# ENSEMBLE LEARNING TECHNIQUES FOR BREAST CANCER PREDICTION

**Abstract**

Breast cancer is predominantly diagnosed in women and remains a leading cause of rising health concerns among females. Manual identification of the disease is often time-consuming and limited in accessibility. To address this, automated diagnostic systems using machine learning (ML) have become increasingly valuable for early detection and classification of cancer. This paper explores the use of machine learning and ensemble learning techniques for classifying tumors. Specifically, it evaluates the performance of Logistic Regression, Support Vector Machines (SVM), Decision Trees, and Random Forests on a breast cancer dataset. The study compares models based on key performance metrics, including False Positive Rate, Accuracy, Precision, and Recall. The effectiveness of ensemble learning methods is also analyzed and benchmarked against individual models. Statistical analysis reveals that the ensemble model combining Decision Tree and Random Forest algorithms achieves an accuracy of 89.3%, while the ensemble of Logistic Regression and SVM reaches an accuracy of 90.4%. These ensemble models outperform their counterparts, demonstrating the advantages of combining multiple algorithms for improved diagnostic accuracy.

**Keywords:** Ensemble Machine Learning, Breast Cancer, Prediction, Accuracy



## 1 INTRODUCTION

Breast Cancer is a disease that causes death among women. It is a more hazardous cancer in public. 2018 statistical report by World Cancer Research, they are projected that over 2 million new cases registered out of 626,679 deaths. These cases continuously increased from 11.6% to 24.2% among women. If any symptoms are found, people visit doctors (oncologists) immediately [1]. The diagnosis of breast cancer, observing the medical past, physical inspection of the breast, and checking for puffiness. Different types of examinations of the breast with Magnetic Resonance Imaging, Ultrasound of the breast, X-ray, and Mammogram. The elimination of tissue from the

breast is examined by a pathologist. If the cancer is confirmed, patients go for a sentinel node biopsy [2]. It helps detect cancerous cells in the breast. If detected, cells oncologist orders additional tests. It diagnoses cancer easily based on tests. These tests are for finding Mammograms, Breast ultrasounds Biopsy. Benign and malignant tumors are classified by Machine Learning techniques [3].

Machine learning techniques can help achieve the optimal outcome for diminishing the price of medicines, helping people, enhancing healthcare value, and saving the lives of people. Machine learning techniques categorize benevolent and malevolent tumors by the finest [5]. Many machine learning models were researched to get the best results on breast cancer datasets. Based on these emerging technologies, a healthcare system for reduce the price of treatment and quality of diagnosis. In real-time systems, doctors make strong decisions by doctors to save people's lives in less time to retrieve accurate outcomes from different fields using classification methods [6]. The following paper is arranged in order of introduction in section one. Section two proposes the system and architecture. Section three discusses materials and methods. Section four states the consequences and analysis. Concludes the paper with the final section.

## 2 PROPOSED SYSTEM AND ARCHITECTURE

This paper provides the enactment of models using Logistic Regression, Support Vectors, Decision Trees, and Random Forest on the dataset. The efficiency of the ensemble machine learning models is measured and equated with other machine learning models [7]. From statistical analysis, the ensemble model predicts that the accuracy rate is 89.3 % for decision tree and random forest models and 90.4% for logistic regression and support vector models. If you compare these two ensemble models with individual models provide a higher accuracy rate. The following algorithm represents the procedure for finding breast cancer [8].

Step 1: Input (dataset of breast cancer)
Step 2: Calculate the feature cost from every cell.

Step 3: Compute the mean of the features' cost.

Step 4: Compute the boundary

Step 5: Calculate the part of the affected cell.

Step 6: Detect the mean difference.

Step 7: Formula: Perimeter^2 / area - 1.0

Step 8: Compute concaveness

Step 9: Total curved parts of the cell outline

Step 10: Identify the symmetry of cells

11: Finally, fractal calculation

Figure 1 illustrates the architectural representation of our projected system. Starting from input to output, so many phases are included in this system [10].

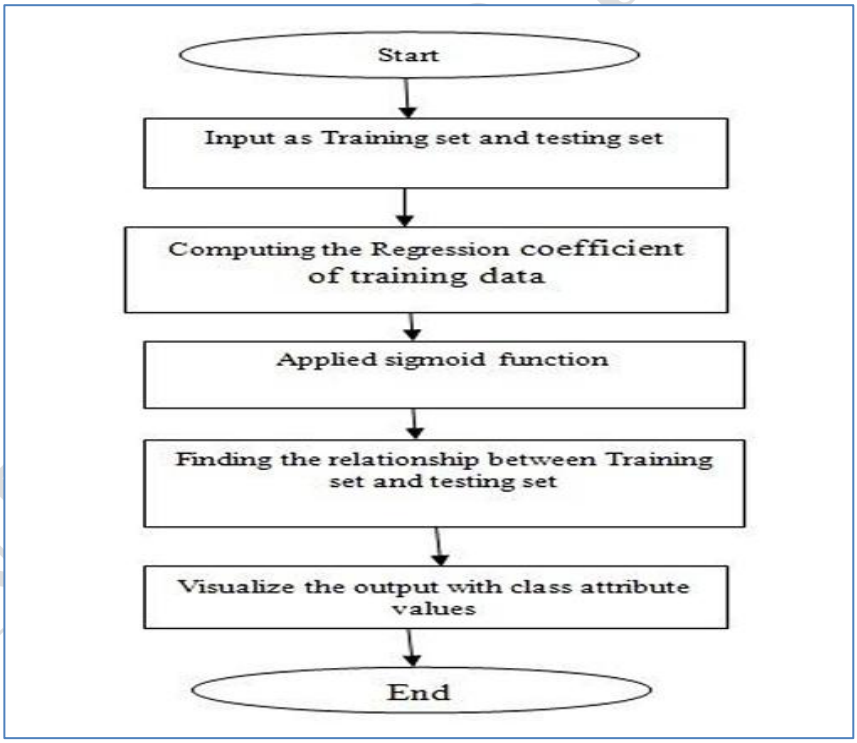


**Figure 1:** Diagrammatic representation of our proposed system [9]

The above figure depicts a flowchart outlining the process of a machine learning classification model, specifically using logistic regression. Here's a step-by-step description of the flowchart:

1. **Start**
   The process begins.
2. Input as Training set and Testing set
   The dataset is divided into two parts: training data for model learning and testing data for evaluation.
3. Computing the Regression coefficient of training data
   The logistic regression model is trained on the training set, resulting in the computation of regression coefficients (weights).
4. Applied sigmoid function
   The sigmoid (logistic) function is applied to map predicted values to a probability between 0 and 1.
5. Finding the relationship between the Training set and the Testing set
   The model uses learned relationships from the training data to make predictions on the testing data.
6. Visualize the output with class attribute values
   The final results are visualized, showing how data points are classified based on the model's predictions.
7. **End**
   The process concludes.

## 3. RESULTS AND ANALYSIS

### 3.1 Dataset Description

The Dataset for health and it is for Social Good: Women Coders' Bootcamp, organized by Artificial Intelligence for Development in collaboration with UNDP Nepal. features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe the characteristics of the cell nuclei present in the image.

**Table 1:** Sample Dataset

| id | diagnosis | radius_m | texture_m | perimeter | area_mea | smoothne | compactn | concavity_ | concave_p | symmetry | fractal_dir | radius_se | texture_se | perimeter | area_se | smoothne | compactn | concavity_co |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 842302 | M | 17.99 | 10.38 | 122.8 | 1001 | 0.1184 | 0.2776 | 0.3001 | 0.1471 | 0.2419 | 0.07871 | 1.095 | 0.9053 | 8.589 | 153.4 | 0.006399 | 0.04904 | 0.05373 |
| 842517 | M | 20.57 | 17.77 | 132.9 | 1326 | 0.08474 | 0.07864 | 0.0869 | 0.07017 | 0.1812 | 0.05667 | 0.5435 | 0.7339 | 3.398 | 74.08 | 0.005225 | 0.01308 | 0.0186 |
| 84300903 | M | 19.69 | 21.25 | 130 | 1203 | 0.1096 | 0.1599 | 0.1974 | 0.1279 | 0.2069 | 0.05999 | 0.7456 | 0.7869 | 4.585 | 94.03 | 0.00615 | 0.04006 | 0.03832 |
| 84348301 | M | 11.42 | 20.38 | 77.58 | 386.1 | 0.1425 | 0.2839 | 0.2414 | 0.1052 | 0.2597 | 0.09744 | 0.4956 | 1.156 | 3.445 | 27.23 | 0.00911 | 0.07458 | 0.05661 |
| 84358402 | M | 20.29 | 14.34 | 135.1 | 1297 | 0.1003 | 0.1328 | 0.198 | 0.1043 | 0.1809 | 0.05883 | 0.7572 | 0.7813 | 5.438 | 94.44 | 0.01149 | 0.02461 | 0.05688 |
| 843786 | M | 12.45 | 15.7 | 82.57 | 477.1 | 0.1278 | 0.17 | 0.1578 | 0.08089 | 0.2087 | 0.07613 | 0.3345 | 0.8902 | 2.217 | 27.19 | 0.00751 | 0.03345 | 0.03672 |
| 844359 | M | 18.25 | 19.98 | 119.6 | 1040 | 0.09463 | 0.109 | 0.1127 | 0.074 | 0.1794 | 0.05742 | 0.4467 | 0.7732 | 3.18 | 53.91 | 0.004314 | 0.01382 | 0.02254 |
| 84458202 | M | 13.71 | 20.83 | 90.2 | 577.9 | 0.1189 | 0.1645 | 0.09366 | 0.05985 | 0.2196 | 0.07451 | 0.5835 | 1.377 | 3.856 | 50.96 | 0.008805 | 0.03029 | 0.02488 |
| 844981 | M | 13 | 21.82 | 87.5 | 519.8 | 0.1273 | 0.1932 | 0.1859 | 0.09353 | 0.235 | 0.07389 | 0.3063 | 1.002 | 2.406 | 24.32 | 0.005731 | 0.03502 | 0.03553 |
| 84501001 | M | 12.46 | 24.04 | 83.97 | 475.9 | 0.1186 | 0.2396 | 0.2273 | 0.08543 | 0.203 | 0.08243 | 0.2976 | 1.599 | 2.039 | 23.94 | 0.007149 | 0.07217 | 0.07743 |
| 845636 | M | 16.02 | 23.24 | 102.7 | 797.8 | 0.08206 | 0.06669 | 0.03299 | 0.03323 | 0.1528 | 0.05697 | 0.3795 | 1.187 | 2.466 | 40.51 | 0.004029 | 0.009269 | 0.01101 |
| 84610002 | M | 15.78 | 17.89 | 103.6 | 781 | 0.0971 | 0.1292 | 0.09954 | 0.06606 | 0.1842 | 0.06082 | 0.5058 | 0.9849 | 3.564 | 54.16 | 0.005771 | 0.04061 | 0.02791 |
| 846226 | M | 19.17 | 24.8 | 132.4 | 1123 | 0.0974 | 0.2458 | 0.2065 | 0.1118 | 0.2397 | 0.078 | 0.9555 | 3.568 | 11.07 | 116.2 | 0.003139 | 0.08297 | 0.0889 |
| 846381 | M | 15.85 | 23.95 | 103.7 | 782.7 | 0.08401 | 0.1002 | 0.09938 | 0.05364 | 0.1847 | 0.05338 | 0.4033 | 1.078 | 2.903 | 36.58 | 0.009769 | 0.03126 | 0.05051 |
| 84667401 | M | 13.73 | 22.61 | 93.6 | 578.3 | 0.1131 | 0.2293 | 0.2128 | 0.08025 | 0.2069 | 0.07682 | 0.2121 | 1.169 | 2.061 | 19.21 | 0.006429 | 0.05936 | 0.05501 |
| 84799002 | M | 14.54 | 27.54 | 96.73 | 658.8 | 0.1139 | 0.1595 | 0.1639 | 0.07364 | 0.2303 | 0.07077 | 0.37 | 1.033 | 2.879 | 32.55 | 0.005607 | 0.0424 | 0.04741 |
| 848406 | M | 14.68 | 20.13 | 94.74 | 684.5 | 0.09867 | 0.072 | 0.07395 | 0.05259 | 0.1586 | 0.05922 | 0.4727 | 1.24 | 3.195 | 45.4 | 0.005718 | 0.01162 | 0.01998 |

### 3.2 Data Collection

After collecting data, we need to know what the shape of this dataset is. Here, we have an attribute(property) called data. The data used in this experiment is 569 rows × 33 columns, which means 33 attributes are used here. Using this method, we can see how many object (categorical) types of features exist in the dataset.

### 3.3 Data Filtering

Now, we have one categorical feature, so we need to convert it into numeric values. Finally, we can see in this output that categorical values are converted into 0 and 1. Find the correlation between other features and mean features only.

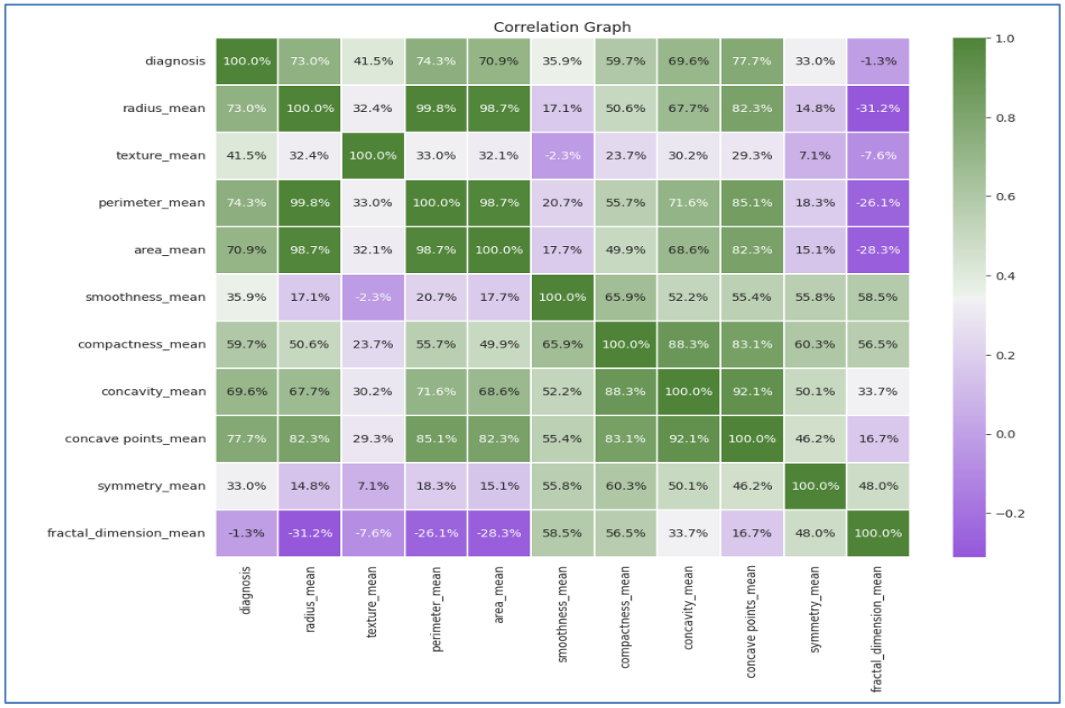


**Figure 2:** correlation graph of attributes

A correlation heatmap is a visual tool used in data analysis to show the strength and direction of relationships (correlations) between multiple variables (features) in a dataset. It's especially useful in exploratory data analysis (EDA) to understand the underlying structure of the data before applying machine learning models.

The correlated attribute is content that is partly confined to another attribute. Most correlated attributes are more interdependent and have more redundancy. The following figure 3 shows the relationship of each attribute. A heatmap is a graphical representation that shows the correlation between numerous variables as a matrix. It's shows that shows how closely connected dissimilar variables. The following Figure 3 shows the correlation representation with the heatmap.

### 3.4 Model Implementation

In model implementation, the part mainly *concentrates* on

- Train Test Splitting.
- Preprocessing and model selection
- Import Machine Learning Models
- Check the Model Accuracy, Errors, and Validations

### 3.4.1 Feature Selection

Select a feature for predictions

- Take the dependent and independent features for prediction

**Table 2.** Selection of features

| | radius_mean | perimeter_mean | area_mean | symmetry_mean | compactness_mean | concave points_mean |
|---|---|---|---|---|---|---|
| **0** | 17.99 | 122.80 | 1001.0 | 0.2419 | 0.27760 | 0.14710 |
| **1** | 20.57 | 132.90 | 1326.0 | 0.1812 | 0.07864 | 0.07017 |
| **2** | 19.69 | 130.00 | 1203.0 | 0.2069 | 0.15990 | 0.12790 |
| **3** | 11.42 | 77.58 | 386.1 | 0.2597 | 0.28390 | 0.10520 |
| **4** | 20.29 | 135.10 | 1297.0 | 0.1809 | 0.13280 | 0.10430 |
| **...** | ... | ... | ... | ... | ... | ... |
| **564** | 21.56 | 142.00 | 1479.0 | 0.1726 | 0.11590 | 0.13890 |
| **565** | 20.13 | 131.20 | 1261.0 | 0.1752 | 0.10340 | 0.09791 |
| **566** | 16.60 | 108.30 | 858.1 | 0.1590 | 0.10230 | 0.05302 |
| **567** | 20.60 | 140.10 | 1265.0 | 0.2397 | 0.27700 | 0.15200 |
| **568** | 7.76 | 47.92 | 181.0 | 0.1587 | 0.04362 | 0.00000 |

Feature selection is a key step in the machine learning pipeline. It involves choosing the most relevant variables (features) from your dataset that contribute the most to your predictive model. Good feature selection improves model performance, reduces overfitting, and leads to faster, more efficient training.

Divided the data into the Training and Testing Set by 33% and established 15 fixed records. Perform Feature Scaling for Strong features by eliminating the mean and pre-processing to a specific unit variance.

### 3.5 ML Model Selection and Model Prediction

#### 3.5.1 Model Building

Now, we are ready to build our prediction model for the I made function for model building and performing prediction, and measuring its prediction accuracy score.

#### 3.5.2 Arguments

1. model => ML Model Object
2. Feature Training Set data
3. Feature Testing Set data
4. Targeted Training Set data
5. Targeted Testing Set data

Let's make a dictionary for multiple models for bulk predictions. Before sending it to the prediction, check the key and values to store its values in the Data Frame below.

#### 3.5.3 Model Implementing

Now, train the models one by one and show the classification report of particular models. The precision, recall, F1 score, support, and accuracy values of classification algorithms. The precision, recall, F1 score, support, and accuracy values of decision tree and support vector tree algorithms. In the confusion metric of classification algorithms, the x label shows the negative, and the y label shows the True Positive. While predicting, we can store the model's score and prediction values in a newly generated data frame.

**Table 3.** Accuracy of four models

| | model_name | score | accuracy_score | accuracy_percentage |
|---|---|---|---|---|
| 0 | LogisticRegression | 0.916010 | 0.909574 | 90.96% |
| 1 | RandomForestClassifier | 0.992126 | 0.925532 | 92.55% |
| 2 | DecisionTreeClassifier | 1.000000 | 0.909574 | 90.96% |
| 3 | SVC | 0.923885 | 0.914894 | 91.49% |

### 3.6 Ensemble models

#### 3.6.1 Decision Tree + Random Forest

Ensemble learning is a machine learning technique where multiple models (often called "base learners" or "weak learners") are combined to produce a more robust and accurate prediction than a single model. It leverages the "wisdom of the crowd" by combining diverse perspectives and potentially compensating for the individual errors of each model.

A Decision Tree is a flowchart-like tree structure. Random Forest is an ensemble learning method that combines multiple Decision Trees to improve performance and robustness. Combines the strengths of individual weak learners (Decision Trees). Reduces overfitting, improves accuracy, and handles missing or unbalanced data better.

```
from sklearn.ensemble import VotingClassifier

dt=DecisionTreeClassifier(max_features='sqrt', min_samples_leaf=3);
rf=RandomForestClassifier(min_samples_leaf=2, min_samples_split=5);

ensemble_model = VotingClassifier(estimators=[('decision_tree', dt), ('random_forest', rf)], voting='hard')

ensemble_model.fit(X_train, y_train)

# Make predictions on the testing data
predictions = ensemble_model.predict(X_test)

# Calculate accuracy
accuracy = accuracy_score(y_test, predictions)
print(f"Accuracy: {accuracy}")
```

```
Accuracy: 0.8936170212765957
```

**Figure 3:** Decision Tree to improve performance and robustness

#### 3.6.2 SVC + Logistic Regression

Ensemble learning is a machine learning technique where multiple models (often called "base learners" or "weak learners") are combined to produce a more robust and accurate prediction than

a single model. It leverages the "wisdom of the crowd" by combining diverse perspectives and potentially compensating for the individual errors of each model.

SVC is a classification algorithm from the Support Vector Machine (SVM) family. Logistic Regression is a linear model used for binary classification. Combining SVC and Logistic Regression in an ensemble model aims to leverage the strengths.

```
from sklearn.ensemble import VotingClassifier

svc=SVC(C=10, gamma=0.001);
lr=LogisticRegression(C=0.001, solver='liblinear');

ensemble_model = VotingClassifier(estimators=[('svc', svc), ('logistic_regression', lr)], voting='hard')

ensemble_model.fit(X_train, y_train)

# Make predictions on the testing data
predictions = ensemble_model.predict(X_test)

# Calculate accuracy
accuracy = accuracy_score(y_test, predictions)
print(f"Accuracy: {accuracy}")
```

```
Accuracy: 0.9042553191489362
```

**Figure 4:** Classification algorithm from the Support Vector Machine

If you observe the 3.5.1 and 3.5.2 ensemble models, the accuracy rate is 89.3 % for the decision tree and random forest models. 90.4% for logistic regression and support vector models. If you compare these two ensemble models with individual models, you will find a higher accuracy rate. (Table 2).

## 4. CONCLUSION

This paper focuses on the application of machine learning techniques—including Logistic Regression, Support Vector Machines (SVM), Decision Trees, and Random Forests—on ancient datasets

that may contain novel types of input data. The models are evaluated based on key performance metrics such as False Positive Rate, Accuracy, Precision, and Recall. Ensemble machine learning methods are also explored, and their efficacy is measured and compared with individual model results. Statistical analysis shows that ensemble models achieve higher accuracy rates: 89.3% for the combination of Decision Tree and Random Forest, and 90.4% for the combination of Logistic Regression and Support Vector Machine. These results indicate that ensemble models outperform individual models in terms of predictive accuracy.

Disclaimer (Artificial intelligence)

Option 1:

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc.) and text-to-image generators have been used during the writing or editing of this manuscript.

Option 2:

Author(s) hereby declare that generative AI technologies such as Large Language Models, etc. have been used during the writing or editing of manuscripts. This explanation will include the name, version, model, and source of the generative AI technology and as well as all input prompts provided to the generative AI technology

Details of the AI usage are given below:

1.
2.
3.



## References

1. Dr. R. Vijaya Kumar Reddy, Dr. Shaik Subhani, Dr. G. Rajesh Chandra, Dr. B. Srinivasa Rao," Breast Cancer Prediction using Classification Techniques", International Journal of Emerging Trends in Engineering Research, Vol. 8, No.9,2020.
2. Ms. Mamatha, Srinivasa Datta and Subhani Shaik," Fake Profile Identification using Machine

Learning Algorithms", International Journal of Engineering Research and Applications (IJERA), Vol 11, Series-2, July 2021.

3. GK Reddy, NR Vullam, GS Sekhar, D Sundaragiri, S Shaik," Ensemble Learning Algorithms based on Road Accident Data Prediction", International Conference on Contemporary Pervasive Computational Intelligence, Sreenidhi University, Hyderabad, 27-28 September 2024.
4. Nagagopiraju Vullam, Subhani Shaik, Gondi Konda Reddy, Korivi Vamshee Krishna, Aseena Babu Shaik, Forest fire prediction using K-Nearest Neighbour Model", International Conference on Contemporary Pervasive Computational Intelligence, Sreenidhi University, Hyderabad, 27-28 September 2024.
5. Shiva Keertan J and Subhani Shaik," Machine Learning Algorithms for Oil Price Prediction", International Journal of Innovative Technology and Exploring Engineering, Volume-8 Issue-8, 2019.
6. Ravi Kumar A, C. Sunil Kumar, Subhani Shaik, K. R. Praneeth," Machine learning techniques to predict and manage knee injury in sports medicine", International Journal of Emerging Trends in Health Sciences Volume 8, Issue 2, (2024) 26-35.
7. Egamamidi Rishika Reddy, Sai Durga Satturi, Medavarapu Harshini, and Subhani Shaik," Rose Plant Leaf Disease Recognition Using Machine Learning Methodologies", Asian Journal of Research in Computer Science, Volume 17, Issue 11, Page 65-72, June 2024.
8. V. Kakulapati, Shaik Subhani, Gowripriya Kukunuri, Pallavi KSL Lakkoju," Demographic Factors Based on Predicting Mental Health of Coronavirus Victims", International Journal of Pharmaceutical Sciences (0975-4725), Vol. 2, Issue 4, April 2024.
9. Shaik Zareena, P Jaya Surya, T Divya Rani, B. Sanjeev, and Subhani Shaik," Machine Learning Algorithms for Finding Credit Score Prediction for Optimal Outcome", Journal of Emerging Technologies and Innovative Research (2349-5162), Vol. 11, Issue 4, April 2024.
10. Gunda Nithin, M Sai Trilochan, G Sangamesh, Vigneswara Reddy, Subhani Shaik" Regression techniques based on weather forecasting prediction", Journal of Emerging Technologies and Innovative Research (2349-5162), Vol. 11, Issue 4, April 2024.