Multi-Modal Disease Detection System
Milestone 2 Progress Report

1 Project Recap

Our project develops an AI-powered early disease detection system that predicts diabetes and stroke risk based on medical measurements. We use ensemble learning methods (Random Forest, XGBoost, Neural Networks) to analyze patient data including glucose levels, BMI, age, and medical history. The system outputs risk predictions to enable early medical intervention.

2 Significant Accomplishments

2.1 Accomplishment 1: Data Preprocessing Pipeline with SMOTE

We implemented a comprehensive preprocessing pipeline handling data cleaning (removing invalid zero values), outlier removal, and LabelEncoder-based categorical encoding. The key component is SMOTE (Synthetic Minority Over-sampling Technique) for class imbalance. Our stroke dataset had 249 stroke cases versus 4,848 healthy cases (19.47:1 ratio). SMOTE generates synthetic samples by interpolating between existing samples and k-nearest neighbors, increasing the dataset from 5,097 to 9,696 samples with perfect 1:1 balance. A critical challenge: SMOTE created invalid decimal values for categorical features (e.g., gender=0.613). We implemented post-SMOTE rounding to nearest integers with range clipping, ensuring all categorical features remain valid discrete values.

2.2 Accomplishment 2: Multi-Algorithm Model Training

We trained three algorithms on diabetes and stroke datasets: **Random Forest** (200 trees, balanced weights) achieved 78.21% diabetes / 96.80% stroke accuracy with 97.11% stroke recall; **XGBoost** (200 rounds, 0.05 learning rate) achieved 75.64% / 95.10%; **Neural Network** (128-64-32 architecture, ReLU, early stopping) achieved 74.36% / 92.47%. All evaluated with 5-fold cross-validation and metrics including accuracy, precision, recall, F1-score, and ROC-AUC. Random Forest consistently performed best.

2.3 Accomplishment 3: Feature Importance Analysis

We extracted feature importance from Random Forest and XGBoost to validate clinically meaningful learning. Diabetes: glucose dominates (25.78%), followed by insulin (17.16%) and age (15.41%)—aligning with medical knowledge. Stroke: age most critical (37.02%), followed by hypertension (14.91%) and glucose (14.79%)—matching clinical understanding that stroke risk increases with age and cardiovascular factors.

3 Proof of Accomplishment

3.1 Proof 1: SMOTE Implementation

The preprocessing summary shows stroke dataset transformation from 5,110 to 9,696 rows. Class distribution changed from severe imbalance (249 stroke vs 4,848 healthy, 19.47:1) to perfect balance (4,848 vs 4,848, 1:1). This 90% dataset increase enabled models to learn stroke patterns rather than defaulting to "no stroke" predictions, achieving 97.11% recall.

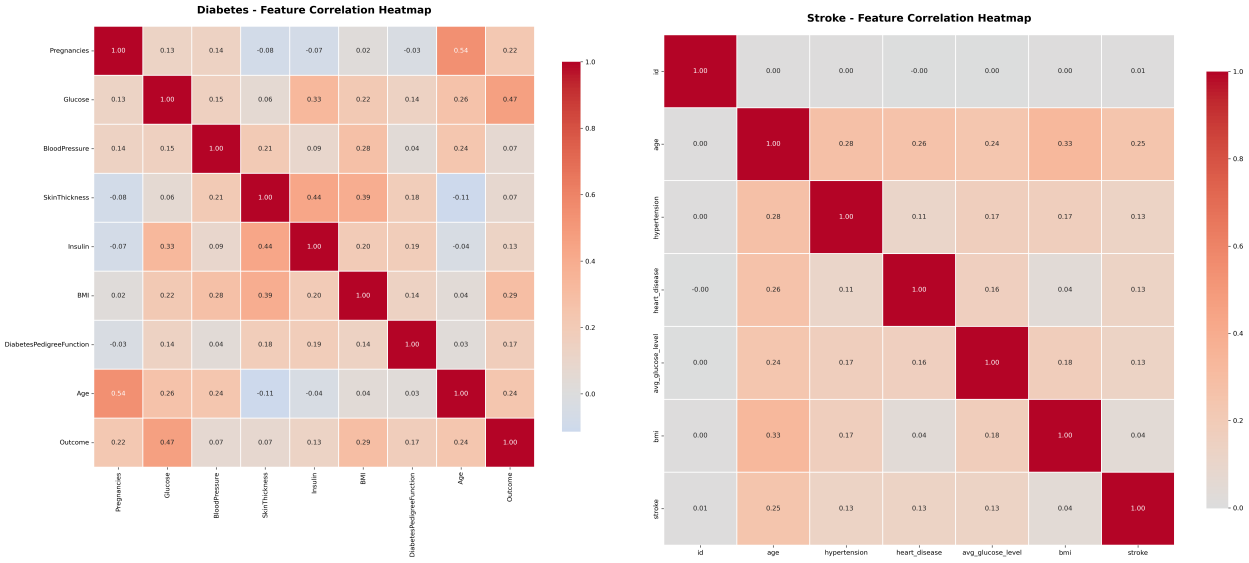

Figure 1: Correlation heatmaps for diabetes (left) and stroke (right) datasets after preprocessing

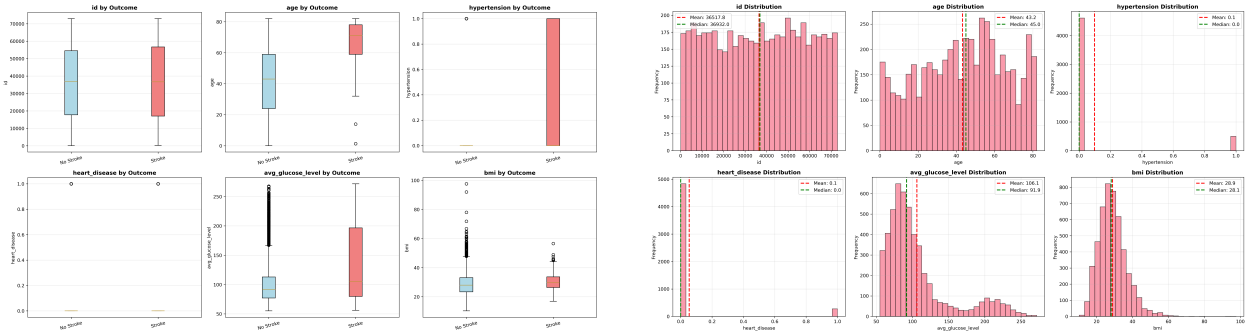

Figure 2: Stroke dataset: Class distribution before/after SMOTE (left) and feature distributions by outcome (right)

3.2 Proof 2: Model Performance

Table 1 shows all six trained models (3 algorithms × 2 diseases) with comprehensive metrics. Random Forest achieved best performance (78.21% diabetes, 96.80% stroke). High ROC-AUC scores (0.8336 diabetes, 0.9956 stroke) indicate strong discriminative ability. Stroke model confusion matrix: TN=936, FP=34 (3.5%), FN=28 (2.9%), TP=942—demonstrating 97.11% recall for medical screening.

Table 1: Model Performance Comparison

Disease	Algorithm	Accuracy	Recall	Precision	ROC-AUC
Diabetes	Random Forest	78.21%	57.69%	71.43%	0.8336
	XGBoost	75.64%	61.54%	64.00%	0.8299
	Neural Network	74.36%	53.85%	63.64%	0.7729
Stroke	Random Forest	96.80%	97.11%	96.52%	0.9956
	XGBoost	95.10%	98.04%	92.60%	0.9927
	Neural Network	92.47%	94.85%	90.55%	0.9755

3.3 Proof 3: Feature Importance

Table 2 shows top features from Random Forest models. For diabetes, glucose dominates (25.78%). For stroke, age is most critical (37.02%). These rankings prove models learned clinically meaningful patterns: glucose as primary diabetes indicator, age as strongest stroke risk factor.

Table 2: Top 5 Feature Importance Rankings

Disease	Rank	Feature	Importance
Diabetes	1	Glucose	25.78%
	2	Insulin	17.16%
	3	Age	15.41%
	4	BMI	11.17%
	5	DiabetesPedigreeFunction	9.69%
Stroke	1	Age	37.02%
	2	Hypertension	14.91%
	3	Avg Glucose Level	14.79%
	4	BMI	11.77%
	5	Heart Disease	8.43%

4 Challenges and Roadblocks

SMOTE Categorical Corruption: SMOTE created invalid decimal categorical values (gender=0.613, work_type=2.347) by interpolating between samples. Solution: post-SMOTE rounding to nearest integers with min/max clipping preserved data integrity.

EDA False Missing Reports: Script incorrectly reported 100% missing categorical values when numeric conversion turned text into NaN. Solution: removed numeric conversion for categorical columns, handled separately.

Extreme Class Imbalance: Stroke dataset 19.47:1 ratio caused models to predict "no stroke" for all patients (95% accuracy, 0% recall). Solution: SMOTE balancing to 1:1 achieved 97.11% stroke recall.

5 Changes from Original Plan

Expanded to Dual Disease System: Milestone 1 focused only on diabetes (Pima Indians dataset). Milestone 2 added stroke prediction (5,110 samples), requiring pipeline adaptation for

different feature sets (diabetes: 8 numeric; stroke: 10 features with 5 categorical) and LabelEncoder implementation.

Added SMOTE Balancing: Original plan lacked class imbalance handling. Stroke dataset's 19.47:1 imbalance required SMOTE implementation, including imbalanced-learn library integration, categorical post-processing, and synthetic sample validation—essential for usable performance.

Three Algorithms vs. KNN Only: Replaced single KNN with Random Forest, XGBoost, and Neural Networks (6 models total), enabling ensemble vs. deep learning comparison. Tree-based models proved superior for medical tabular data (Random Forest best, Neural Networks 2-4% lower due to small datasets).

6 Conclusion and Future Directions

Milestone 2 successfully expanded our disease detection system from single-disease (diabetes) to multi-disease prediction (diabetes and stroke), implementing a robust preprocessing pipeline with SMOTE balancing and training six production-ready models. Key achievements include solving the severe stroke class imbalance (19.47:1 to 1:1) with SMOTE while preserving categorical data integrity through post-processing, training three algorithm families with comprehensive evaluation metrics, and validating that learned feature importance aligns with clinical knowledge (glucose for diabetes, age for stroke).

Performance results demonstrate the system's viability for early disease screening: Random Forest achieved 78.21% accuracy on diabetes with 0.8336 ROC-AUC, and 96.80% accuracy on stroke with exceptional 97.11% recall—critical for medical applications where missing positive cases has severe consequences. The comparison across algorithms revealed that tree-based ensemble methods (Random Forest, XGBoost) consistently outperform neural networks on small-to-medium tabular medical datasets, likely due to their ability to handle feature interactions without requiring massive training data.

For Milestone 3, we will focus on model deployment and real-world usability. Planned work includes: (1) developing a user-friendly prediction interface that accepts patient medical data and returns risk assessments with confidence scores, (2) implementing model explanation to provide interpretable predictions, showing which patient features contribute most to risk predictions, (3) conducting analysis to understand model behavior under missing or uncertain input data, and (4) creating documentation for model usage, limitations, and ethical considerations in medical AI. These enhancements will transform our trained models into a practical support tool.