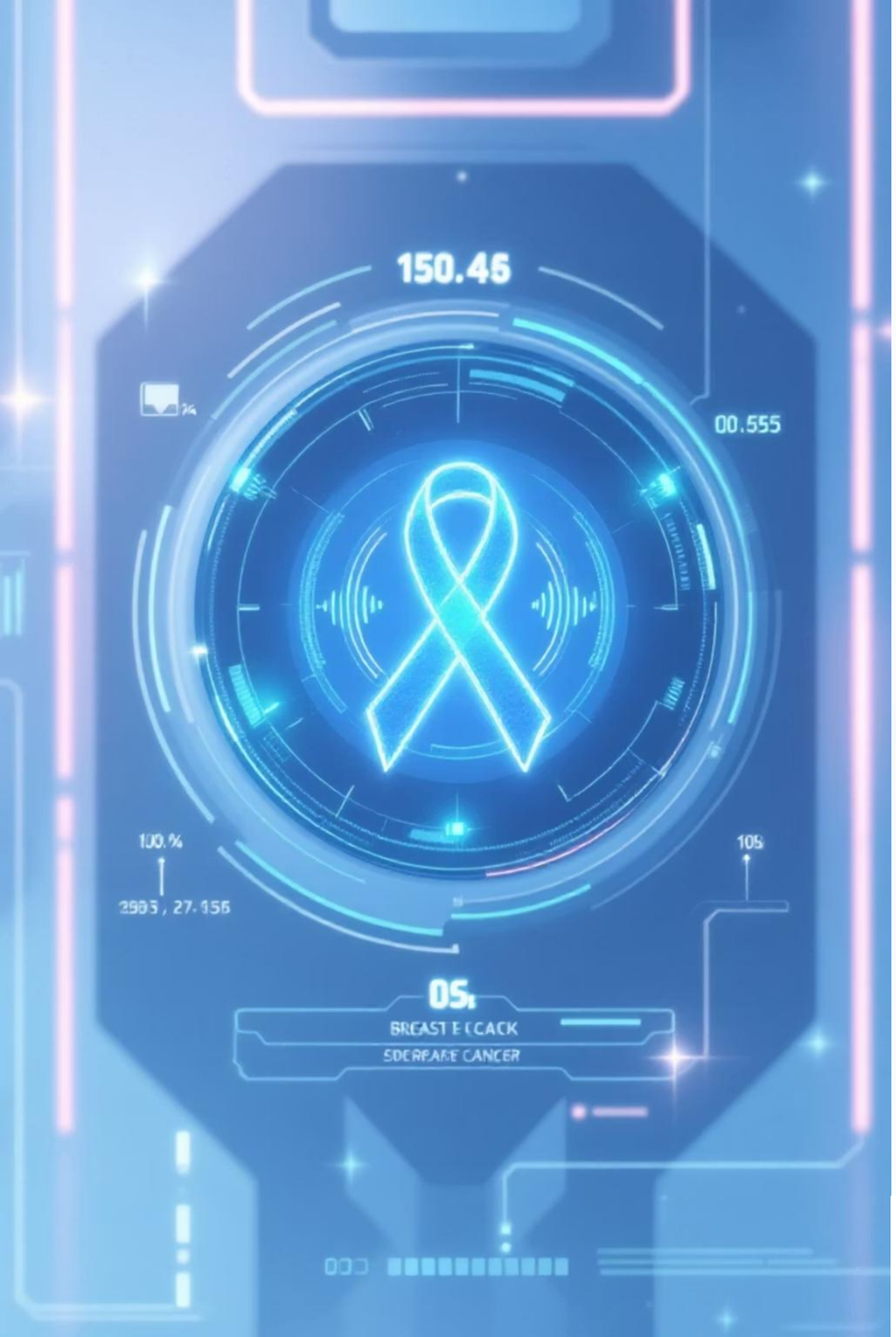


AI-Powered Breast Cancer Detection

An Advanced Machine Learning System for Enhanced Diagnostics

This project presents an advanced machine learning system for automated breast cancer detection using a Random Forest classifier, achieving high diagnostic accuracy. It features data preprocessing with SMOTE and a user-friendly web interface built with Flask to assist healthcare professionals.

Chapter 1: The Challenge

Enhancing Early Detection

Breast cancer is one of the most common and life-threatening cancers globally. Early and accurate detection is paramount to improving survival rates and treatment outcomes.

Traditional diagnostic methods can be time-consuming and subject to human variability. We developed an intelligent, machine-learning-based system to provide an AI-powered diagnostic tool, enhancing accuracy and efficiency.

Project Objectives

High-Accuracy Model

Develop a machine learning model for binary classification of breast tumors (benign vs. malignant).

Intuitive Web Interface

Create a Flask-based interface for single and batch data prediction entries.

Robust Preprocessing

Implement feature scaling and handle class imbalance using SMOTE.

Batch Capabilities

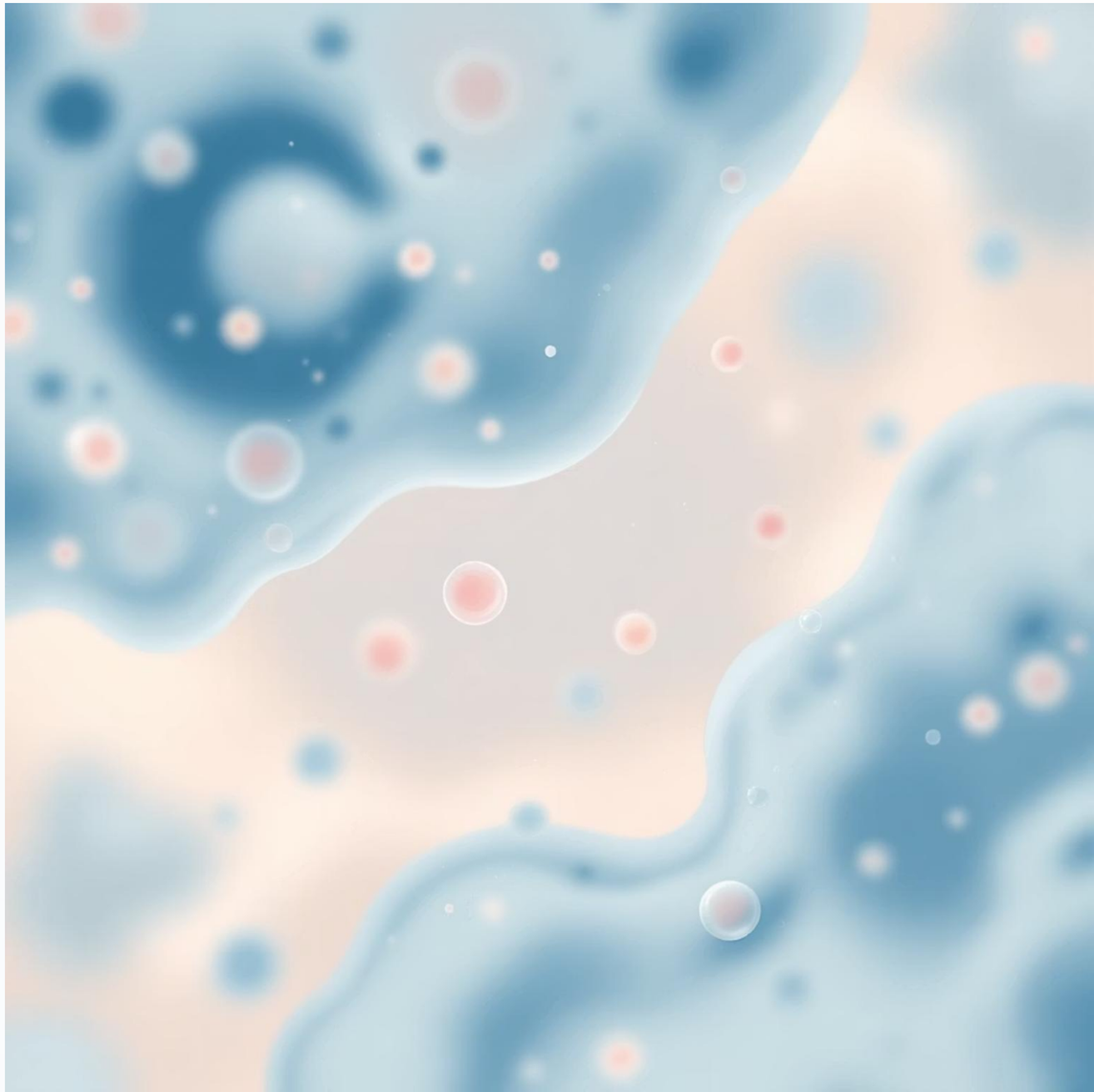
Allow prediction of multiple samples from an uploaded file for practical use.

Chapter 2: Data & Methodology

The Wisconsin Breast Cancer Dataset

The model was trained and tested using the well-known **Wisconsin Breast Cancer Dataset**, a reliable benchmark for classification tasks.

- Source: UCI Machine Learning Repository
- Total Samples: 569 cases
- Features: 30 numerical measurements (e.g., radius, texture, smoothness) from a digitized FNA image.
- Classes: Binary (Malignant vs. Benign)



Data Preprocessing Pipeline

Raw data requires critical steps to ensure model fairness and performance.

Feature Scaling

Applied **StandardScaler** to normalize the 30 numerical features, ensuring equal contribution to the model's decisions.

Handling Imbalance

Used **SMOTE** (Synthetic Minority Over-sampling Technique) to generate synthetic samples for the minority (malignant) class, balancing the training set.

Model Selection

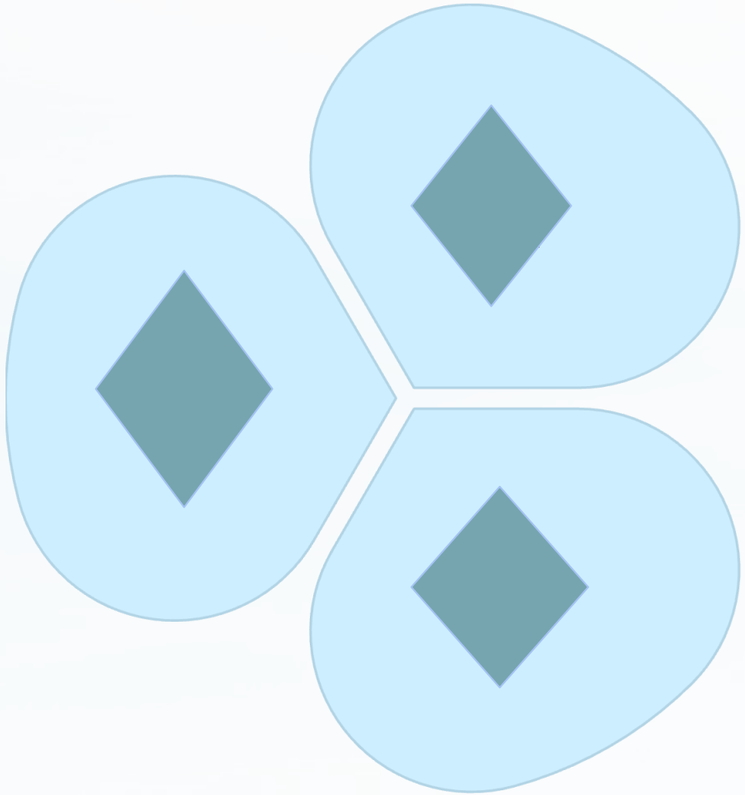
The **RandomForestClassifier** was chosen for its robustness, high performance, and ability to handle non-linear relationships.

System Architecture: Three-Tier Design

The web application is built on a simple and effective architecture for separation of concerns and maintainability.

Presentation Layer

Front-End built with HTML5, CSS3, and JavaScript. Provides interactive forms and displays results.



Application Layer

Flask-based server (`app.py`) handles requests, processes data, and exposes REST API endpoints (`/predict`, `/predict_batch`).

Data Layer

Contains the pre-trained Random Forest model (`arya_best_cancer.joblib`) and uses Pandas/NumPy for data formatting

Exceptional Diagnostic Accuracy

The final model was evaluated on the unseen 20% test split, demonstrating excellent performance and validating the chosen methodology.

97.4%

Accuracy

Overall effectiveness at distinguishing between benign and malignant tumors.

99.64%

ROC AUC Score

Near-perfect score indicating high separability between the two classes.

98.6%

Precision (Malignant)

High confidence in positive predictions (malignant cases).

96.5%

Recall (Malignant)

Effectiveness at identifying all actual malignant cases.

Visualizing Data Relationships

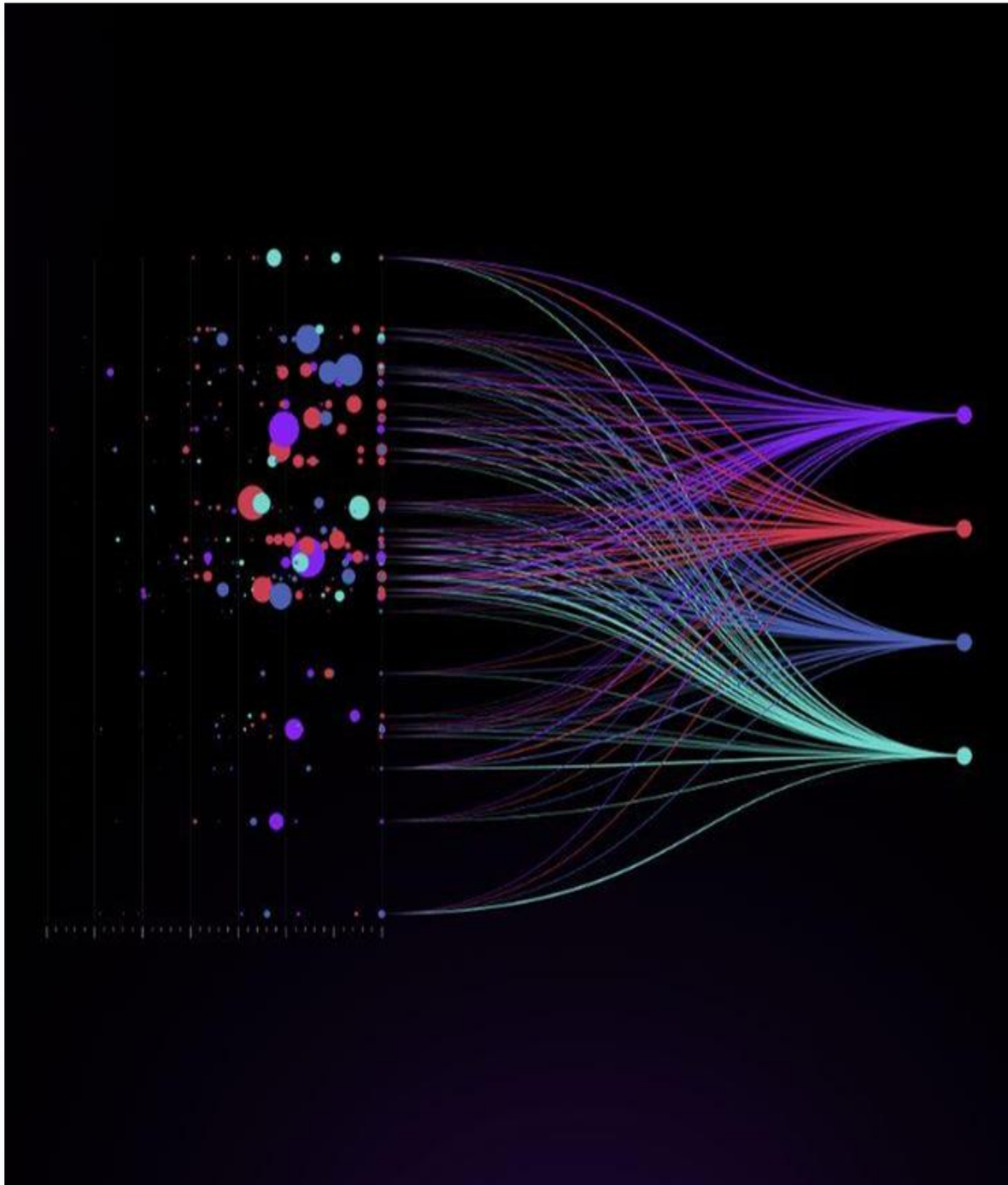
Exploratory Data Analysis (EDA) was crucial for understanding the dataset and guiding feature selection.

Correlation Heatmap

Generated to show correlations between all 30 features, helping identify multicollinearity and features strongly related to the 'diagnosis' target.

Pairplot Analysis

Used to visualize relationships between key features (e.g., radius_mean, texture_mean). This visually confirmed the clear separation between the 'benign' and 'malignant' classes.



Chapter 4: Future Direction

Expanding the Scope of AI Diagnostics

Imaging Integration

Extend the system to accept medical images (mammograms) and use Deep Learning (CNNs) for direct image-based classification.

Multi-class Models

Adapt the model to classify different sub-types of breast cancer, moving beyond the binary benign/malignant distinction.

Real-time Deployment

Deploy the model on a cloud platform (AWS/GCP) for a scalable, real-time API integrated into hospital systems.

Explainable AI (XAI)

Implement XAI techniques (LIME or SHAP) to provide explanations for each prediction, increasing trust for medical staff.

Project Team & Mentorship

This project was made possible by the dedication and expertise of our team.

Mentor

M.R SAURABH

Project Mentor

Team Leaders

MILLI SRIVASTAVA (BCS2023004)

ADITYA RAJ (BCS2023014)

Team Members

1. SNEHIL SINGH
2. AISHWARY MAHESHWARI
3. PRATHAM KUMAR
4. SANJEEV MAURYA
5. JATIN SHARMA
6. AYUSH GANGWAR
7. ABHISHEK YADAV
8. AYUSH PARASHARI