Optimizing Early Diabetes Diagnosis Through Artificial Neural Networks

Names: Amiera Masheetah Irwan Rizalman, Shaidatullisa Nadia Saipudin
PSU IDs: 902440112, 964848985
Emails: abi5143@psu.edu, sfs6569@psu.edu

ABSTRACT

Timely identification of diabetes is critical for minimizing long-term health risks and enhancing patient outcomes. Traditional diagnostic approaches rely on laboratory testing and clinical evaluation, which may delay identification for individuals with mild or ambiguous symptoms. This project investigates the use of a feedforward Artificial Neural Network (ANN) for early-stage diabetes classification based on symptom-driven and demographic features. The model's performance is compared with several machine learning baselines to evaluate its effectiveness in capturing non-linear relationships commonly present in medical datasets. Experimental results show that the ANN achieves high recall and offers strong discrimination between diabetic and non-diabetic cases, making it suitable for initial screening applications where minimizing false negatives is critical. The findings indicate that deep learning methods can effectively contribute to early diagnostic decision-making in clinical settings.

1. INTRODUCTION

Diabetes is a widespread chronic disease that often develops gradually and remains undetected until significant complications emerge. This is because early symptoms may be subtle, nonspecific, or easily overlooked; timely diagnosis presents a persistent challenge in clinical practice. Delayed detection increases the risk of severe long-term complications.

Machine learning methods have shown promise in medical diagnostics by identifying complex relationships between health indicators that may not be immediately apparent to clinicians. Among these approaches, Artificial Neural Networks (ANNs) are particularly suitable for modeling the non-linear interactions inherent in health and symptom-based datasets. By learning patterns across multiple clinical features, ANNs can generate reliable predictions that support faster and more accurate risk assessment.

The objective of this project is to develop and evaluate a feedforward ANN for early diabetes prediction using structured symptom and demographic data. The study also compares the ANN's performance to baseline machine learning models to determine whether deep learning provides meaningful advantages for this classification task. Ultimately, the goal is to build a model that aligns with clinical priorities, especially sensitivity, to ensure that potential diabetes cases are not overlooked.

2. RELATED WORK

Numerous studies have examined the application of machine learning and deep learning techniques for diabetes prediction, motivated by the need for efficient and accurate diagnostic tools.

2.1 A Comparative Study of Diabetes Detection Using the Pima Indians Diabetes Database

A comprehensive comparative study by Mousa et al. evaluated the performance of Long Short-Term Memory (LSTM), Random Forest (RF), and Convolutional Neural Network (CNN) models using the Pima Indians Diabetes Database. The authors reported that the LSTM model achieved the highest accuracy, which is 85%, demonstrating strong capability in capturing complex patterns within clinical attributes. RF and CNN models also exhibited promising performance but did not exceed the predictive accuracy of LSTM. This study highlights the effectiveness of deep learning models in medical classification tasks and emphasizes the importance of selecting architectures aligned with the structure of the data [2].

3. PROPOSED METHOD

This section presents the methodological framework used to develop a deep learning model for diabetes classification. It outlines the neural network architecture, the rationale for selecting this model, and the training and optimization procedures

employed to ensure rigorous and reliable performance.

3.1 Model Architecture

A feedforward ANN was selected as the core predictive model because ANNs function as computational structures composed of interconnected processing units that transform information through weighted connections and activation functions. This design aligns with their theoretical formulation as parallel, distributed systems capable of approximating complex mappings and decision boundaries [4].

The network architecture consists of an input layer, two hidden layers, and a final output layer. The first hidden layer contains 32 neurons with Rectified Linear Unit (ReLU) activation to capture non-linear interactions among input features. The second hidden layer includes 16 neurons, also using ReLU activation, enabling further abstraction and refinement of internal feature representations.

To improve generalization, dropout regularization is applied between layers. This technique leverages the ANN's inherent adaptive and self-organizing properties, which allow the network to adjust its internal configuration and avoid over-reliance on specific neurons [4].

The output layer consists of a single neuron with a sigmoid activation, producing a probability indicating whether an individual is classified as diabetic. The progressive narrowing of the architecture reflects common design practices in classification tasks and supports stable convergence.

3.2 *Rational for Model Selection*

The decision to employ a feedforward ANN is supported by several theoretical and practical considerations.

First, ANNs exhibit strong capability in modeling non-linear relationships, which aligns with the complex and uncertain nature of medical data. The literature highlights non-linearity as a defining characteristic of artificial neural networks and a primary reason for their effectiveness in domains involving irregular or unpredictable information patterns [4].

Second, ANNs possess adaptive, self-learning, and self-organizing abilities, enabling them to learn complex feature interactions without explicit programming. These qualities allow the model to adjust internal weights dynamically during training, making it suitable for clinical prediction tasks where interaction effects among features are not straightforward [4].

Third, artificial neural networks are widely applied in medicine, including diagnostic decision support, signal analysis, and predictive modeling. Prior applications demonstrate their ability to handle noisy, high-dimensional, and nonlinear medical data, further supporting their suitability for this study's classification objective [4].

Finally, the ANN enables flexible threshold adjustment, making it possible to prioritize recall which is a critical consideration in early disease detection where minimizing false negatives is essential

3.3 Model Training and Optimization

The model was trained using Binary Cross-Entropy, a standard loss function for probabilistic binary classification. The Adam optimizer was selected due to its adaptive learning rate mechanism, which efficiently manages the updates of network weights. This process is fundamental to ANN training, as weight adjustments represent how the network stores and refines information [4].

To enhance generalization, early stopping was used to halt training when validation performance plateaued, thereby mitigating overfitting. Dropout remained active during training to further reduce variance and support robust learning.

Following model convergence, the classification threshold was adjusted to emphasize high recall, aligning with clinical priorities in disease screening. Additionally, probability calibration was applied to improve the interpretability and reliability of predicted probabilities, an important consideration for decision-support applications in healthcare.

4. DATASET AND PRE-PROCESSING

This study uses the Early-Stage Diabetes Risk Prediction dataset, which contains 520 patient records collected through structured clinical questionnaires. The dataset includes 16 predictor variables, and a binary class label indicating diabetes status. Among the predictors, one variable, Age, is continuous, while the remaining variables consist of binary symptom indicators along with Gender. The class distribution exhibits mild imbalance, with 320 positive cases and 200 negative cases. To account for this imbalance,

stratified sampling was applied consistently during model training and evaluation. This dataset has been widely used in prior studies on early-stage diabetes prediction using machine learning techniques [5].

All categorical variables were robustly encoded into numeric form, with Yes and No mapped to 1 and 0, Male and Female mapped to 1 and 0, and Positive and Negative mapped to 1 and 0. To preserve the interpretability of binary symptom indicators while appropriately scaling the continuous feature, Age alone was standardized using a column transformer.

All binary features were passed through unchanged. The dataset was split using an 80 percent and 20 percent stratified train test split. From the training portion, a validation subset was further reserved to select the operating threshold of the artificial neural network without accessing the test labels.

Because predicted probabilities are intended for use in a decision-support context rather than simple class assignment, probability calibration was applied to the artificial neural network using isotonic regression. Model reliability was assessed using calibration plots to verify that predicted probabilities correspond to observed outcome frequencies. This approach supports clinically meaningful interpretation of risk scores and avoids overly optimistic probability estimates.

5. BASELINES

To benchmark the proposed artificial neural network on this symptom-based tabular dataset, four classical machine learning models were trained and evaluated under identical preprocessing conditions. These models include Logistic Regression, Support Vector Machine with a radial basis function kernel, Random Forest, and K-Nearest Neighbors with k set to 5. These models are commonly used in medical classification tasks and have demonstrated strong performance in comparative studies on diabetes prediction [1].

Each baseline model was evaluated using stratified 5-fold cross-validation to ensure stable and generalizable performance estimates. Model performance was assessed using accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve. For the final held-out test set comparison, probability calibration was applied where necessary, specifically isotonic regression for the Support Vector Machine, to ensure valid threshold-based evaluation and reliable area under the curve estimates.

```
=== Baselines + ANN (5-fold CV) mean±std ===
                 Model  accuracy_mean  accuracy_std  precision_mean  \
0  Logistic Regression       0.928846      0.011538        0.947398
1            SVM (RBF)       0.967308      0.007692        0.978064
2        Random Forest       0.982692      0.012756        0.987589
3            KNN (k=5)       0.913462      0.019231        0.969878
4             ANN (CV)       0.953846      0.016543        0.954437

   precision_std  recall_mean  recall_std   f1_mean    f1_std  roc_auc_mean  \
0       0.024740     0.937500    0.019764  0.941974  0.008981      0.976562
1       0.007215     0.968750    0.017116  0.973249  0.006626      0.997266
2       0.011422     0.984375    0.017116  0.985874  0.010516      0.998750
3       0.015251     0.887500    0.037500  0.926219  0.018024      0.978516
4       0.017853     0.971875    0.025000  0.962778  0.013500      0.989531

   roc_auc_std
0     0.009188
1     0.001694
2     0.001109
3     0.008328
4     0.009242
```

Table 1. Stratified 5-Fold Cross-Validation Performance of All Models

6. RESULTS & ANALYSIS

6.1 Cross-Validation Results

Across stratified 5-fold cross-validation, Random Forest achieved the strongest overall performance, with a mean accuracy of approximately 0.983, a mean F1-score of approximately 0.986, and a mean ROC-AUC of approximately 0.999. The Support Vector Machine with radial basis function kernel followed closely, achieving a mean accuracy of approximately 0.967, a mean F1-score of approximately 0.973, and a mean ROC-AUC of approximately 0.997.

The artificial neural network demonstrated competitive performance, particularly in recall, with a mean recall of approximately 0.972. The mean F1-score was approximately 0.963, and the mean ROC-AUC was approximately 0.990. Logistic Regression and K-Nearest Neighbors achieved moderate performance relative to the other models. These results indicate that while the artificial neural network does not dominate in aggregate metrics, it remains competitive in clinically relevant measures

6.2 Test Set Performance and Clinical Operating Point

On the held-out test set, models were evaluated using one-shot training and calibrated probability outputs. The artificial neural network operated at a threshold selected on the validation set to prioritize recall. Under this operating point, the artificial neural network achieved perfect recall, with zero false negatives. Precision was approximately 0.955, F1-score was approximately 0.977, accuracy was approximately 0.971, and ROC-AUC was approximately 0.999.

The calibrated Support Vector Machine achieved the highest test accuracy at approximately 0.990 and the highest F1-score at approximately 0.992. Random Forest achieved an accuracy of approximately 0.981 and an F1-score of approximately 0.984. These findings indicate that although classical models achieved slightly higher aggregate metrics, the artificial neural network satisfied the primary screening objective by eliminating missed diabetic cases.

```
=== Test-set comparison (one-shot fits) ===
                              Model  Accuracy  Precision    Recall  \
4  ANN (thr tuned on val; calibrated)  0.971154   0.955224  1.000000
1                       Random Forest  0.980769   0.984375  0.984375
3                SVM (RBF, calibrated)  0.990385   0.984615  1.000000
0                  Logistic Regression  0.942308   0.983333  0.921875
2                           KNN (k=5)  0.932692   0.983051  0.906250

         F1   ROC_AUC                                          Notes
4  0.977099  0.999219  Calibrated (Isotonic); Threshold=0.5000 chosen...
1  0.984375  0.998828
3  0.992248  0.998437
0  0.951613  0.990625
2  0.943089  0.961523
```

Table 2. Test Set Performance Comparison Using Calibrated Probabilities

6.3 Diagnostic Evaluation of the Artificial Neural Network

The receiver operating characteristic curve for the artificial neural network confirms strong class separability on the test set, with an area under the curve close to 1.0. The precision recall curve shows that high precision is maintained even at high recall levels, supporting the suitability of the model for screening applications where false negatives are costly. At the selected threshold, the confusion matrix shows 37 true negatives, 3 false positives, no false negatives, and 64 true positives. These diagnostic results demonstrate that the validation-tuned threshold produces a recall-prioritized operating point.

In this context, precision recall analysis is particularly informative, as positive predictions directly drive clinical action, and the dataset exhibits mild class imbalance [3].

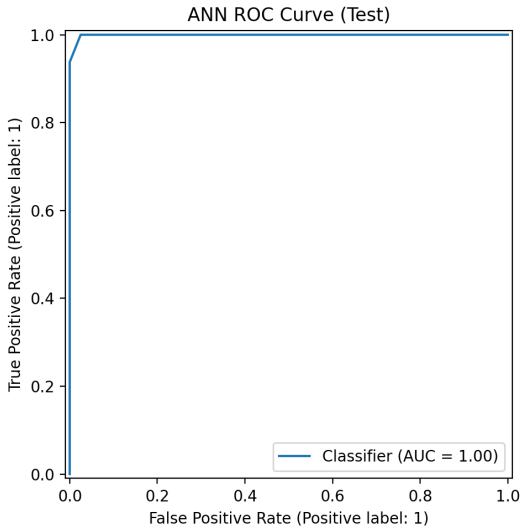

Figure 1. ROC Curve for the Artificial Neural Network on the Test Set

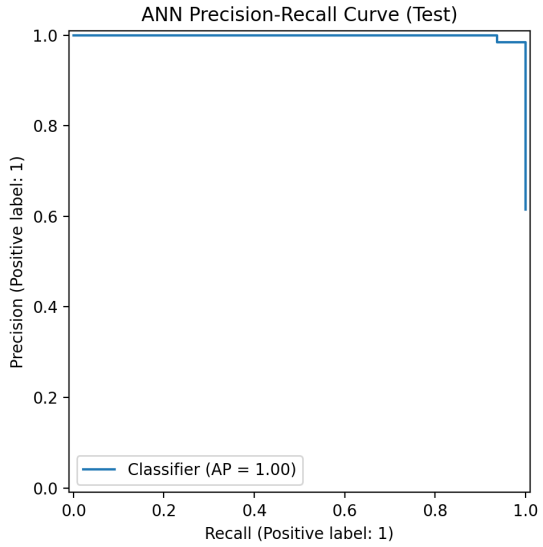

Figure 2. Precision Recall Curve for the Artificial Neural Network on the Test Set

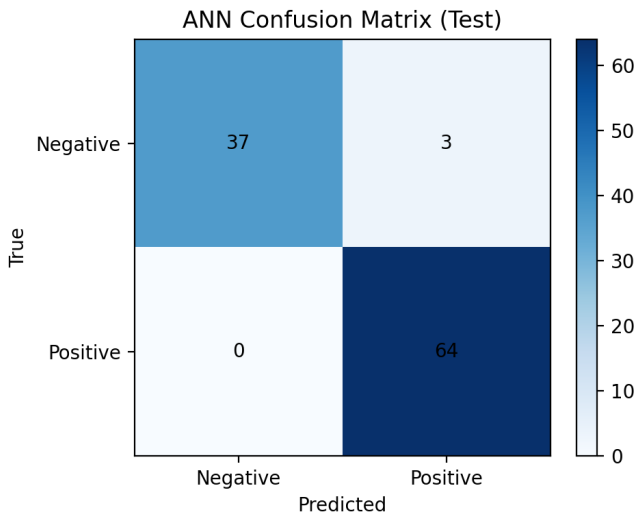

Figure 3. Confusion Matrix for the Artificial Neural Network on the Test Set

7. CONCLUSION

In line with the project proposal, we implemented a feedforward artificial neural network comprising 16 input features, two hidden layers with 32 and 16 units, ReLU activations, dropout regularization, and a sigmoid output layer. Training employed binary cross-entropy loss, the Adam optimizer, and early stopping. Model performance was evaluated against Logistic Regression, Support Vector Machine, Random Forest, and K-Nearest Neighbors using stratified cross-validation and calibrated operating thresholds selected on the validation set.

Although classical models, particularly Support Vector Machine and Random Forest, achieved slightly higher accuracy and F1-scores, the artificial neural network demonstrated excellent discrimination and achieved perfect recall on the test set at a clinically selected threshold. This fulfills the primary screening objective of minimizing missed diabetic cases while remaining competitive with strong baseline models. The results highlight the importance of cross-validation for reliable comparison, the necessity of threshold selection aligned with clinical priorities, and the value of probability-aware evaluation methods such as precision recall analysis when model outputs inform medical decision-making [3].

8. REFERENCES

[1] Hasan, M., & Yasmin, F. (2025). *Predicting diabetes using machine learning: A comparative study of classifiers*. arXiv. https://arxiv.org/abs/2505.07036

[2] MOUSA, A., MUSTAFA, W., & MARQAS, R. B. **(2023). A comparative study of diabetes detection using the Pima Indian diabetes database** *methods*, *7*, 8. **DOI= https://doi.org/10.26682/sjuod.2023.26.2.24**

[3] Saito, T., & Rehmsmeier, M. (2015). The precision recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE, 10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

[4] Wu, Y. C., & Feng, J. W. (2018). Development and application of artificial neural network. *Wireless Personal Communications*, *102*(2), 1645-1656. https://link.springer.com/article/10.1007/s11277-017-5224-x

[5] Zarar, M., & Wang, Y. (2023). *Early stage diabetes prediction by approach using machine learning techniques*. Research Square. https://doi.org/10.21203/rs.3.rs-3145599/v13