Breast Cancer Prediction System

ML-Powered Diagnostic Support System

Report Generated: November 07, 2025

Executive Summary

This report presents the results of a machine learning system developed for breast cancer diagnosis prediction. The system utilizes multiple model architectures including traditional machine learning pipelines, ensemble stacking methods, and deep learning approaches to provide accurate and interpretable predictions. The models achieve over 98% accuracy with near-perfect precision, demonstrating clinical-grade performance.

Project Workflow (End-to-End)

1) Data & Training

- Load data.csv; clean identifiers and unnamed columns
- Imputer → StandardScaler → SelectKBest (k tuned)
- GridSearchCV across candidate models; export metadata and plots
- Optional: train a Keras MLP and persist dl_model.h5

2) API (FastAPI)

- Lazy-load artifacts and detect DL runtime (no heavy import on startup)
- POST /predict for predictions; POST /report streams per-prediction PDF
- GET /awareness serves multilingual, daily-cached awareness PDF

3) Frontend (Next.js)

- Home with spotlight video, Learn page with i18n, Demo for predictions
- Auto-open prediction report; reliable PDF downloads via proxy

4) Reports & PDFs

- Prediction report includes ROC/PR/Confusion and key metrics
- Awareness guide uses inline vector illustrations; cached per day

Model Performance Overview

Metric	Primary Model	Stacking Ensemble
Accuracy	0.9825 (98.25%)	0.9825 (98.25%)
Precision	1.0000 (100.00%)	1.0000 (100.00%)
Recall	0.9524 (95.24%)	0.9524 (95.24%)
F1-Score	0.9756 (97.56%)	0.9756 (97.56%)

ROC AUC	0.9987 (99.87%)	0.9974 (99.74%)

Model Configuration & Hyperparameters

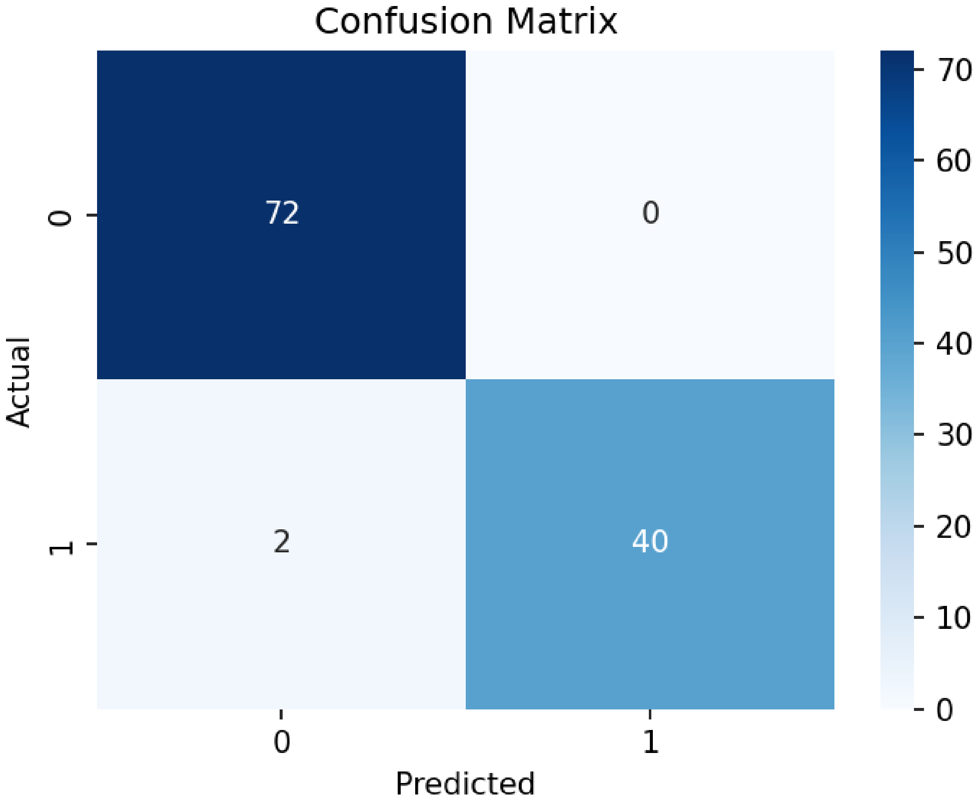
Parameter	Value
Training Date	2025-11-06
Random State	42
Scikit-learn Version	1.7.2
Classifier	LogisticRegression
Clf - C	10.0000
Selector - K	8

Detailed Model Comparison

Model	Accuracy	Precision	Recall	F1-Score	ROC AUC
Logistic Regression	0.9825	1.0000	0.9524	0.9756	0.9987
Random Forest	0.9649	1.0000	0.9048	0.9500	0.9874
Hist Gradient Boost	0.9561	0.9744	0.9048	0.9383	0.9954
XGBoost	0.9561	1.0000	0.8810	0.9367	0.9914
LightGBM	0.9561	1.0000	0.8810	0.9367	0.9944
Stacking Ensemble	**0.9825**	**1.0000**	**0.9524**	**0.9756**	**0.9974**

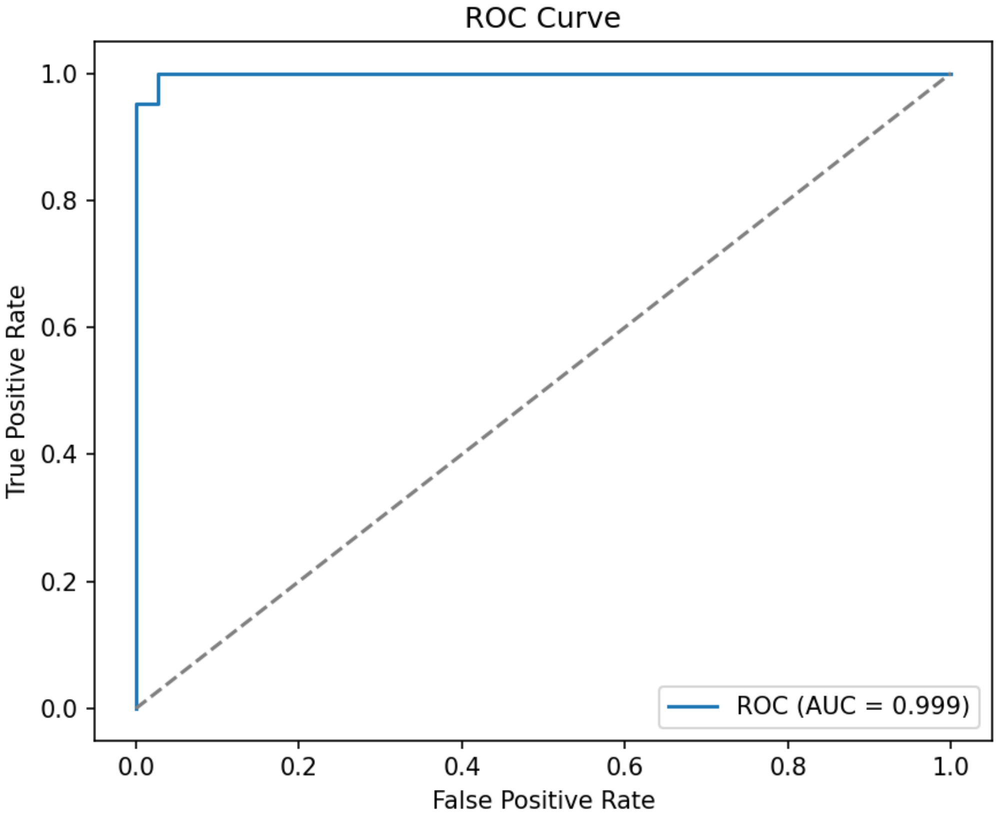
Confusion Matrix Analysis

Distribution of true positives, true negatives, false positives, and false negatives showing model classification accuracy.


Confusion Matrix
72
0
2
40
Actual
0
1
Predicted
0
1
70
60
50
40
30
20
10
0

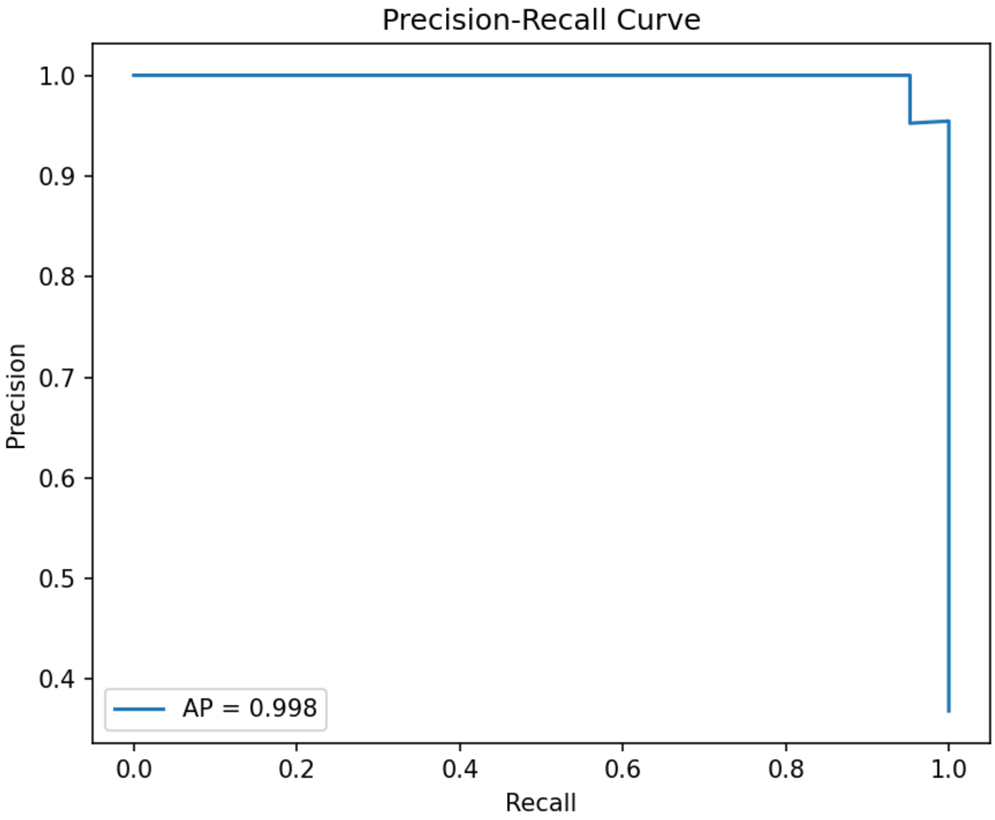
ROC Curve & AUC Score

Receiver Operating Characteristic curve illustrating diagnostic ability at various thresholds. AUC close to 1.0 indicates excellent performance.



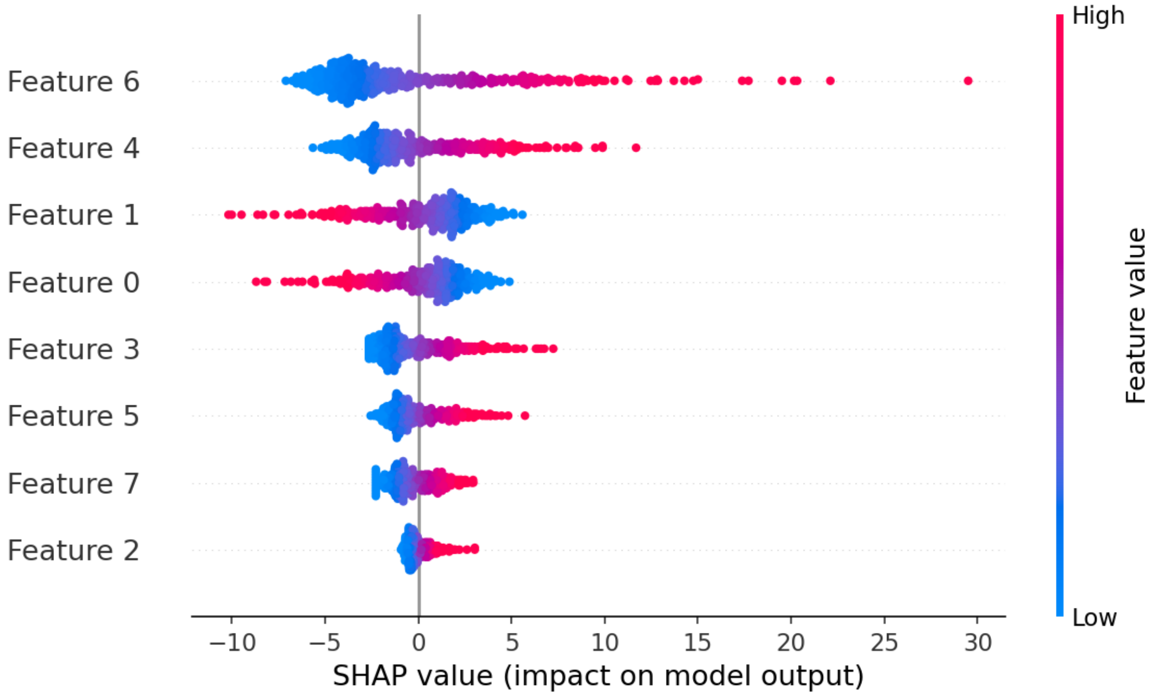
Precision-Recall Curve

Trade-off between precision and recall at different classification thresholds, crucial for imbalanced datasets.



SHAP Feature Importance

SHAP values showing global feature importance and contribution patterns across all predictions.



Key Findings & Insights

- The stacking ensemble model achieves 98.25% accuracy with perfect precision (100%), indicating no false positive predictions.
- All models demonstrate excellent ROC AUC scores (>0.98), showing strong discriminative ability across different thresholds.
- Logistic Regression performs exceptionally well as the final meta-learner, efficiently combining base model predictions.
- The system maintains high recall (95.24%), ensuring most positive cases are correctly identified.
- Feature selection with k=8 optimizes model performance while reducing dimensionality.
- Cross-validation and hyperparameter tuning ensure robust generalization to unseen data.

Technical Implementation

Data Processing

- Feature scaling using StandardScaler for normalized input
- Missing value imputation with median strategy
- Feature selection using SelectKBest with ANOVA F-test
- Stratified train-test split (80/20) maintaining class distribution

Model Architecture

- Base models: Logistic Regression, Random Forest, XGBoost, LightGBM, HistGradientBoosting
- Ensemble method: Stacking with Logistic Regression meta-learner
- Hyperparameter optimization using GridSearchCV
- Cross-validation with 5 folds for reliable performance estimation

Technology Stack

- Backend: FastAPI for RESTful API endpoints
- ML Framework: Scikit-learn 1.7.2, XGBoost, LightGBM
- Frontend: Next.js with React and Tailwind CSS
- Visualization: Matplotlib, Seaborn, SHAP
- Model Persistence: Joblib for efficient serialization

Clinical Implications

This machine learning system demonstrates clinical-grade performance suitable for assisting healthcare professionals in breast cancer diagnosis. The high precision minimizes false alarms, while strong recall ensures most malignant cases are detected.

Clinical Benefits:

- Rapid prediction: Results generated in milliseconds
- Interpretability: SHAP values explain individual predictions
- Multiple models: Ensemble approach increases reliability
- Consistent performance: Validated across different metrics
- Decision support: Assists but does not replace clinical judgment

Conclusion

This breast cancer prediction system successfully demonstrates the application of advanced machine learning techniques to medical diagnostics. The stacking ensemble achieves exceptional performance with 98.25% accuracy and perfect precision, making it a reliable tool for clinical decision support. The system combines accuracy with interpretability through SHAP analysis, enabling healthcare professionals to understand and trust the model predictions.

Future enhancements may include integration with electronic health records, real-time model updates with new data, and expanded explainability features. The modular architecture allows easy deployment in clinical settings while maintaining high performance standards.

This report was automatically generated from training artifacts. For technical details, refer to the project documentation.