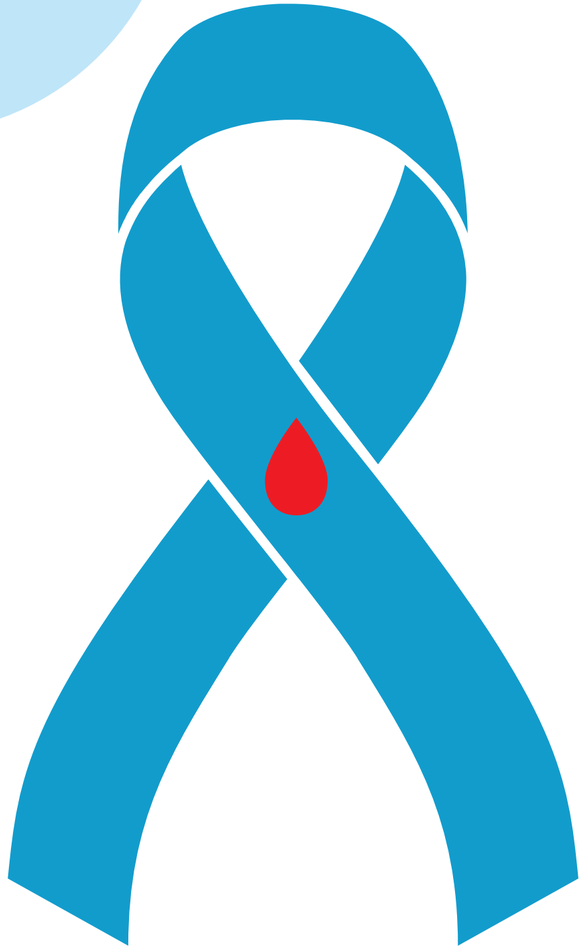
Optimizing Early Diabetes Diagnosis Through Artificial Neural Networks



PRESENTED BY:
Amiera Masheetah & Shaidatullisa Nadia

Introduction

- Diabetes is often described as a “silent disease” because many individuals do not realize they are affected until serious complications appear
- Traditional diagnosis relies on lab tests and clinical evaluations, which may delay detection, especially for patients with mild or unclear symptoms
- We propose using an Artificial Neural Network (ANN) to classify individuals as diabetic or non-diabetic

Proposed Approach

- **Model Type:** Feedfoward Artificial Neural Network (ANN)
- **Input Layer:** 16 features
- **Hidden Layers:**
  - Layer 1 → 32 neurons (ReLU)
  - Layer 2 → 16 neurons (ReLU)
- **Output Layer:** 1 neuron (Sigmoid)
- **Loss Function:** Binary Cross-Entropy
- **Optimizer:** Adaptive Moment Estimation (Adam)
- **Regularization:** Dropout + Early Stopping
- **Goal:** Learn non-linear symptom relationships for accurate diabetes prediction

Datasets

Early Stage Diabetes Risk Prediction Dataset

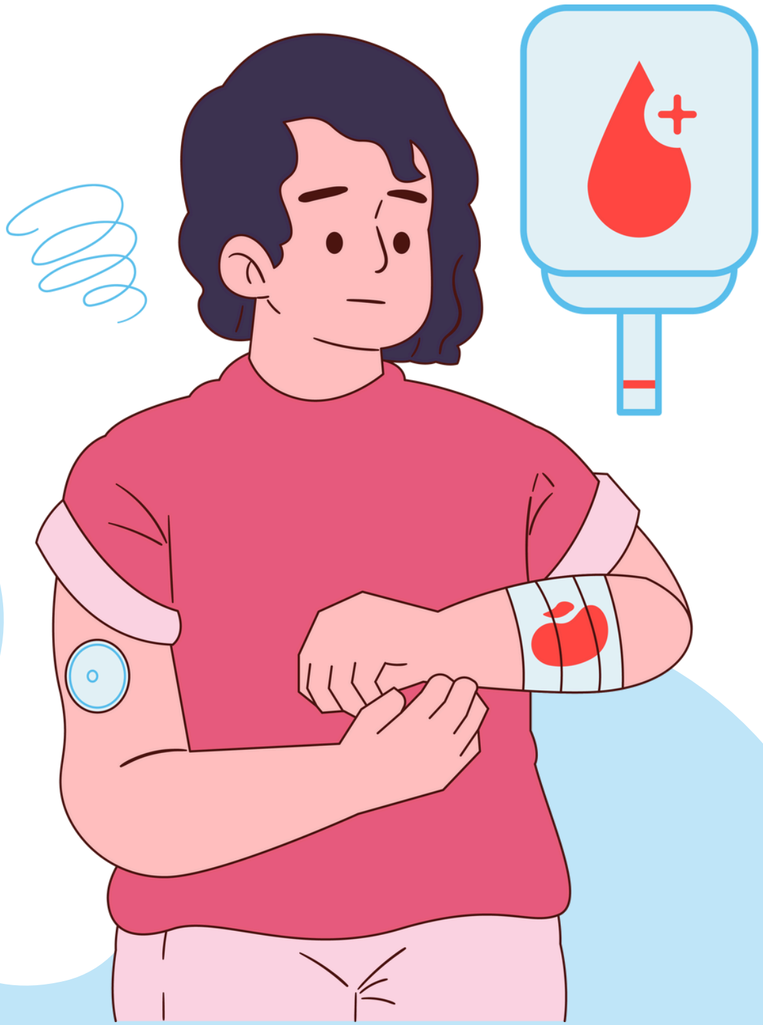
Total Records: 520 patients

Features: 16 demographic & symptom-based variables

Target: Diabetic/Non-diabetic

Early stage diabetes risk prediction dataset

Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itching	Irritability	delayed healing	partial paresis	muscle stiffness	Alopecia	Obesity	class
40	Male	No	Yes	No	Yes	No	No	No	Yes	No	Yes	No	Yes	Yes	Yes	Positive
58	Male	No	No	No	Yes	No	No	Yes	No	No	No	Yes	No	Yes	No	Positive
41	Male	Yes	No	No	Yes	Yes	No	No	Yes	No	Yes	No	Yes	Yes	No	Positive
45	Male	No	No	Yes	Yes	Yes	Yes	No	Yes	No	Yes	No	No	No	No	Positive
60	Male	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Positive
55	Male	Yes	Yes	No	Yes	Yes	No	Yes	Yes	No	Yes	No	Yes	Yes	Yes	Positive
57	Male	Yes	Yes	No	Yes	Yes	Yes	No	No	No	Yes	Yes	No	No	No	Positive
66	Male	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	Yes	Yes	No	No	Positive
67	Male	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	No	Yes	Yes	No	Yes	Positive
70	Male	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	No	No	No	Yes	No	Positive
44	Male	Yes	Yes	No	Yes	No	Yes	No	No	Yes	Yes	No	Yes	Yes	No	Positive
38	Male	Yes	Yes	No	No	Yes	Yes	No	Yes	No	Yes	No	Yes	No	No	Positive
35	Male	Yes	No	No	No	Yes	Yes	No	No	Yes	Yes	No	No	Yes	No	Positive
61	Male	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	No	Yes	Yes	Positive
60	Male	Yes	Yes	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes	No	No	No	Positive
58	Male	Yes	Yes	No	Yes	Yes	No	No	No	No	Yes	Yes	Yes	No	No	Positive
54	Male	Yes	Yes	Yes	Yes	No	Yes	No	No	No	Yes	No	Yes	No	No	Positive
67	Male	No	Yes	No	Yes	Yes	No	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Positive
66	Male	Yes	Yes	No	Yes	Yes	No	Yes	No	No	No	Yes	Yes	No	No	Positive
43	Male	Yes	Yes	Yes	Yes	No	Yes	No	No	No	No	No	No	No	No	Positive
62	Male	Yes	Yes	No	Yes	Yes	No	Yes	No	Yes	No	Yes	Yes	No	No	Positive
54	Male	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	No	Yes	Yes	No	Positive
39	Male	Yes	No	Yes	No	No	Yes	No	Yes	Yes	No	No	No	Yes	No	Positive
48	Male	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes	No	No	No	No	Positive
58	Male	Yes	Yes	Yes	Yes	Yes	No	Yes	No	No	Yes	Yes	Yes	No	Yes	Positive
32	Male	No	No	No	No	No	Yes	No	No	Yes	Yes	No	No	No	Yes	Positive
42	Male	No	No	No	Yes	Yes	No	No	No	Yes	No	No	Yes	No	No	Positive
52	Male	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	No	Yes	Yes	Yes	No	No	Positive
38	Male	No	Yes	No	No	No	Yes	No	No	No	No	No	No	Yes	No	Positive
53	Male	Yes	Yes	Yes	Yes	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Positive



Data Processing

- **Label encoding for categorical data**
  - Many features in the dataset are recorded as text value like Yes/No. Since neural networks can only work with numbers, we convert in into 1/0.
- **Normalization of numerical features**
  - Rescales these values into a similar range so that the model trains faster
- **Train-test split**
  - 80% Training - used to teach the model
  - 20% Testing - used to evaluate how well the model performs on new data

Baselines

Models Introduced

- Logistic Regression → interpretable linear model.
- SVM (RBF) → strong non-linear classifier for small datasets.
- Random Forest → robust tree-based ensemble for binary features.
- KNN → simple distance-based classifier.

How We Used Them

- Implemented using Stratified 5-fold Cross-Validation for generalizability.
- Metrics: Accuracy, Precision, Recall, F1, ROC-AUC.

Result Comparison

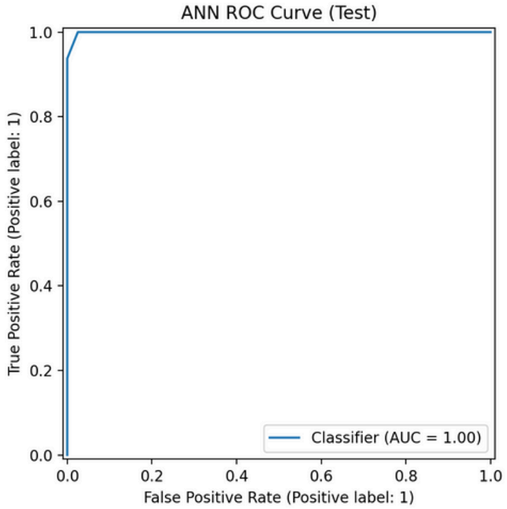
A	B	C	D	E	F
Model	Accuracy	Precision	Recall	F1	ROC_AUC
ANN (thr tuned on val; calibrated)	0.971154	0.955224	1	0.977099	0.999219
Random Forest	0.980769	0.984375	0.984375	0.984375	0.998828
SVM (RBF, calibrated)	0.990385	0.984615	1	0.992248	0.998438
Logistic Regression	0.942308	0.983333	0.921875	0.951613	0.990625
KNN (k=5)	0.932692	0.983051	0.90625	0.943089	0.961523

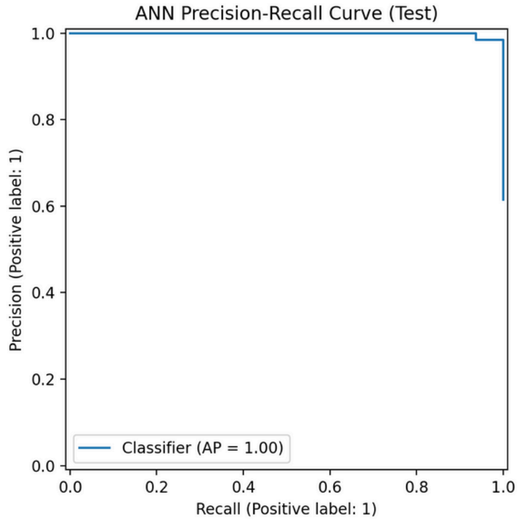
ANN Performance Metrics
Artificial Neural Network reached high recall and balanced precision, fitting clinical goals with zero false negatives.

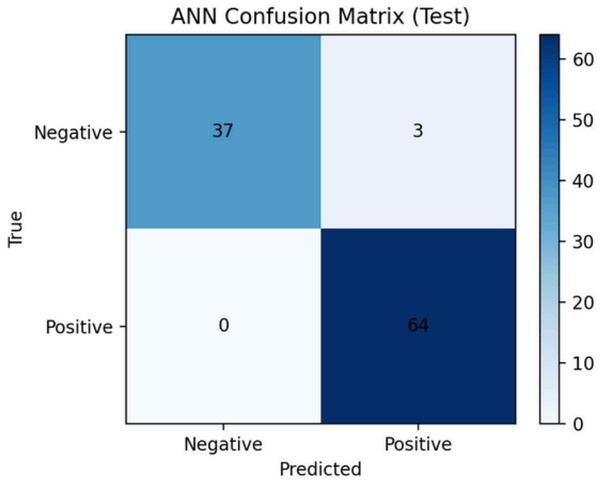
SVM and Random Forest Results
Support Vector Machine showed the highest accuracy and F1-score, while Random Forest closely followed in performance.

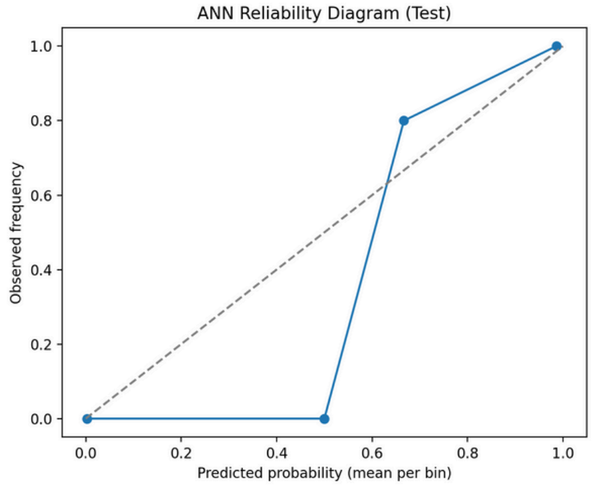
Clinical Screening Priorities
Emphasis on recall minimization of missed diabetic cases achieved by threshold tuning and calibration.

ANN performance









ROC Curve Analysis
ROC curve shows AUC near 1.0, demonstrating excellent classdiscrimination for the ANN model.

Precision-Recall Curve
Precision-Recall curve confirms strong precision at high recall, critical for effective screening applications.

Confusion Matrix Results
Confusion matrix shows zero false negatives and minimal false positives,validating threshold tuning strategy.

Calibration Improvement
Reliability diagram illustrates improved calibration after isotonic regression,ensuring trustworthy probability estimates.

Discussion

Why ANN Didn’t Outperform All Baselines

- Tabular data with binary features often favors tree-based models and SVM.
- ANN requires more data to fully leverage non-linear interactions.
- Despite this, ANN:
  - Matches baselines on AUC (≈0.999).
  - Achieves zero false negatives (critical for early detection).
  - Provides calibrated probabilities for decision-support systems.

Conclusion

ANN Model Effectiveness
The ANN model achieves perfect recall and high AUC,effectively minimizing missed diabetic cases in clinical settings.

Comparison with Other Models
SVM and Random Forest slightly outperform ANN insome metrics but ANN meets the key goal of reducing missed cases.

Future Development Focus
Future efforts include dataset expansion, fairness analysis due to gender influence, and integrating ANN into telehealth tools.

DS402: Group 4

THANK YOU!

sfs6569@psu.edu
abi5143@psu.edu