Jurnal Teknik Indonesia

Volume 5 Nomor 3 July 2026
https://jti.publicascientificsolution.com/index.php/rp

Explainable Ensemble XGBoost Model for Diabetic Peripheral Neuropathy Risk Prediction with Android Deployment

Muhammad Vito Riano
Universitas Pembangunan Nasional Veteran Jakarta, Indonesia
Email: 2110511071@mahasiswa.upnvj.ac.id

Keywords:
XGBoost; Diabetic Peripheral Neuropathy; Machine Learning; Android; Risk Prediction

Abstract

Diabetic Peripheral Neuropathy (DPN) is a common chronic complication of diabetes mellitus that can reduce quality of life and increase the risk of ulcers and amputation if not detected early. Limited access to clinical screening tools highlights the need for accessible preventive risk assessment systems, while existing machine learning approaches have primarily focused on improving predictive performance with limited attention to model interpretability and practical deployment in real-world clinical settings. This study proposes an explainable weighted ensemble XGBoost model integrated into an Android-based mobile application for early DPN risk prediction using the Diabetes 130-US Hospitals dataset. The proposed framework employs Hyperopt to optimize three complementary XGBoost models, whose predictions are combined through validation-based weighted averaging according to their R^2 performance. The proposed model achieved strong predictive performance, with an RMSE of 0.0189, MAE of 0.0122, MAPE of 7.5234%, and R^2 of 0.7269. SHAP analysis identified key risk factors, including insulin usage, Body Mass Index (BMI), and HbA1c levels, highlighting the importance of metabolic control in DPN progression. The trained model was deployed in an Android application using ONNX Runtime to enable offline prediction. User Acceptance Testing demonstrated a satisfaction rate of 90.56%, indicating good usability. The results demonstrate that the proposed system provides an interpretable and practical decision-support tool for early DPN risk assessment while improving the accessibility of mobile-based screening solutions.

INTRODUCTION

Diabetes mellitus is one of the fastest-growing chronic diseases worldwide, with the number of people living with diabetes currently exceeding 537 million (International Diabetes Federation, 2021). Diabetes is associated with multiple long-term complications, one of the most common being Diabetic Peripheral Neuropathy (DPN), a disorder affecting peripheral nerves that may cause numbness, pain, sensory loss, foot ulcers, and reduced quality of life. Approximately 50% of individuals with diabetes experience diabetic neuropathy complications, and the average prevalence of DPN among adults with diabetes is 40.3% (Feldman et al., 2019). Many patients with DPN may remain asymptomatic during the early stages, and without timely recognition and treatment, they are at risk of sustaining injuries without awareness (Pop-Busui et al.,

2016). Early detection of DPN remains challenging because diagnosis relies on physical examinations of the peripheral nervous system, electromyography (EMG), ankle-brachial index (ABI) assessments, and other diagnostic tests (Firnhaber & Powell, 2019). These approaches require trained endocrinologists and specialized diagnostic equipment, which are often limited in resource-constrained regions (Christ et al., 2022; Kumar et al., 2025).

Epidemiological research has provided important insights into the prevalence and underlying mechanisms of DPN (Hansen et al., 2023; Tang et al., 2024; Zhu et al., 2024). A study involving 1,200 patients from Norway, the Netherlands, and the United States identified diabetes as a major cause of peripheral neuropathy (Lehmann et al., 2020). Supporting these findings, Pop-Busui et al. (2016) emphasized that approximately half of all patients with diabetes experience this neurological complication, highlighting its importance in diabetes healthcare management. The neurological impact of DPN is substantial, primarily affecting sensory and motor nerves in the lower and upper extremities, with symptoms developing gradually and resulting in diverse clinical manifestations. The progression of DPN is associated with multiple risk factors. Research conducted by the TODAY Study Group (2021) identified three key determinants associated with neuropathy development: patient age, diabetes duration, and glycemic control as measured by HbA1c levels. These findings demonstrate the relationship between metabolic dysfunction and neurological complications, emphasizing the importance of comprehensive diabetes management in reducing neuropathy risk.

Recent advances in machine learning have created opportunities for disease risk prediction using routinely collected clinical data. Lian et al. (2023) demonstrated that machine learning models trained on clinical risk factors could effectively predict DPN, while Liu et al. (2024) improved model interpretability by incorporating SHAP analysis to identify influential predictors. Previous research by Kazemi et al. (2016) also reported that a support vector machine (SVM)-based model achieved approximately 76% accuracy in predicting DPN severity using patient characteristics and neuropathy disability score questionnaires. Among various machine learning algorithms, XGBoost has demonstrated strong performance on structured healthcare datasets due to its ability to capture nonlinear relationships and complex feature interactions efficiently (Chen & Guestrin, 2016). Despite these advances, previous studies have primarily focused either on improving predictive performance or enhancing model interpretability, with limited attention given to integrating explainable machine learning models into accessible mobile applications for practical early DPN risk screening. Therefore, this study proposes an explainable weighted ensemble XGBoost model for DPN risk prediction using the Diabetes 130-US Hospitals dataset. Furthermore, the trained model was deployed in an Android application to provide accessible offline risk estimation.

METHOD

Data Preparation

The dataset used in this study was the Diabetes 130-US Hospitals for Years 1999–2008 dataset. The data were obtained from the Health Facts database (Cerner Corporation, Kansas City, MO), which contains clinical records collected from hospitals across the United States. The database was part of the Health Facts program provided to healthcare organizations using the Cerner Electronic Medical Record System.

The dataset represented 10 years (1999–2008) of clinical care from 130 hospitals and integrated delivery networks across four regions of the United States: Midwest (18 hospitals), Northeast (58 hospitals), South (28 hospitals), and West (16 hospitals). Each record contained patient information, including diabetes diagnosis, laboratory test results, medication administration, and hospitalization data for stays of up to 14 days. The initial database consisted of 41 tables within a fact-dimension schema and included 117 features. Overall, the database contained 74,036,643 unique encounters, representing 17,880,231 unique patients and 2,889,571 healthcare providers.

The dataset contains information extracted from the database for encounter records that meet the following criteria:

a. They are inpatient encounters (patient admissions to a hospital).
b. They are "diabetes" encounters, meaning any type of diabetes is entered into the system as a diagnosis.
c. The length of stay is at least one day and at most fourteen days.
d. Laboratory tests are performed during the encounter.
e. Medications are administered during the encounter.

The dataset extracted from the database includes 101,766 rows and 50 attribute columns consisting of administration data, demographic data, weight, time in hospital, number of hospital visits, hospital payment information, numbers of lab procedures taken, lab test results for blood sugar (HbA1c) and serum glucose, diagnosis of complications, diabetes medications taken, etc.

Feature Selection

Following the initial analysis of the dataset, feature selection will be conducted in this study. Features to assess the risk level of DPN will be selected based on previous research done on this topic. According to a comparative study on research related to DPN using ML, demographic variables (age, gender, race), laboratory test results (HbA1c, serum glucose), and clinical records (diagnosis, treatment) have shown good results in the DPN prediction model (Wu et al., 2024). Therefore, the features selected from the initial dataset are focused on features that have been strong predictors in previous DPN using ML studies. Features removed from the initial dataset were those deemed to have low impact and correlation in DPN risk factor analysis based on previous studies (Kazemi et al., 2016; Lian et al., 2023; Liu et al., 2024; Wu et al., 2024), alongside features with excessive missing values. These included hospital administrative data (encounter id, patient number, admission type, discharge disposition, admission source, time in hospital, payer code, medical specialty),

medications not directly associated with DPN patients (repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, citoglipton, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone, metformin-rosiglitazone, metformin-pioglitazone), and the Weight feature which has more than 95% missing values.

A total of 20 features were used in the final dataset. Including demographic variables (race, gender, age), patient diagnosis ICD9 codes, number of diagnoses, glucose serum test results, HbA1c test results, number of lab procedures taken, number of medications taken, number of inpatient, outpatient and emergency visits, changes in medications, and readmitted status.

Data Preprocessing

Data preprocessing is a crucial step that prepares raw data to be suitable for analysis and modeling. In this study, the preprocessing stage involves cleaning, transforming, and organizing the raw data obtained from the Diabetes 130-US Hospitals dataset. This stage aims to refine the initial dataset prior to model training. Missing values were handled by deletion, or by using statistical imputation methods. Depending on the data distribution, missing values were replaced using the median, mean, or mode to preserve the underlying characteristics of the dataset.

The next step involves transforming categorical variables into numerical representations to enable processing by the XGBoost model. All features with object data types were first converted into categorical types before encoding. For nominal features such as race and gender, One-Hot Encoding or Label Encoding was applied. For ordinal features, such as A1C result, Ordinal Encoding was used to preserve the inherent order based on severity levels. After encoding, all numerical features were standardized using StandardScaler to achieve distribution with zero mean and unit variance.

Feature Engineering

The feature engineering process aims to enhance the representational quality of the dataset by transforming existing variables and constructing new features that better capture underlying patterns related to Diabetic Peripheral Neuropathy (DPN) risk. This step is essential for improving model performance, as it enables the learning algorithm to identify more informative and clinically relevant relationships within the data. In this study, feature engineering consists of two main components: feature construction and target variable definition.

First, several new features are generated based on domain knowledge and relationships among existing variables. Anthropometric features such as height and weight are estimated using patient demographic information and derived metrics. Body Mass Index (BMI) is calculated by integrating height with multiple clinical indicators, including glycemic measurements and treatment-related variables, to better reflect the patient's physiological condition. In addition, composite clinical indicators are constructed to capture more complex health conditions. A Diagnosis Score is computed

from ICD-9 diagnosis codes, emphasizing conditions related to diabetes and neuropathy. Furthermore, a Risk Score is formulated using key clinical variables such as HbA1c level, serum glucose, medication usage (e.g., insulin and metformin), and hospital visit frequency. These composite features aim to summarize the patient's overall clinical profile and improve the model's ability to detect patterns associated with DPN risk.

Second, the target variable, denoted as DPN_Risk, is defined as a continuous score ranging from 0.1 to 1.0, representing the estimated likelihood of Diabetic Peripheral Neuropathy. Since the dataset does not explicitly provide a direct DPN diagnosis label, the target variable is engineered by integrating diagnosis severity, clinical risk factors, age adjustment, and diagnostic confidence. The Diagnosis Score and Risk Score are first normalized as:

$$D_n = \min\left(\frac{D}{15}, 1\right) \tag{1}$$

$$R_n = \min\left(\frac{R}{25}, 1\right) \tag{2}$$

where D and R denote the Diagnosis Score and Risk Score, respectively. The preliminary DPN risk is then calculated as:

$$Risk = 0.1 + A + 0.4D_n + 0.4R_n \tag{3}$$

where A represents the age adjustment factor. To capture the interaction between diagnosis severity and clinical risk factors, the preliminary risk is further adjusted by:

$$Risk^* = Risk(1 + 0.10D_nR_n) \tag{4}$$

The confidence factor is defined as:

$$C = \min\left(1, \frac{\text{Number of Diagnoses}}{5}\right) \tag{5}$$

and the final target variable is obtained using:

$$DPN\_Risk = C \cdot Risk^* + (1 - C)(0.1 + A) \tag{6}$$

where the resulting value is constrained to the interval $[0.1, 1.0]$. The interaction term enables the model to capture nonlinear relationships between diagnosis severity and clinical risk factors, while the confidence factor reduces the influence of incomplete patient records by assigning lower weight to observations with fewer recorded diagnoses. Consequently, the engineered target variable provides a continuous representation of DPN risk that is suitable for supervised regression.

Overall, this feature engineering pipeline enhances the dataset by constructing clinically meaningful composite variables and defining a continuous target variable that incorporates both clinical knowledge and patient-specific characteristics. These engineered features enable the proposed ensemble XGBoost model to learn more informative, robust, and interpretable patterns for DPN risk prediction.

Data Sampling

Prior to model training, the dataset is divided into training and testing sets using a stratified sampling approach. This method is employed to preserve the distribution of the target variable (Tanha et al., 2020), namely the DPN risk score (DPN_Risk), across all data splits. Since the target variable is continuous and exhibits a wide range of values, a binning strategy is applied to enable stratification. The continuous DPN_Risk values are grouped into several bins, where each bin represents a specific range of risk levels. Values that occur infrequently are treated as rare cases and are assigned to the nearest bin to ensure sufficient representation during the sampling process. This approach helps maintain a balanced distribution of the target variable and ensures that each subset reflects the overall data distribution.

After defining the stratification bins, the dataset is iteratively split into multiple subsets using a stratified train-test splitting procedure. In this study, ten stratified splits are generated as part of a cross-validation framework to support robust model evaluation. In cases where certain bins contain insufficient samples for proper stratification, a random sampling strategy is applied as a fallback to ensure all data points are included. This sampling strategy improves the reliability of model evaluation by reducing bias caused by imbalanced target distributions and ensuring that both training and testing sets are representative of the original dataset.

Modeling

This phase focuses on developing an XGBoost regression model to analyze and assess Diabetic Peripheral Neuropathy (DPN) risk factor scores. The proposed approach employs a multi-stage training strategy that integrates hyperparameter optimization, cross-validation, and ensemble learning to improve predictive performance. The training process is conducted on multiple stratified data samples generated from the original training dataset to ensure balanced representation of the target variable. For each data sample, the model development consists of three main stages.

In the first stage, hyperparameter optimization is performed using the Tree-structured Parzen Estimator (TPE) algorithm implemented in the Hyperopt library. Three distinct hyperparameter search spaces are defined, each designed with different objectives: (1) precision-focused configuration, (2) complex interaction modelling, and (3) generalization-oriented configuration. This strategy enables the exploration of diverse model behaviors and reduces the risk of suboptimal parameter selection. In the second stage, the optimized hyperparameters are used to train three separate XGBoost models. Model performance is evaluated using ten-fold cross-validation, where each fold produces validation predictions. This process ensures that the trained models are capable of generalizing across different data partitions. In the third stage, an ensemble

learning approach is applied by combining the predictions of the three optimized XGBoost models. The contribution of each model in the ensemble is determined using a weighted scheme based on the coefficient of determination (R^2) obtained from validation results. Models with higher predictive performance are assigned greater weights, while models with lower performance contribute less to the final prediction. If all models perform poorly, uniform weighting is applied as a fallback mechanism.

The final prediction for each fold is obtained through weighted aggregation of individual model outputs, and overall performance is evaluated using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and R^2 score. After cross-validation, the best performing model configuration is selected based on the lowest RMSE and evaluated on the hold-out test dataset. Additionally, model interpretability is incorporated using SHAP (SHapley Additive exPlanations) to analyze feature importance and understand the contribution of each feature to the predicted DPN risk scores (Lundberg et al., 2020). Error analysis is also performed to identify patterns in prediction errors and assess model reliability.

System Deployment

To enhance the practical applicability of the proposed model, the trained XGBoost DPN risk prediction model is deployed into a mobile application developed using Android Studio with Java. The application is designed to provide accessible and real-time risk assessment for users without requiring continuous internet connectivity. The deployed model is converted into ONNX (Open Neural Network Exchange) format to enable efficient inference within the mobile environment. This conversion allows the model to be executed using ONNX Runtime, ensuring compatibility, lightweight performance, and reduced computational overhead on mobile devices. The application operates in an offline mode, enabling users to perform risk analysis without reliance on external servers. This design choice will improve usability in environments with limited or unstable internet access and enhances data privacy by ensuring that user data is not stored anywhere.

From a user interaction perspective, the application is designed to accept clinical input data, including demographic information and relevant health indicators. Users are required to input the necessary data through a structured interface, after which the system processes the data using the embedded prediction model. Upon completion of data entry, the application generates a DPN risk score, representing the estimated likelihood of Diabetic Peripheral Neuropathy. This result is presented directly to the user, providing an intuitive and immediate assessment that can support early awareness and preventive decision-making.

RESULTS AND DISCUSSION

Data Preprocessing and Feature Engineering

The dataset used in this study was derived from the Diabetes 130-US Hospitals dataset and consisted of 101,766 patient records with 20 features after preprocessing. The preprocessing stage involved cleaning and handling of missing values by deletion,

removing features that are not correlated with DPN risk analysis and features with excessive missing values. Categorical variables were transformed into numerical representations using appropriate encoding techniques, while numerical features were standardized to ensure uniform scale across all variables. These steps were essential to improve model stability and prevent bias caused by differing feature magnitudes.

Feature engineering was conducted to enhance the representational quality of the dataset by constructing clinically meaningful variables derived from existing attributes. In this study, several key features were generated, including Height, Weight, Body Mass Index (BMI), Diagnosis Score, Risk Score, and the target variable DPN_Risk. Height and Weight were estimated to enable the calculation of BMI, which serves as an important indicator of obesity and is widely associated with diabetes-related complications. BMI was computed to capture the relationship between body composition and metabolic health. A Diagnosis Score was constructed based on ICD-9 diagnosis codes (diag_1, diag_2, diag_3), where specific conditions related to diabetes and neuropathy were assigned weighted contributions. This feature aims to summarize the overall clinical severity of the patient's diagnosed conditions. In addition, a Risk Score was developed using a combination of clinical indicators, including HbA1c levels, blood glucose measurements, medication usage (such as insulin and metformin), and hospital visit frequencies. This score reflects the overall health condition and potential complication risk of the patient. Finally, the target variable DPN_Risk was formulated as a continuous score ranging from 0 to 1, derived from a combination of Diagnosis Score, Risk Score, BMI, age, and healthcare utilization variables. This formulation enables the modeling of DPN risk as a regression problem, allowing the model to capture varying levels of risk rather than performing binary classification.

The distribution of the target variable (DPN_Risk) exhibited moderate imbalance, which was addressed using a stratified sampling approach. Continuous target values were grouped into bins to preserve the underlying distribution during sampling, ensuring that both low-risk and high-risk cases were adequately represented in the training and evaluation process.

Model Performance Evaluation

The performance of the proposed ensemble XGBoost model was evaluated using 10-fold cross-validation across multiple stratified samples. The results demonstrate consistent performance, with an average RMSE of 0.0181 ± 0.0005 (95% CI: 0.0180–0.0182), MAE of 0.0117 ± 0.0003 (95% CI: 0.01164–0.01176), MAPE of 7.1760% ± 0.1490% (95% CI: 7.146–7.206%), and R^2 of 0.7529 ± 0.0127 (95% CI: 0.7504–0.7554). The relatively narrow confidence intervals indicate stable predictive performance across repeated stratified cross-validation folds. On the hold-out test dataset, the model achieved an RMSE of 0.0189, MAE of 0.0122, MAPE of 7.5234%, and R^2 of 0.7269, indicating strong generalization capability. The comparison between cross-validation training and test results are presented in Table 1. The most influential features, identified using SHAP analysis, are presented in Figure 1. and Figure 2.

Tabel 1. Performance Comparison between Cross-Validation Training and Test Data

	Cross Validation Data Training				Data Testing			
	RMSE	R^2	MAE	MAPE	RMSE	R^2	MAE	MAPE
Mean	0.0181	0.7529	0.0117	7.1760	0.0189	0.7269	0.0122	7.5234
Std	0.0005	0.0127	0.0003	0.1490	0.0001	0.0020	0.0001	0.0617
Min	0.0173	0.7321	0.0112	6.9325	0.0188	0.7234	0.0121	7.4169
Max	0.0189	0.7754	0.0121	7.3937	0.0190	0.7302	0.0123	7.6209

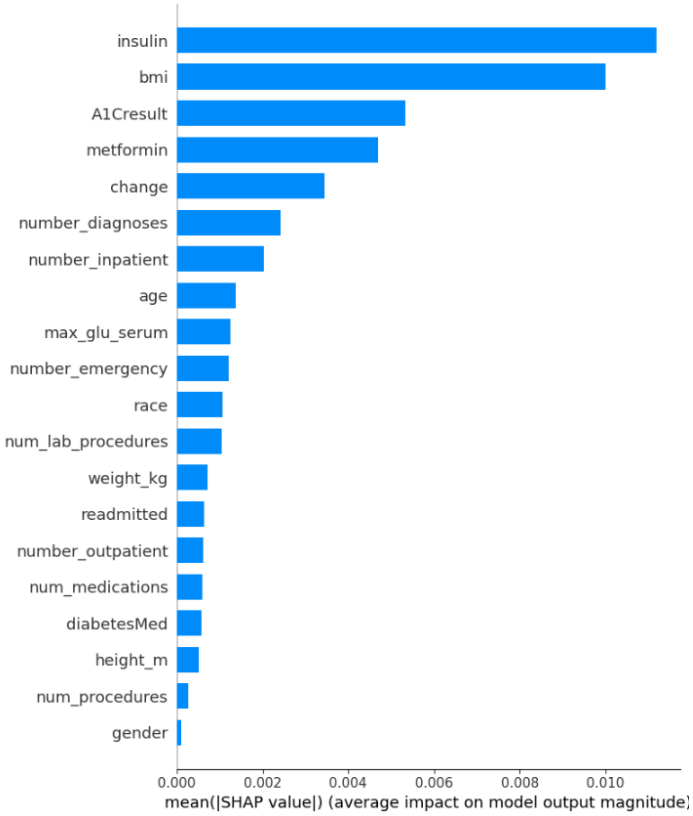

Figure 1. SHAP-based feature importance showing the average impact of each feature on the predicted DPN risk score on best model

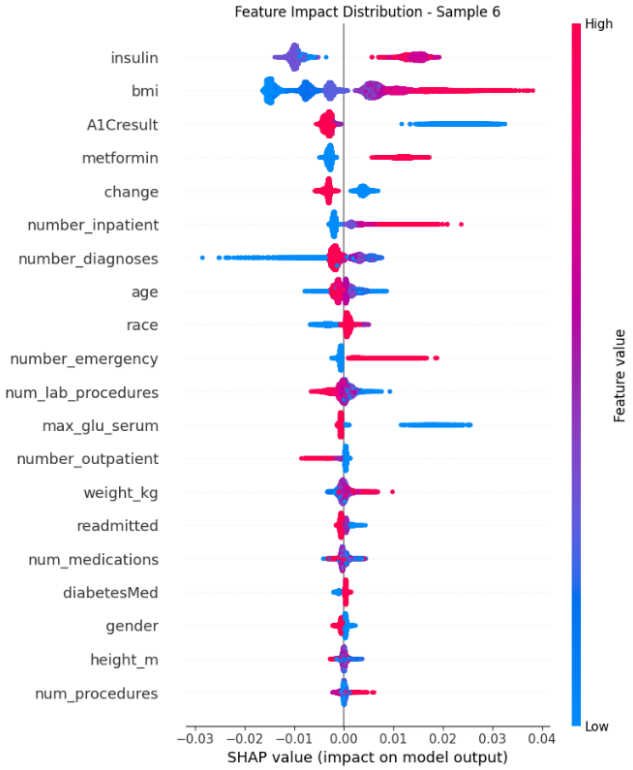

Figure 2. Features impact on model output using SHAP value on best model

Error Analysis and Model Behavior

Error analysis indicates that the majority of prediction errors are centered near zero, as shown in Figure 3. The error distribution shows a strong central tendency with a slight right-skew, suggesting that most predictions are close to the true values with a small number of larger positive deviations. This is further supported by the residual plot, where errors are generally evenly dispersed around zero without a clear systematic pattern, indicating that the model does not suffer from significant bias across the prediction range.

The relationship between actual and predicted values demonstrates a strong linear alignment, although slight underestimation is observed at higher DPN risk values, as indicated by the dispersion of points below the ideal diagonal line. This suggests that the model performs well in lower to moderate risk ranges but becomes less precise for higher-risk cases.

Quantitatively, the model achieved 80.94% of predictions within 10% of the actual values and 95.94% within 20%, indicating a high level of prediction accuracy and reliability. Further analysis of feature-error correlations reveals that certain variables, such as age, A1C result, and diabetes medication status, exhibit stronger relationships with prediction errors. These findings suggest that variability in these features may contribute to increased prediction uncertainty, highlighting potential areas for future model refinement and feature optimization.

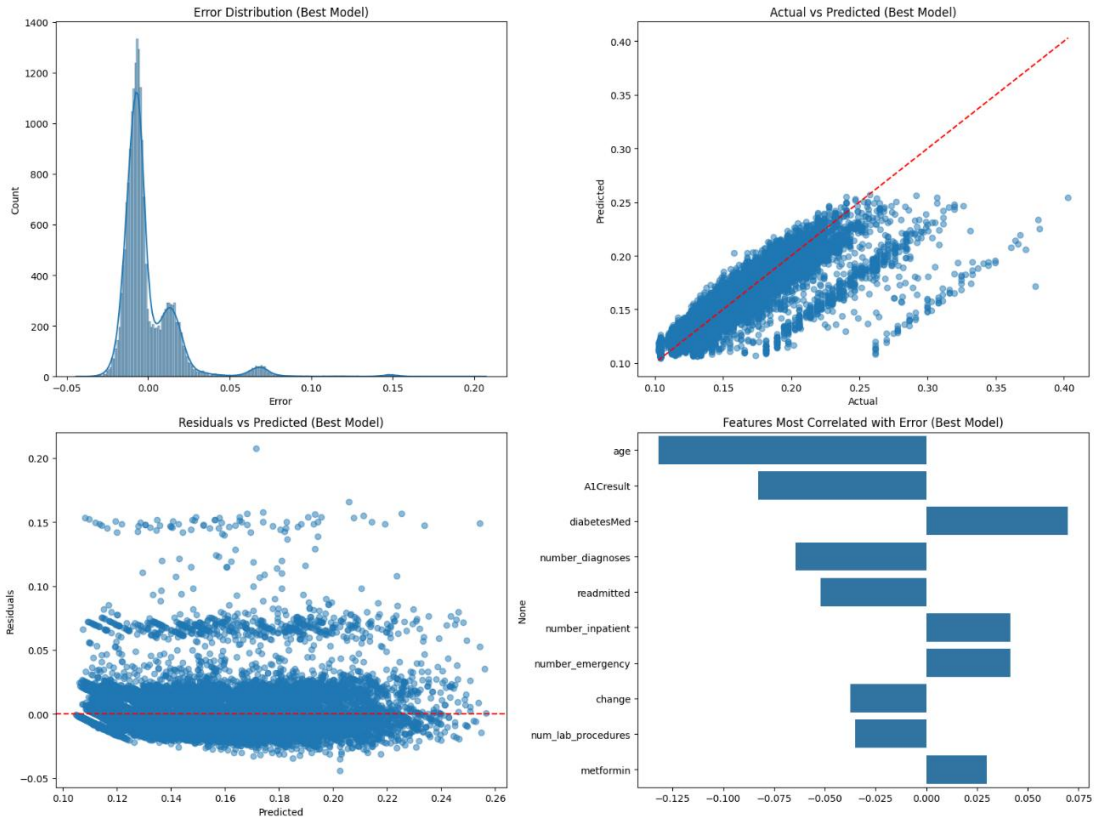

Figure 3. Error distribution, prediction performance, residual analysis, and feature-error correlation for the best-performing model

Model Deployment on Android Application

The trained model was successfully deployed into an Android-based application, enabling offline DPN risk prediction using user-provided clinical data. The application processes input data and generates a risk score in real time, demonstrating the practical applicability of the proposed system. Figure 4. displays the input form, where users are required to enter relevant information such as demographic characteristics (age, gender), physiological indicators (height, weight, BMI), medical history (diagnoses, medications), and laboratory results (e.g., HbA1c and blood glucose levels). The input fields are organized into multiple sections to improve usability and ensure that all required features for the prediction model are properly captured. Figure 5. shows the output screens of the application, presenting the predicted DPN risk levels categorized into Low Risk, Medium Risk, and High Risk. Each result includes a corresponding risk score, key contributing factors, reported symptoms, and personalized recommendations.

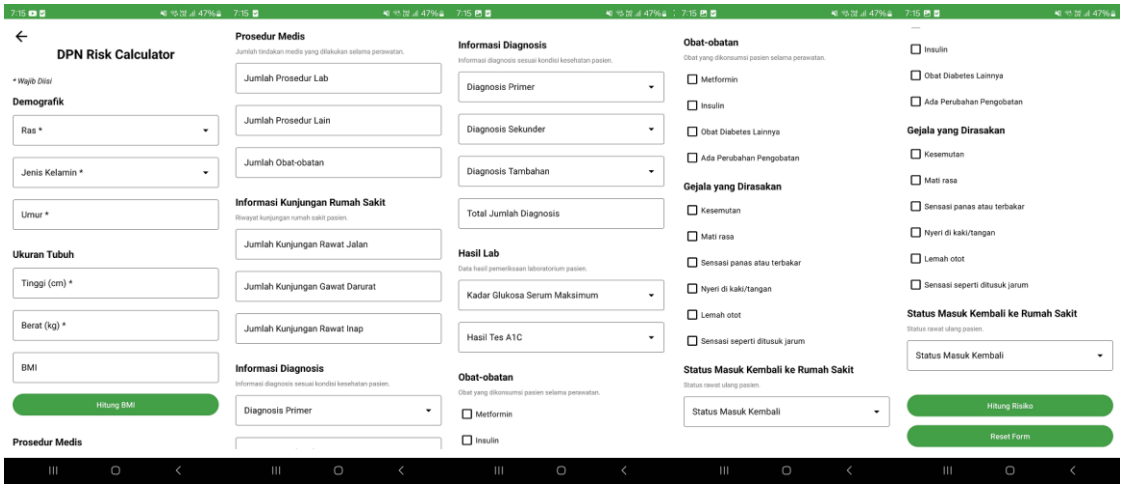

Figure 4. DPN Risk Score Calculator App Input Screen

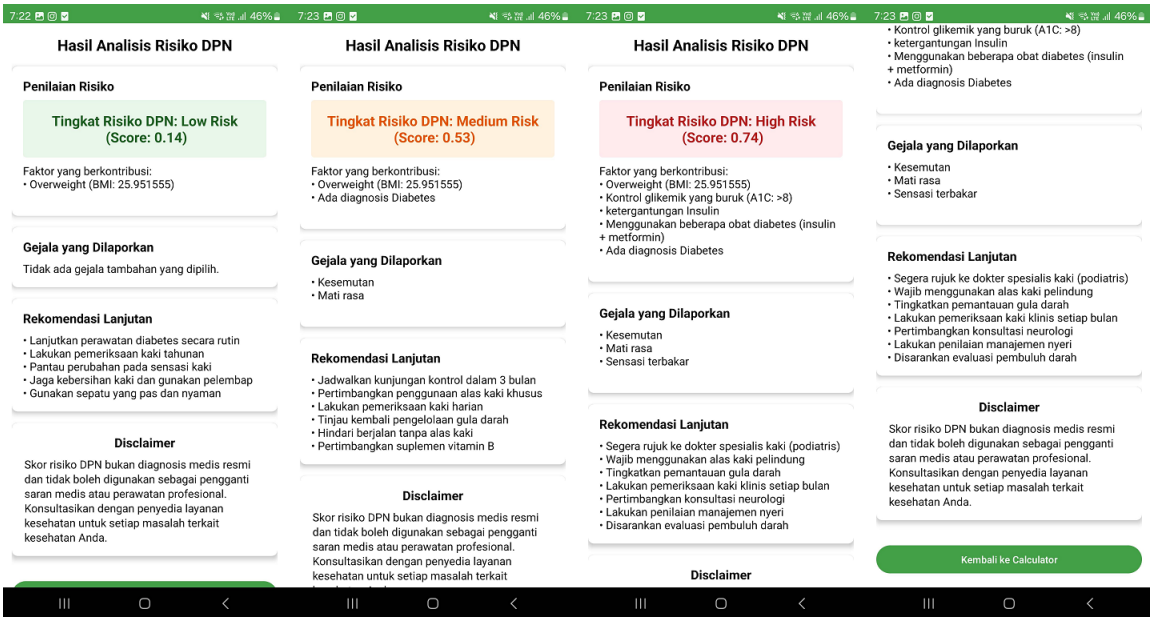

Figure 5. DPN Risk Score Analysis Results Screen

The results of this study demonstrate that the proposed ensemble XGBoost model is capable of effectively predicting Diabetic Peripheral Neuropathy (DPN) risk with strong and consistent performance. The model achieved stable evaluation metrics across multiple stratified samples, indicating good generalization ability and robustness. The relatively low variation in cross-validation results further suggests that the proposed sampling and ensemble strategy successfully mitigates issues related to data imbalance and variability.

Previous studies have explored various approaches for predicting Diabetic Peripheral Neuropathy (DPN) using machine learning techniques. Wu et al. (2024) developed multiple predictive nomographs and selected the optimal model; however, the use of the Toronto Clinical Neuropathy Scoring System may have introduced potential false positives within the diagnosis groups. Similarly, Kazemi et al. (2016)

proposed a DPN prediction model based on a multiclass Support Vector Machine (SVM). While their model demonstrated promising results, the absence of comparative evaluation with alternative methods may have limited the overall optimization of model performance. Lian et al. (2023) developed a DPN prediction model using multiple machine learning algorithms, including logistic regression, k-nearest neighbor, decision tree, naïve Bayes, random forest, and XGBoost, based on clinical and biochemical data from diabetic patients. Their results showed that XGBoost outperformed other models in terms of accuracy and AUC, demonstrating its effectiveness in capturing complex relationships within structured healthcare data. Furthermore, SHAP analysis identified key risk factors such as age, disease duration, and glycemic indicators, reinforcing the importance of metabolic control in DPN progression.

These findings suggest that the development of an effective DPN prediction model requires not only appropriate algorithm selection but also well-structured, sufficiently large, and balanced datasets. In line with previous research, established risk factors such as age, glycemic control indicators (e.g., HbA1c), and clinical measurements remain significant predictors of DPN (Kazemi et al., 2016; Lian et al., 2023; Liu et al., 2024; Wu et al., 2024).

Consistent with previous studies, the proposed model confirms the importance of metabolic and clinical variables for DPN risk prediction. However, unlike Kazemi et al. (2016), which evaluated only a multiclass SVM model, the proposed framework employs a weighted ensemble of multiple XGBoost models optimized through Hyperopt, enabling the model to capture more complex nonlinear relationships within structured clinical data. Compared with Lian et al. (2023), which demonstrated the superiority of XGBoost over several conventional machine learning algorithms, this study further extends the approach by incorporating ensemble learning, performance-based model weighting, and SHAP-based explainability. In contrast to Wu et al. (2024), whose work primarily focused on interpretable prediction models, the present study additionally demonstrates practical deployment through an Android-based application for offline DPN risk assessment. The integration of ensemble learning, SHAP-based explainability, and Android deployment extends previous DPN prediction studies beyond algorithm development by providing an interpretable decision-support system that can be used for practical offline risk assessment.

From a feature importance perspective, the SHAP analysis provides valuable insights into the factors influencing DPN risk prediction. The results indicate that insulin usage and Body Mass Index (BMI) are among the most influential features in the model. This highlights the importance of metabolic control and obesity-related factors in the development of diabetic complications. Additionally, clinical indicators such as HbA1c levels and blood glucose measurements contribute significantly to the model's predictions, which is consistent with the well-established understanding that prolonged hyperglycemia plays a major role in nerve damage among diabetic patients (Vincent et al., 2010). Furthermore, demographic and healthcare utilization features, including age, number of inpatient visits, and number of emergency visits, also show

notable influence. These variables may act as proxies for disease progression and overall patient condition, suggesting that patients with more frequent hospital interactions are likely to have more severe or advanced complications. This aligns with previous findings in the literature, where disease duration and healthcare utilization are often associated with increased risk of diabetic neuropathy (Tesfaye et al., 2010; TODAY Study Group, 2021).

The ensemble modeling approach also plays a key role in achieving strong predictive performance. By combining multiple XGBoost models with different hyperparameter configurations, the system is able to capture diverse patterns within the data. The use of R^2-based weighting further enhances this approach by assigning higher importance to better-performing models, resulting in more accurate and stable predictions compared to a single-model approach.

Despite the overall strong performance, the error analysis reveals several important characteristics of the model. The distribution of prediction errors is centered near zero, indicating that most predictions are close to the actual values. However, a slight tendency to underestimate higher DPN risk values is observed. This suggests that while the model performs well for low to moderate risk cases, it may have limited sensitivity in identifying extreme or high-risk conditions. This limitation is likely influenced by the distribution of the dataset, where high-risk cases may be less represented. In addition, the correlation analysis between input features and prediction errors shows that certain variables, such as age, HbA1c levels, and medication-related features, have a stronger association with prediction uncertainty. This indicates that variability in these features may introduce complexity that is not fully captured by the current model, highlighting potential areas for further improvement.

The practical applicability of the proposed model is demonstrated through its deployment in an Android-based application. The system enables users to input clinical data and obtain DPN risk predictions in real time, without requiring internet connectivity. This offline capability is particularly valuable in resource-limited settings, where access to specialized diagnostic tools and healthcare professionals may be limited. By providing an accessible risk estimation tool, the application has the potential to support early screening and increase awareness of diabetic complications.

From a clinical perspective, the proposed application is intended to support early risk assessment rather than replace clinical diagnosis. The system may assist healthcare providers by identifying individuals who could benefit from further neurological examination, particularly in primary healthcare facilities or resource-limited settings where access to specialized diagnostic equipment and endocrinologists is limited. Consequently, the application has the potential to facilitate earlier referral and preventive intervention while complementing existing clinical decision-making processes.

However, this study has several limitations that should be considered. First, the model is developed using a secondary dataset, which may not fully represent the diversity of real-world patient populations. Second, the DPN risk score is constructed

based on proxy indicators rather than direct clinical diagnosis, which may introduce some degree of uncertainty in the ground truth labels. Finally, while the ensemble approach improves performance, it also increases model complexity, which may impact interpretability and computational efficiency.

CONCLUSION

This study presents an ensemble XGBoost-based approach for predicting Diabetic Peripheral Neuropathy (DPN) risk using structured clinical data. The proposed method integrates feature engineering, hyperparameter optimization, stratified sampling, and ensemble learning to improve predictive performance and robustness. The model achieved an RMSE of 0.0181 on the training data and 0.0189 on the testing data, with R^2 values of 0.7529 and 0.7269, respectively. The MAE and MAPE values were 0.0117 and 7.1760 for the training data and 0.0122 and 7.5234 for the testing data, respectively.

The results demonstrated that the model achieved stable and reliable performance across multiple evaluation metrics, indicating strong generalization capability. The ensemble strategy with performance-based weighting further improved prediction accuracy compared with single-model approaches. In addition, the integration of SHAP-based interpretability provided meaningful insights into key risk factors associated with DPN, while deployment of the model in an Android-based application demonstrated its practical applicability for real-time, offline risk assessment.

Despite these promising results, several limitations remained regarding dataset representation and the use of proxy-based target variables. Future research should focus on validating the model using real-world clinical datasets and improving predictive performance for high-risk cases. Theoretically, this study contributed to the application of explainable ensemble machine learning for continuous DPN risk prediction by demonstrating that predictive performance and model interpretability could be effectively integrated within a unified framework. Methodologically, this study introduced a framework that combined feature engineering, Hyperopt-based hyperparameter optimization, weighted ensemble XGBoost regression, SHAP-based explainability, and an end-to-end explainable clinical decision-support framework, bridging predictive modeling with real-world mobile deployment.

REFERENCES

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Christ, J. P., Vu, M., Mehr, H., Jackson-Bey, T., & Herndon, C. N. (2022). See one, do one, teach one: reimagining reproductive endocrinology and infertility training programs to expand access to care. *F&S Reports*, *3*(2), 114–121.

Feldman, E. L., Callaghan, B. C., Pop-Busui, R., Zochodne, D. W., Wright, D. E., Bennett, D. L., Bril, V., Russell, J. W., & Viswanathan, V. (2019). Diabetic neuropathy. *Nature Reviews Disease Primers,* *5*(1), 1–18. https://doi.org/10.1038/s41572-019-0092-1

Firnhaber, J. M., & Powell, C. S. (2019). Lower extremity peripheral artery disease: Diagnosis and treatment. *American Family Physician, 99*(6), 362–369.

Hansen, C. S., Määttä, L. L., Andersen, S. T., & Charles, M. H. (2023). The epidemiology of diabetic neuropathy. In *Diabetic Neuropathy: Advances in Pathophysiology and Clinical Management* (pp. 5–36). Springer.

International Diabetes Federation. (2021). *IDF diabetes atlas* (10th ed.). https://www.diabetesatlas.org

Kazemi, M., Moghimbeigi, A., Kiani, J., Mahjub, H., & Faradmal, J. (2016). Diabetic peripheral neuropathy class prediction by multicategory support vector machine model: A cross-sectional study. *Epidemiology and Health, 38*, e2016011. https://doi.org/10.4178/epih.e2016011

Kumar, P., Lionis, C., Andoko, D., Rahman, Z., Anastasaki, M., & Awankem, B. (2025). Evaluation of diagnostic services in rural and remote areas: bottlenecks, success stories, and solutions. *Journal of Surgical Specialties and Rural Practice*, *6*(1), 32–37.

Lehmann, H. C., Wunderlich, G., Fink, G. R., & Sommer, C. (2020). Diagnosis of peripheral neuropathy. *Neurological Research and Practice, 2*(1), 20. https://doi.org/10.1186/s42466-020-00064-2

Lian, X., Qi, J., Yuan, M., Li, X., Wang, M., Li, G., Yang, T., & Zhong, J. (2023). Study on risk factors of diabetic peripheral neuropathy and establishment of a prediction model by machine learning. *BMC Medical Informatics and Decision Making, 23*(1), 146. https://doi.org/10.1186/s12911-023-02232-1

Liu, L., Bi, B., Cao, L., Gui, M., & Ju, F. (2024). Predictive model and risk analysis for peripheral vascular disease in type 2 diabetes mellitus patients using machine learning and shapley additive explanation. *Frontiers in Endocrinology, 15*. https://doi.org/10.3389/fendo.2024.1320335

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Pop-Busui, R., Boulton, A. J. M., Feldman, E. L., Bril, V., Freeman, R., Malik, R. A., Sosenko, J. M., & Ziegler, D. (2016). Diabetic neuropathy: A position statement by the American Diabetes Association. *Diabetes Care, 40*(1), 136–154. https://doi.org/10.2337/dc16-2042

Tang, F., Zhao, F., Jiang, Y., Zhang, T., & Wang, B. (2024). Global hotspots and trends in diabetic peripheral neuropathy research from 2011 to 2023. *Medicine*, *103*(32), e39295.

Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N., & Asadpour, M. (2020). Boosting methods for multi-class imbalanced data classification: An experimental review. *Journal of Big Data, 7*(1), 70. https://doi.org/10.1186/s40537-020-00349-y

Tesfaye, S., Boulton, A. J. M., Dyck, P. J., Freeman, R., Horowitz, M., Kempler, P., Lauria, G., Malik, R. A., Spallone, V., Vinik, A., Bernardi, L., Valensi, P., & on behalf of the Toronto Diabetic Neuropathy Expert Group. (2010). Diabetic neuropathies: Update on definitions, diagnostic criteria, estimation of severity, and treatments. *Diabetes Care, 33*(10), 2285–2293. https://doi.org/10.2337/dc10-1303

TODAY Study Group. (2021). Risk factors for diabetic peripheral neuropathy in adolescents and young adults with type 2 diabetes: Results from the TODAY study. *Diabetes Care, 45*(5), 1065–1072. https://doi.org/10.2337/dc21-1074

Vincent, A. M., Edwards, J. L., McLean, L. L., Hong, Y., Cerri, F., Lopez, I., Quattrini, A., & Feldman, E. L. (2010). Mitochondrial biogenesis and fission in axons in cell culture and animal models of diabetic neuropathy. *Acta Neuropathologica, 120*(4), 477–489. https://doi.org/10.1007/s00401-010-0697-7

Wu, Y., Dong, D., Zhu, L., Luo, Z., Liu, Y., & Xie, X. (2024). Interpretable machine learning models for detecting peripheral neuropathy and lower extremity arterial disease in diabetics: An analysis of critical shared and unique risk factors. *BMC Medical Informatics and Decision Making, 24*(1), 200. https://doi.org/10.1186/s12911-024-02595-z

Zhu, J., Hu, Z., Luo, Y., Liu, Y., Luo, W., Du, X., Luo, Z., Hu, J., & Peng, S. (2024). Diabetic peripheral neuropathy: pathogenetic mechanisms and treatment. *Frontiers in Endocrinology*, *14*, 1265372.