Breast Cancer Detection and Prevention using Machine Learning

By Akash Gitty

Contents

Introduction

- Breast cancer is one of the leading causes of death among women.
- Early detection and treatment significantly improves survival rates.
- Machine learning (ML) can aid in early diagnosis.

Objective

Problem Statement

- Develop a machine learning classification model that can accurately predict whether a breast tumor is benign or malignant using patient data.
- Dataset: Wisconsin Breast Cancer Dataset
- Goal: High accuracy, precision, and recall with a robust and interpretable model.

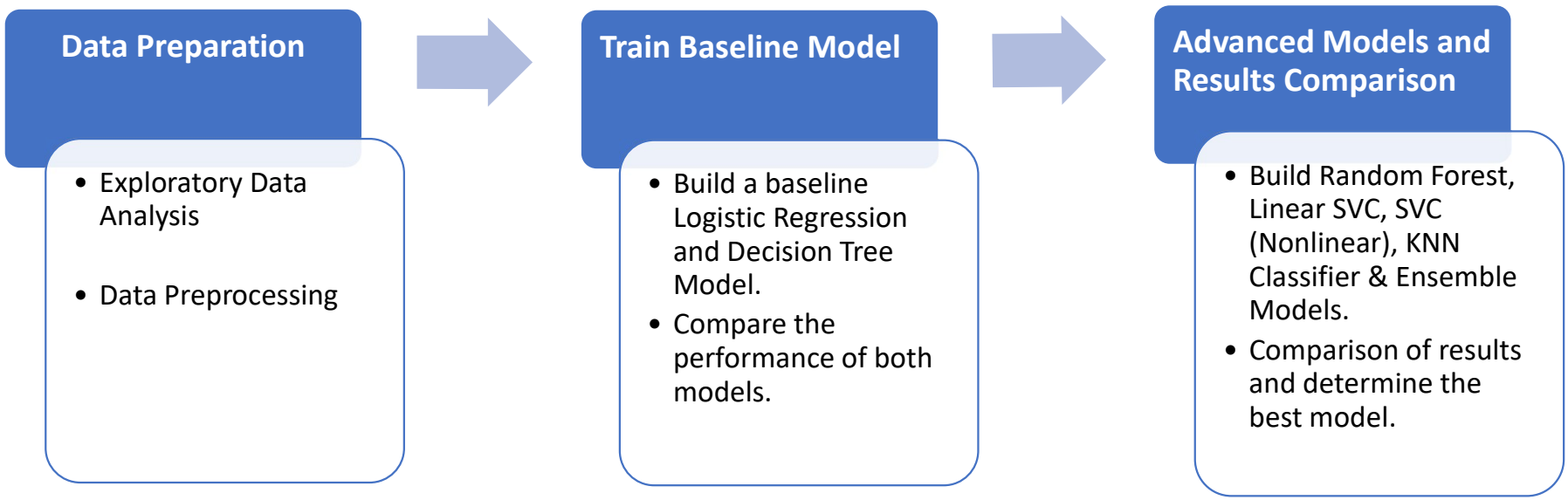
Stakeholders

- Healthcare professionals (doctors, radiologists, pathologists)
- Hospital administrators and IT departments
- Patients and their families
- Health tech startups and researchers

Business Use Case

- Integrate ML models into diagnostic tools for faster and more accurate screenings.
- Reduce diagnostic workload for physicians
- Support rural healthcare centers lacking specialist access
- Enable predictive healthcare systems and preventive treatment planning

Proposed Methodology



Exploratory Data Analysis

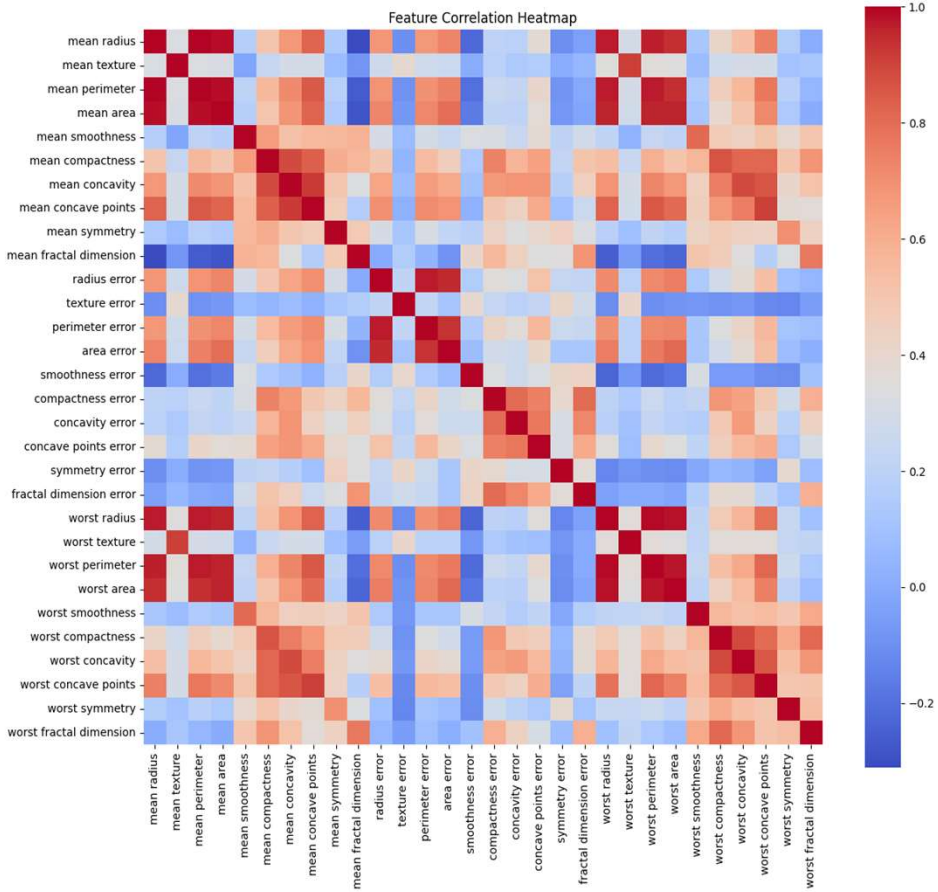
Sample Dataset

	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension	...	worst texture	worst perimeter	worst area	worst smoothness	worst compactness	worst concavity	worst concave points	worst symmetry	worst fractal dimension	diagnosis
0	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	0.2419	0.07871	...	17.33	184.60	2019.0	0.1622	0.6656	0.7119	0.2654	0.4601	0.11890	0
1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	...	23.41	158.80	1956.0	0.1238	0.1866	0.2416	0.1860	0.2750	0.08902	0
2	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	0.2069	0.05999	...	25.53	152.50	1709.0	0.1444	0.4245	0.4504	0.2430	0.3613	0.08758	0
3	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	0.2597	0.09744	...	26.50	98.87	567.7	0.2098	0.8663	0.6869	0.2575	0.6638	0.17300	0
4	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	0.1809	0.05883	...	16.67	152.20	1575.0	0.1374	0.2050	0.4000	0.1625	0.2364	0.07678	0

5 rows × 31 columns

- Samples: 569
- Target: Diagnosis (0 = Malignant, 1 = Benign)
- Features: 30 numerical features
- No Missing Values Found in the Dataset.
- All are numerical Features.

Correlation Heatmap



Key Inferences for Modelling:

- Multicollinearity: The strong positive correlations between certain groups of features (like radius, perimeter, area; and compactness, concavity, concave points).
- Importance of Size and Shape: Features related to the size (radius, perimeter, area) and shape/contour (compactness, concavity, concave points) of the cell nuclei appear to be highly interconnected.
- Potential for Feature Reduction: Due to the high correlations.
- Independent Information: Features like smoothness and symmetry (especially their mean and error versions) might provide more unique information due to weaker correlations with other features.

Data Preprocessing

- **Step 1:Feature Scaling using Min Max Scaling.**

- **Step 2: Feature Selection into 12 important features**

  -> Method used: KBest Feature Selection

Top Selected Features by SelectKBest:

	Feature	Score
11	worst concave points	964.385393
7	worst perimeter	897.944219
5	mean concave points	861.676020
6	worst radius	860.781707
1	mean perimeter	697.235272
8	worst area	661.600206
0	mean radius	646.981021
2	mean area	573.060747
4	mean concavity	533.793126
10	worst concavity	436.691939
3	mean compactness	313.233079
9	worst compactness	304.341063

- **Step 3: Feature Filtering based on Correlation matrix**

```
 Features Removed Due to Correlation > 0.99:
['mean perimeter', 'worst perimeter']

Final Selected Features After Correlation Filtering:
['mean radius', 'mean area', 'mean compactness', 'mean concavity', 'mean concave points', 'worst radius', 'worst area', 'worst compactness', 'worst concavity', 'worst concave points']
```

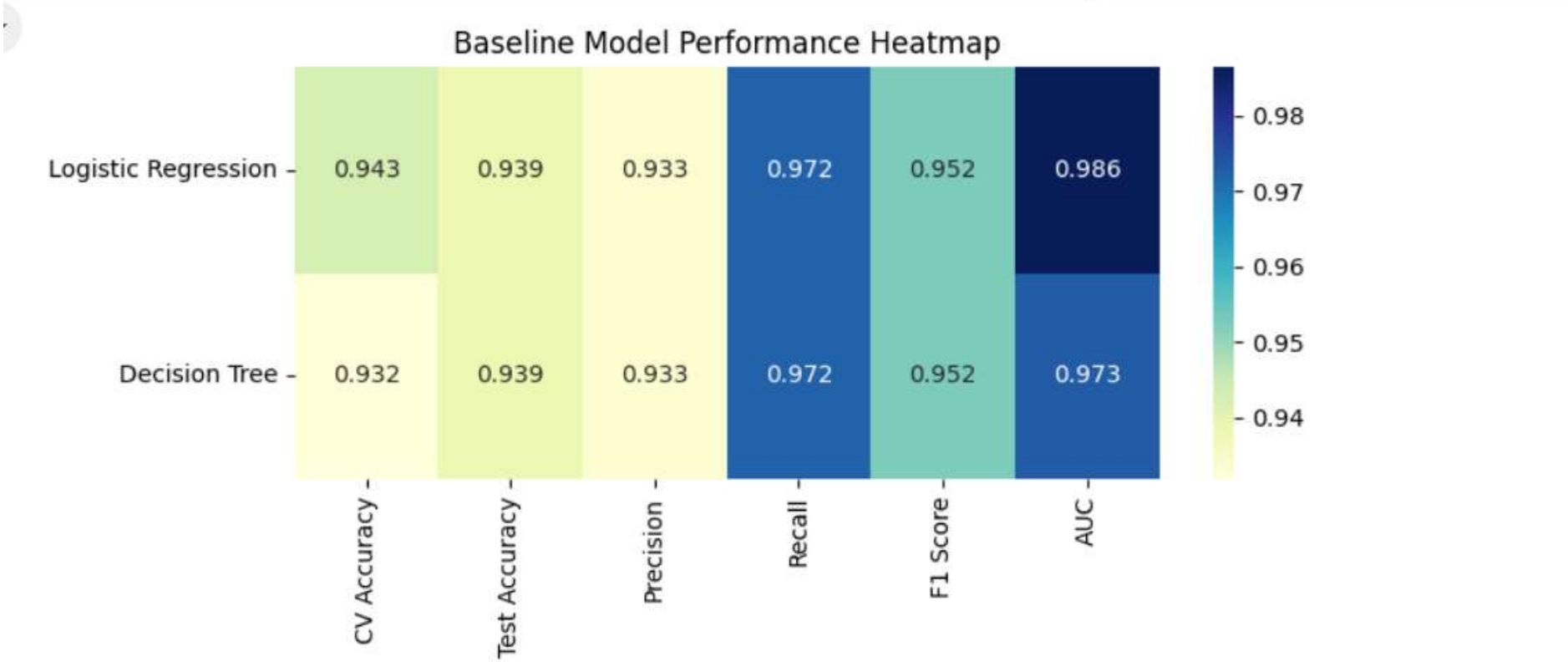
- **Step 4: Data Splitting:**

  -> Stratified Sampling in the ratio 80:20.

  -> Random state=42 to ensure reproducibility of data.

Baseline Model Training

- **Models used:** Logistic Regression, Decision Tree.
- **Metrics used:** CV & Test Accuracy,Precision,Recall,F1 score, AUC.



Baseline Model Results and Inferences

- Overall, **both Logistic Regression and Decision Tree models demonstrated high performance.**

- **Key Observations:**

-> Logistic Regression slightly outperformed Decision Tree in CV Accuracy and AUC.

-> Both models showed consistency across metrics (high accuracy, precision, recall, and F1 scores).

-> Cross-Validation results indicate that both models are likely to generalize well.

Advanced Model Training and Evaluation

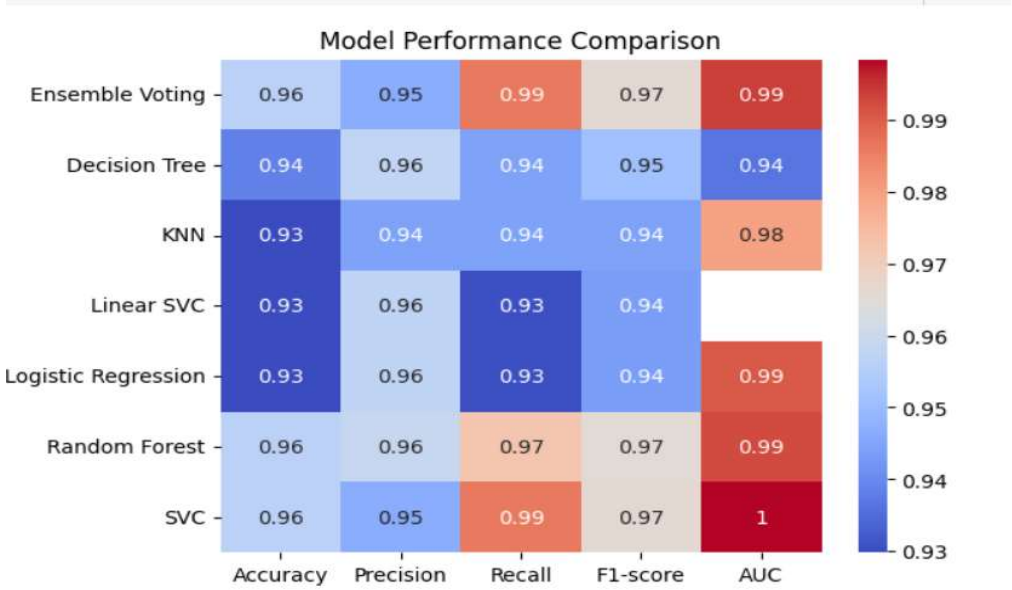
- **Purpose:** To investigate if alternative machine learning models could enhance the prediction of breast tumors.

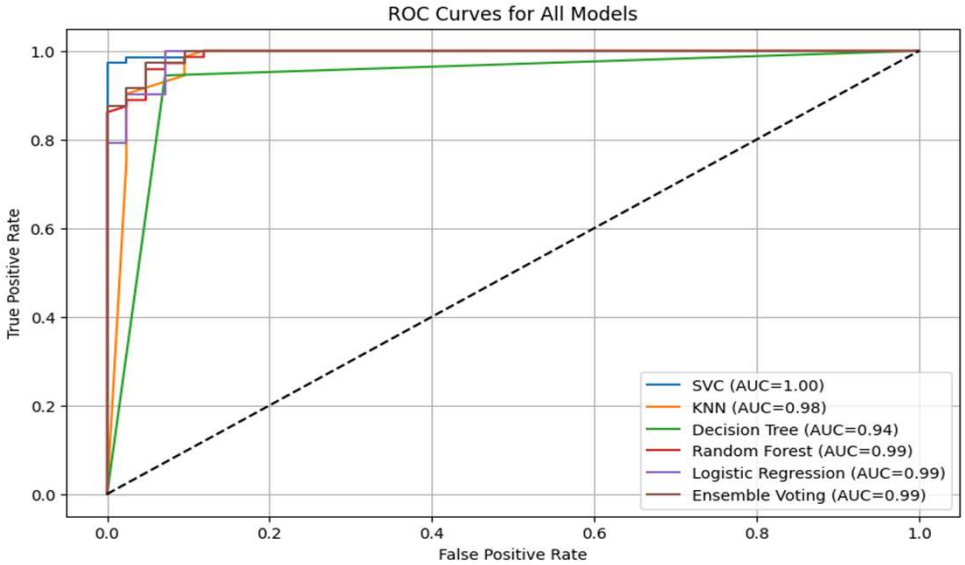
- **Models Trained:**
  -> Linear SVC
  -> SVC (Polynomial)
  -> Random Forest Classifier
  -> KNN Classifier
  -> Ensemble (SVC, Random Forest, Logistic Regression).

- **Hyperparameter tuning:** 5 Fold Cross Validation.

- **Metrics Evaluated:** Accuracy, Precision, Recall, F1 score, AUC.

Results


Model Performance Comparison
Ensemble Voting
Decision Tree
KNN
Linear SVC
Logistic Regression
Random Forest
SVC
0.96 0.95 0.99 0.97 0.99
0.94 0.96 0.94 0.95 0.94
0.93 0.94 0.94 0.94 0.98
0.93 0.96 0.93 0.94
0.93 0.96 0.93 0.94 0.99
0.96 0.96 0.97 0.97 0.99
0.96 0.95 0.99 0.97 1
Accuracy
Precision
Recall
F1-score
AUC
0.99
0.98
0.97
0.96
0.95
0.94
0.93


ROC Curves for All Models
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
SVC (AUC=1.00)
KNN (AUC=0.98)
Decision Tree (AUC=0.94)
Random Forest (AUC=0.99)
Logistic Regression (AUC=0.99)
Ensemble Voting (AUC=0.99)

Model Comparison

Model	Accuracy	Precision	Recall	AUC	Interpretability	Robustness
KNN	0.93	0.94	0.94	0.98	Black-Box	✅
Linear SVC	0.93	0.96	0.93	0.94	Limited Interpretability	✅
SVC	**0.96**	**0.95**	**0.99**	**1.00**	**Black-box**	✅
Ensemble Voting	**0.96**	**0.95**	**0.99**	**0.99**	**Difficult to Interpret**	✅
Random Forest	**0.96**	**0.96**	**0.97**	**0.99**	**Interpretable**	✅
Logistic Regression	0.93	0.96	0.93	0.99	Very Interpretable	✅
Decision Tree	0.94	0.96	0.94	0.94	Highly Interpretable	❌ (can overfit)

The Random Forest model was chosen as the best model due to its high accuracy, precision, recall, and AUC, combined with its good interpretability and generalization ability.

Literature Review

Papers Compared:

- **Paper 1:**Khalid et al., 2024 – “Breast Cancer Detection and Prevention Using Machine Learning, Diagnostics” (MDPI).

- **Paper 2:** Almarri et al., 2024 – “The BCPM method: decoding breast cancer with machine learning”.

- **Paper 3:** Naji et al., 2021 – “Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis”, Procedia CS.

Results Comparison

ML Techniques Used	Reference	Accuracy of Existing Model	AUC score of Existing Model	Proposed Model Accuracy	Proposed Model AUC score
Random Forest	Paper 1	0.96	-	0.96	**0.99**
	Paper 2	0.92	-		
	Paper 3	0.96	0.96		
SVC	Paper 1	0.88	-	0.96	**1**
	Paper 2	0.91	-		
	Paper 3	0.96	0.96		
Logistic Regression	Paper 1	0.93	-	0.93	**0.99**
	Paper 2	0.9	-		
	Paper 3	0.95	0.94		
Decision Tree	Paper 1	0.94	-	0.94	0.94
	Paper 2	0.9	-		
	Paper 3	0.95	0.94		
KNN	Paper 1	0.92	-	0.93	**0.98**
	Paper 2	0.91	-		
	Paper 3	0.93	0.95		

Future Scope

- **Model Enhancement**:

  -> Integrate deep learning models (e.g., CNNs) for image-based diagnostics.

  -> Fine-tune models with larger and more diverse datasets.
- **Real-time Detection and Clinical Integration**: Develop mobile/web applications and integrate into clinical workflows for instant predictions.
- **Data Expansion**: Combine clinical, genetic, and imaging data for better accuracy.
- **Explainability:** Implement explainable AI to increase clinician trust.
- **Continuous Learning:** Enable models to learn from new patient data.

Conclusion

- Machine learning can be a valuable tool for the early and accurate diagnosis of breast cancer.
- The Random Forest model demonstrated strong performance in this study, achieving high accuracy, precision, and recall, while also offering interpretability and robustness.
- These findings support the potential integration of machine learning into clinical practice to improve patient outcomes.

References

- **Code & Dataset**
  - Code Link: https://github.com/AkashGitty97/Cloudxlab-Project-by-Akash
  - Dataset: Dua, D. & Graff, C. (2019). UCI Machine Learning Repository.

- **Books**
  - Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning.
  - Bishop, C. M. (2006). Pattern Recognition and Machine Learning, Springer.

- **Framework & Documentation**
  - Scikit-learn Documentation.

- **Research Papers**
  - Naji et al., 2021 – Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis, Procedia CS, Vol. 191, pp. 487–492.
  - Almarri et al., 2024 – The BCPM method: decoding breast cancer with machine learning, BMC Med Imaging.
  - Khalid et al., 2024 – Breast Cancer Detection and Prevention Using Machine Learning, Diagnostics (MDPI).