BREAST CANCER PREDICTION

TEAM -D10

VADA GOURI HANSIKA REDDY-CB.SC.U4AIE23304

MALAVIKA S PRASAD-CB.SC.U4AIE23315

KATIKALA DEDEEPYA-CB.SC.U4AIE23349

GESHNA B-CB.SC.U4AIE23360

OVERVIEW

- *Utilized CBIS-DDSM and Wisconsin Breast Cancer datasets for mammographic images and clinical features.*
- *Implemented Convolutional Neural Network (CNN) model for image data classification.*
- *Used Logistic Regression, Decision Tree, K-Nearest Neighbors (KNN), and Random Forest for tabular data classification.*
- *Preprocessed image data with techniques like resizing, normalization, and augmentation.*

- *Standardized tabular data for training and testing sets.*
- *Tabular data models outperformed CNN model in prediction accuracy.*
- *Future work includes hyperparameter tuning, advanced image processing, genomic data integration, and user-friendly interface.*

IMPLEMENTATION

Data Collection and Preprocessing:

Data collection and preprocessing involve resizing, normalizing, and augmenting image data from CBIS-DDSM and histopathology datasets, and loading the Wisconsin Breast Cancer Dataset for training and testing sets.

Model Development:

The study utilized Convolutional Neural Network (CNN) and Tabular Data Models to classify images as cancerous or non-cancerous, using Logistic Regression, Decision Tree, KNN, and Random Forest algorithms.

IMPLEMENTATION

- ***Training and Validation:***

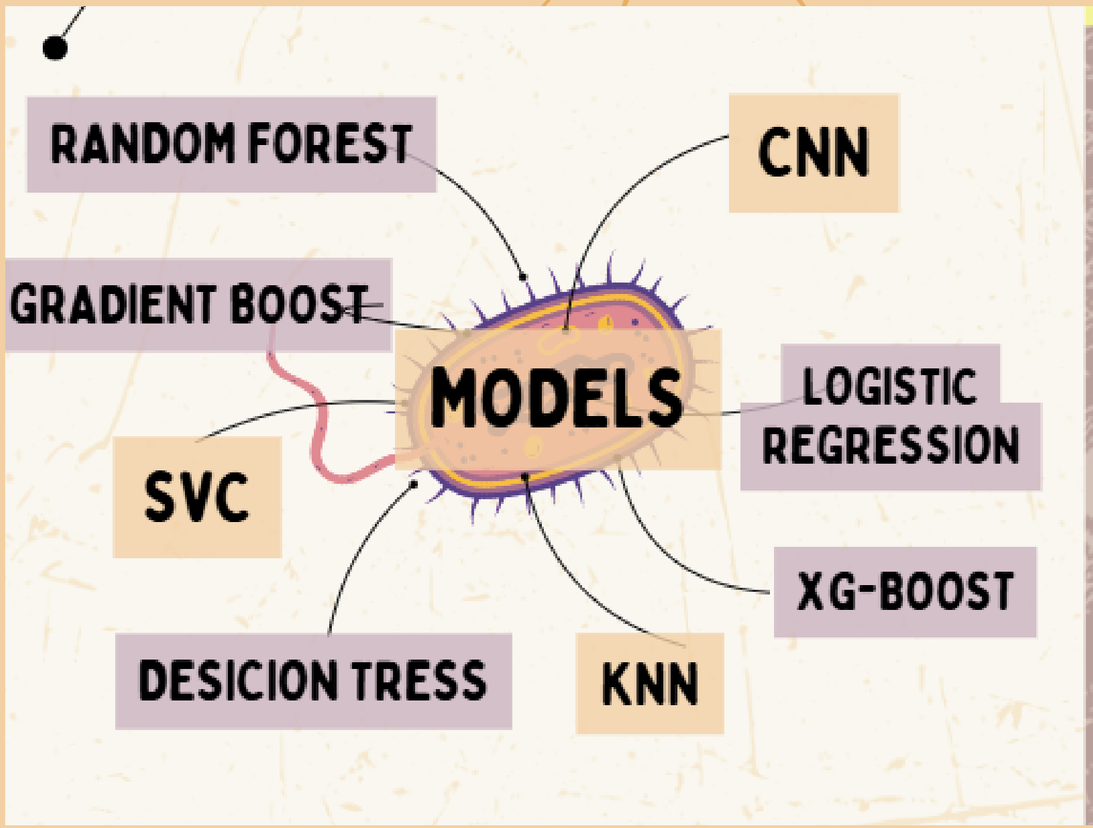
The CNN model was trained on image data and cross-validated for improved performance, while each model was trained and cross-validated for tabular data to maximize accuracy.

- ***Evaluation and Comparison:***

Evaluated each model using accuracy, precision, recall, and F1-score. Compared the performance between the CNN model and traditional ML models on CSV data, observing that tabular data models provided higher accuracy.

- ***Result Analysis and Future Recommendations:***

The study analyzed model performance, finding structured tabular data yielded better results than image data, and suggested future improvements like advanced image preprocessing and ensemble learning.



LOGISTIC REGRESSION

- ***Purpose:*** *Logistic Regression is a statistical model primarily used for binary classification tasks.*
- ***Probability Estimation:*** *It estimates the probability that an input belongs to a specific category.*
- ***Key Mechanism:*** *Uses the sigmoid function to map outputs to a probability range between 0 and 1.*
- ***Decision Boundary:*** *Assumes a linear relationship between input features and the log-odds of the outcome.*
- ***Strengths:*** *Easy to implement, interpret, and performs well with linearly separable data.*
- ***Limitations:*** *Limited by its assumption of linearity and may underfit on more complex relationships.*

DECISION TREES

- ***Purpose**: Decision Trees are a non-parametric method used for classification and regression tasks.*
- ***Structure:** They make decisions based on feature values, creating a tree-like structure.*
- ***Splitting Criteria:** Nodes split using criteria like Gini impurity or entropy to maximize information gain.*
- ***Terminal Nodes:** Leaf nodes represent the final classification or regression outcome for data points.*
- ***Strengths:** Easy to interpret, visualize, and suitable for both numerical and categorical data.*
- ***Limitations:** Prone to overfitting, particularly with deep trees, and sensitive to noisy data.*

RANDOM FOREST

- ***Purpose:*** *Random Forest is an ensemble learning method that builds multiple decision trees and combines their outputs for classification or regression.*
- ***Bootstrap Aggregation:*** *Each tree is trained on a random subset of the data, increasing diversity and reducing variance.*
- ***Feature Randomness:*** *At each split, a random subset of features is selected, improving the model's generalization.*
- ***Tuning and Parallelization:*** *Key hyperparameters like the number of trees and tree depth can be tuned, with trees trained independently for parallel processing.*
- ***Strengths:*** *Reduces overfitting and performs well with large feature sets, retaining high accuracy.*
- ***Limitations:*** *Less interpretable than individual trees and more resource-intensive in terms of computation and memory.*

SUPPORT VECTOR CLASSIFICATION (SVC)

- ***Purpose:*** *SVC aims to find the optimal hyperplane that maximally separates data points from different classes.*
- ***Support Vectors:*** *Key data points, known as support vectors, help define and position this separating hyperplane.*
- ***Hyperparameters:*** *Important parameters include C (regularization) for balancing margin and errors, and Gamma for adjusting decision surface complexity, especially in non-linear kernels.*
- ***Kernels:*** *SVC uses different kernels (e.g., linear, polynomial, RBF) to handle both linear and non-linear separable data by mapping it into higher dimensions.*
- ***Advantages:*** *Effective for high-dimensional data, versatile due to kernel options, and resists overfitting with appropriate parameter tuning.*
- ***Limitations:*** *Computationally intensive on large datasets, sensitive to outliers, and less interpretable compared to simpler models like logistic regression.*
-

GRADIENT BOOST

Definition:

- *An ensemble learning method that builds a series of decision trees, each one focusing on reducing the errors of the previous trees.*

Working Principle:

- *Starts with a weak initial model (like a shallow tree).*
- *Each subsequent tree is trained on the residual errors (differences between actual and predicted values) of prior models.*
- *Trees are added sequentially, improving the model step-by-step.*

Key Components:

- ***Loss Function:*** *Guides the model by measuring prediction error; common losses are Log Loss (classification) and Mean Squared Error (regression).*
- ***Learning Rate:*** *Controls the contribution of each tree to prevent overfitting.*
- ***Number of Trees:*** *More trees generally increase accuracy but at the cost of training time.*

Advantages:

- *Highly accurate, captures complex patterns, reduces bias.*
- *Effective for structured/tabular data like the Wisconsin Breast Cancer dataset.*

Disadvantages:

- *Can be computationally expensive.*
- *Prone to overfitting without careful parameter tuning.*

XG-BOOST

Definition:

- *XGBoost is an optimized gradient-boosting algorithm designed for performance and efficiency. It improves predictive accuracy and speeds up model training.*

Objective Function:

- *Includes both a loss function and a regularization term.*

Key Features:

- ***Parallel Processing:*** *Utilizes CPU cores to speed up computation.*
- ***Regularization Terms:*** *L1 (Lasso) and L2 (Ridge) regularization reduce model complexity and prevent overfitting.*
- ***Handling Missing Data:*** *Automatically learns the best direction to handle missing values in the data.*

Advantages:

- *Highly flexible and can handle structured/tabular data very well.*
- *Outperforms many other algorithms in accuracy for various ML problems.*

Disadvantages:

- *More complex than traditional boosting, requiring careful tuning of multiple hyperparameters.*
- *Resource-intensive, potentially consuming high memory on large datasets.*

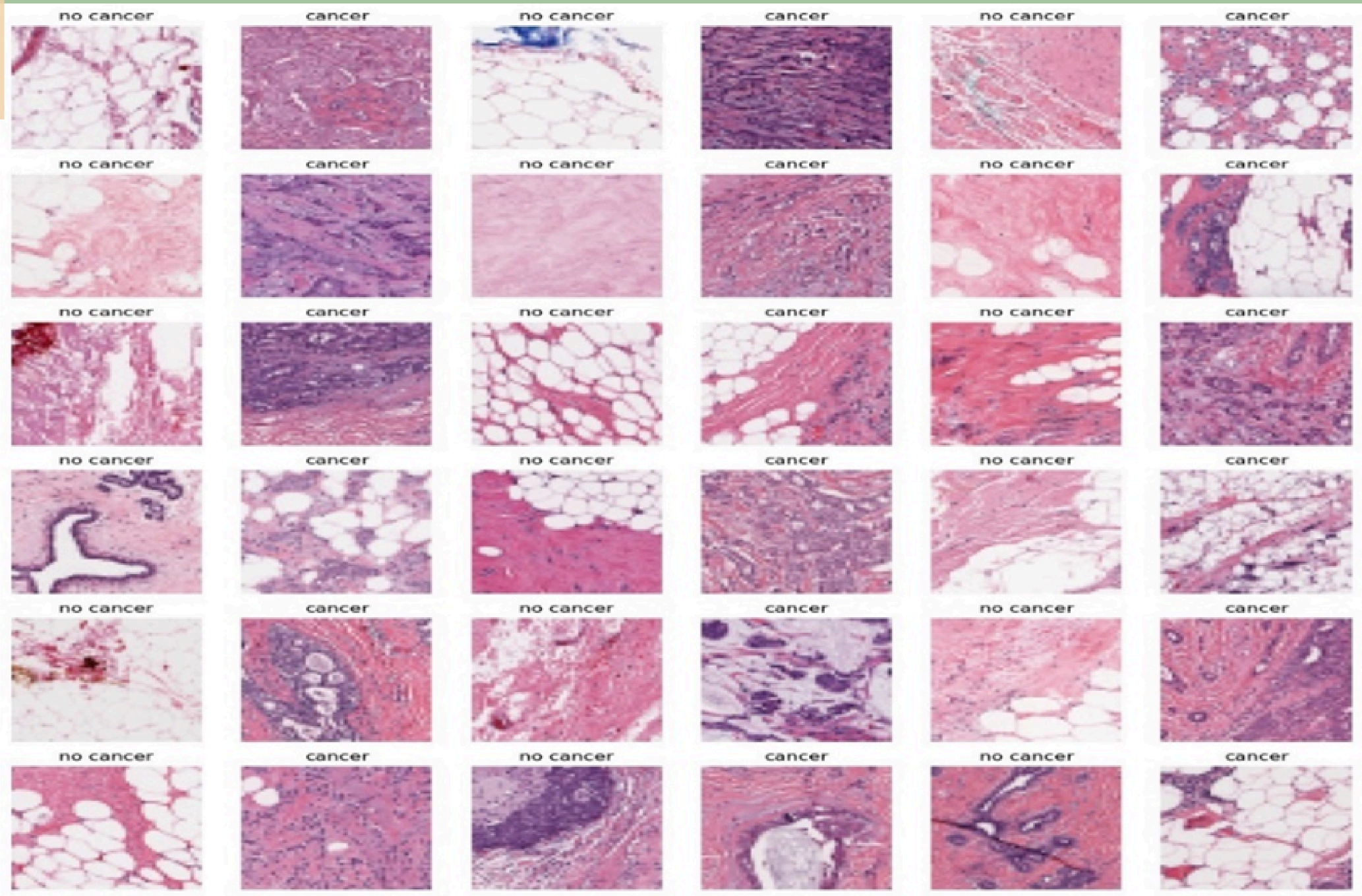
K-NEAREST NEIGHBORS (KNN)

- ***Purpose**: KNN is an **instance-based learning** algorithm used for **classification** and **regression.***
- ***Classification Method:** It classifies instances based on the majority label of the k-nearest neighbors in the feature space.*
- ***Distance Metrics:** Commonly uses **Euclidean, Manhattan, or Minkowski** distances to measure similarity between instances.*
- ***Parameter Selection:** The choice of k (number of neighbors) is crucial and typically optimized through **cross-validation.***
- ***Strengths:** Simple to understand and implement, with no explicit training phase.*
- ***Limitations:** Computationally intensive at prediction time, and sensitive to irrelevant features and data scaling.*

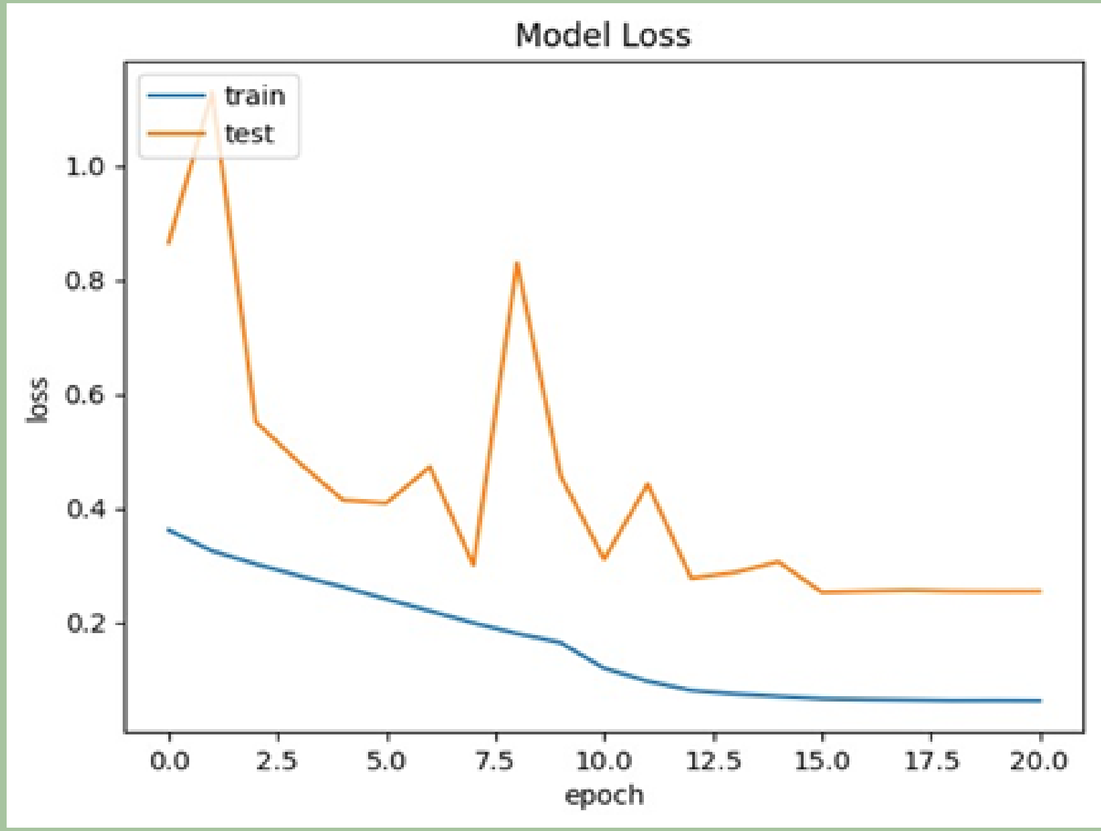
CNN

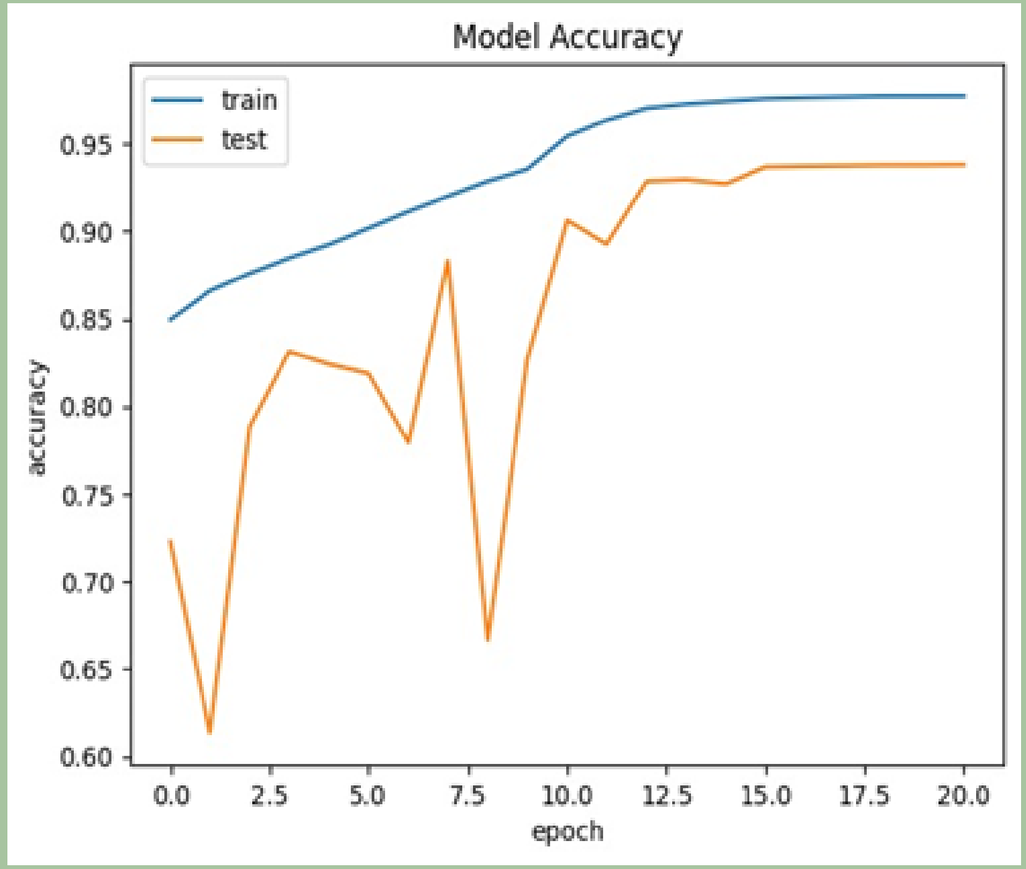
- ***Convolutional Neural Network*** *(CNN) is a deep learning model for processing grid-like data like* ***images.***
- *It learns spatial hierarchies of features through convolutional layers.*
- *Architecture components include* ***convolutional layers, pooling layers, flattening layer, fully connected layers,*** *and* ***output layer****.*
- *Implementation details include* ***ReLU*** *for* ***hidden layers*** *and* ***sigmoid*** *for* ***output layer.***
- ***Loss Function*** *measures the error between predicted probabilities and actual labels.*
- ***Optimizers*** *like* ***Adam*** *or* ***RMSprop*** *are commonly used for efficient model training.*
- ***Strengths:*** *CNNs extract* ***relevant features*** *and* ***capture spatial relationships*** *in images.*
- ***Weaknesses:*** *Requires large datasets and is sensitive to image variations.*

RESULTS

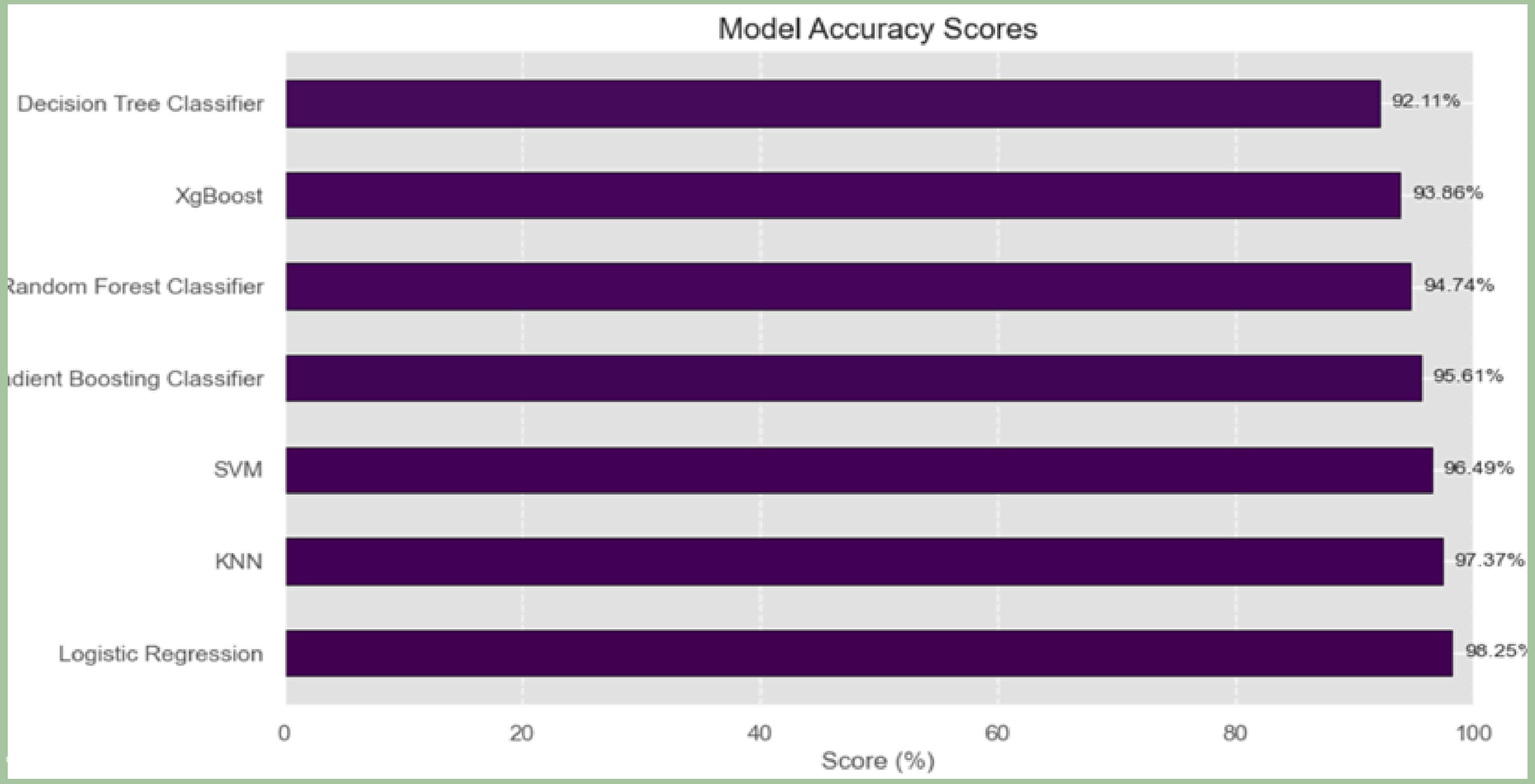

no cancer
cancer
no cancer
cancer
no cancer
cancer
no cancer
cancer
no cancer
cancer
no cancer
cancer
no cancer
cancer
no cancer
cancer
no cancer
cancer
no cancer
cancer
no cancer
cancer
no cancer
cancer
no cancer
cancer
no cancer
cancer
no cancer
cancer
no cancer
cancer
no cancer
cancer
no cancer
cancer
no cancer
cancer
no cancer
cancer
no cancer
cancer
no cancer
cancer
no cancer
cancer
no cancer
cancer
no cancer
cancer

RESULTS


Model Loss
train
test
loss
epoch
1.0
0.8
0.6
0.4
0.2
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0


Model Accuracy
train
test
accuracy
epoch
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0

RESULTS


Model Accuracy Scores
Decision Tree Classifier
XgBoost
Random Forest Classifier
dient Boosting Classifier
SVM
KNN
Logistic Regression
92.11%
93.86%
94.74%
95.61%
96.49%
97.37%
98.25%
0
20
40
60
80
100
Score (%)

ACCURACY SUMMARY

MODEL NAME	ACCURACY	F1 SCORE
Logistic Regression	98.25	0- 0.99 1- 0.98
KNN	97.31	0- 0.98 1- 0.97
SVM	96.46	0- 0.97 1- 0.95
Gradient Boosting	95.61	0- 0.96 1- 0.95

ACCURACY SUMMARY

MODEL NAME	ACCURACY	F1 SCORE
Random Forest	94.74	0- 0.96 1- 0.93
XG-Boosting	93.86	0- 0.95 1- 0.92
Decision Tree	92.11	0- 0.93 1- 0.90
CNN	93.72	

CONCLUSION

- **Dual Approaches:** This study explored two methods for breast cancer prediction: CNN-based image analysis and machine learning models on structured CSV data.
- **Superior Accuracy with Structured Data:** Machine learning models on CSV data, containing cell nucleus features, outperformed the CNN model, demonstrating the advantage of structured data for prediction accuracy.
- **Top-Performing Model:** Logistic Regression, applied to the CSV data, yielded the highest accuracy, surpassing more complex models like Random Forest, XGBoost, and SVC.
- **Structured Data Advantage:** The well-organized, feature-rich CSV format enabled simpler models to identify patterns effectively, leading to better performance.
- **CNN Model Insights:** Optimizing CNNs through enhanced image preprocessing and feature extraction methods could improve their effectiveness for breast cancer prediction on image data
- **Future Directions:** For well-organized data in structured formats simpler models (like Logistic Regression) may be preferable over complex models.


Thank
you