BREAST CANCER WISCONSIN DIAGNOSTIC DATASET ANALYSIS

A Comprehensive Machine Learning Approach for Early Tumor Detection

---

Course: Math for Data Science (AID311)
Advisor: Dr. Ahmed Anter
Project Phase: I & II
Date: December 22, 2025
Name: AbdulRahman Essam 320230120

ABSTRACT

Breast cancer remains one of the most prevalent and life-threatening diseases affecting women globally. Early and accurate diagnosis is critical for improving survival rates and treatment outcomes. This project presents a comprehensive analysis of the Breast Cancer Wisconsin Diagnostic (WBCD) dataset using advanced statistical techniques and machine learning algorithms. We implemented and evaluated eight different classification models, including Naive Bayes, Bayesian Belief Network, Decision Tree, Linear Discriminant Analysis, K-Nearest Neighbors, Neural Networks (Feed Forward and LSTM), and both Linear and Logistic Regression. Our preprocessing pipeline included extensive statistical analysis (covariance, correlation, t-tests, ANOVA, chi-square), feature reduction techniques (PCA, LDA, SVD), and rigorous model evaluation using cross-validation. The Linear Discriminant Analysis classifier achieved the highest accuracy of 96.49% on the test set, demonstrating excellent potential for clinical deployment. This study validates the effectiveness of classical machine learning methods for breast cancer classification while providing insights into feature importance and model interpretability.

Keywords: Breast Cancer Classification, Machine Learning, Feature Reduction, Statistical Analysis, Clinical Decision Support

---

TABLE OF CONTENTS

---

1. INTRODUCTION

1.1 Background

Breast cancer is characterized by the uncontrolled growth of abnormal cells in breast tissue and represents a significant global health challenge. According to the World Health Organization, breast cancer accounts for approximately 25% of all cancer cases in women worldwide. The key to improving survival rates lies in early detection and accurate diagnosis, which enables timely intervention and appropriate treatment planning.

Traditional diagnostic methods, including mammography, clinical breast examination, and biopsy analysis, while effective, are subject to human interpretation variability and can be time-consuming. The integration of machine learning algorithms with medical diagnostics offers a promising avenue for enhancing diagnostic accuracy, reducing interpretation time, and supporting clinical decision-making.

1.2 Motivation

The motivation for this project stems from several key factors:

1. **Clinical Impact:** Improving diagnostic accuracy directly translates to better patient outcomes through early intervention
2. **Computational Efficiency:** Machine learning models can process large volumes of data rapidly, enabling faster diagnosis
3. **Consistency:** Automated systems reduce variability in interpretation across different practitioners
4. **Accessibility:** ML-based diagnostic tools can potentially extend expertise to underserved areas

1.3 Objectives

The primary objectives of this project are:

1. Perform comprehensive statistical analysis of the WBCD dataset to understand data characteristics and feature relationships
2. Implement multiple feature reduction techniques (PCA, LDA, SVD) to identify optimal feature subsets
3. Develop and evaluate eight different machine learning models for binary classification
4. Compare model performance using standardized metrics (accuracy, precision, recall, F1-score, AUC)
5. Analyze overfitting/underfitting patterns to ensure model generalizability
6. Provide clinical interpretability through feature importance analysis

---

2. PROBLEM STATEMENT

2.1 Clinical Problem

The primary challenge is to develop an automated system that can accurately classify breast tumors as malignant (cancerous) or benign (non-cancerous) based on cell nuclei characteristics extracted from digitized images of fine needle aspirate (FNA) samples.

2.2 Technical Problem

Given 30 numerical features representing statistical properties of cell nuclei and a binary target variable (0=Malignant, 1=Benign), the task is to:

- Identify the most discriminative features through statistical analysis
- Reduce dimensionality while preserving classification performance
- Train multiple classification models and compare their effectiveness
- Evaluate models using rigorous cross-validation techniques
- Ensure models generalize well to unseen data (avoid overfitting)

2.3 Success Criteria
A successful solution should achieve:

- Classification accuracy > 90%
- High precision and recall for both classes (balanced performance)
- AUC-ROC > 0.95 (excellent discrimination capability)
- Low overfitting (< 5% gap between training and test accuracy)
- Interpretable feature importance for clinical validation

---

3. DATASET DESCRIPTION

3.1 Dataset Overview

Name: Breast Cancer Wisconsin Diagnostic (WBCD) Dataset
Source: UCI Machine Learning Repository
Original Authors: W.H. Wolberg, W.N. Street, and O.L. Mangasarian (1995)
Instances: 569 patients
Features: 30 numerical features
Target Variable: Binary (0=Malignant, 1=Benign)

3.2 Class Distribution

- **Malignant Cases:** 212 (37.3%)
- **Benign Cases:** 357 (62.7%)

The dataset exhibits mild class imbalance (ratio: 1:1.68), which is within acceptable limits and does not require resampling techniques like SMOTE or undersampling.

3.3 Feature Categories

The 30 features are organized into three groups, each containing 10 measurements:

Mean Features (10): Average values computed from cell nuclei

- radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, fractal dimension

Standard Error Features (10): Standard error of measurements

- radius error, texture error, perimeter error, area error, etc.

Worst Features (10): Mean of the three largest values

- worst radius, worst texture, worst perimeter, worst area, etc.

3.4 Data Quality

- **Missing Values:** None (100% complete data)
- **Data Type:** All features are float64
- **Outliers:** Present in several features (area, perimeter, concavity error), but retained as they represent clinically meaningful extreme cases
- **Feature Scaling Required:** Yes, due to varying magnitudes (e.g., area: 143-2501 vs. smoothness: 0.05-0.16)

---

4. METHODOLOGY

4.1 Overall Framework

Our methodology follows a systematic pipeline:

```
Data Loading → Preprocessing → Statistical Analysis → Feature Reduction →
Model Training → Evaluation → Comparison → Interpretation
```

4.2 Software and Libraries

Programming Language: Python 3.12

Environment: Google Colab / Jupyter Notebook

Key Libraries:

- **Data Manipulation:** pandas, numpy
- **Visualization:** matplotlib, seaborn
- **Statistical Analysis:** scipy.stats
- **Machine Learning:** scikit-learn (v1.6.1)
- **Bayesian Networks:** pgmpy (v1.0.0)
- **Deep Learning:** TensorFlow/Keras (v2.9.0)

4.3 Experimental Setup

- **Train-Test Split:** 80% training (455 samples), 20% testing (114 samples)
- **Stratification:** Maintained class distribution in both sets
- **Cross-Validation:** 5-fold stratified cross-validation
- **Random Seed:** 42 (for reproducibility)
- **Scaling Method:** StandardScaler (zero mean, unit variance)

---

5. PREPROCESSING AND DATA ANALYSIS

5.1 Data Visualization

Initial exploratory analysis included:

- **Histograms:** Distribution of all 30 features revealed right-skewed patterns for most features
- **Boxplots:** Identified outliers across multiple features, particularly in area and perimeter measurements
- **Correlation Heatmap:** Showed strong correlations between radius-perimeter-area features (r > 0.9)

5.2 Missing Value Analysis

Result: Zero missing values detected across all 569 samples and 30 features.

Implication: No imputation required, ensuring data integrity and reliability for subsequent analysis.

5.3 Statistical Analysis

5.3.1 Descriptive Statistics

Mean Values:

- Features exhibit wide ranges: mean radius (14.13), mean area (654.89)
- Standard deviations indicate considerable variability: area std = 351.91

Skewness Analysis:

- **Highly Skewed Features (|skew| > 1):** mean area, mean compactness, mean concavity, mean concave points, mean fractal dimension, radius error, texture error, perimeter error, area error, smoothness error, concavity error, concave points error, symmetry error, fractal dimension error, worst area, worst compactness, worst concavity, worst symmetry, worst fractal dimension
- **Right-skewed features:** 30 out of 30 (100%)
- **Left-skewed features:** 0

Kurtosis Analysis:

- **Heavy-tailed features (kurtosis > 3):** mean area, mean fractal dimension, radius error, texture error, perimeter error, area error, smoothness error, concavity error, concave points error, symmetry error, fractal dimension error, worst area, worst compactness, worst symmetry, worst fractal dimension

5.3.2 Covariance Matrix

Key Findings:

- High covariance between radius-perimeter-area features (cov > 1000)
- Positive covariance dominates, indicating features tend to increase together
- Justifies the use of PCA for dimensionality reduction

5.3.3 Correlation Analysis

High Correlations (r > 0.9):

- mean radius ↔ mean perimeter (r = 0.998)
- mean radius ↔ mean area (r = 0.987)
- worst radius ↔ worst perimeter (r = 0.994)
- worst radius ↔ worst area (r = 0.984)

Interpretation: Strong multicollinearity exists, suggesting redundancy in feature space.

5.4 Hypothesis Testing

5.4.1 T-Test (Independent Samples)

Objective: Test if mean feature values differ significantly between malignant and benign classes.

Top 5 Features Tested:

1. worst concave points: t = 31.05, p = 1.97e-124 ✓
2. worst perimeter: t = 29.97, p = 5.77e-119 ✓
3. mean concave points: t = 29.35, p = 7.10e-116 ✓
4. worst radius: t = 29.34, p = 8.48e-116 ✓
5. mean perimeter: t = 26.41, p = 8.44e-101 ✓

Conclusion: All p-values << 0.05, strongly rejecting null hypothesis. Significant mean differences exist between classes.

5.4.2 ANOVA (One-Way)

Results: F-statistics ranging from 697.24 to 964.39, all with $p < 0.05$.

Interpretation: High F-statistics confirm strong between-group variance, validating excellent class separability.

5.4.3 Chi-Square Test

Binning Strategy: Features discretized into Low/Medium/High categories.

Tested Features:

- mean radius: $\chi^2 = 280.14$, p = 1.47e-61, DOF = 2 ✓ mean
- concave points: $\chi^2 = 309.18$, p = 7.30e-68, DOF = 2 ✓ worst
- perimeter: $\chi^2 = 383.01$, p = 6.77e-84, DOF = 2 ✓

Conclusion: Strong association between binned features and target class, validating categorical relationships.

5.5 Data Scaling

Method: StandardScaler (Z-score normalization)

Formula: $X_{scaled} = (X - \mu) / \sigma$
Rationale:

- PCA and LDA are variance-based algorithms requiring scaled data
- KNN uses distance metrics sensitive to feature magnitudes
- Neural networks converge faster with normalized inputs

---

6. FEATURE REDUCTION AND SELECTION

6.1 Principal Component Analysis (PCA)

Objective: Reduce dimensionality while preserving maximum variance.

6.1.1 Variance Analysis

Components vs. Variance Explained:

- 2 components: 63.36%
- 5 components: 85.14%
- 10 components: 95.27%
- 15 components: 98.68%
- 20 components: 99.58%

Optimal Selection: 10 components (captures 95% variance with 67% dimensionality reduction)

6.1.2 PCA as Classifier

Combined PCA with Logistic Regression:

Results:

Components	Accuracy	Variance Explained
2	94.74%	63.36%
5	95.61%	85.14%
10	97.37%	95.27%
15	96.49%	98.68%
20	98.25%	99.58%

Best Performance: 20 components (98.25% accuracy)

6.2 Linear Discriminant Analysis (LDA)

Objective: Maximize class separability by finding optimal linear discriminants.

6.2.1 LDA for Dimensionality Reduction

Reduction: 30 dimensions → 1 discriminant axis (for binary classification)

Visualization: Histogram of LDA projections showed clear separation between malignant and benign classes with minimal overlap.

6.2.2 LDA as Classifier

Performance:

- Training Accuracy: 96.92%
- Test Accuracy: 95.61%
- Precision: 0.97 (Malignant), 0.95 (Benign)
- Recall: 0.90 (Malignant), 0.99 (Benign)
- F1-Score: 0.94 (Malignant), 0.97 (Benign)
- AUC: 0.992

Interpretation: LDA outperforms PCA for classification tasks because it uses class labels (supervised) rather than just variance (unsupervised).

6.3 Singular Value Decomposition (SVD)

Implementation: TruncatedSVD from scikit-learn

Comparison with PCA:

Method	15 Components Accuracy	Variance Explained
SVD	96.49%	98.68%
PCA	96.49%	98.68%

Note: SVD and PCA yield identical results for centered data, but SVD is computationally more efficient for sparse matrices.

---

7. MODEL IMPLEMENTATIONS

7.1 Naive Bayes Classifier

Algorithm: Gaussian Naive Bayes

Assumption: Features follow Gaussian distribution and are conditionally independent given the class label.

Results:

- Test Accuracy: 93.86%
- Precision: 0.93 (M), 0.95 (B)
- Recall: 0.90 (M), 0.96 (B)
- F1-Score: 0.92 (M), 0.95 (B)
- AUC: 0.988

Confusion Matrix:

```
          Predicted
Actual    M    B
  M      38    4
  B       3   69
```

Interpretation: Strong baseline performance despite independence assumption. High recall for benign class (96%) is clinically desirable.

7.2 Bayesian Belief Network (BBN)

Implementation: pgmpy with Discrete Bayesian Network

Network Structure: Naive Bayes topology (target → all features)

Feature Selection: 6 most important features:

- mean radius, mean texture, mean smoothness, mean compactness, mean concavity, mean concave points

Discretization: KBinsDiscretizer (3 bins: Low, Medium, High)

Estimator: Bayesian Estimator with BDeu prior

Results:

- Test Accuracy: 90.35%
- Precision: 0.86 (M), 0.93 (B)
- Recall: 0.88 (M), 0.92 (B)
- F1-Score: 0.87 (M), 0.92 (B)

Baseline Comparisons:

- Random Classifier: 55.26%
- Majority Class: 63.16%
- BBN: 90.35% (+63% vs. Random, +43% vs. Majority)

Interpretation: BBN with discretized features achieved excellent performance, demonstrating robustness to feature preprocessing.

7.3 Decision Tree

Algorithm: CART with Entropy criterion

Hyperparameters:

- Criterion: Entropy (information gain)
- Max Depth: 5 (prevents overfitting)
- Min Samples Leaf: 5 (ensures statistical significance)

Results:

- Test Accuracy: 92.98%
- Precision: 0.89 (M), 0.96 (B)
- Recall: 0.93 (M), 0.93 (B)
- F1-Score: 0.91 (M), 0.94 (B)
- AUC: 0.961

Overfitting Analysis:

- Training Accuracy: 98.24%
- Test Accuracy: 92.98%
- Gap: 5.26% (slight overfitting detected)

Feature Importance: Decision tree identified concave points and radius as top discriminative features.

7.4 K-Nearest Neighbors (KNN)

Implementation: Tested with 3 distance metrics

Hyperparameters:

- k = 5 neighbors
- Distances: Euclidean, Manhattan, Minkowski

Results:

Distance Metric	Accuracy	Precision	Recall	F1-Score	AUC

Euclidean	91.23%	0.90	0.91	0.91	0.956
Manhattan	92.98%	0.92	0.92	0.92	0.966
Minkowski	91.23%	0.90	0.91	0.91	0.956

Best Performer: Manhattan distance (92.98% accuracy)

Interpretation: Manhattan distance works better for this dataset, likely due to the high-dimensional feature space where L1 norm is more robust.

7.5 Neural Networks

7.5.1 Feed Forward Neural Network

Architecture:

```
Input (30) → Dense(128, ReLU) → Dropout(0.3) →
Dense(64, ReLU) → Dropout(0.3) →
Dense(32, ReLU) → Dropout(0.2) →
Output(1, Sigmoid)
```

Training:

- Optimizer: Adam
- Loss: Binary Crossentropy
- Epochs: 200 (with early stopping)
- Batch Size: 32
- Validation Split: 20%

Results:

- Test Accuracy: 96.49%
- Precision: 0.97
- Recall: 0.96
- F1-Score: 0.97
- AUC: 0.991

Training Curves: Validation loss stabilized after epoch 30, indicating good convergence without overfitting.

7.5.2 Recurrent Neural Network (LSTM)

Architecture:

```
Input (30, 1) → LSTM(64, return_sequences=True) → Dropout(0.3) →
LSTM(32) → Dropout(0.3) →
Dense(16, ReLU) → Dropout(0.2) →
Output(1, Sigmoid)
```

Data Reshaping: Features treated as sequence (30 timesteps, 1 feature per step)

Results:

- Test Accuracy: 92.11%
- Precision: 0.92
- Recall: 0.92
- F1-Score: 0.92

Interpretation: LSTM underperformed compared to Feed Forward network, which is expected for nonsequential data. The tabular nature of the dataset doesn't benefit from recurrent connections.

7.5.3 Multi-Layer Perceptron (scikit-learn)

Architecture: (128, 64, 32) hidden layers

Results:

- Test Accuracy: 93.86%
- Precision: 0.94
- Recall: 0.94
- F1-Score: 0.94

Comparison: Similar performance to Gaussian Naive Bayes, demonstrating that simpler models can match complex architectures on this dataset.

7.6 Linear Regression

Note: Applied to binary classification (treating 0/1 as continuous output)

Results:

- Test Accuracy: 95.61% (with 0.5 threshold)
- MSE: 0.065
- RMSE: 0.255
- MAE: 0.210
- R^2: 0.720

Interpretation: Linear regression surprisingly effective for this binary problem, achieving 95.61% accuracy despite not being designed for classification.

7.7 Logistic Regression

Algorithm: Binary Logistic Regression with L2 regularization

Hyperparameters:

- Solver: lbfgs
- Max Iterations: 1000
- Random State: 42

Results:

- Training Accuracy: 97.80%
- Test Accuracy: 97.37%
- Precision: 0.98 (M), 0.97 (B)

- Recall: 0.95 (M), 0.99 (B)
- F1-Score: 0.97 (M), 0.98 (B)
- AUC: 0.995

Cross-Validation (5-fold):

- Mean Accuracy: 97.02%
- Std Dev: 0.014

Feature Coefficients (Top 5):

1. worst radius: -0.856
2. mean radius: +0.687
3. mean perimeter: -0.611
4. worst area: +0.573
5. mean compactness: +0.271

Overfitting Analysis:

- Train-Test Gap: 0.43% (excellent generalization)

Interpretation: Logistic regression achieved near-perfect performance, making it the best single model for this task due to its simplicity, interpretability, and accuracy.

---

8. RESULTS AND EVALUATION

8.1 Model Performance Summary

Comprehensive Comparison Table:

Model	Accuracy	Precision	Recall	F1-Score	AUC	Train-Test Gap
Logistic Regression	**97.37%**	**0.976**	**0.974**	**0.975**	**0.995**	**0.43%**
Feed Forward NN	96.49%	0.966	0.965	0.965	0.991	2.10%
LDA Classifier	95.61%	0.960	0.956	0.958	0.992	1.31%
Linear Regression	95.61%	0.957	0.956	0.956	N/A	1.31%
Naive Bayes	93.86%	0.940	0.939	0.939	0.988	0.21%
MLP Classifier	93.86%	0.938	0.939	0.938	N/A	2.71%
Decision Tree	92.98%	0.925	0.930	0.927	0.961	5.26%*
KNN (Manhattan)	92.98%	0.920	0.920	0.920	0.966	1.75%
LSTM	92.11%	0.923	0.921	0.922	N/A	N/A
KNN (Euclidean)	91.23%	0.900	0.910	0.905	0.956	3.50%
BBN	90.35%	0.904	0.904	0.904	N/A	N/A

*Note: Decision Tree shows slight overfitting despite regularization

8.2 Cross-Validation Results

5-Fold Stratified Cross-Validation:

Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean ± Std
LDA	95.61%	96.49%	94.74%	96.49%	96.46%	95.96% ± 0.70%
Naive Bayes	92.11%	92.11%	94.74%	94.74%	95.58%	93.85% ± 1.46%
Decision Tree	89.47%	92.98%	96.49%	94.74%	94.69%	93.67% ± 2.38%
KNN (Euclidean)	88.60%	93.86%	93.86%	94.74%	92.92%	92.79% ± 2.18%

Interpretation: Low standard deviations (< 2.5%) indicate stable performance across folds, confirming model robustness.

8.3 Overfitting/Underfitting Analysis

Classification:

Model	Status	Evidence
Logistic Regression	✓ Good Fit	Gap = 0.43%
Feed Forward NN	✓ Good Fit	Gap = 2.10%
LDA	✓ Good Fit	Gap = 1.31%
Naive Bayes	✓ Good Fit	Gap = 0.21%
KNN (Manhattan)	✓ Good Fit	Gap = 1.75%
MLP Classifier	✓ Good Fit	Gap = 2.71%
KNN (Euclidean)	✓ Good Fit	Gap = 3.50%
Decision Tree	⚠ Slight Overfitting	Gap = 5.26%

Criteria:

- Good Fit: Train-Test Gap < 5%
- Overfitting: Gap > 5%
- Underfitting: Both accuracies < 85%

8.4 ROC Curve Analysis

AUC Rankings:

1. Logistic Regression: 0.995 (Excellent)
2. LDA: 0.992 (Excellent)
3. Feed Forward NN: 0.991 (Excellent)
4. Naive Bayes: 0.988 (Excellent)
5. KNN (Manhattan): 0.966 (Very Good)
6. Decision Tree: 0.961 (Very Good)
7. KNN (Euclidean): 0.956 (Very Good)

Interpretation: All models achieve AUC > 0.95, indicating excellent discrimination between malignant and benign cases.

8.5 Confusion Matrix Analysis

Best Model (Logistic Regression):

```
          Predicted
Actual    Malignant  Benign
Malignant     40       2
Benign         1      71
```

Key Metrics:

- True Positives (Benign correctly identified): 71/72 = 98.6%
- True Negatives (Malignant correctly identified): 40/42 = 95.2%
- False Positives (Benign misclassified as Malignant): 1 (1.4%)
- False Negatives (Malignant misclassified as Benign): 2 (4.8%)

Clinical Significance:

- High recall for benign (98.6%): Minimizes unnecessary biopsies
- High recall for malignant (95.2%): Ensures cancer detection
- Low false negative rate (4.8%): Critical for patient safety

---

9. COMPARISON WITH RELATED WORK

9.1 Literature Review

Comparison Table:

Study	Year	Method	Accuracy	Features	Dataset
Our Work	**2024**	**Logistic Regression**	**97.37%**	**30 (original)**	**WBCD**
Our Work	2024	Feed Forward NN	96.49%	30 (original)	WBCD
Our Work	2024	LDA	95.61%	30 (original)	WBCD
Akay [2]	2009	SVM + Feature Selection	99.51%	Selected	WBCD
Zheng et al. [4]	2014	K-means + SVM	97.38%	Extracted	WBCD
Sarkar & Leong [3]	2000	K-NN	96.70%	30 (original)	WBCD
Salama et al. [5]	2012	Multi-classifier	96.70%	30 (original)	WBCD

9.2 Analysis Key

Observations:

1. **Competitive Performance:** Our Logistic Regression (97.37%) ranks 2nd among published works, outperforming most studies
2. **Feature Engineering:** Akay (2009) achieved highest accuracy (99.51%) using feature selection, suggesting that dimensionality reduction can improve performance
3. **Simple vs. Complex Models:** Our results demonstrate that classical algorithms (Logistic Regression, LDA) match or exceed complex methods (Neural Networks, ensemble methods)

4. **Consistency:** Multiple studies (including ours) achieve 96-97% accuracy using original features, validating dataset quality
5. **Practical Implications:** High accuracy across different methods suggests WBCD is a well-curated dataset suitable for benchmarking

9.3 Our Contributions

1. **Comprehensive Analysis:** Implemented 8 different models with rigorous evaluation
2. **Statistical Rigor:** Extensive hypothesis testing (t-test, ANOVA, chi-square)
3. **Multiple Dimensionality Reduction:** Compared PCA, LDA, and SVD
4. **Overfitting Analysis:** Systematic evaluation of generalization capability
5. **Clinical Interpretability:** Feature importance analysis for medical validation

---

10. DISCUSSION

10.1 Key Findings

10.1.1 Model Selection Insights

Best Overall Model: Logistic Regression (97.37% accuracy)

Rationale:

- Highest accuracy with excellent generalization (0.43% train-test gap)
- Interpretable coefficients enable clinical validation
- Fast training and prediction (< 1 second)
- Minimal hyperparameter tuning required
- Probabilistic outputs support threshold adjustment

Runner-Up Models:

- Feed Forward Neural Network (96.49%): Superior for non-linear relationships
- LDA (95.61%): Excellent for visualization and feature reduction

10.1.2 Feature Importance

Most Discriminative Features (from Logistic Regression):

1. worst radius (coefficient = -0.856)
2. mean radius (coefficient = +0.687)
3. mean perimeter (coefficient = -0.611)
4. worst area (coefficient = +0.573)
5. mean compactness (coefficient = +0.271)

Clinical Interpretation:

- Larger worst radius/perimeter → Higher malignancy probability
- Features related to cell size and shape are most informative
- Validates clinical intuition that irregular, larger cells indicate cancer

10.1.3 Dimensionality Reduction Impact

PCA Performance:

- 10 components (67% reduction): 97.37% accuracy
- 20 components (33% reduction): 98.25% accuracy

Conclusion: Significant dimensionality reduction possible with minimal accuracy loss, reducing computational complexity and overfitting risk.

10.2 Statistical Validation

Hypothesis Testing Results:

- All statistical tests (t-test, ANOVA, chi-square) confirmed significant differences between classes
- p-values < 1e-100 provide overwhelming evidence for class separability
- High F-statistics (> 600) indicate excellent discrimination potential

Cross-Validation Stability:

- Standard deviations < 2.5% across all models
- Confirms robustness to data partitioning
- Validates generalization capability

10.3 Clinical Implications

Deployment Considerations:

1. **High Sensitivity Required:** Logistic Regression achieves 95.2% sensitivity for malignant detection, minimizing missed cancer cases
2. **Specificity Trade-off:** 98.6% specificity for benign cases reduces unnecessary procedures
3. **Decision Threshold:** Default 0.5 threshold can be adjusted based on clinical priorities:
    - Lower threshold (e.g., 0.3): Increase sensitivity (fewer missed cancers)
    - Higher threshold (e.g., 0.7): Increase specificity (fewer false alarms)
4. **Real-time Applicability:** Fast prediction times (< 100ms) enable integration into clinical workflows

10.4 Limitations

1. **Dataset Size:** 569 samples is relatively small for deep learning; larger datasets may improve neural network performance
2. **Single Institution Data:** WBCD from one institution may not generalize to other populations/imaging protocols
3. **Feature Engineering:** All features are pre-computed; raw image analysis might capture additional patterns
4. **Class Imbalance:** 37:63 ratio is mild but could affect minority class performance in extreme cases
5. **Temporal Validation:** No longitudinal validation; models not tested on data collected years later
6. **External Validation:** Not tested on independent external datasets from other institutions

10.5 Future Work

Recommended Extensions:

1. **Ensemble Methods:** Combine top models (Logistic Regression + LDA + NN) for improved accuracy
2. **External Validation:** Test on independent breast cancer datasets (e.g., MIAS, DDSM)
3. **Deep Learning on Images:** Apply CNNs directly to raw FNA images rather than extracted features
4. **Explainable AI:** Implement SHAP or LIME for interpretable predictions at instance level
5. **Multi-class Classification:** Extend to subtype classification (e.g., invasive vs. in-situ carcinoma)
6. **Cost-Sensitive Learning:** Assign higher penalty to false negatives (missed cancers)
7. **Uncertainty Quantification:** Implement Bayesian neural networks for prediction confidence intervals

---

11. CONCLUSION

This comprehensive study successfully developed and evaluated multiple machine learning approaches for breast cancer classification using the WBCD dataset. Through rigorous statistical analysis, feature engineering, and model comparison, we achieved the following:

11.1 Summary of Achievements

1. **High Accuracy:** Logistic Regression achieved 97.37% test accuracy, competitive with state-of-theart methods
2. **Robust Generalization:** Low train-test gaps (< 5%) across most models confirm excellent generalization
3. **Statistical Rigor:** Hypothesis testing validated class separability and feature significance
4. **Dimensionality Reduction:** PCA, LDA, and SVD successfully reduced features while maintaining performance
5. **Comprehensive Evaluation:** 8 models evaluated using standardized metrics, cross-validation, and overfitting analysis
6. **Clinical Interpretability:** Feature importance analysis provides actionable insights for medical validation

11.2 Best Practices Demonstrated

- Stratified train-test splitting to maintain class distribution
- Standardization for distance-based and variance-based algorithms
- Cross-validation for robust performance estimation
- Multiple evaluation metrics (accuracy, precision, recall, F1, AUC)
- Overfitting detection through train-test comparison
- Statistical hypothesis testing for feature validation

11.3 Practical Impact

The developed models, particularly Logistic Regression and Feed Forward Neural Networks, demonstrate strong potential for clinical deployment as decision support tools. The high accuracy (> 95%) and excellent AUC scores (> 0.99) suggest these systems could:

- Assist radiologists in FNA sample analysis
- Reduce diagnostic time and variability
- Provide second opinions for challenging cases
- Enable early cancer detection in resource-limited settings

11.4 Final Recommendation

For Clinical Deployment: Logistic Regression is recommended due to:

- Highest accuracy (97.37%)
- Best generalization (0.43% gap)
- Full interpretability (coefficient analysis)
- Fast inference (< 100ms)
- Minimal computational requirements

For Research Purposes: Feed Forward Neural Network offers:

- Comparable accuracy (96.49%)
- Ability to capture non-linear patterns
- Scalability to larger datasets
- Potential for ensemble methods

11.5 Concluding Remarks

This project validates the effectiveness of classical machine learning methods for medical diagnosis while demonstrating best practices in data analysis, model development, and evaluation. The comprehensive approach—from statistical hypothesis testing to deep learning—provides a template for future medical ML applications. With continued refinement and external validation, these models have the potential to meaningfully impact breast cancer diagnosis and patient outcomes.

---

12. REFERENCES

[1] W.H. Wolberg, W.N. Street, and O.L. Mangasarian, "Breast Cancer Wisconsin (Diagnostic) Data Set",UCI Machine Learning Repository, 1995.

[2] M.F. Akay, "Support vector machines combined with feature selection for breast cancer diagnosis",*Expert Systems with Applications*, vol. 36, pp. 3240-3247, 2009.

[3] M. Sarkar and T.Y. Leong, "Application of K-nearest neighbors algorithm on breast cancer diagnosisproblem", *Proceedings of the AMIA Symposium*, pp. 759-763, 2000.

[4] B. Zheng et al., "Breast cancer diagnosis based on feature extraction using a hybrid of K-means andsupport vector machine algorithms", *Expert Systems with Applications*, vol. 41, pp. 1476-1482, 2014.
[5] G.I. Salama, M. Abdelhalim, and M.A. Zeid, "Breast cancer diagnosis on three different datasets using multi-classifiers", *International Journal of Computer and Information Technology*, vol. 1, no. 1, pp. 3643, 2012.

End of Report