease of the person. Under machine learning models Random forest, decision tree, gradient boosting, KNN(K nearest neighbour), logistic regression have been used. Under ensemble methods, voting classifier using maximum voting, average voting, Blending, Bagging and boosting, stacking of models have been used.

Index Terms machine learning models, Random forest, decision tree, gradient boosting, KNN(K nearest neighbor), Ensemble techniques, voting classifier using maximum voting, average voting, Blending, Bagging and boosting.

I. INTRODUCTION

Early detection of liver disease is essential for effective treatment. Advanced imaging methods, such as elastography (including FibroScan), aid in assessing liver stiffness, which may indicate fibrosis or cirrhosis. Blood biomarkers like ALT, AST, and specialized liver function tests are commonly used to evaluate liver health. Recently, researchers have explored non invasive biomarkers and genetic markers for their potential in predicting disease progression, especially in conditions like NAFLD and nonalcoholic steatohepatitis (NASH).

2. LITERATURE REVIEW.

Ruhul Amin [1] has employed projection-based feature extraction techniques to reduce data redundancy. The Indian Liver Patient Dataset (ILPD) from the University of California, Irvine (UCI) repository has been used to classify chronic liver disease.

Peter IfeoluwaAdegbola [2] has utilized data collected from rats exposed to environmentally concerning chemicals. Linear regression techniques were applied to extract significant features for predicting the probability of disease occurrence, followed by the use of random forest (RF) classification to determine disease likelihood. Similarly, Osama Mohareb Khaled [3] has implemented three machine learning models—K-nearest neighbors (KNN), Gaussian Naïve Bayes, and random forest—on a dataset comprising 32,000 records. Their findings indicated that the random forest model outperformed the other algorithms.

Trupti M. Kodinariya [4] has analyzed ultrasonic images, computerized tomography (CT) scans, and magnetic resonance imaging (MRI) data, applying both machine learning and deep learning techniques for liver disease detection. Likewise, SrilathaTokala [5] has employed logistic regression, support vector machines, K-nearest neighbors, and random forest algorithms, with random forest demonstrating superior performance compared to the others.

Engy A. El-Shafeiy [6] has used KNN, Gaussian Naïve Bayes (Gaussian NB), and random forest (RF) classifiers, reporting that the random forest model yielded the best performance. Similarly, Barinderjit Kaur [7] has worked with a dataset of 416 patients from an Indian hospital and developed a hybrid model combining KNN and random forest for liver disease prediction.

Additionally, Neha Tanwar [8] has utilized a liver patient dataset to predict liver disorders using machine learning models such as support vector machines, logistic regression, K-nearest neighbors, and random forest. Among these, random forest exhibited the highest accuracy. Shahid Mohammad Ganie [9] has analyzed a dataset containing 23 attributes and 7,000 patient records collected from the Egyptian Liver Research Institute and Mansoura Central Hospital in Egypt. He has employed support vector machines (SVM), Boosted C5.0, and Naïve Bayes (NB) data mining techniques for liver disease prediction.

Finally, R. Kalaiselvi [10] has conducted a comprehensive survey on the application of various machine learning algorithms for liver disease prediction. The study explored both supervised and unsupervised learning techniques for improving disease diagnosis.The proposed approach by the authors integrates Long Short-Term Memory (LSTM) for sequence learning and Convolution Neural Networks (CNNs) for extracting non-linear features [12].Mammography has demonstrated promising outcomes with the application of deep learning technologies in the quantitative evaluation of parenchymal density, categorization, detection, diagnosis, and breast cancer risk prognosis, enabling more precise patient management [13]. Additionally, deep learning has streamlined the interpretation process, reducing both interpretation time and workload. This paper presents a comparative analysis of segmentation and feature extraction methods for detecting lung cancer [14]. Various segmentation techniques, such as Thresholding, global Thresholding, and watershed segmentation, have been implemented and assessed. Additionally, feature extraction has been applied to improve the performance of these segmentation techniques. The study focuses on medical image analysis and classification using a convolution+ReLU algorithm, which integrates convolutional techniques with ReLU optimization [15]. The research employs a robust and efficient convolution+ReLU approach on the BraTS 2020 dataset, significantly reducing segmentation time compared to other optimization methods. Researchers process extensive and complex healthcare data using various deep learning techniques, enabling medical professionals to predict diseases effectively [16].The authors have concentrated on machine learning (ML) algorithms for cancer prediction, which is impacted by various performance metrics [17]. By utilizing widely used ML techniques such as Support Vector Machines (SVM), K-Nearest Neighbours (KNN), Linear Regression, Decision Tree, and Naive Bayes, the study evaluates the accuracy of cancer prediction.

In this study, the authors propose enhancing a MobileNet base model by fine-tuning it with additional features to improve brain tumour detection [18].The model's precision and accuracy are increased by restructuring its layers.Pre-processing techniques are applied to MRI images to enhance their quality, and data augmentation is employed to expand the dataset size, thereby improving the model's training process. Since insulin plays a crucial role in regulating various properties of plasma, including water, enzymes, proteins, vitamins, and minerals, its imbalance can lead to diabetes [19]. Existing methods [20] for medical image feature extraction have proven insufficient in effectively addressing the challenges of early brain tumour detection.To overcome this limitation, a novel model leveraging the Inception-v3 convolution neural network has been proposed.The authors introduce an automated identification method that leverages deep learning and visual analysis technology [21]. Their approach classifies fundus images using a convolution neural network (CNN) based on the severity of diabetic retinopathy (DR). The article[22] presents a fuzzy distance-based ensemble approach integrating deep learning models for cervical cancer detection in Pap smear images. The

methodology employs three transfer learning models—Inception V3, MobileNet V2, and Inception ResNet V2—enhanced with additional layers to capture data-specific features. This study explores the effectiveness of machine learning and deep learning models in detecting heart murmurs from audio recordings. Utilizing the PhysioNet Challenge 2016 dataset [23], the authors compare traditional machine learning models—Support Vector Machine, Random Forest, AdaBoost, and Decision Tree—with a Fully Convolution Neural Network (FCNN). This study introduces a novel approach utilizing Deep Separable Convolution Neural Networks (DS-CNNs) to enhance Chronic Kidney Disease (CKD) prediction [24]. Using the Chronic Kidney Disease Dataset from Kaggle, the proposed model integrates DS-CNNs with advanced optimization techniques to improve predictive accuracy. The authors introduce and evaluate an innovative method for heart disease prediction by integrating deep learning models with bioinspired algorithms [25]. Deep learning techniques facilitate automatic feature extraction and the recognition of complex patterns from raw data, while bioinspired algorithms enhance optimization, improving model accuracy and generalization.

3. MOTIVATION.

A lot of research work ([1]-[9], [12]-[25]) have been done in the area of healthcare prediction with an objective to detect sickness of various organs. Machine learning algorithms have also been proposed([1], [2], [4]-[6], [12], [17], [19]]). Deep Learning models have been proposed in {[9], [16], [22], ([24]-[25]), Segmentation has been done in ([9], [14],).However, no author has worked on the same data set and not evaluated several evaluation measures. That is the reason for this proposed work which has been written inthis paper.Here machine learning models and ensemble methods of machine learning models have been applied for selecting a proper model for prediction of lever disease of the person. Under machine learning models Random forest, decision tree, gradient boosting, KNN(K nearest neighbour), logistic regression have been used. Under ensemble methods, voting classifier using maximum voting, average voting, Blending, Bagging and boosting, stacking of models have been used.

4. DATA SET.

The lever data set has been collected from UCI Machine Repository ([11]). The data set comprises of 7 attributes. These are furnished below:-

Table 1
Lever Data Set

No	Attribute Code	Attribute Information
1	mcv	mean corpuscular volume
2	alkphos	alkaline phosphotase
3	sgpt	alamine aminotransferase
4	sgot	aspartate aminotransferase
5	gammagt	gamma-glutamyltranspeptidase
6	drinks	number of half-pint equivalents of alcoholic beveragesdrunk per day
7	selector	field used to split data into two sets viz. healthy or diseased.

5. METHODOLOGY.

The hierarchy of simple ensemble techniques and advanced ensemble techniques has been furnished in figure 1. The hierarchy of Machine Learning models has been furnished in figure 2.

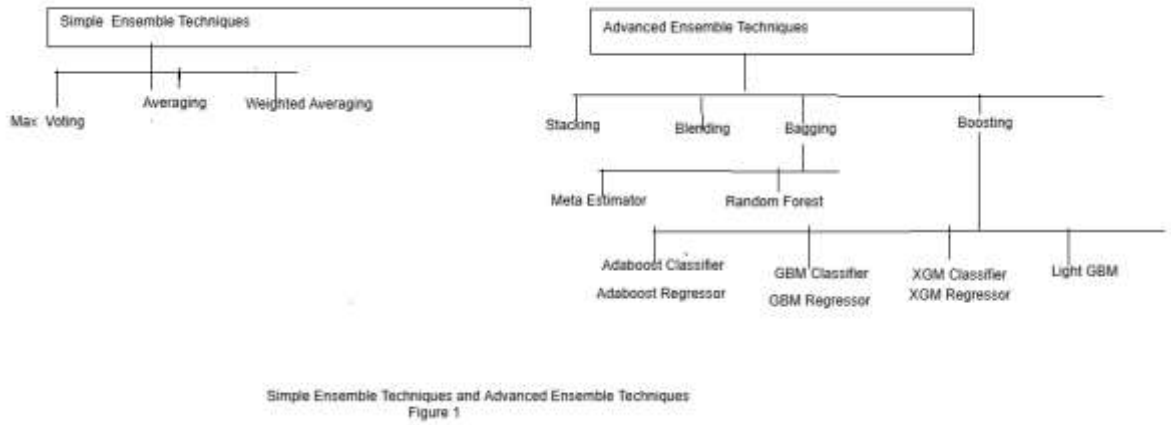

Figure 1: The hierarchy of Simple Ensemble and Advanced Ensemble Techniques

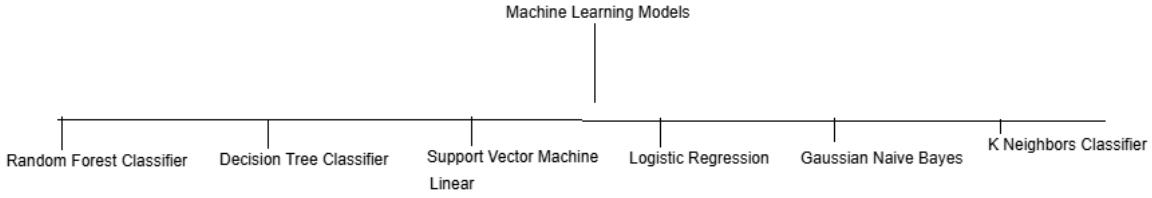

Figure 2 :The hierarchy of Machine Learning Models

6. CONTRIBUTION.

Proposed Flow of Work

Step 1. Data Collection. Lever data set Data have been collected from[11]
Step 2. The dataset containing relevant information for the prediction of disease of lever have been taken.
Step 3. The data have been entered accurately and completely.
Step 4. Data Pre-processing: Cleaning of data and Removal of Outlier have been done.
Step 5. Taking care of missing data by input certain concerned data or removal of that data based on the nature and quantity of missing values.
Step 6. Cleaning the data by tackling inconsistencies, errors, and anomalies.
Step 7. Detection and removal of outliers that may affect the analysis.
Step 8. Implementation of Machine Learning models and Performance Evaluation:
Step 9. Under machine learning classifier models, Random forest classifier, Decision Tree classifier, Support Vector Machine (Linear), Logistic Regression, Gaussian Naïve Bayes, K Nearest Neighbours have been used.
Step 10. Train the classifiers using the pre-processed data.
Step 11. Evaluate the performance of each classifier using evaluation metrics like accuracy, precision, recall, and F1 score.
Step 12. Application of Ensemble Methods and Performance Evaluation:
Step 13. Ensemble methods include Simple Ensemble Methods and advanced Ensemble Techniques.
Step 14. Under Simple Ensemble Methods Max Voting, Averaging and Weighted Average techniques have been used.
Step 15. Evaluate each simple ensemble model based on accuracy, precision, recall, and F1 score to assess the effectiveness of each approach.
Step 16. Under Advanced technique Stacking, Blending, Bagging, Boosting have to be used.
Step 16. Bagging includes Meta Estimator and Random Forest.
Step 17. Boosting includes Adaboost classifier, AdaboostRegressor, GBM Classifier, GBM Regressor, XGBoost Classifier, XGBoostRegressor, Light GBM.
Step 18. Evaluate each advanced ensemble model based on accuracy, precision, recall, and F1 score to assess the effectiveness of each approach.
Step 19. Train each ensemble models using the pre-processed data.

Step 20. Evaluate the performance of each ensemble method and compare the results.
Step 21. Compare and analyse the performance of all the different techniques employed in previous steps (Machine Learning models and Ensemble Techniques).
Step 22. Evaluation metrics such as accuracy, precision, recall, and F1 score to assess the effectiveness of each approach
Step 23. Identify the most accurate and reliable method for predicting the lever disease based on the results as obtained.

6.1. CLASSIFICATION.

6.2. Application of Machine Learning Models.

The application of machine learning models has been applied on input data [11]. The application of machine learning models has been furnished in figure 2.

6.2.1. Random Forest.

Input data set as available has to be applied to random forest algorithm. It has used 10 estimators that means 10 decision trees have been constructed and finally the average of these tree values has to be taken. The value of accuracy has been found as 88.89% for criterion as gini index as well as for entropy.

6.2.2. Decision Tree.

Input data set as available has to be applied to decision tree algorithm. Criterion as gini index has been used. The value of accuracy has been found as 89%.

6.2.3.KNN(K Nearest Neighbours) algorithm:

Input data set as available has to be applied to KNN(K Nearest Neighboursalgorithm. Number of neighbours has been used as 5. Distance function is used as 'minkowski'. The value of accuracy has been found as 77.77%.

6.2.4. Support Vector Machine algorithm.

Input data set as available has to be applied to support vector machine algorithm. The value of accuracy has been found as 91.66 % based on kernel function as linear.The value of accuracy has been found as 94.44 % based on kernel function as radial basis.

6.2.5. Linear Regression.

Input data set as available has to be applied to logistic regression algorithm.The value of accuracy has been found as 100%.

6.2.6. Logistic Regression.

Input data set [11] has to be applied to logistic regression algorithm. The value of accuracy has been found as 86.66 %.

6.2.7. Gaussian Naïve Bayes Algorithm.

Input data set [11] has to be applied to Gaussian Naïve Bayes algorithm. The value of accuracy has been found as 94.44 %.

6.2.8. Gradient Boosting Classifier Algorithm.

Input data set [11] has to be applied to Gradient Boosting Classifier algorithm. The value of accuracy has been found as 94.44 %.

6.3. Ensemble learning Techniques.

The application of ensemble learning techniques has been furnished in figure 3.

6.3.1. Simple Ensemble Techniques.

Under simple ensemble based models, max voting, averaging and weighted averaging models have been used. Here decision tree classifier, KNearest NeighbourClassifier and logistic regression models have been used. Under Max voting accuracy is 90.04 % on training data set and 69.23 % on tested data set, under averaging method, accuracy is 90.04 % on training data set and 25.00 % on tested data set and based on weighted average method, accuracy is 90.04 % on training data set and 26.92 % on tested data set.

6.3.2. Advanced Ensemble Techniques.

6.3.2.1. Blending.

Under blending, decision tree classifier, K Nearest NeighbourClassifier and logistic regression models have been used. Based on blending technique, accuracy is found to be 87.14 % on training data set and 66.35 % on tested data set.

6.3.2.2. Bagging Meta Estimator.

6.3.2.2.1. Bagging Classifier.

Under bagging classifier, decision tree classifier has been used. The accuracy is found as 99.59 % on training data set and 74.04 % on tested data set.

Random Forest Classifier.

Under Random Forest Classifier,the accuracy is found as 100.00 % on training data set and 75.00 % on tested data set.

6.3.2.2.2. Bagging Regressor.

Under bagging regressor, decision tree regressor has been used. The accuracy is found as 99.17 % on training data set and 68.27 % on tested data set.

6.3.2.2. Boosting.

6.3.2.2.1. Adaboost(Adaptive Boosting) Classifier.

The accuracy is found as 82.99 % on training data set and 80.77 % on tested data set.

6.3.2.2.2. Adaboost(Adaptive Boosting)Regressor.

The accuracy is found as 85.48 % on training data set and 68.27 % on tested data set.

6.3.2.2.3. Gradient Boosting Classifier.

The accuracy is found as 75.93 % on training data set and 75.96 % on tested data set.

6.3.2.2.4.Gradient Boosting Regressor.

The accuracy is found as 97.51 % on training data set and 75.96 % on tested data set.

6.3.2.2.5. XGB(Extreme Gradient Boosting) Classifier.

The accuracy is found as 7.88 % on training data set and 16.34 % on tested data set.

6.3.2.2.6. XGB(Extreme Gradient Boosting)Regressor.

The accuracy is found as 100.00 % on training data set and 66.35 % on tested data set.

6.3.2.2.7. LightGBM(Light Gradient Boosting Machine).

The accuracy is found as 57.26 % on training data set and 59.62 % on tested data set.

6.3.2.3. Stacking.

Under blending, decision tree classifier, K Nearest NeighbourClassifier as base model and logistic regression as final model has been used. Based on decision tree as base model, the accuracy is found as 99.59 % on training data set and 53.85 % on tested data set.

Based on K nearest neighbour as base model, the accuracy is found as 99.58 % on training data set and 62.5 % on tested data set.

Based on logistic regression as final model, the accuracy is found as 99.58 % on training data set and 63.46 % on tested data set.

6.4. Training and Test data.

The number of training date has been used as 70% and that of test data as 30% for the above models for better performance of the models.

7. RESULTS.

The comparative study of all machine learning models on the basis of accuracy have been furnished in table 2 as furnished below. The classification report, confusion matrix of all the models have been furnished in table 3, table 4 respectively.

Table 2

Machine Learning models versus accuracy

No	Name of Machine Learning Model	Accuracy(%)
1.	Random Forest Algorithm	88.89 %
2.	Decision Tree Algorithm	89 %
3.	KNN(K-Nearest Neighbor) Algorithm	77.77 %
4.	Support Vector Machine with Linear Kernel	100 %
5.	Linear Regression Algorithm	100 %
6.	Gaussian Naïve Bayes Algorithm	94.44 %
7.	Logistic Regression	86.66 %
8.	Support Vector Machine with Radial Basis Kernel	94.44 %
9.	Gradient Boosting Classifier	94.44 %

Table 3
Classification Report Item wise based on Machine Learning Models

Model	Precision	Recall	f1-score	Support
Random Forest 0	0.91	0.91	0.91	11
Random Forest 1	0.86	0.86	0.86	7
Decision Tree 0	1.00	0.82	0.9	11
Decision Tree 1	0.78	1.00	0.88	7
KNN Classifier 0	0.82	0.82	0.82	11
KNN Classifier 1	0.71	0.71	0.71	7
SVM(Kernel=Linear) 0	1.00	1.00	1.00	11
SVM(Kernel=Linear) 1	1.00	1.00	1.00	7
Linear Regression 0	1.00	1.00	1.00	11
Linear Regression 1	1.00	1.00	1.00	7
Gaussian Naïve Bayes Algorithm 0	0.92	1.00	0.96	11
Gaussian Naïve Bayes Algorithm 1	1.00	0.86	0.92	7
Logistic Regression 0	0.90	0.90	0.90	11
Logistic Regression 1	0.80	0.80	0.80	7
SVM(Kernel=Radial Basis) 0	0.92	1.00	0.96	11
SVM(Kernel=Radial Basis) 1	1.00	0.86	0.92	7
Gradient Boosting Classifier 0	0.92	1.00	0.96	11
Gradient Boosting Classifier	11.00	0.86	0.92	7

Table 4
Confusion Matrix based on Machine Learning Models

Model Item	True Positive	False Positive	False Negative	True Negative
Random Forest 0	10	1	1	6
Random Forest 1	6	1	1	10
Decision Tree 0	9	2	0	7
Decision Tree 1	7	0	2	9
KNN Classifier 0	9	2	2	5
KNN Classifier 1	5	2	2	9
SVM(Kernel=Linear) 0	11	0	0	7
SVM(Kernel=Linear) 1	7	0	0	11
Linear Regression 0	10	1	1	6
Linear Regression 1	6	1	1	10
Gaussian Naïve Bayes 0	11	0	1	6
Gaussian Naïve Bayes 1	6	1	0	11
Logistic Regression 0	10	1	1	6
Logistic Regression 1	6	1	1	10
Gradient Boosting 0	11	0	1	6
Gradient Boosting 1	6	1	0	11

The preference of model has to be decided on the more value of accuracy. From table 2 it has been observed that the accuracy of support vector machine with linear kernel, linear regression is 100.00 % which is the maximum value among all models. Therefore, support vector machine with linear kernel, linear regression has to be preferred. Considering theoretical concept, support vector machine with linear kernel has to be considered as compared to linear regression.

The Change of values of accuracy has been furnished in graph named figure 3. The change of values of precision, recall, f1-score (classification report) and confusion matrix has been furnished in figure 4, figure 5 respectively.

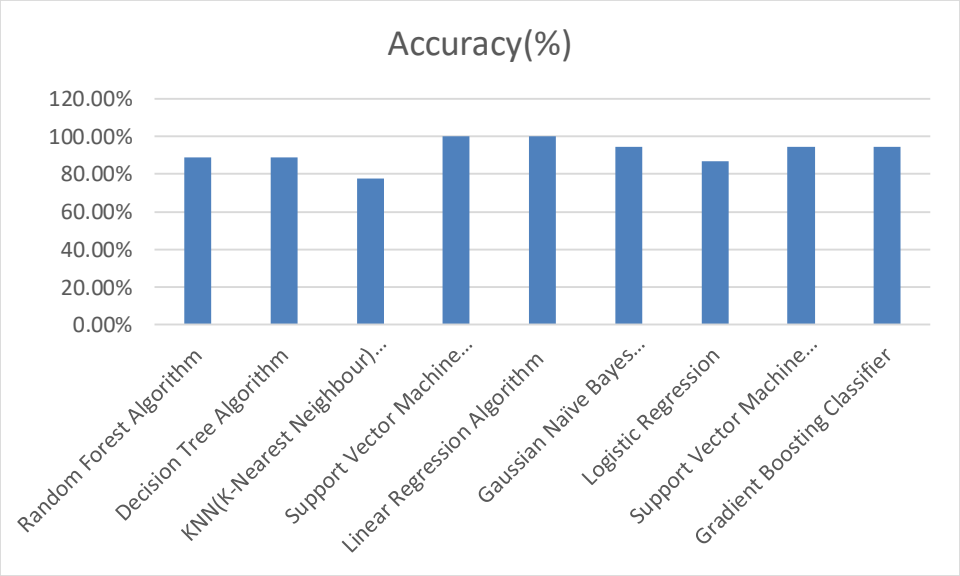

Figure 3: Machine Learning model wise Accuracy

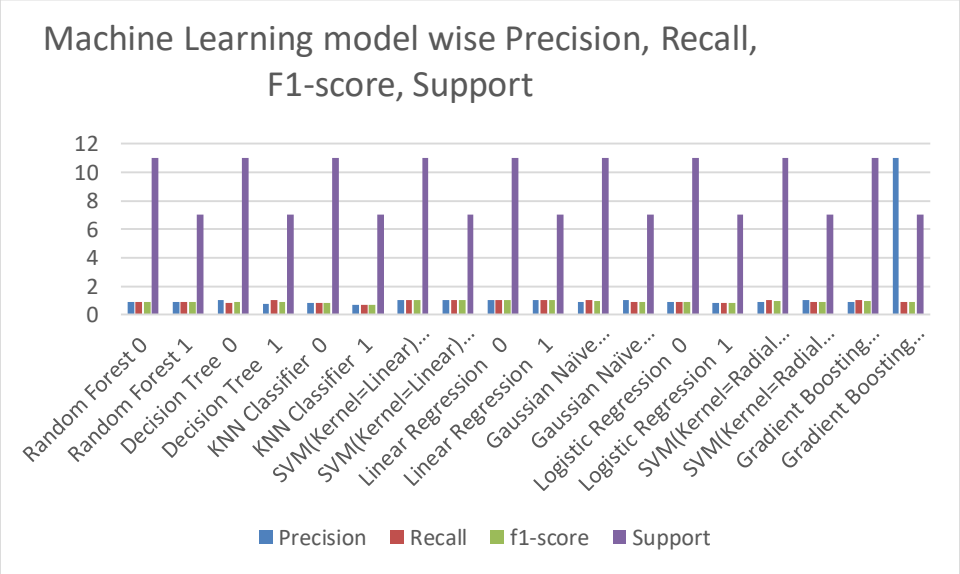

Figure 4: Machine Learning model wise Precision, Recall, F1-score and Support

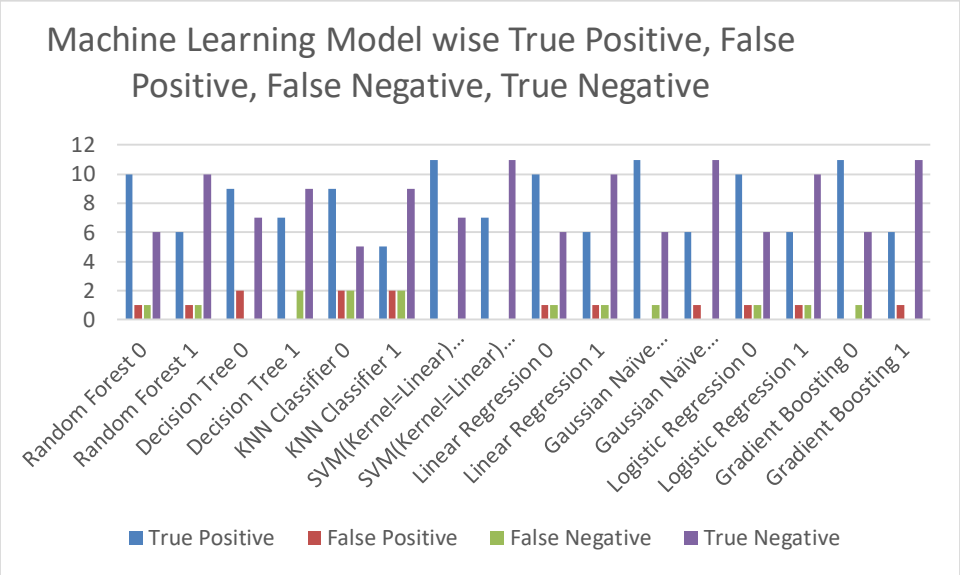

Figure 5: Machine Learning model wise Classification Report

The comparative study of simple ensemble techniques on the basis of accuracy has been furnished in table 5 as furnished below:-

Table 5
Simple Ensemble Techniques versus Accuracy

No	Name of Simple EnsembleTechnique	Accuracy on Training data(%)	Accuracy on Tested data(%)
1	Max Voting	90.04	69.23
2	Averaging	90.04	25.00
3	Weighted Averaging	90.04	26.92

Advanced ensemble techniques include Stacking, Blending, Bagging and Boosting. Bagging includes Meta Estimator, Random Forest. Boosting includes Adaboost classifier, AdaboostRegressor, GBM classifier, GBM Regressor, XGBoost classifier, XGBoostRegressor, Light GBM.
The comparative study of advanced ensemble techniques on the basis of accuracy has been furnished in table 6 as furnished below:-

Table 6
Advanced Ensemble Techniques versus Accuracy

No	Name of Simple EnsembleTechnique	Accuracy on Training data(%)	Accuracy on Tested data(%)
1	Blending	87.14	66.35
2	Bagging Classifier	99.59	74.04
3	Random Forest Classifier	100.00	75.00
4	Bagging Regressor	99.17	68.27
5	Adaboost Classifier	82.99	80.77
6	AdaboostRegressor	85.48	68.27
7	Gradient Boosting Classifier	75.93	75.96
8	Gradient Boosting Regressor	97.51	75.96
9	XGB Classifier	7.88	16.34
10	XGB Regressor	100.00	66.35
11	Light GBM	57.26	59.62
12	Stacking using KNN as base classifier	99.58	62.5
13	Stacking using DTree as base classifier	99.00	53.84
14	Stacking using Logistic Regression as final Classifier	99.58	63.46

The classification report, confusion matrix of simpleensemble techniqueshave been furnished in table 7, table 8 respectively.

Table 7
Classification Report Item wise based on Simple ensemble techniques

Model	Item	Precision	Recall	f1-score	Support
Max Voting	0	0.64	0.70	0.67	46
Max Voting	1	0.74	0.69	0.71	58
Averaging	0	0.36	0.57	0.44	46
Averaging	1	0	0	0	58
Weighted Averaging	0	0.36	0.61	0.45	46
Weighted Averaging	1	0	0	0	58

Table 8
Confusion Matrix based on Simple Ensemble Techniques

Model Item	True Positive	False Positive	False Negative	True Negative
Max Voting 0	32	14	18	40
Max Voting 1	40	18	14	32
Averaging 0	20	26	11	47
Averaging 1	47	11	26	20
Weighted Average 0	18	28	8	50
Weighted Average 1	50	8	28	18

The classification report, confusion matrix of simple ensemble techniques have been furnished in table 9, table 10 respectively.

The Change of values of accuracy based on simple ensemble techniques has been furnished in figure6. The change of values of precision, recall, f1-score (classification report) has been furnished in figure 7, the change of values in confusion matrix has been furnished in figure 8 respectively. The Change of values of accuracy based on advanced ensemble techniques has been furnished in graph named figure 9. The change of values of precision, recall, f1-score (classification report) of advanced ensemble techniques has been furnished in figure 10, the change of value in confusion matrix of these has been furnished in figure 11 respectively.

Table 9
Classification Report Item wise based on Advanced ensemble techniques

Model	Item	Precision	Recall	f1-score	Support
Blending	0	0.36	0.61	0.45	46
Blending	1	0	0	0	58
Bagging Classifier	0	0.73	0.65	0.69	46
Bagging Classifier	1	0.75	0.81	0.78	58
Random Forest Classifier	0	0.74	0.6	0.66	42
Random Forest Classifier	1	0.76	0.85	0.80	62
Bagging Regressor	0	0.62	0.57	0.59	42
Bagging Regressor	1	0.72	0.76	0.74	62
Adaboost Classifier	0	0.78	0.74	0.76	42
Adaboost Classifier	1	0.83	0.85	0.84	62
AdaboostRegressor	0	0.65	0.48	0.55	42
AdaboostRegressor	1	0.7	0.82	0.76	62
Gradient Boosting Classifier	0	0.79	0.55	0.65	42
Gradient Boosting Classifier	1	0.75	0.9	0.82	62
Gradient Boosting Regressor	0	0.71	0.69	0.7	42
Gradient Boosting Regressor	1	0.79	0.81	0.8	62
XGB Classifier	0	0.25	0.4	0.31	42
XGB Classifier	1	0	0	0	62
XGB Regressor	0	0.58	0.62	0.6	42
XGB Regressor	1	0.73	0.69	0.71	62
Light GBM	0	0	0	0	42

Light GBM	1	0.6	1.0	0.75	62
Stacking using KNN as base classifier	0	0	0	0	42
Stacking using KNN as base classifier	1	0.6	1.0	0.75	62
Stacking using DTree as base classifier	0	0	0	0	42
Stacking using DTree as base classifier	1	0.6	1.0	0.75	62
Stacking using Logistic Regression as final Classifier	0	0	0	0	42
Stacking using Logistic Regression as final Classifier	1	0.6	1.0	0.75	62

Table 10
Confusion Matrix based on Advanced Ensemble Techniques

Model Item	True Positive	False Positive	False Negative	True Negative
Blending 0	18	28	8	50
Blending 1	50	8	28	18
Bagging Classifier 0	30	16	11	47
Bagging Classifier 1	47	11	16	30
Random Forest Classifier 0	25	17	9	53
Random Forest Classifier 1	53	9	17	25
Bagging Regressor 0	24	18	15	47
Bagging Regressor 1	47	15	18	24
Adaboost Classifier 0	31	11	9	53
Adaboost Classifier 1	53	9	11	31
AdaboostRegressor 0	20	22	11	51
AdaboostRegressor 1	51	11	22	20
Gradient Boosting Classifier 0	29	13	12	50
Gradient Boosting Classifier 1	50	12	13	29
XGB Classifier 0	25	17	10	52
XGB Classifier 1	52	10	17	25
XGB Regressor 0	26	16	19	43
XGB Regressor 1	43	19	16	26
Light GBM 0	0	42	0	62
Light GBM 1	62	0	42	0
Stacking using KNN as base classifier 0	0	42	0	62
Stacking using KNN as base classifier 1	62	0	42	0
Stacking using DTree as base classifier 0	0	42	0	62
Stacking using DTree as base classifier 1	62	0	42	0

Stacking using Logistic Regression as final Classifier 0	0	42	0	62
Stacking using Logistic Regression as final Classifier 1	62	0	42	0

The comparative study of final advanced ensemble techniques and final machine learning model on the basis of accuracy has been furnished in table 11 as furnished below. The Change of values of accuracy based on final ensemble techniques and final machine learning model has been furnished in figure 12.

Table 11

Comparison of Random Forest Classifier with Max voting in terms of accuracy on Training data set and Tested data set

No	Name of EnsembleTechnique	Accuracy on Training data(%)	Accuracy on Tested data(%)
1	Max Voting	90.04	69.23
2	Random Forest Classifier	100.00	75.00

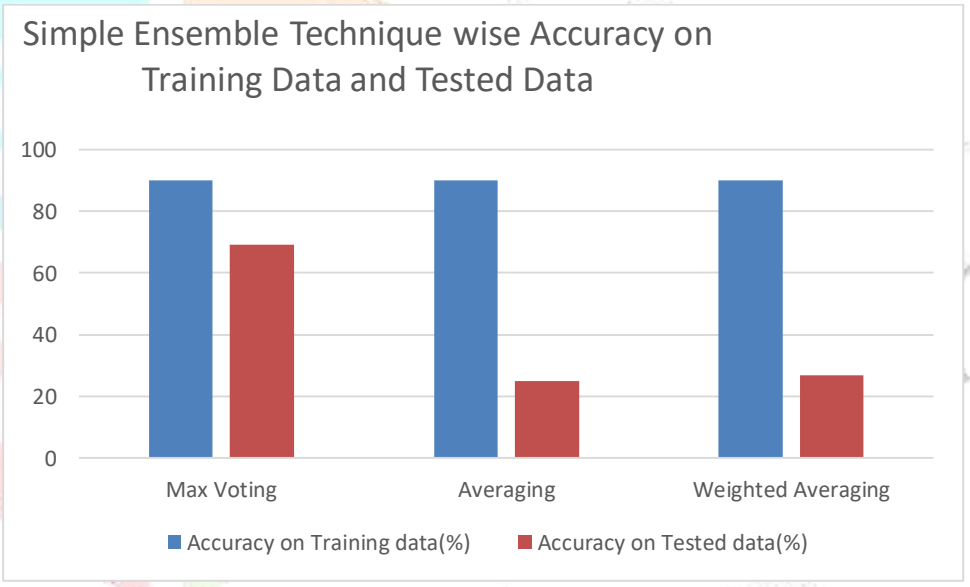

Figure 6: Simple Ensemble Technique wise Accuracy

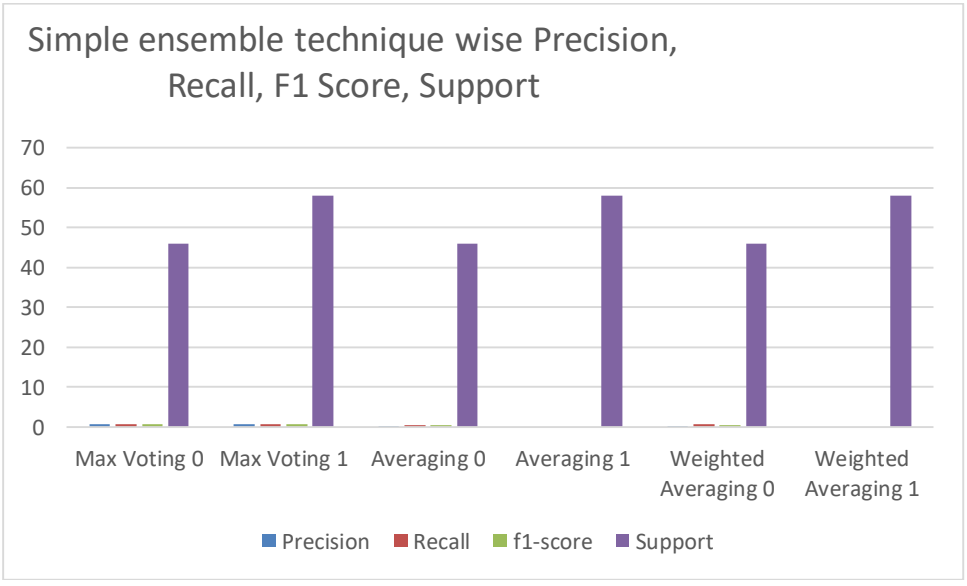

Figure 7: Simple Ensemble Technique wise Precision, Recall, F1 Score and Support

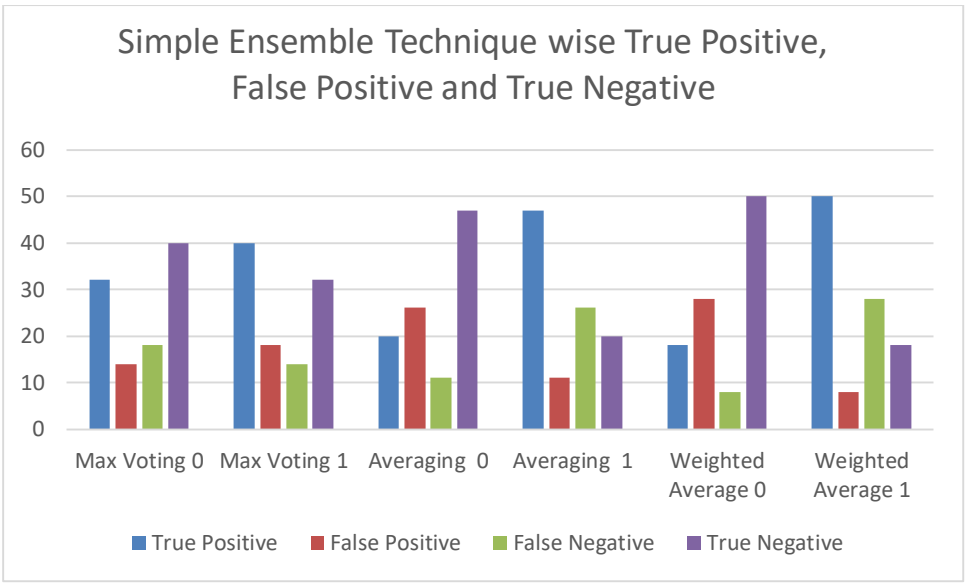

Figure 8: Simple Ensemble Technique wise Classification Report

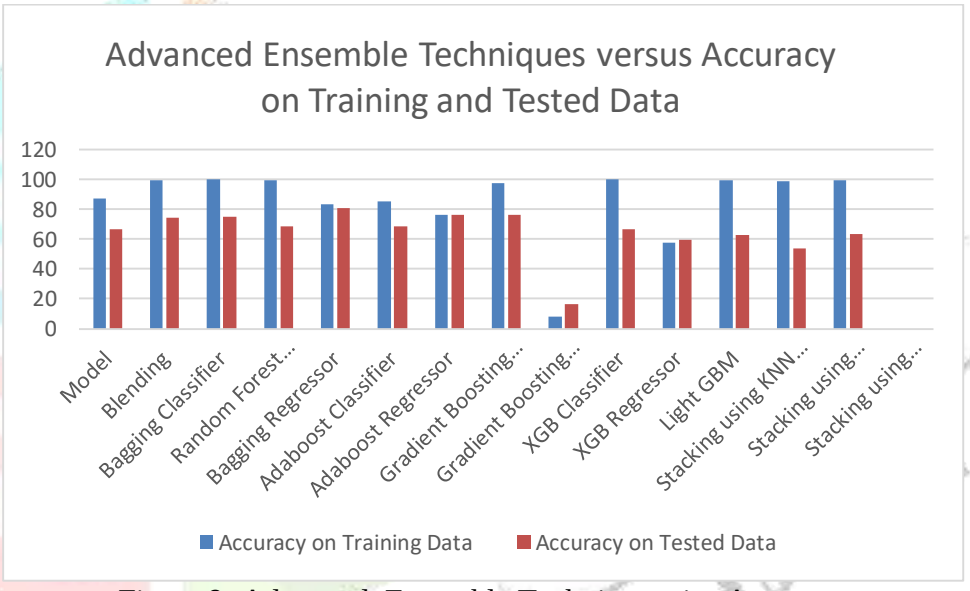

Figure 9: Advanced Ensemble Technique wise Accuracy

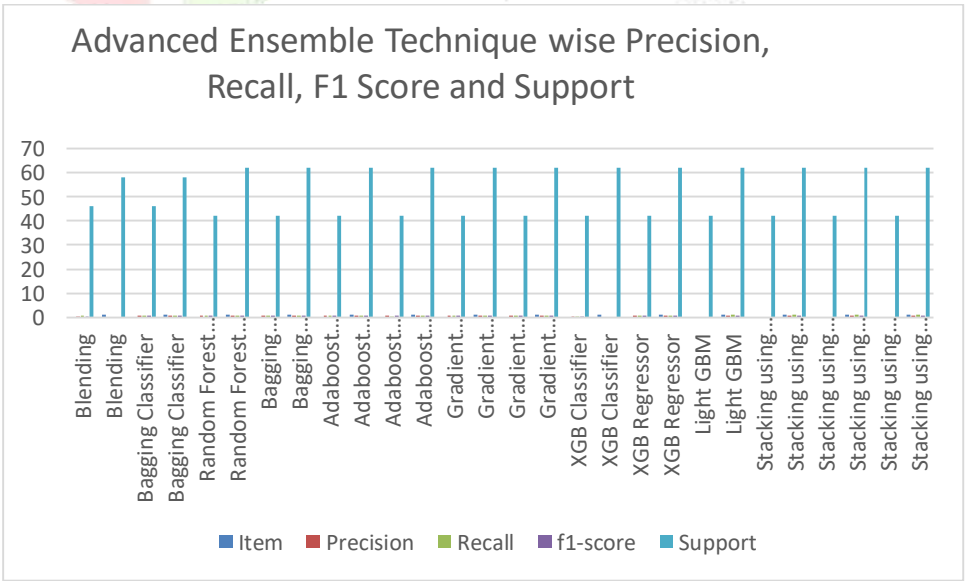

Figure 10: Advanced Ensemble Technique wise Precision, Recall, F1 Score and Support

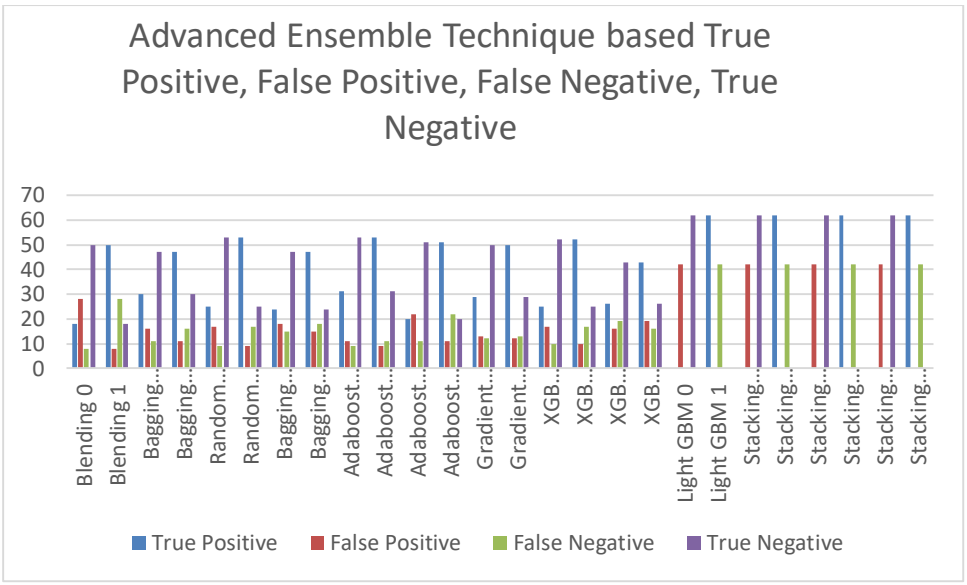

Figure 11: Advanced Ensemble Technique wise Classification Report

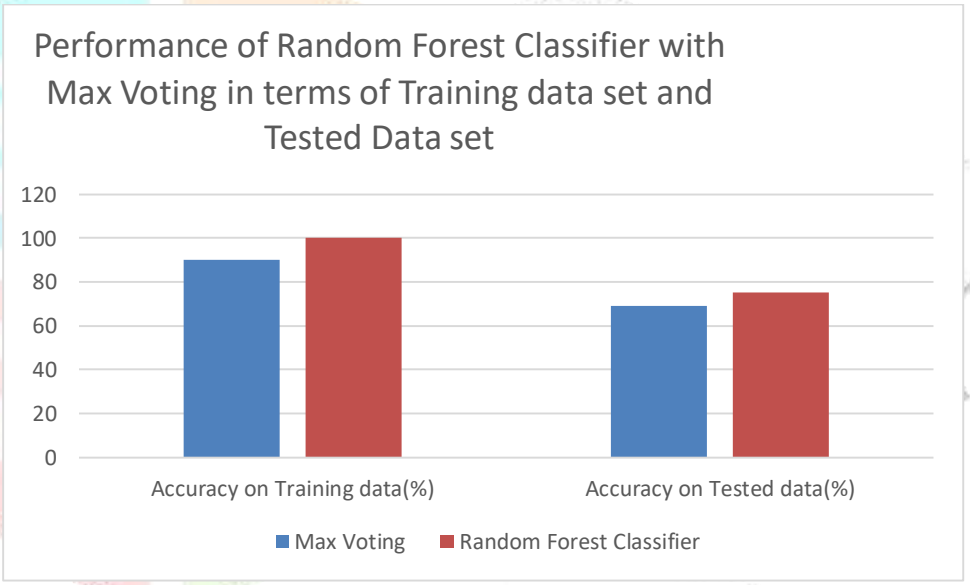

Figure 12: Performance of Random Forest Classifier with Max Voting in terms of Accuracy

8. Conclusion.

From table 2, it has been observed that performance of Support vector machine with linear kernel and linear regression model is excellent as compared to other models. The advantage of Support vector machine with linear kernel is that it can perform well at classifying non-linear data, it can reduce the overfitting of data, it can learn without a local minima, it can perform well on data sets that have many attributes.

The advantage of linear regression is simple in implementation, performs best on Linear Data, overfitting can be reduced by regularization. Comparing these, the preference of Support vector machine with linear kernel is more than linear regression.

From table 5, it has been observed that the performance of max voting is better as compared to other models. From table 6, it has been observed that the performance of random forest classifier is better as compared to other models. The accuracy of random forest classifier on training data set is 100 % in training data and 75 % in tested data. However, the accuracy of XGB regressor is 100 % on training data set and 66.35 % on tested data set. Considering the performance of training data set and tested data set, random forest classifier is preferable than XGB regressor. The performance of Max voting classifier is 90.04 % on training data and 69.23 % on tested data. Comparing with of random forest classifier, it has been found that the random forest classifier is the best ensemble technique among others. The advantage of random forest classifier is that it is robust to overfitting, it can handle missing Values, it can give feature importance i.e. It can provide insights

into those features which are most influential in predictions, it can help in aiding feature selection. Random Forest classifier can handle large datasets with numerous features and data points, making it versatile for various applications. Considering all these points, it is preferable to choose random forest classifier for prediction of lever disease.

References

[1] Ruhul Amin and Rubia Yasmin and Sabba Ruhi and Md Habibur Rahman and Md Shamim Reza,"Prediction of chronic liver disease patients using integrated projection based statistical feature extraction with machine learning approaches Infor- matics in Medicine Unlocked, https://doi.org/10.1016/j.imu.2022.101155

[2]TerIeoluwa Adegbola, Abiodun Bukunni Aborisade,AdelwaleAdetutu,"Liver Disease Prediction in rat sex posedto environment toxicantsusing Machine Learning Approaches" Informatics in Medicine Unlocked, https://doi.org/10.1016/j.imu.2023.101369 url: http://www.ijpam.eu.

[3] Srilatha Tokala, Koduru Hajarathaiah, Sai Ram Praneeth Gunda, Srinivasaro Botla, Lakshmikanth Nalluri, Pathipati Nagamahohar, Satish Anamalamudi, Murali Krishna Enduri,"Liver Disease Prediction and Classification using Machine Learn- ing Techniques", International Journal Of Advanced Computer Science and Applications, Volume 14,No 2 , 2023 https://thesai.org/Publications/IJACSA

[4] Osama Mohareb Khaled, Ahmed Zakareia Elsherif, Ahmed Salama, Mostafa Her- ajy,Elsayed Elsedimy,"Evaluating Machine Learning Models For Predictive Analytics Of Liver Disease Detection using Healthcare Big ",“International Journal of Electrical and Computer Engineering (IJECE)Vol. 15, No. 1, February 2025, pp.1162 1174 ISSN: 2088-8708, DOI: 10.11591/ijece.v15i1.pp1162-1174

[5] Engy A. El. Shafeiy, Ali I. El-Desouky , Sally M. Elghamrawy, " Prediction Of Liver Disease Based On Machine Learning Technique For Machine Data", Chapter in Advances in Intelligent Systems and Computing January 2018, Springer Interna- tional Publishing AG, part of Springer Nature 2018 A. E. Hassanien et al. (Eds.): AMLTA 2018, AISC 723, pp. 1–13, 2018 https://doi.org/10.1007/978-3-319-74690-636

[6] Riya, Barinderjit Kaur," Liver Disease Prediction Using Machine Learning Algorithms", International Journal Of Computer Applications(0975-8887),Volume 185- No 27,August 2023

[7] Neha Tanwar, Khandakar Faridar Rahman,”Machine Learning In Liver Disease Diagnosis: Current Progress And Future Opportunities”, IOP Conference Se- ries:MaterialsScienceandEngineering,IOPConf.Series:MaterialsScienceandEn- gineering1022 (2021)012029https://doi.org/10.1088/1757-899X/1022/1/012029

[8] Madhab Paul Choudhury, J. Paul Choudhury, “A Framework for Developing Correct Software Programs through Software Defect Prediction and Elimination using Machine Learning Models”, International Journal for Research in Applied Science Engineering Technology (IJRASET), ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538, Volume 12 Issue IX Sep 2024, page 402-411

[9] Madhab Paul Choudhury, J. Paul Choudhury, “A comparative study on the performance of Soft Computing models in the prediction of Orthopedic disease in the environment of Internet of Things” . Proceedings of ICCM 2022 Department of Electronics and Communication Engineering and the Department of Electrical Engineering, North Eastern Regional Institute of Science and Technology (NERIST), Arunachal Pradesh, India in collaboration with Emlyon Business School France, July 2022 Applications of Computational Intelligence in Management Mathematics Page247-258.https://link.springer.com/chapter/10.1007/978-3-031-25194-820

[10]https://www.analyticsvidhya.com/blog/2018/06/comprehensive-guide-for-ensemble-models/

[11] Lever Data Set: https://www.kaggle.com/datasets/uciml/indian-liver-patient-records.

[12] Saruchi Kukkar, Japreet Singh, “ Breast Cancer Detection and Classification by Features Non-Linear Mapping with Random Forest Classifier”, International Journal of Intelligent Systems and Applications in Engineering, ISSN:2147-6799, IJISAE, 2024, 12(1), 193–202.

[13] Saruchi Kukkar, Japreet Singh, “Breast Cancer Detection and Classification by Features Non-Linear Mapping with Random Forest Classifier”, International Journal of Intelligent Systems and Applications in Engineering, ISSN:2147-6799 2147-6799 IJISAE, 2024, 12(1), 193–202.

[14] Ms. Seema B. Rathod, Dr. Lata L. Ragha, “The Detection of Lung Tumors Using CT scan Images with Feature Extraction and Segmentation Techniques.”,International Journal of Intelligent Systems and Applications in Engineering ISSN:2147-6799214, IJISAE, 2024, 12(1), 628–638.

[15] Dr.Pallavi Hallappanavar Basavaraja, Dr. Nandeeswar Sampigehalli Basavaraju ,Dr. Pooja Nayak S., Dr.Anusha Preetham, Dr. Ramya R. S.,Prof. Shravya S. “A Framework for Brain Tumor Image Analysis using Convolution with RELU”, International Journal of Intelligent Systems and Applications in Engineering, ISSN:2147-67992, IJISAE, 2024, 12(3), 312–321.

[16] Abdulrahman Arishi , Suma Alex Kanjramni lkunathil , Sudha K. Rajan, Afshan Kausar, Arshia Arjumand and Fred Torres-Cruz, "Cardiac Abnormalities Classification Model Using Improved Deep Learning Approach", International Journal of Intelligent Systems and Applications in Engineering, ISSN:2147-679921, IJISAE, 2024, 12(1), 360–367.
[17] Srikanth R, Tamil Priya D., Jagadeesan S., Savita P. Patil, Anupama K. Ingale, Manojkumar Vivekanandan, Venkadeshan Ramalingam, "A Comprehensive Review on Cancer Prediction Using Machine Learning Techniques", International Journal of Intelligent Systems and Applications in Engineering, ISSN:2147-679921, IJISAE, 2024, 12(3), 115–127.
[18] Archana J. Jadhav, Dipali. H. Patil, G. S. Mate, R. A. Deshmukh, Anjali S. More, Chandan Prasad, "Detection of Brain Tumor using Fine-Tuned Pre-Trained Mobile Net Deep Learning Model", International Journal of Intelligent Systems and Applications in EngineeringISSN:2147-6799214, IJISAE, 2024, 12(3), 361–368.
[19] Aditya Gupta, Angad Singh, Maharshi Jani, Yuvraj Salaria, Dr. Vani Hiremani, Dr.Sudhanshu Gonge, Dr.Ketan Kotecha, "Diabetes Prediction and Apprehension with Focus Both on Clinical and Non-Clinical Factors", International Journal of Intelligent Systems and Applications in Engineering, ISSN:2147-679921, IJISAE, 2024, 12(1), 746–755.
[20] V. Kavitha, K. Ulagapriya, "Comparative Evaluation for Brain Tumor Detection Using Inception-V3 Architecture", International Journal of Intelligent Systems and Applications in Engineering,ISSN:2147-6799214, IJISAE, 2024, 12(1), 277–283.
[21] Sujeeth Babu Kolli, Vinay Kalisetti, Rakesh Varaparla, Vamsi Sai Chandra Vasarla, Veeraswamy Ammisetty, "Automated Diagnosis of Diabetic Retinopathy using Deep Learning and Image Analysis", International Journal of Intelligent Systems and Applications in Engineering,ISSN:2147-679921, IJISAE, 2024, 12(3), 174–179.
[22] Hajer Sayed Hussein, Hussein AlBazar, Roxane Elias Mallouhy, Fatima Al-Hebshi, "Deep Learning in Heart Murmur Detection: Analyzing the Potential of FCNN vs. Traditional Machine Learning Models", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 16, No. 2, 2025, 1296-1304.
[23] Mohammed A M Abueed1, Danial Md Nor2, Nabilah Ibrahim3, Jean-Marc Ogier4, "Pneumonia Detection Using Transfer Learning: A Systematic Literature Review", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 16, No. 2, 2025, 1032-1041.
[24] Janjhyam Venkata Naga Ramesh, P N S Lakshmi, Dr. Thalakola Syamsundararao, Elangovan Muniyandy, Linginedi Ushasree, Prof.Ts. Dr. Yousef A. Baker El-Ebiary, Dr. David Neels Ponkumar Devadhas, "Enhancing Chronic Kidney Disease Prediction with Deep Separable Convolutional Neural Networks", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 16, No. 2, 2025, 1011-1023.
[25] Nanda kumar Pandiyan, Subhashini Narayan, "Comparative Analysis of Cardiac Disease Classification Using a Deep Learning Model Embedded with a Bio-Inspired Algorithm", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 16, No. 2, 2025, 976-986.