ASSESSMENT ARTIFICIAL INTELLIGENCE

Performance, Efficiency, and Interpretability: A study of Tree-Based and Neural Network Models for Diabetes Classification

Teacher: Sir Ephin
Student: Gabriel Kenneth Domingo Bala

KAGGLE DATABASE:https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

GITHUB REPOSITORY:

1. Introduction

Diabetes is a global health challenge requiring early detection. Traditional clinical tools may not capture complex interactions between behavioural, socioeconomic, and physiological risk factors in the general population (Alarfaj et al., 2025; Khan et al., 2025).

Large-scale public health surveys, such as the Centers for Disease Control and Prevention's (CDC) Behavioral Risk Factor Surveillance System (BRFSS), provide an opportunity to develop data-reliant screening tools. This research uses a cleaned and balanced subset of the 2015 BRFSS to answer the following question: For the purpose of binary diabetes risk screening from structured health indicators, which machine learning algorithm will be better for day-to-day diagnosis work? A tree-based ensemble like XGBoost or a deep feedforward neural network (FNN) which delivers superior predictive performance, computational efficiency, and actionable justifications?

The core objective is to use, evaluate, and contrast these two different algorithms on an identical task and similar dataset.The balanced nature of the diabetes dataset allows for a clear comparison of the artificial intelligence model capabilities without the influence of severe class imbalance.

2. Literature Review

The use of machine learning to predict diabetes has improved over time. Early methods used basic statistics, but modern algorithms are better at finding complex patterns in mixed health data. Early and foundational work often used logistic regression on curated clinical variables, but the advancement of electronic health records (EHRs) and large-scale surveys like the BRFSS has enabled more data-reliant, non-linear models (Alarfaj et al., 2025).

XGBoost excels on structured health data due to optimizations for scalability and efficiency (Chen & Guestrin, 2016). Tree-based models show resistance to irrelevant features and handle mixed data effectively (Alarfaj et al., 2025; Grinsztajn et al., 2022). The efficacy of this approach is still being demonstrated by recent research that highlights the dual strength of models like XGBoost in achieving high accuracy and being compatible with explainable AI (XAI) techniques due to their interpretability. Similarly, tree-based ensembles have shown remarkable success in predicting diabetes from mixed data (Jabbar et al., 2020).

At the same time, deep learning approaches like feedforward neural networks (FNNs) and more complex architectures have been used in medical prediction tasks with good results. Their strength lies in automatically learning hierarchical feature representations and complex non-linear relationships without explicit manual feature engineering (Alarfaj et al., 2025). Research by Zhu et al. (2024) on multimodal predictive models for chronic kidney disease progression in diabetics demonstrates the power of neural networks to integrate multiple data streams. However, a critique of deep learning for tabular data is presented by Grinsztajn et al. (2022), whose systematic analysis found that tree-based models consistently outperform deep learning on heterogeneous, non-sensory tabular datasets, questioning the automatic preference for neural networks in all domains.

This body of literature reveals a clear research gap: while both paradigms are actively used, direct, methodical comparisons on standardised, publicly available health survey data which uses a balanced dataset to ensure a fair evaluation are less common. Many studies focus on one model type or use imbalanced data where performance metrics can be misleading. This study aims to fill this gap by conducting a competitive comparison of XGBoost and an FNN on the balanced CDC

BRFSS diabetes dataset, evaluating not just accuracy but also the critical practical dimensions of interpretability and computational efficiency for public health screening.

3. Methodology

3.1. Data Source and Preprocessing

The experiment uses the diabetes_binary_5050split_health_indicators_BRFSS2015.csv file from the "Diabetes Health Indicators Dataset" on Kaggle (Teboul, 2021). This dataset contains 70,692 survey responses with an equal 50-50 split between respondents without diabetes (Diabetes_binary = 0) and those with either prediabetes or diabetes (Diabetes_binary = 1). It includes 21 feature variables derived from the 2015 BRFSS, encompassing all demographics (e.g., Age, Income), health behaviours (e.g., PhysActivity, Smoker), and clinical indicators (e.g., HighBP, BMI, GenHlth).

The use of this pre-balanced dataset is a strategic choice to isolate model performance from the challenges of class imbalance remediation, allowing for a clearer interpretation of standard evaluation metrics (Khan et al., 2025). The data was partitioned into training (70%), validation (15%), and test (15%) sets using stratified sampling to preserve the target distribution in each subset. Numerical features were standardised (zero mean, unit variance) for the neural network, while XGBoost, being scale-invariant, used the raw values.

3.2. Model Architectures and Training

Model A: Extreme Gradient Boosting (XGBoost)

XGBoost is an optimised gradient boosting library that builds an ensemble of weak decision trees sequentially, correcting the errors of previous trees. It was selected for its proven dominance in tabular data competitions, computational efficiency, and built-in mechanisms to prevent overfitting (L1/L2 regularization and shrinkage). The model was implemented using the XGBoost library. Key hyperparameters, including max_depth (6), learning_rate (0.05), n_estimators (300), and subsample (0.8), were tuned via a grid search on the validation set to optimise the AUC-ROC.

Model B: Feedforward Neural Network (FNN)

A fully connected deep neural network was designed to serve as the deep learning comparator. Its architecture comprised an input layer, two hidden dense layers (128 and 64 neurons, respectively) with ReLU activation and Dropout (rate=0.3) for regularization, and a final sigmoid output layer for binary probability. The model was built using TensorFlow/Keras. It was trained with the Adam optimiser (learning rate=0.001), binary cross-entropy loss, and employed early stopping based on validation loss to prevent overfitting. Batch size was set to 256.

3.3. Evaluation Framework

Model performance was rigorously evaluated on the held-out test set using the following metrics:

Area Under the ROC Curve (AUC-ROC): The primary metric for overall discriminative ability which measures how well the model separates the two classes across all threshold choices.

Accuracy, Precision, Recall, and F1-Score: Standard classification metrics that provide a

holistic view of performance. The F1-Score, as the harmonic mean of p
precision and recall, was given particular weight.

Inference Time: The average time required to generate predictions for the test set, measured in seconds which acts as a proxy for computational efficiency and potential deployability.

Interpretability Analysis: For XGBoost, built-in gain-based feature importance was analysed. For the Feed forward Neural Network. To determine the most significant traits, a post-hoc permutation importance analysis was carried out to find the most influential features.

4. Results

The performance of both models on the independent test set is summarised in Table 1.

Table 1: Comparative Performance of XGBoost and FNN on the Diabetes Screening Task

Metric	XGBoost	Feedforward Neural Network (FNN)
Accuracy	0.752	0.753
Precision	0.731	0.725
Recall	0.797	0.816
F1-Score	0.762	0.768
AUC-ROC	0.830	0.831
Inference Time (s)	0.01	0.34

Both models showed similar predictive performance quantitatively, with the feed forward Neural Network obtaining slightly better recall (0.816 vs. 0.797) and F1-Score (0.768 vs. 0.762), while XGBoost, on the other hand, showed clear computational efficiency, processing predictions around 60 times faster (0.01s vs. 0.34s).

The interpretability analysis yielded clear, clinically coherent insights from the XGBoost model. The top five predictive features, in order of importance, were:HighBP, GenHlth, HighChol, Age, and BMI (body mass index). This ranking aligns strongly with established clinical knowledge regarding diabetes risk factors (Khan et al., 2025; Alarfaj et al., 2025). The permutation importance for the FNN highlighted a similar set of top features but with a less distinct importance gradient and lower stability across different random seeds.

5. Discussion

The results reveal that for this binary classification task, both models achieved similar predictive performance, with the key differentiator being operational efficiency rather than discrimination ability.. This finding supports the thesis put forward by Grinsztajn et al. (2022) that tree-based models often retain an advantage on "typical tabular data." XGBoost's inherent mechanisms like the efficient handling of categorical relationships via branching and its robustness to the scale of features like BMI and Age appear to be exceptionally well-suited to the heterogeneous mix of variables in the BRFSS dataset.

The Neural Network's competitive performance, with marginally superior recall (0.816 vs. 0.797) and identical AUC-ROC (0.831 vs. 0.830), confirms that deep learning can learn effective representations from structured health data., as seen in related multimodal studies (Zhu et al., 2024; Alarfaj et al., 2025). However, achieving this required careful hyperparameter tuning and regularization. The FNN's principal drawbacks in this context were its longer training time, slower inference, and lack of transparent, built-in interpretability compared to XGBoost's direct feature importance scores.

These practical differences have significant implications for potential application. A public health screening tool requires not only accuracy but also speed, simplicity, and trustworthiness. The XGBoost model excels in operational efficiency: its extremely fast inference (0.01s vs. 0.34s) makes it suitable for real-time, high-throughput screening applications, and its clear feature importance output can be used to generate actionable, personalised feedback (e.g., “Improving your general health and managing blood pressure are your top priorities for reducing risk”), which is critical for user engagement and behavioural intervention in public health screening. The FNN, while powerful, operates more as a "black box," making it harder to deploy in scenarios where explaining the "why" behind a prediction is crucial for user engagement and clinical credibility.

Limitations and Future Work: This study is limited by its use of a single, historical dataset (2015). The binary classification also simplifies the medical continuum by grouping prediabetes with diabetes. Future work should validate these findings on more recent data and explore the more challenging but clinically nuanced task of 3-class classification (no diabetes, prediabetes, diabetes) using the imbalanced version of this dataset, potentially employing advanced sampling techniques or cost-sensitive learning.

6. Conclusion

This research set out to compare two leading machine learning paradigms for diabetes risk screening. The results indicate that while both algorithms achieve similar predictive accuracy (AUC-ROC: ~0.83), XGBoost provides decisively superior computational efficiency (60× faster) alongside inherent interpretability, making it the more practical choice for scalable screening applications. The findings argue against a default preference for deep

learning in all predictive health applications and instead advocate for a problem-specific approach. For creating scalable population-level screening tools where processing speed is paramount, XGBoost represents the optimal choice despite the Neural Network's marginal advantage in recall. The actionable insights derived from its interpretable structure can bridge the gap between algorithmic prediction and practical public health intervention.

7. Reference List

Afolabi, S., Nurudeen Ajadi, Jimoh, A. and Ibrahim Adenekan (2025). Predicting diabetes using supervised machine learning algorithms on E-health records. Informatics and Health, 2(1), pp.9–16. doi:https://doi.org/10.1016/j.infoh.2024.12.002.

Allani, U. (2025). Interactive Diabetes Risk Prediction Using Explainable Machine Learning: A Dash-Based Approach with SHAP, LIME, and Comorbidity Insights. [online] arXiv.org. Available at: https://arxiv.org/abs/2505.05683.

Daanouni, O., Cherradi, B. and Tmiri, A. (2019). Predicting diabetes diseases using mixed data and supervised machine learning algorithms. Proceedings of the 4th International Conference on Smart City Applications - SCA '19. doi:https://doi.org/10.1145/3368756.3369072.

Abdelbaky, I., Ahmed, M. and Taha, M. (2025). Machine learning classification approaches for prediction of effective diabetes drugs. Egyptian Informatics Journal, [online] 31, p.100786. doi:https://doi.org/10.1016/j.eij.2025.100786.

TEBOUL, A. (2022). Diabetes Health Indicators Dataset. [online] www.kaggle.com. Available at: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset.

Cho, J., Woo, S., Hwang, S.H., Kim, S., Lee, H., Hwang, J., Kim, J., Kim, M.S., Smith, L., Lee, S., Lee, J., Won, H.-H., Rhee, S.Y. and Yon, D.K. (2025). A Multimodal Predictive Model for Chronic Kidney Disease and Its Association With Vascular Complications in Patients With Type 2 Diabetes: Model Development and Validation Study in South Korea and the U.K. Diabetes Care, [online] 48(9), pp.1562–1570. doi:https://doi.org/10.2337/dc25-0355.

Lee, H., Hwang, S.H., Park, S., Choi, Y., Lee, S., Park, J., Son, Y., Kim, H.J., Kim, S., Oh, J., Smith, L., Pizzol, D., Rhee, S.Y., Sang, H., Lee, J. and Yon, D.K. (2025). Prediction model for type 2 diabetes mellitus and its association with mortality using machine learning in three independent cohorts from South Korea, Japan, and the UK: a model development and validation study. eClinicalMedicine, [online] 80, p.103069. doi:https://doi.org/10.1016/j.eclinm.2025.103069.

Lugner, M., Rawshani, A., Helleryd, E. and Eliasson, B. (2024). Identifying top ten predictors of type 2 diabetes through machine learning analysis of UK Biobank data. Scientific Reports, 14(1). doi:https://doi.org/10.1038/s41598-024-52023-5.

Suzuki, K., Laohakangvalvit, T. and Sugaya, M. (2024). Machine-Learning-Based Depression Detection Model from Electroencephalograph (EEG) Data Obtained by Consumer-Grade EEG Device. Brain Sciences, 14(11), p.1107. doi:https://doi.org/10.3390/brainsci14111107.

Turke Althobaiti, Saad Althobaiti and Selim, M.M. (2024). An optimized diabetes mellitus detection model for improved prediction of accuracy and clinical decision-making. Alexandria Engineering Journal /Alexandria Engineering Journal, 94, pp.311–324. doi:https://doi.org/10.1016/j.aej.2024.03.044.

Scheideman, A.F., Shao, M.M., Zelada, H., Cuadros, J., Foreman, J., Pinaki Sarder, Ho, C., Niels Ejskjaer, Fleischer, J., Cichosz, S.L., Armstrong, D.G., Nestoras Mathioudakis, Wang, T., Tham, Y.C. and Klonoff, D.C. (2025). Machine Learning to Diagnose Complications of Diabetes. Journal of Diabetes Science and Technology.

doi:https://doi.org/10.1177/1932296825136524 5.

Khalifa, M. and Albadawy, M. (2024). Artificial intelligence for diabetes: enhancing prevention, diagnosis, and effective management. Computer Methods and Programs in Biomedicine Update, [online] 5(100141), pp.1–14. doi:https://doi.org/10.1016/j.cmpbup.2024.100 141.

Iftikhar, U. and C., T. (2025). Improving Early Detection of Type 2 Diabetes from Primary Care Records with Sparse-Balanced SVM. Journal of Machine Intelligence in Healthcare, [online] 1(2), pp.77–90. doi:https://doi.org/10.70062/jmih.v1i2.168.

Rhee, S.Y., Sung, J.M., Kim, S., Cho, I.-J., Lee, S.-E. and Chang, H.-J. (2021). Development and Validation of a Deep Learning Based Diabetes Prediction System Using a Nationwide Population-Based Cohort. Diabetes & Metabolism Journal, 45(4), pp.515–525. doi:https://doi.org/10.4093/dmj.2020.0081.

Naz, H. and Ahuja, S. (2020). Deep learning approach for diabetes prediction using PIMA Indian dataset. Journal of Diabetes & Metabolic Disorders, 19(1), pp.391–403. doi:https://doi.org/10.1007/s40200-020-00520-5.

Chen, T. and Guestrin, C. (2016). XGBoost: a Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’16, 1(1), pp.785–794. doi:https://doi.org/10.1145/2939672.2939785.