HYBRID MODEL FOR DIABETES PREDICTION FEATURE ANALYSIS

S. deepika[1], S. Amsa[2]

[1]PG student, Department of Computer Application, Jaya College of Arts and Science, Tiruninravur.
[2]Assistant Professor, Department of Computer Application, Jaya College of Arts and Science, Tiruninravur.

Abstract - *Predicting diabetes with high accuracy remains a significant challenge in medical informatics. Existing models, whether traditional machine learning or deep learning, often hit a performance ceiling due to issues like overfitting or an inability to capture the full complexity of patient data. Our work confronts this limitation head-on by proposing a novel hybrid architecture that strategically integrates a Deep Neural Network (DNN) with a Random Forest (RF) classifier. The DNN acts as a powerful feature extractor, uncovering subtle, non-linear patterns within the data, which are then passed to the robust RF for final classification. We trained and tested this model on the Pima Indians Diabetes Dataset from the UCI Machine Learning Repository [1], applying SMOTE to ensure class balance. The results are compelling: our hybrid model achieved a* ***96.4% accuracy****, outperforming both standalone Random Forest (91.2%) and Deep Learning (93.0%) models. This demonstrates that the synergy between these two algorithms creates a more reliable and accurate tool for diabetes risk assessment, holding real promise for clinical decision-support systems.*

Key Words**: Diabetes Prediction, Machine Learning, Random Forest, Deep Learning, Hybrid Model, Feature Importance, Clinical Analytics.**

1. INTRODUCTION

Diabetes mellitus represents a significant worldwide health burden, a persistent metabolic disorder marked by abnormally high concentrations of sugar in the blood. Left unmanaged, it can lead to severe complications, including heart disease, kidney failure, and neuropathy. The cornerstone of mitigating these outcomes is early and accurate detection. Traditional diagnostic methods, while effective, often rely on multiple tests and clinical visits, creating a barrier to rapid screening.

The emergence of machine learning (AI/ML) offers a paradigm shift, enabling the analysis of complex patient datasets to identify at-risk individuals efficiently. However, single-model approaches frequently possess inherent weaknesses. For instance, Random Forest models might overlook complex feature interactions, while Deep Learning models can be data-hungry and prone to overfitting on smaller clinical datasets.

This research is driven by a simple yet powerful question: can we combine the strengths of different algorithms to create a more potent predictor? We answer this by developing a novel hybrid model that fuses the feature extraction prowess of Deep Learning with the classification stability of Random Forest. The key novelty of our approach lies in its architecture: using the DNN not as a classifier, but as an intelligent feature engineering layer that transforms raw input into a higher-order representation optimized for the Random Forest.

To ensure rigorous, reproducible, and comparable results, this study is benchmarked on the Pima Indians Diabetes Dataset from the UCI Machine Learning Repository [1]. The results demonstrate that our hybrid model achieves superior performance, offering a promising path toward reliable, AI-powered clinical decision-support systems.

1.1 LITERATURE REVIEW

The application of machine learning to diabetes prediction has evolved significantly, beginning with traditional statistical models. Algorithms such as **Logistic Regression (LR)** and **Support Vector Machines (SVM)** were valued for their interpretability and efficiency on smaller clinical datasets [5]. While these methods provided a foundational baseline, their performance is often limited by an inherent inability to model the complex, non-linear interactions between key risk factors such as glucose, BMI, and insulin levels, which are critical for accurate diagnosis.

To overcome these limitations, ensemble methods like **Random Forest (RF)** became the benchmark for this task [3]. By aggregating predictions from multiple decision trees, RF effectively reduces overfitting and captures non-linear relationships more reliably than single models, typically achieving accuracies between 85-90% on standard benchmarks. However, its performance can plateau, as it may struggle to infer highly complex, abstract patterns directly from the raw feature space without advanced feature engineering.

The recent shift towards **deep learning (DL)** promised a solution through automatic feature representation. Deep Neural Networks (DNNs) can learn intricate hierarchies of features, pushing accuracy boundaries further. However, their "black-box" nature and propensity to overfit on small, tabular clinical datasets remain significant hurdles. This has spurred interest in **hybrid models** that combine the strengths of different algorithms. While recent ensemble techniques like stacking have shown promise, our work introduces a novel sequential pipeline that uniquely leverages a DNN as a dedicated feature extractor for a

Random Forest classifier. This synergy is designed to harness the representational power of deep learning while retaining the robustness and stability of an ensemble method, addressing the core limitations of each standalone approach.

1.2 NOVELTY OF THE PROPOSED WORK

While hybrid models have been explored in medical informatics, this study introduces a specific and novel architecture that strategically repurposes a Deep Neural Network (DNN) solely as an advanced, non-linear feature extractor for a tree-based ensemble model. Unlike traditional hybrid approaches that might use complex weighting or parallel models, our sequential DNN-to-RF pipeline is designed to directly address the weaknesses of each component: it leverages the DNN's ability to create high-level, abstract feature representations from raw clinical data, which then provides a more discriminative input space for the inherently robust but sometimes limited Random Forest classifier. This synergistic fusion, to the best of our knowledge, has not been extensively applied and evaluated for diabetes prediction on a clinical dataset. Furthermore, our work provides a granular feature importance analysis derived from this hybrid framework, offering unique insights into the risk factors as perceived by the combined model, thereby enhancing both predictive performance and clinical interpretability.

1.3 METHODOLOGY

Data Source and Preprocessing This study utilized the Pima Indians Diabetes Dataset from the UCI Machine Learning Repository, containing 768 patient records with 8 clinical features: Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, and Age. The dataset was preprocessed through median imputation for missing values and Min-Max normalization to scale features between 0-1. To address class imbalance, SMOTE was applied exclusively to training data after an 80-20 train-test split, ensuring no data leakage into the test set.

Model Architecture and Training

Our proposed hybrid model employs a sequential two-stage architecture:

Deep Feature Extraction:

A Deep Neural Network with architecture [8-64-32]nodes using ReLU activation and dropout (0.3) was trained for 150 epochs using Adam optimizer (learning rate=0.001) with early stopping.

Random Forest Classification:

The 32-dimensional features extracted from the DNN's final hidden layer were used as input to a Random Forest classifier with 200 trees and maximum depth of 10.

Experimental Setup:

Performance was compared against two baseline models: a standalone DNN with sigmoid output layer and a standalone Random Forest trained on original features. All models were evaluated on the same held-out test set using accuracy, precision, recall, and F1-score metrics.

2. EXPERIMENTAL RESULTS AND DISCUSSION

We implemented all models in Python using an 80:20 train-test split. Performance was evaluated using standard classification metrics.

Table 1: Performance comparison of models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Random Forest (RF)	91.2	89.5	90.1	89.8
Deep Learning (DNN)	93.0	92.1	91.4	91.7
Hybrid (DNN + RF)	96.4	95.8	96.1	95.9

Our hybrid model achieved the best performance across all metrics, showing statistically significant improvement over both baseline models ($p < 0.05$, McNemar's test). The 96.4% accuracy represents a substantial 5.2% improvement over RF and 3.4% over DNN alone.

The performance gain demonstrates the effectiveness of our sequential architecture. The DNN successfully creates higher-level feature representations that enable the Random Forest to make more accurate classifications. The balanced high scores in both Precision (95.8%) and Recall (96.1%) are particularly valuable for medical applications, where both false positives and false negatives carry significant consequences.

This synergy addresses key limitations of both individual models: the DNN's feature extraction compensates for RF's limited pattern recognition, while RF's ensemble approach provides robustness against overfitting.

3. FEATURE ANALYSIS

To ensure model interpretability and validate clinical relevance, we conducted a comprehensive feature importance analysis using SHAP (SHapley Additive exPlanations) on our hybrid model. This method consistently explains the output of complex models by calculating the contribution of each feature to individual predictions.

3.1 Feature Importance Evaluation

The SHAP analysis revealed a clear hierarchy of risk factors that aligns with clinical knowledge, building confidence in the model's decision-making process:

Feature	Mean SHAP Value
Glucose	0.28
BMI	0.22
Blood Pressure	0.16
Insulin	0.14
Age	0.10
Diabetes Pedigree Function	0.07

The dominance of **Glucose** and **BMI** as the primary predictive factors strongly corroborates established medical understanding of diabetes risk. This alignment demonstrates that our hybrid model, while complex, learns biologically plausible relationships. The DNN's role in creating higher-level feature representations particularly enhanced the model's ability to capture non-linear interactions between these key clinical markers, which the final Random Forest classifier then leveraged effectively.

4. FUTURE WORK

While our hybrid model demonstrates strong performance, several avenues exist for further development. Our immediate priority is to validate the model's generalizability on larger, multi-institutional datasets to ensure robustness across diverse patient demographics.

To enhance clinical adoption, we plan to integrate Explainable AI (XAI) techniques such as SHAP plots, providing physicians with case-specific insights into the model's predictions. We are also developing a prototype web-based interface to allow real-time diabetes risk assessment in clinical settings.

Longer-term, we aim to adapt this hybrid architecture for predicting other metabolic conditions, such as hypertension and cardiovascular disease, potentially enabling a comprehensive AI-powered diagnostic system for metabolic syndrome management.

5. CONCLUSIONS

This research has conclusively demonstrated the superiority of a hybrid DNN-RF model for diabetes prediction. By leveraging a Deep Neural Network as an intelligent feature extractor for a Random Forest classifier, the proposed architecture achieves a notable accuracy of 96.4%, significantly outperforming both individual models. The performance gain underscores the synergistic effect of combining deep learning's capacity for identifying complex, non-linear patterns with the robust classification capabilities of an ensemble tree method. Furthermore, the feature importance analysis derived from the model aligns strongly with established clinical knowledge, validating its reasoning process and reinforcing its potential for real-world clinical application. This work provides a reliable, accurate, and interpretable tool for early diabetes risk assessment, paving the way for more effective AI-powered clinical decision-support systems.

6. REFERENCES

[1]. Dua, D., & Graff, C. (2019). *UCI Machine Learning Repository*. University of California, Irvine, School of Information and Computer Sciences. http://archive.ics.uci.edu/ml

[2]. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

[3]. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357.

[4]. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (pp. 4765–4774).

[5]. Abdollahi, J., & Nouri-Moghaddam, B. (2022). A stacked ensemble model for breast cancer diagnosis. *Computers in Biology and Medicine, 151*, 106299. https://doi.org/10.1016/j.compbiomed.2022.106299

[6]. Swapna, G., Vinayakumar, R., & Soman, K. P. (2021). Diabetes detection using deep learning algorithms. *ICT Express, 7*(4), 457–461. https://doi.org/10.1016/j.icte.2021.02.004