CYBER SHADE

Machine Learning project: Breast Cancer Detection

Presented By

Ahmed Rayen Aloui
Ahmed Tounsi
Farah Boubaker
Farah Derbel
Mahdi Saoudi
Mohamed Yassine Janfaoui

Supervised By :
Dr. Wiem Trabelsi

Table of contents

01

Introduction

Problem Statement

Our Goals

Save Lives Through Early Detection

Protect At-Risk Patients Before It's Too Late

Reduce Healthcare Costs and Patient Stress

Business Understanding

Breast cancer is one of the most life-threatening diseases among women worldwide.
Early detection is essential to improve survival and reduce costs.

Traditional diagnosis relies on manual interpretation—slow and error-prone.

This project uses machine learning to support healthcare professionals in detection, diagnosis, and prevention.

01 Confirm the presence of a breast tumor

02 Characterize the tumor to guide clinical decisions

03 Detect early risk factors in healthy patients and recommend ap-propriate preventive actions

Data Understanding

- **DSO**
  - **DSO1: Classify Tumor Type**
    - Dataset — Wisconsin Diagnostic Breast Cancer Dataset
    - Characteristics
      - Moderate class imbalance
      - No missing values
      - Few outliers
      - Concave point features
    - Goal — Classify tumors as benign or malignant
  - **DSO2: Analyze Histopathology Images**
    - Dataset — IDC Detection from Tissue Images (277,524 patches)
    - Characteristics
      - Significant class imbalance
      - All filenames valid
      - Patient-level data split to prevent leakage
      - Visual patterns distinguish IDC from non-IDC tissue
    - Goal — Detect Invasive Ductal Carcinoma (IDC) in breast tissue images
  - **DSO3: Identify Risk Factors**
    - Dataset — Clinical & Biochemical Data Analysis
    - Characteristics
      - Moderate class imbalance
      - No missing values
      - Few outliers
    - Goal — Identify early risk factors to support prevention

Data Preparation

Transforming Raw Data into Model-Ready Features

DSO1: Classify tumor type

- Clean: Encode M→1/B→0, remove empty columns & ID
- Split: 70/30 stratified (398 train, 171 test)
- Scale: Z-score normalization

DSO2: Analyze Histopathology Images

- Preprocess: Resize 50×50, normalize [0,1], augment
- Split: Patient-wise 70/10/20 (prevent leakage)
- Pipeline: TF.data (parallel, batch 64, prefetch)

DSO3: Identify risk factors

- Feature Engineering: 4 composite features (BMI×Glucose, etc.)
- Balance: SMOTETomek (synthetic + denoise)
- Scale: StandardScaler on all features

Modeling

From Algorithms to Clinical Decision Support

DSO1: Classify tumor type

Model 1: Linear Regression :	Hyperparameters: Batch=128, LR=0.001, Epochs=3000, Loss=MSE, Optimizer=SGD **Why? Simple baseline, high interpretability**
Model 2: k-Nearest Neighbors (k=1):	Distance Metrics: Manhattan (L1) & Euclidean (L2) **Why? Non-parametric, no training required, geometric approach**
Model 3: Softmax Regression:	Hyperparameters: Batch=128, LR=0.001, Epochs=3000, Loss=Cross-entropy, Optimizer=SGD **Why? Outputs probability distribution, multi-class ready, convex optimization (guaranteed convergence)**
Model 4: Support Vector Machine (SVM):	Hyperparameters: C=5, Norm=L2, LR=0.001, Batch=128, Epochs=3000, Optimizer=Adam **Why? Optimal decision boundary, robust to outliers, effective in high dimensions**

Model 5: Multilayer Perceptron (MLP)	Architecture: 3 hidden layers of 500 neurons each, ReLU Hyperparameters: Batch=128, LR=0.01, Epochs=3000, Loss=Cross-entropy, Optimizer=SGD **Why? Learns complex non-linear patterns, hierarchical feature extraction**

Model 6: GRU-SVM Hybrid	Architecture: GRU(128 cells) → Dropout(0.5) → SVM layer Hyperparameters: C=5, LR=0.001, Batch=128, Epochs=3000,Optimizer=Adam **Why? GRU learns hierarchical representations, SVM provides optimal classification boundary**

DSO2: Analyze Histopathology Images

Model 1: Custom CNN	Architecture: 3 Conv layers (32→64→128) + Dense(128) + Dropout(0.4) + Sigmoid Parameters: 683,329 trainable Training: 10 epochs, Adam optimizer **Why? Progressive feature extraction (edges→textures→ structures), lightweight, tailored for 50×50 images**

Model 2: ResNet50 Transfer Learning	Base: ResNet50 pre-trained on ImageNet (frozen) Custom Head: Dense(128) + Dropout(0.4) + Sigmoid Parameters: 23.8M total (262K trainable) Training: 5 epochs, Adam optimizer **Why? Leverage pre-trained features from 14M images, less data needed, proven architecture**

DSO3: Identify risk factors in healthy patients

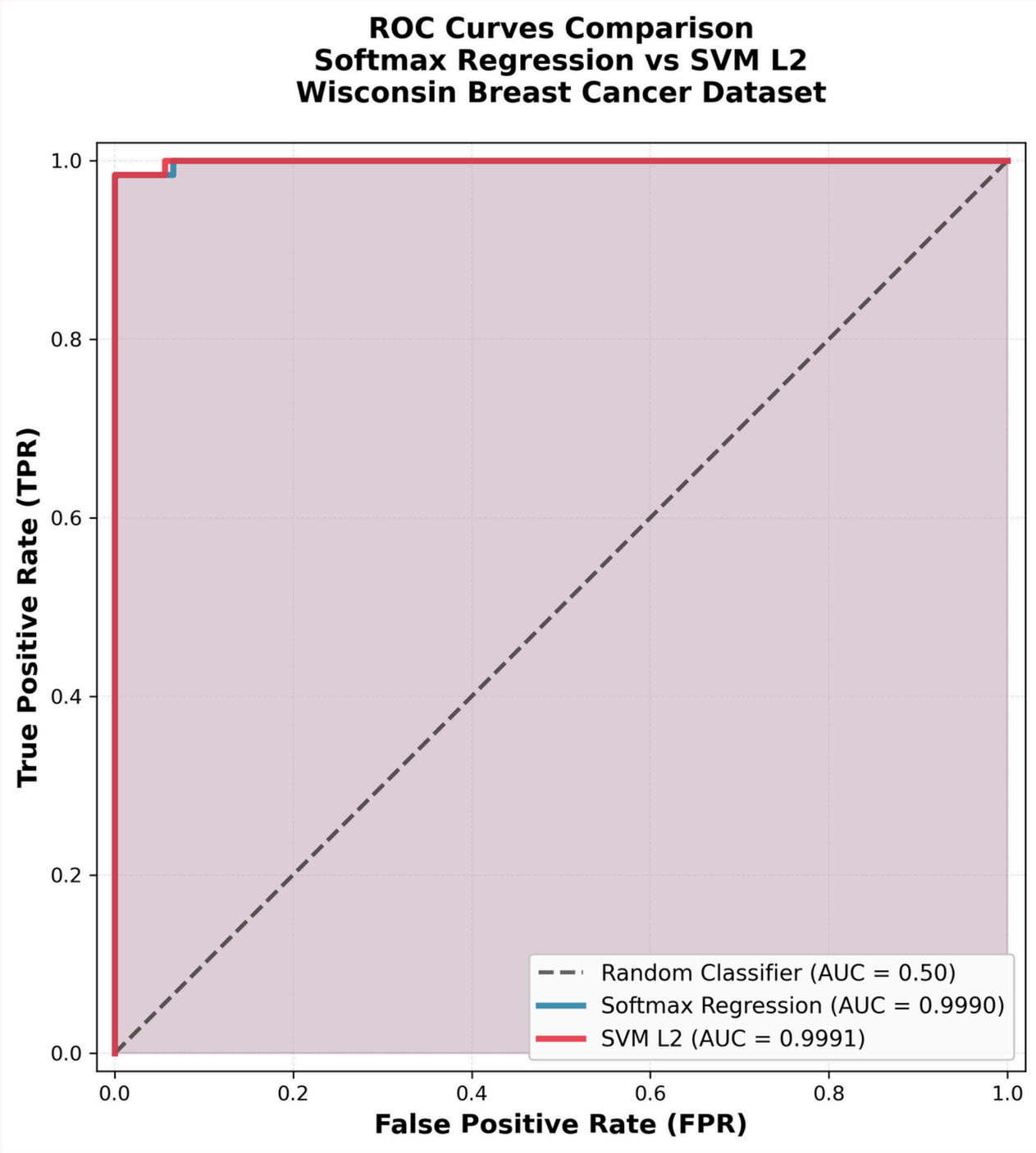
Model 1: Random Forest	Hyperparameters: n_estimators: 100 ,class_weight: balanced, max_depth: unconstrained , random_state: 42 **Why? Handles non-linearity, feature importance for clinical interpretation, robust to noise**
Model 2: Support Vector Machine (SVM)	Hyperparameters: kernel: RBF (Radial Basis Function), class_weight: balanced, probability: enabled **Why? Captures complex relationships, flexible decision boundaries, maps to infinite dimensions**
Model 3: Gradient Boosting	Hyperparameters: n_estimators: 100 ,learning_rate: default max_depth: default **Why? Corrects mistakes iteratively, high accuracy on tabular data, handles complex interactions**
Model 4: XGBoost	Hyperparameters: n_estimators: 100 ,learning_rate: 0.1 , eval_metric: log-loss ,random_state: 42 **Why? State-of-the-art for tabular data, fast training, built-in regularization, industry standard**

Evaluation

DSO1: Classify tumor type

Algorithm	Accuracy	TPR	TNR	Article Comparison
Linear Regression	96.49%	90.62%	100.00%	+0.40%
Nearest Neighbor (L1)	95.91%	93.75%	97.20%	+2.34%
Nearest Neighbor (L2)	94.15%	90.62%	96.26%	-0.58%
Softmax Regression	98.83%	96.87%	100.00%	+1.17%
SVM (L2)	98.25%	95.31%	100.00%	+2.15%
MLP	96.49%	92.19%	99.07%	-2.55%
GRU-SVM	96.49%	90.62%	100.00%	+2.74%

ROC CURVE

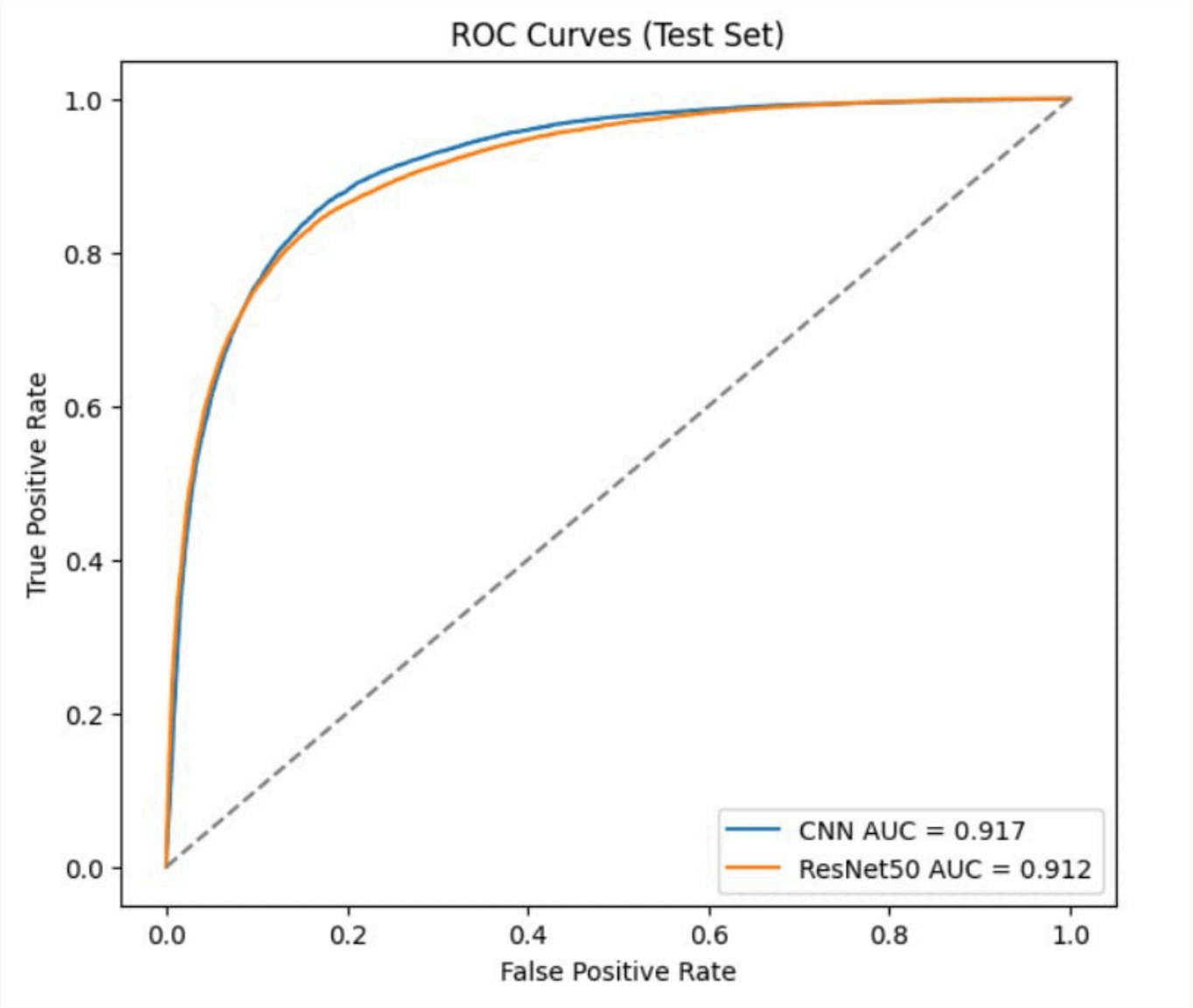


'Softmax Regression was chosen for production'

DSO2: Analyze Histopathology Images

Algorithm	*Accuracy*	Loss	*Training Time*
CNN	85.69%	0.3918	500 seconds per epoch
ResNet50	85.78%	0.3516	700 seconds per epoch (slower due to larger architecture)

ROC CURVE

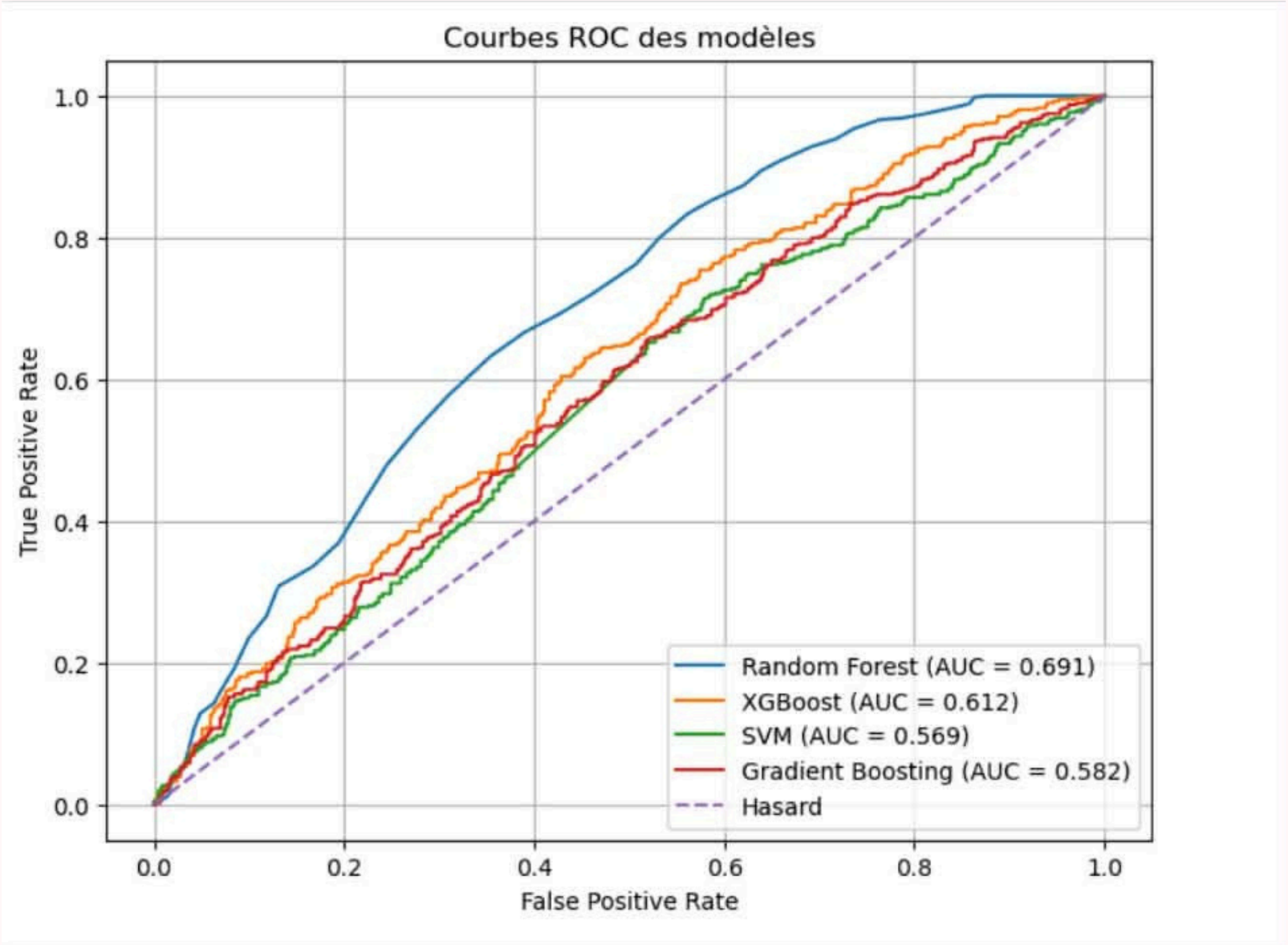


'CNN was chosen for production'

DSO3: Identify risk factors in healthy patients

Algorithm	Accuracy	F1-Score
Random Forest	63,92%	0.636
SVM	56,85%	0.569
XGBoost	56,43%	0.569
Gradient Boosting	55,8%	0.533

ROC CURVE



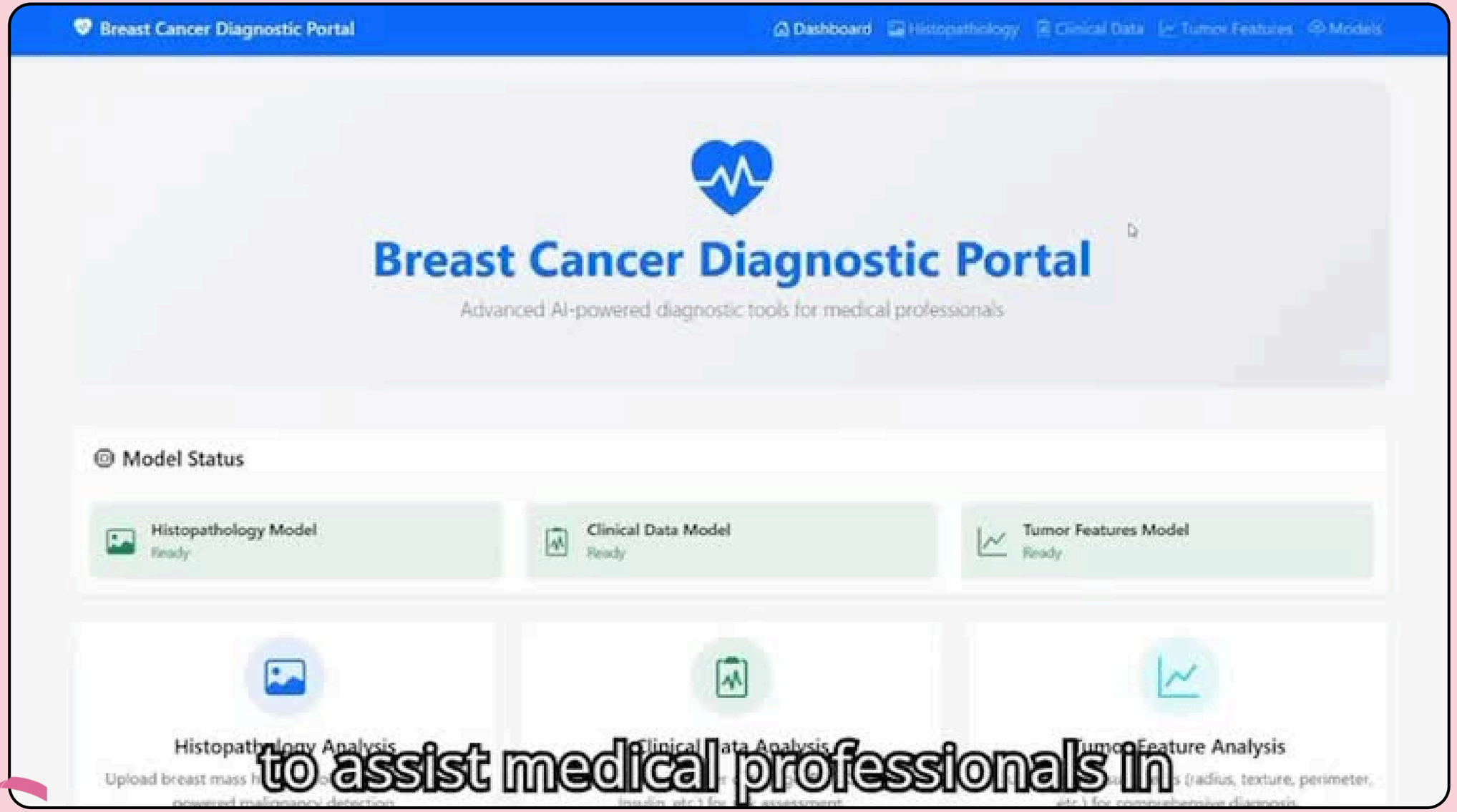
'Random Forest outperformed other models and was chosen for production'

Deployment

a web-based deployment was implemented

- **Flask** framework
- **HTML, CSS, and JavaScript**

Demo



to assist medical professionals in

Why Choose Our Solution?

Faster Decisions, Better Outcomes:

Deliver diagnoses in minutes, not weeks—giving patients the care they need when they need it.

Stay Ahead of the Disease:

Catch cancer risk before it becomes cancer—protecting more patients with early intervention

Reduce Costs, Increase Efficiency:

Cut down on unnecessary procedures and save valuable resources while improving accuracy

Conclusion

Breast cancer doesn't wait—and neither should we.

Our project delivers faster, smarter diagnosis through machine learning that empowers doctors and saves lives. We're transforming early detection from a challenge into an opportunity.

The technology works. The need is real. The time is now. Let's catch cancer earlier and give patients the future they deserve.

Perspectives

Expand to Other Cancers:

Apply our proven approach to lung, prostate, and colon cancer detection for broader impact

Patient Portal

Integrate a patient portal counterpart to facilitate access to diagnosis results and streamline doctor - patient communication

Real-Time Clinical Integration:

Connect directly with hospital systems and electronic health records for seamless, automated screening

THANK YOU