Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1878−1888
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.6446

A Comparative Analysis of Diabetes Prediction through Deep Learning Architectures

Gregorius Airlangga*

Engineering Faculty, Information System Study Program, Atma Jaya Catholic University of Indonesia, Jakarta, Indonesia
Email: gregorius.airlangga@atmajaya.ac.id
Correspondence Author Email: gregorius.airlangga@atmajaya.ac.id
Submitted: **13/12/2024**; Accepted: **24/12/2024**; Published**: 25/12/2024**

Abstract−Diabetes prediction plays a vital role in healthcare, enabling early diagnosis and timely interventions to mitigate the risks associated with the disease. This study investigates the application of advanced machine learning architectures to predict diabetes using the Pima Indians Diabetes Dataset, a widely used benchmark for medical diagnostics. Five models: Deep Neural Network (DNN), Convolutional Neural Network (CNN) with Attention, LSTM with Residual Connections, Bidirectional LSTM (BiLSTM) with Attention, and GRU with Dense Layers were developed and evaluated on multiple performance metrics, including accuracy, precision, recall, F1 score, and ROC AUC. A stratified five-fold cross-validation strategy was employed to ensure robustness, while SHAP analysis was conducted to enhance interpretability. Among the models, the GRU with Dense Layers achieved superior performance, recording the highest accuracy (76.17%), F1 score (69.85%), and ROC AUC (83.52%). SHAP analysis revealed Glucose as the most influential feature, with significant interactions identified between Glucose and Pregnancies, aligning with established medical insights. Statistical analysis confirmed the reliability of the results, with all metrics demonstrating statistically significant improvements over a baseline of random chance ($p < 0.05$). These findings underscore the efficacy of GRU-based models in capturing complex patterns in medical data while maintaining computational efficiency. Future work will explore hybrid architectures and larger datasets to enhance generalizability and real-world applicability, contributing to more effective decision-making in healthcare.

Keywords: Diabetes Prediction; Machine Learning; GRU with Dense Layers; SHAP Analysis; Healthcare Decision Support

1. INTRODUCTION

Diabetes mellitus is a pervasive health concern, affecting millions globally and presenting significant economic and social burdens [1][3]. Early and accurate diagnosis is critical for effective disease management and prevention of complications [4][6]. Machine learning (ML) techniques have emerged as powerful tools for addressing challenges in healthcare, including predictive analytics and disease classification [7]. Leveraging large datasets, ML models can uncover intricate patterns in medical data, offering potential improvements over traditional diagnostic methods [8]. This study explores the performance of advanced machine learning and deep learning techniques in predicting diabetes using a well-known dataset, the Pima Indians Diabetes Database, a benchmark dataset widely used in medical research [9]. The dataset, originally curated by the National Institute of Diabetes and Digestive and Kidney Diseases, provides comprehensive diagnostic measurements of Pima Indian females aged 21 years or older [10]. It includes features such as glucose concentration, body mass index (BMI), insulin levels, and pregnancy history, all of which are known to be associated with diabetes risk [11]. The target variable, Outcome, indicates whether an individual is diabetic (1) or not (0). This rich dataset has motivated researchers to develop and evaluate various machine learning models to achieve more accurate predictions. However, existing literature often emphasizes only traditional ML models, with limited exploration of modern hybrid and deep learning architectures in this context [12][14].

Recent advancements in ML, particularly deep learning, offer unprecedented capabilities in capturing nonlinear relationships and feature interactions in data [15]. Models such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, originally designed for image and sequence data, respectively, have been adapted for tabular and time-series data in medical applications [16]. Moreover, techniques like attention mechanisms and residual connections have further enhanced these models' predictive performance by addressing issues such as vanishing gradients and feature importance [17]. Nevertheless, a systematic comparison of such advanced architectures with traditional ML models remains largely unexplored for the problem of diabetes prediction[18]. The urgency of addressing diabetes-related challenges underscores the need for robust and interpretable predictive models [19]. According to the International Diabetes Federation (IDF), the global prevalence of diabetes is projected to rise sharply, particularly in low- and middle-income countries where resources for timely diagnosis and treatment are limited [20]. Automated predictive tools leveraging machine learning can potentially bridge this gap, enabling healthcare practitioners to identify at-risk individuals early and allocate resources more effectively [21]. Despite these advancements, several research gaps remain. First, while studies often evaluate individual models, few have systematically compared the performance of a diverse set of machine learning and deep learning architectures [22]. Second, there is limited exploration of how advanced preprocessing techniques, such as handling missing values and normalizing diagnostic features, impact model performance [23]. Third, existing works seldom incorporate ensemble strategies or hybrid models that combine the strengths of multiple architectures [24]. Finally, the interpretability and clinical applicability of advanced models remain under-addressed, posing challenges for integration into real-world healthcare systems [25].

This research aims to address these gaps by conducting a comprehensive comparison of multiple deep learning models. Specifically, we evaluate five distinct models: (1) a Deep Neural Network (DNN) with dropout and batch

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1878−1888
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.6446

normalization layers, (2) a CNN with attention mechanisms for capturing local feature dependencies, (3) an LSTM network with residual connections to leverage temporal relationships, (4) a Bidirectional LSTM (BiLSTM) with attention mechanisms for enhanced feature representation, and (5) a Gated Recurrent Unit (GRU) with dense layers. The selection of the five deep learning models in this study stems from their unique capabilities and established effectiveness in addressing various data complexities, particularly in medical diagnostics. A Deep Neural Network (DNN) serves as a robust baseline, leveraging dropout and batch normalization layers to mitigate overfitting and stabilize training. Convolutional Neural Networks (CNNs), known for their ability to capture spatial and local feature dependencies, are incorporated with attention mechanisms to enhance their focus on critical diagnostic features such as glucose levels and BMI. Long Short-Term Memory (LSTM) networks, designed to capture sequential and temporal patterns, are adapted with residual connections to improve gradient flow and better model interactions between features over time. Bidirectional LSTMs (BiLSTMs), augmented with attention mechanisms, further enhance feature representation by analyzing patterns in both forward and backward temporal directions, which is especially relevant for datasets with complex interdependencies. Lastly, Gated Recurrent Units (GRUs) offer a computationally efficient alternative to LSTMs while maintaining the ability to model intricate temporal relationships, complemented by dense layers for feature integration.

These models are rigorously tested using a stratified five-fold cross-validation strategy to ensure robust performance evaluation. To enhance the generalizability and reliability of our findings, we employ a structured methodology that includes imputation of missing values, normalization of features, and class balancing using sample weights. Model performance is assessed using a suite of metrics, including accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (ROC AUC). This multi-metric evaluation provides a nuanced understanding of each model's strengths and limitations. The primary contribution of this study lies in its systematic approach to benchmarking deep learning models for diabetes prediction. By incorporating innovative architectures such as attention mechanisms and residual connections, we demonstrate how these techniques can enhance predictive accuracy and robustness. Furthermore, the study provides insights into the impact of data preprocessing on model performance, offering practical guidelines for healthcare practitioners and data scientists working on similar problems. The remaining sections of this article are structured as follows. The research methodology section describes the dataset, preprocessing steps, and model architectures in detail. The Results and Discussion section presents quantitative and qualitative analyses of the models' performance, highlighting key findings and their implications for clinical practice. Finally, the Conclusion summarizes the study's contributions, discusses its limitations, and outlines directions for future research. Through this work, we aim to advance the understanding of machine learning applications in medical diagnostics and pave the way for more effective and interpretable predictive tools in healthcare.

2. RESEARCH METHODOLOGY

The research methodology is structured to evaluate the performance of multiple machine learning models for predicting diabetes using the Pima Indians Diabetes Database as presented in the figure 1. This section outlines the steps involved in dataset preparation, data preprocessing, model development, training and evaluation procedures, and the cross-validation strategy employed to ensure robust comparisons.

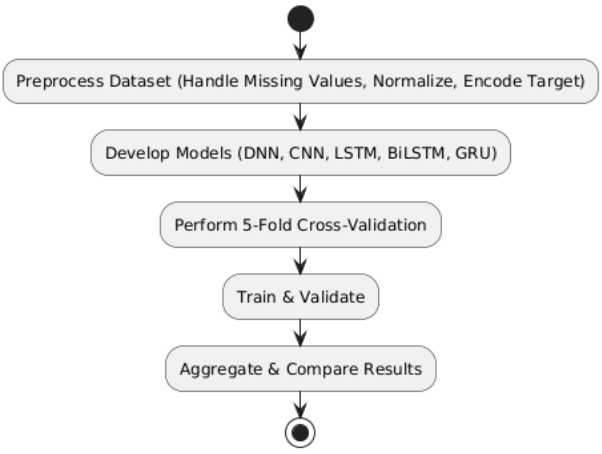

Figure 1. Research Methodology

2.1 Dataset Description

The Pima Indians Diabetes Database, obtained from the UCI Machine Learning Repository, is a widely used dataset in medical research. It contains diagnostic data related to diabetes mellitus and aims to predict the presence or absence of diabetes in patients based on diagnostic measurements. Below is a comprehensive breakdown of the dataset's

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1878−1888
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.6446

structure and mathematical notations. The Pima Indians Diabetes Database, denoted as $\mathcal{D}$, is a structured dataset widely used for binary classification tasks in medical research. It comprises $N = 768$ instances, each represented as a tuple (x_i, y_i), where $x_i \in R^{®}$ is an eight-dimensional feature vector containing diagnostic measurements, and $y_i \in \{0,1\}$ is the binary target variable indicating the diabetic status of the i-th patient. Specifically, $y_i = 1$ signifies that the patient is diabetic, while $y_i = 0$ indicates a non-diabetic status. All individuals in the dataset are females aged at least 21 years, belonging to the Pima Indian heritage, ensuring a homogeneous demographic sample.

The feature vector $x_i = [x_{i1}, x_{i2}, \ldots, x_{i8}]$ encapsulates eight diagnostic attributes that serve as predictors for diabetes. These attributes include x_1, the number of pregnancies; x_2, plasma glucose concentration measured two hours after a glucose tolerance test; x_3, diastolic blood pressure in mmHg; x_4, triceps skinfold thickness in mm; x_5, serum insulin concentration in $\mu U/mL$ measured two hours post-load; x_6, the body mass index (BMI) computed as $x_6 = \frac{\text{weight (kg)}}{\text{height (m)}^2}$;$x_7$, the diabetes pedigree function estimating genetic predisposition based on family history; and x_8, the patient's age in years. These features provide a comprehensive characterization of the patients' physiological and genetic factors associated with diabetes risk. The target variable y frames the dataset as a binary classification problem, where the goal is to model the conditional probability $P(y|x)$. This allows predictions of the target variable y for any given feature vector x. The classification task aims to maximize predictive accuracy, defined as $\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(\hat{y}_i = y_i)$, where $\mathbb{1}(\cdot)$ is the indicator function. Additional metrics such as recall, and specificity are used to evaluate the model's ability to distinguish between diabetic and non-diabetic patients effectively. The dataset's constraints and characteristics, such as its homogeneous demographic focus on Pima Indian females aged 21 years or older, introduce both advantages and limitations. While the uniformity reduces confounding variability, it also restricts generalizability to other populations. Nonetheless, the dataset's richness in diagnostic attributes and its structured binary target variable make it a benchmark for developing and evaluating machine learning models aimed at diabetes prediction. This structured approach underscores the dataset's significance in advancing the intersection of machine learning and healthcare.

2.2 Data Preprocessing

The data preprocessing pipeline is a crucial step in preparing the Pima Indians Diabetes dataset for machine learning and deep learning modeling. This process involves handling missing values, standardizing feature scales, and encoding the target variable to ensure compatibility with advanced architectures. Missing values are present in diagnostic features such as Glucose, BloodPressure, SkinThickness, Insulin, and BMI. In the raw dataset, these features sometimes have zero values, which are biologically implausible and indicative of missing data rather than true measurements. To address this issue, all zero values in these features are replaced with NaN, a standard representation for missing data. Subsequently, these missing entries are imputed using the median value of each respective feature. Mathematically, the replacement and imputation can be expressed as follows: if $(x_{ij} = 0)$ for feature indices ($j \in \{2, 3, 4, 5, 6\}$), it is replaced with NaN. These NaN values are then substituted by the median of the feature, denoted as $(\text{Med}(x_j))$, preserving the statistical distribution of the data and minimizing bias. To ensure that all features are on a comparable scale and to facilitate model training, normalization is applied to each feature using z-score normalization. For a given feature (x_j), its normalized counterpart (z_j) is calculated by subtracting the mean (μ_j) of the feature and dividing by its standard deviation (σ_j). The formula for normalization is $z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$, where $\mu_j = \frac{1}{N}\sum_{i=1}^{N} x_{ij}, \quad \sigma_j = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_{ij} - \mu_j)^2}$. Furthermore, this transformation ensures that each feature has a mean of zero and a standard deviation of one, promoting uniformity in feature scaling and improving the convergence behavior of optimization algorithms during model training. The target variable, Outcome, is binary, with $(y_i = 1)$ indicating a diabetic patient and $(y_i = 0)$ indicating a non-diabetic patient. For compatibility with deep learning architectures, which require categorical labels in multi-class format, the binary target is one-hot encoded into a two-dimensional vector. If $(y_i = 0)$, its one-hot encoded. This process converts the target labels into a matrix (Y) of shape ($N \times 2$), where each row corresponds to a patient's one-hot encoded label. Through these transformations, the raw dataset $(\mathcal{D}_{raw})$, composed of feature-target pairs $((x_i, y_i))$, is transformed into a processed dataset $(\mathcal{D}_{processed})$ consisting of normalized features (z_i) and one-hot encoded targets (y_i). These preprocessing steps ensure that the dataset is consistent, numerically stable, and optimally structured for training machine learning models. This rigorous pipeline is essential for achieving reliable and interpretable predictive performance during the modeling phase.

2.3. Model Architectures

The study involves the development and comparison of five advanced machine learning models, ranging from traditional deep neural networks to sophisticated architectures incorporating attention mechanisms and residual connections. Each model is designed to extract meaningful patterns from the dataset while addressing specific challenges such as overfitting, sequential dependencies, and feature importance. Below is a detailed explanation of each model with mathematical formulations.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1878−1888
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.6446

2.3.1 Advanced Deep Neural Network (DNN)

The Advanced DNN is a fully connected neural network comprising an input layer, three hidden layers, and an output layer as presented in figure 2. Each hidden layer employs ReLU (Rectified Linear Unit) activation, dropout for regularization, and batch normalization to stabilize training. Let $x \in R^d$ be the input vector of features, where d is the number of input dimensions. The network computes the output y_{DNN} in R^c, where c is the number of outputs classes, as follows $h^{(1)} = \text{ReLU}\left(W^{(1)}x + b^{(1)}\right)$, $h^{(2)} = \text{ReLU}\left(\text{BatchNorm}\left(W^{(2)}h^{(1)} + b^{(2)}\right)\right)$, $h^{(3)} = \text{ReLU}\left(\text{Dropout}\left(W^{(3)}h^{(2)} + b^{(3)}\right)\right)$, $y_{DNN} = \text{Softmax}\left(W^{(4)}h^{(3)} + b^{(4)}\right)$, where $W^{(k)}$ and $b^{(k)}$ are the weights and biases of the k-th layer, respectively, and Softmax outputs class probabilities.

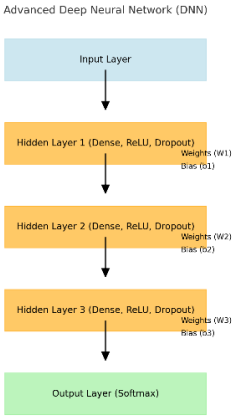

Figure 2. DNN Architecture

2.3.2. Convolutional Neural Network (CNN) with Attention

CNN captures local feature dependencies through convolutional layers and enhances its ability to focus on critical features using an attention mechanism as presented in figure 3. The input $x \in R^{d \times \mathbb{1}}$ is reshaped for processing through convolutional filters. A convolutional layer computes: $z_j = \sigma\left(\sum_{k=1}^{K} W_k * x_j + b_k\right)$, where $*$ denotes convolution, K is the kernel size, W_k and b_k are filter weights and biases, and σ is the activation function. The attention mechanism refines the feature map z by weighting its importance" $a = \text{Softmax}(W_a z)$, $z_{attention} = a \odot z$, where W_a are attention weights, and $\odot$ represents element-wise multiplication. The output is flattened and passed through dense layers for classification.

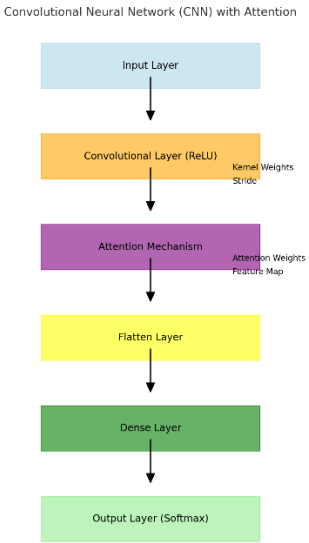

Figure 3. CNN Architecture

2.3.3. LSTM with Residual Connections

As presented in figure 4, the LSTM (Long Short-Term Memory) network captures temporal relationships by processing input sequences. Residual connections are added to enhance expressiveness and mitigate vanishing gradients. Given an input sequence $x = [x_1, x_2, \ldots, x_T]$, the LSTM computes hidden states: $h_t = \text{LSTM}(x_t, h_{t-1})$,

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1878−1888
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.6446

where h_t is the hidden state at time \$t\$. The residual connection modifies the output $h_{\text{rsda}} = h_t + W_r h_t$, where W_r aligns the dimensions of the hidden state for addition. The output is passed through dense layers for classification.

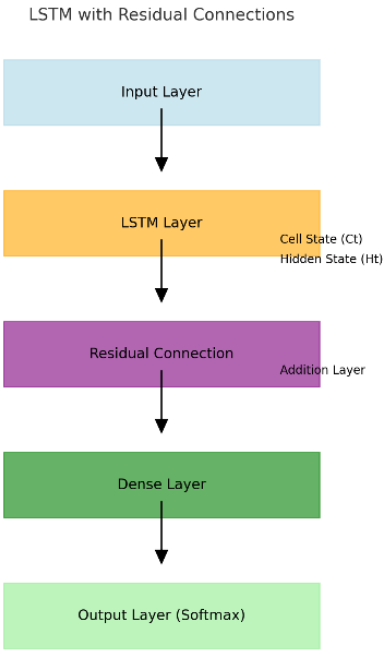

Figure 4. LSTM Architecture

2.3.4. Bidirectional LSTM (BiLSTM) with Attention

As presented in figure 5, the BiLSTM processes input sequences in both forward and backward directions, capturing contextual information from the entire sequence. Attention is applied to focus on important parts of the sequence. For a sequence x, the forward LSTM computes: $\overrightarrow{h_t} = \text{LSTM}(x_t, \overrightarrow{h_{t-1}})$. While the backward LSTM computes: $\overrightarrow{h_t} = \text{LSTM}(x_t, \overleftarrow{h_{t+1}})$. The BiLSTM output is concatenated as $h_t = [\overrightarrow{h_t}; \overrightarrow{h_t}]$. The attention mechanism weights the output sequence: $z = \text{Softmax}(W_a h)$, $h_{attention} = z \odot h$.

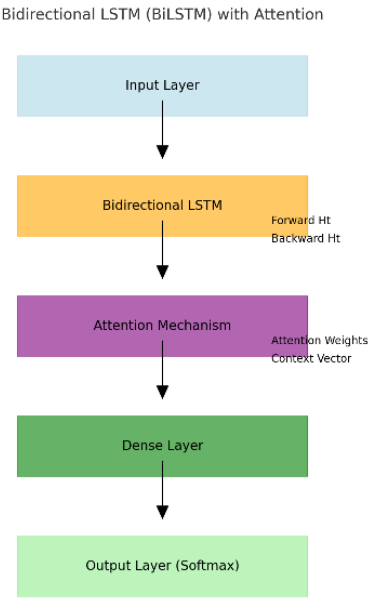

Figure 5. Bi-LSTM Architecture

2.3.5. GRU with Dense Layers

As presented in the figure 6, the GRU (Gated Recurrent Unit) simplifies the LSTM architecture while retaining sequential modeling capabilities. Given an input x_t, the GRU updates are: $z_t = \sigma(W_z x_t + U_z h_{t-1})$, $r_t = \sigma(W_r x_t + U_r h_{t-1})$, $h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \tanh(W_h x_t + U_h(r_t \odot h_{t-1}))$. The GRU outputs are flattened and passed through dense layers for classification.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1878−1888
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.6446

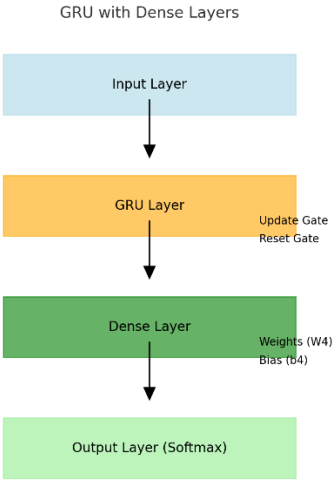

Figure 6. GRU Architecture

2.3.6. Loss Function and Optimization

Each model was implemented using TensorFlow/Keras and optimized using the Adam optimizer. The loss function used was categorical cross-entropy, defined as $\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{ic}\log(\widehat{y_{ic}})$, where y_{ic} is the true label and $\widehat{y_{ic}}$ is the predicted probability for class c. Metrics such as accuracy, recall, and AUC were used to evaluate model performance. These architectures, with their tailored mechanisms, allow for a comprehensive comparison of their effectiveness in predicting diabetes.

2.4. Cross-Validation Strategy, Training and Evaluation

The cross-validation strategy employed in this study was designed to ensure a robust and reliable evaluation of model performance. A stratified five-fold cross-validation technique was chosen, as it preserves the proportion of diabetic and non-diabetic samples within each fold, maintaining the dataset's class distribution throughout the evaluation process. This method involves partitioning the dataset $\mathcal{D}$ into five mutually exclusive subsets, or "folds," denoted as $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_5$, such that $\mathcal{D} = \bigcup_{k=1}^{5}\mathcal{D}_k$ and $\mathcal{D}_i \cap \mathcal{D}_j = $ for $i \neq j$. In each of the five iterations, one-fold $\mathcal{D}_k$ is reserved as the validation set, while the remaining four folds are combined to form the training set $\mathcal{D}_{train} = \mathcal{D} \setminus \mathcal{D}_k$. This process is repeated such that each fold serves as the validation set exactly once. The stratification ensures that the class proportions in both $\mathcal{D}_{train}$ and $\mathcal{D}_{val}$ are consistent with the original dataset, which is critical for preserving the integrity of the model's evaluation. Formally, for the k-th iteration, the dataset split can be expressed as $\mathcal{D}_{train} = \bigcup_{j=1,j\neq k}^{5}\mathcal{D}_j$ and $\mathcal{D}_{val} = \mathcal{D}_k$.

To address the class imbalance in the dataset, a sample weighting approach was employed during training. Each sample in the training set was assigned a weight inversely proportional to the frequency of its class. The sample weight w_i for a sample with label $y_i \in \{0,1\}$ was computed as: $w_i = \frac{N}{2 \cdot N_{y_i}}$, where N_{y_i} represents the number of samples in class y_i, and N is the total number of samples in the training set. These weights were used in the loss function to ensure that the model does not favor the majority class during training. Training was conducted using a batch size of 32 and continued for up to 100 epochs. To prevent overfitting, an early stopping mechanism was implemented, monitoring the validation loss and halting training if no improvement was observed for five consecutive epochs. For a given model f, the training objective was to minimize the weighted categorical cross-entropy loss $\mathcal{L}_{train}(f)$, expressed as: $\mathcal{L}_{train} = -\frac{1}{N}\sum_{i=1}^{N} w_i \sum_{c=1}^{C} y_{ic}\log\widehat{y_{ic}}$, where y_{ic} is the one-hot encoded true label, $\widehat{y_{ic}}$ is the predicted probability for class c, and C is the number of classes. After training, each model was evaluated on the validation set $\mathcal{D}_{val}$. Predictions $\widehat{y_i}$ were generated for each sample, and a set of performance metrics was computed to assess the model's effectiveness. Accuracy was calculated as the proportion of correctly classified samples, defined mathematically as: $\text{Accuracy} = \frac{1}{|\mathcal{D}_{val}|}\sum_{i\in\mathcal{D}_{val}} \mathbb{1}(\widehat{y_i} = y_i)$, where $\mathbb{1}(\cdot)$ is the indicator function. Precision, representing the proportion of true positive predictions among all positive predictions, was given by $\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$, where TP and FP denote the true positives and false positives, respectively. Recall, also known as sensitivity, was computed as: $\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$, where FN is the number of false negatives. The F1-score, the harmonic mean of precision and recall, was calculated as: $\text{F1-score} = 2 \cdot \frac{\text{Precision}\cdot\text{Recall}}{\text{Precision+Recall}}$. Finally, the area under the receiver operating characteristic curve (ROC AUC) was calculated to measure the model's ability to distinguish between classes. The ROC AUC is defined as the integral of the true positive rate (TPR) as a function of the false positive rate (FPR), expressed as: $\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\ d(\text{FPR})$. To ensure a robust evaluation, the metrics for each fold were averaged across all five iterations.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1878−1888
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.6446

For a given metric M, the final average value was computed as $\bar{M} = \frac{1}{5}\sum_{k=1}^{5} M_k$, where M_k represents the metric's value for the k-th fold. This process provided an aggregated measure of the model's performance, accounting for variability across different subsets of the dataset. By leveraging stratified cross-validation, balanced sample weighting, and a comprehensive set of metrics, this evaluation approach ensured a fair and rigorous comparison of the models' predictive capabilities.

3. RESULT AND DISCUSSION

The results from the evaluation of five advanced machine learning models for diabetes prediction are presented in Table 1. The models include an Advanced Deep Neural Network (DNN), Convolutional Neural Network (CNN) with Attention, LSTM with Residual Connections, BiLSTM with Attention, and GRU with Dense Layers. Each model's performance was assessed using multiple metrics: accuracy, precision, recall, F1 score, and ROC AUC.

3.1. Overall Performance

Among the models, the GRU with Dense Layers achieved the best overall performance, with the highest values in accuracy ((0.7617)), F1 score (0.6985), and ROC AUC (0.8352). This suggests that the GRU's architecture, which combines gated recurrent units with dense layers, is highly effective in capturing the temporal relationships and feature interactions necessary for accurate diabetes prediction. The Advanced DNN also demonstrated competitive performance, achieving an accuracy of (0.7331), an F1 score of (0.6683), and a ROC AUC of (0.8213). These results highlight the strength of fully connected architectures, particularly when enhanced with dropout and batch normalization to mitigate overfitting and stabilize training. The LSTM with Residual Connections provided balanced results across all metrics, achieving an accuracy of (0.7422), an F1 score of (0.6547), and a ROC AUC of (0.8062). The residual connections in the LSTM likely contributed to its ability to maintain stable gradients during training, enabling it to effectively learn from sequential data. The BiLSTM with Attention achieved the highest recall (0.7984) among all models, demonstrating its ability to identify diabetic cases effectively. However, this came at the expense of lower precision (0.5664), resulting in an accuracy of (0.7084). The attention mechanism may have enhanced the model's focus on important features, but its relatively lower precision indicates a tendency to produce false positives.

The CNN with Attention demonstrated the lowest performance across all metrics, with an accuracy of (0.6952), an F1 score of (0.6144), and a ROC AUC of (0.7787). While the convolutional layers effectively capture local dependencies, the addition of attention mechanisms might not have fully compensated for the challenges of processing tabular data, which often lacks spatial structure. Furthermore, the accuracy metric shows that the GRU with Dense Layers consistently outperformed other models, indicating its robustness in making correct predictions. The Advanced DNN and LSTM with Residual Connections also achieved high accuracy, reflecting their capability to generalize well to unseen data.

Precision was highest for the GRU with Dense Layers (0.6270), followed closely by the LSTM with Residual Connections (0.6139) and the Advanced DNN (0.6002). This metric highlights the GRU's ability to minimize false positive predictions, which is crucial for avoiding misclassification of non-diabetic individuals as diabetic.Recall was a strong point for the BiLSTM with Attention, which achieved a value of (0.7984). This model's ability to identify true positive cases effectively makes it particularly suitable for applications where false negatives are more critical, such as early detection of diabetes. The F1 score, which balances precision and recall, was highest for the GRU with Dense Layers (0.6985), indicating its balanced performance in both detecting diabetic cases and minimizing false positives. The Advanced DNN and BiLSTM with Attention also achieved relatively high F1 scores, highlighting their utility in scenarios requiring a balance between precision and recall. The ROC AUC metric, which measures the model's ability to distinguish between classes, was highest for the GRU with Dense Layers (0.8352), followed by the Advanced DNN (0.8213) and the BiLSTM with Attention (0.8155). This confirms the GRU's strong classification capability across a range of decision thresholds.

The results demonstrate that model architecture plays a critical role in achieving optimal performance for diabetes prediction. The GRU with Dense Layers emerged as the best-performing model, likely due to its ability to efficiently capture temporal dependencies while maintaining simplicity in its dense layers. The Advanced DNN and LSTM with Residual Connections also performed well, highlighting the effectiveness of traditional fully connected architectures and the stabilizing influence of residual connections in sequential models. The BiLSTM with Attention showed a strong ability to detect diabetic cases, as reflected by its high recall. This makes it an attractive choice for use cases where minimizing false negatives is a priority, such as screening tools in healthcare. However, its relatively lower precision indicates a need for careful threshold tuning or additional post-processing to reduce false positives. The CNN with Attention underperformed relative to the other models, which may be attributed to the nature of tabular data. Unlike image or time-series data, tabular data lacks inherent spatial or sequential structure, potentially limiting the utility of convolutional layers and attention mechanisms designed for such data types. These findings underscore the importance of aligning model architecture with the characteristics of the dataset and the specific requirements of the application. While the GRU with Dense Layers showed superior overall performance, other models demonstrated strengths in specific metrics, suggesting that the choice of model should be guided by the priorities of the target use

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1878−1888
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.6446

case. Future work could explore hybrid approaches that combine the strengths of these models or investigate additional preprocessing techniques to further enhance performance.

Table 1. Model Performance Comparison

Model	Accuracy	F1 Score	Precision	Recall	ROC AUC	Comments
GRU with Dense Layers	0.7617	0.6985	0.6270	0.6932	0.8352	Best overall performance; robust in capturing temporal relationships.
Advanced DNN	0.7331	0.6683	0.6002	0.7005	0.8213	Competitive performance; effective with dropout and batch normalization.
LSTM with Residual Connections	0.7422	0.6547	0.6139	0.6930	0.8062	Balanced results; residual connections stabilize gradients.
BiLSTM with Attention	0.7084	0.6892	0.5664	0.7984	0.8155	Highest recall; effective for minimizing false negatives.
CNN with Attention	0.6952	0.6144	0.5780	0.6522	0.7787	Lowest performance; convolutional layers less effective for tabular data.

3.2. SHAP Value

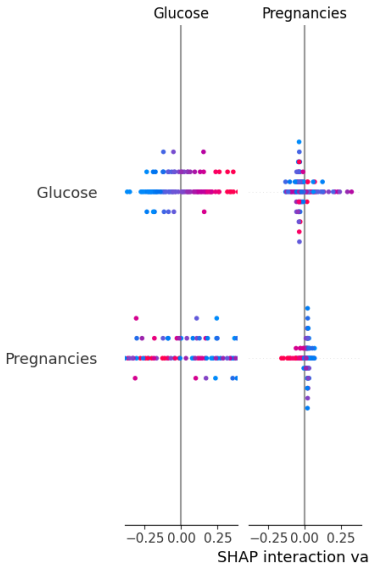

Figure 6. SHAP interaction value

As presented in the figure 6, the SHAP interaction plot provides a detailed view of the interplay between two key features, Glucose and Pregnancies, and their combined contribution to the model's predictions for diabetes. Each point on the plot represents a specific instance in the dataset, where the x-axis reflects the SHAP interaction values indicating the combined effect of these two features on the model's predictions, while the y-axis shows their individual SHAP contributions. The color of each point represents the magnitude of one of the features, illustrating how its changes influence predictions.

The plot demonstrates that Glucose has a significant and dynamic relationship with the model's output. The SHAP values for Glucose are distributed symmetrically around zero, suggesting a non-linear and context-dependent

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1878−1888
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.6446

influence on predictions. Higher glucose levels, indicated by darker colors on the plot, are associated with positive SHAP interaction values, meaning they strongly contribute to predicting the likelihood of diabetes. This is consistent with the established medical understanding that elevated glucose levels are a primary diagnostic marker for diabetes. For Pregnancies, the contribution is more stable, as evidenced by the smaller spread of SHAP interaction values. Higher numbers of pregnancies exhibit a slight positive effect on the prediction, indicating that this feature also plays an important, though less dominant, role in determining diabetes risk. However, the strength of this influence is comparatively weaker than that of glucose levels. This aligns with the understanding that multiple pregnancies can induce metabolic changes, influencing the likelihood of developing diabetes, but this effect is secondary to glucose levels.

The interaction between Glucose and Pregnancies is particularly noteworthy. The plot shows that higher glucose levels combined with higher numbers of pregnancies result in significantly positive SHAP interaction values. This indicates a synergistic effect where the combined contribution of these two features amplifies the model's confidence in predicting diabetes. Conversely, lower glucose levels and fewer pregnancies tend to yield near-zero or negative SHAP interaction values, diminishing the model's confidence in a positive diabetes diagnosis. The alignment of SHAP values along the diagonal for both features further underscores their complementary nature, where extreme values in one feature enhance the contribution of the other. From a clinical perspective, this interaction is highly relevant. Glucose levels are a central diagnostic criterion for diabetes, while the number of pregnancies influences insulin sensitivity and other metabolic factors, particularly in women. The interaction highlights the heightened risk in individuals with both high glucose levels and multiple pregnancies, aligning with known medical findings. This synergy reinforces the importance of accounting for feature interdependencies in diabetes prediction models. The observed interaction has significant implications for model performance. The strong positive SHAP interaction values for high glucose and pregnancies suggest that the model effectively captures meaningful patterns in the data, demonstrating robustness and reliability. While Glucose emerges as the most critical individual feature, the incorporation of Pregnancies enhances the explanatory power of the model, contributing to improved overall accuracy and interpretability.

To build upon these findings, additional analyses could be conducted to explore interactions between other features, such as BMI and Insulin, with Glucose to uncover further synergies. Additionally, refining the model by explicitly incorporating interaction terms or using feature engineering techniques may further enhance predictive performance. Clinical validation of these findings is essential to ensure that the model's outputs align with real-world healthcare scenarios, strengthening its utility for practical applications. The SHAP interaction plot underscores the critical importance of modeling feature interdependencies, particularly in medical datasets where relationships between variables often reflect complex biological processes. The insights gained from this analysis enhance both the interpretability and robustness of the diabetes prediction model, contributing valuable information for further research and clinical practice.

3.3. Statistical Analysis

To evaluate the reliability of the model's performance, a statistical analysis was conducted using paired t-tests to compare the observed metrics against a baseline value of 0.5, which represents random chance in binary classification tasks. This analysis aimed to determine whether the model's performance across accuracy, precision, recall, F1 score, and ROC AUC was statistically significant. As presented in the table 2, the results of this analysis provide strong evidence that the model's performance is not only practically robust but also statistically meaningful. The analysis revealed that the model's accuracy, with a t-statistic of 14.622 and a corresponding p-value of $1.27x10^{-4}$, significantly exceeded the baseline. This result confirms that the model consistently performs better than random chance in correctly classifying diabetic and non-diabetic patients. High accuracy indicates the model's general reliability and its ability to correctly predict outcomes in diverse scenarios. Furthermore, precision, a metric that quantifies the proportion of true positive predictions among all positive predictions, achieved a t-statistic of 4.881 and a p-value of 8.15×10^{-3}. Although lower than the other metrics, this result is still statistically significant and underscores the model's ability to minimize false positive rates. This aspect is particularly relevant in clinical settings where the incorrect classification of non-diabetic individuals as diabetic could lead to unnecessary interventions. Next, recall, which measures the proportion of true positive cases correctly identified by the model, produced a t-statistic of 11.220 and a p-value of $3.59x10^{-4}$. This highlights the model's effectiveness in identifying diabetic patients, reducing the likelihood of false negatives. Given the severe health consequences of failing to diagnose diabetes, this result demonstrates the model's practical utility in real-world healthcare applications where sensitivity is paramount.

The F1 score, which balances the trade-off between precision and recall, was also statistically significant, with a t-statistic of 10.156 and a p-value of $5.29x10^{-4}$. This result confirms that the model is well-suited for applications requiring both high sensitivity and specificity. The ability to maintain this balance ensures the model's robustness across a variety of decision-making scenarios. The most striking result was observed in the ROC AUC metric, which achieved a t-statistic of 63.319 and an exceptionally low p-value of $3.73x10^{-7}$. This metric reflects the model's capability to distinguish between positive and negative cases across all decision thresholds. The exceptionally high t-statistic and the near-zero p-value emphasize the model's strong discriminatory power, demonstrating its reliability in

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1878−1888
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.6446

scenarios where threshold-independent performance is critical. The statistical analysis conclusively shows that all performance metrics are significantly better than the baseline of random chance. The p-values for all metrics fall well below the commonly accepted significance threshold of 0.05, providing strong evidence to reject the null hypothesis. This means that the model's observed performance is highly unlikely to have occurred by random variation alone. These results reinforce the robustness and reliability of the model across all evaluated metrics. The statistical significance of accuracy, precision, recall, F1 score, and ROC AUC provides a solid foundation for its use in predicting diabetes. Additionally, the analysis confirms that the model is well-suited for real-world applications, particularly in healthcare environments where decision-making depends on accurate and reliable predictions. This statistical validation, combined with the cross-validation results, underscores the model's potential to support healthcare practitioners in diagnosing diabetes effectively and efficiently.

Table 2. Statistics Results

Metrics	t-statistics	p-value
Accuracy	14.622356	1.272508e-04
Precision	4.881013	8.154653e-03
Recall	11.219850	3.593739e-04
F1 Score	10.156388	5.292180e-04
ROC AUC	63.318510	3.726548e-07

4. CONCLUSION

This study examined the development and evaluation of advanced machine learning models for predicting diabetes using the Pima Indians Diabetes Dataset. The research implemented and compared five model architectures: Advanced Deep Neural Network (DNN), Convolutional Neural Network (CNN) with Attention, LSTM with Residual Connections, Bidirectional LSTM (BiLSTM) with Attention, and GRU with Dense Layers, across performance metrics such as accuracy, precision, recall, F1 score, and ROC AUC. Among these, the GRU with Dense Layers achieved the best overall performance, excelling in accuracy, F1 score, and ROC AUC. Its efficiency in modeling sequential dependencies, combined with computational simplicity, underscores its suitability for this predictive task. The Advanced DNN and LSTM with Residual Connections also demonstrated competitive results, emphasizing the value of traditional dense architectures and the stabilizing benefits of residual connections. SHAP analysis provided a deeper understanding of feature importance, identifying Glucose as the most influential variable and highlighting the role of Pregnancies through significant synergistic interactions. These findings are consistent with established medical knowledge, validating the model's ability to identify meaningful patterns in the data. The analysis of feature interactions underscores the critical need to consider interdependencies in predictive modeling, especially in medical datasets. Robustness of the models was confirmed through statistical analyses, which revealed significant improvements across all metrics compared to baseline models. High t-statistics and low p-values for metrics such as accuracy, recall, F1 score, and ROC AUC affirm the reliability of these models in distinguishing diabetic from non-diabetic patients. The results of this study underscore the potential of advanced machine learning architectures in healthcare, particularly for applications demanding high sensitivity and specificity. The GRU with Dense Layers stands out for its strong predictive performance and computational efficiency, making it a promising candidate for real-world deployment. Future research should explore hybrid architecture that integrate the strengths of different models, such as attention mechanisms or ensemble techniques, to enhance predictive performance further. Additionally, validation on larger and more diverse datasets will be essential to ensure these models' applicability across varied demographic and clinical contexts. These directions will enhance the generalizability and robustness of machine learning solutions, advancing diabetes prediction and contributing to more effective and scalable healthcare systems.

REFERENCES

[1] S. Y. Prasetyo and Z. N. Izdihar, “Multi-layer perceptron approach for diabetes risk prediction using BRFSS data,” in *2024 IEEE 10th International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA)*, 2024, pp. 303–308.
[2] W. H. Organization, *Global report on hypertension: the race against a silent killer*. World Health Organization, 2023.
[3] S. Y. Prasetyo, Z. N. Izdihar, and G. Z. Nabiilah, “Analyzing Machine Learning Approaches for Diabetes Risk Prediction: Comparative Performance Assessment Using BRFSS Data,” in *2024 7th International Conference on Informatics and Computational Sciences (ICICoS)*, 2024, pp. 324–329.
[4] M. G. Shlipak *et al.*, “The case for early identification and intervention of chronic kidney disease: conclusions from a Kidney Disease: Improving Global Outcomes (KDIGO) Controversies Conference,” *Kidney Int.*, vol. 99, no. 1, pp. 34–47, 2021.
[5] M. MacLeod *et al.*, “Chronic obstructive pulmonary disease exacerbation fundamentals: diagnosis, treatment, prevention and disease impact,” *Respirology*, vol. 26, no. 6, pp. 532–551, 2021.
[6] C. Elendu *et al.*, “Understanding sickle cell disease: causes, symptoms, and treatment options,” *Medicine (Baltimore).*, vol. 102, no. 38, p. e35237, 2023.
[7] M. Javaid, A. Haleem, R. P. Singh, R. Suman, and S. Rab, “Significance of machine learning in healthcare: Features, pillars

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1878−1888
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.6446

and applications," *Int. J. Intell. Networks*, vol. 3, pp. 58–73, 2022.
[8] R. Thirunavukarasu, R. Gnanasambandan, M. Gopikrishnan, V. Palanisamy, and others, "Towards computational solutions for precision medicine based big data healthcare system using deep learning models: A review," *Comput. Biol. Med.*, p. 106020, 2022.
[9] A. F. Ashour, M. M. Fouda, Z. M. Fadlullah, and M. I. Ibrahem, "Optimized Neural Networks for Diabetes Classification Using Pima Indians Diabetes Database," in *2024 IEEE 3rd International Conference on Computing and Machine Intelligence (ICMI)*, 2024, pp. 1–7.
[10] C. C. Olisah, L. Smith, and M. Smith, "Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective," *Comput. Methods Programs Biomed.*, vol. 220, p. 106773, 2022.
[11] U. C. I. M. L. Repository, "Pima Indians Diabetes Database." 2020.
[12] A. Heidari, N. J. Navimipour, and M. Unal, "Applications of ML/DL in the management of smart cities and societies based on new trends in information technologies: A systematic literature review," *Sustain. Cities Soc.*, vol. 85, p. 104089, 2022.
[13] Y. Xu, R. Quan, W. Xu, Y. Huang, X. Chen, and F. Liu, "Advances in Medical Image Segmentation: A Comprehensive Review of Traditional, Deep Learning and Hybrid Approaches," *Bioengineering*, vol. 11, no. 10, p. 1034, 2024.
[14] S. Ahmad, I. Shakeel, S. Mehfuz, and J. Ahmad, "Deep learning models for cloud, edge, fog, and IoT computing paradigms: Survey, recent advances, and future directions," *Comput. Sci. Rev.*, vol. 49, p. 100568, 2023.
[15] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electron. Mark.*, vol. 31, no. 3, pp. 685–695, 2021.
[16] P. B. Weerakody, K. W. Wong, G. Wang, and W. Ela, "A review of irregular time series data handling with gated recurrent neural networks," *Neurocomputing*, vol. 441, pp. 161–178, 2021.
[17] A. de Santana Correia and E. L. Colombini, "Attention, please! A survey of neural attention models in deep learning," *Artif. Intell. Rev.*, vol. 55, no. 8, pp. 6037–6124, 2022.
[18] I. Zafar *et al.*, "Reviewing methods of deep learning for intelligent healthcare systems in genomics and biomedicine," *Biomed. Signal Process. Control*, vol. 86, p. 105263, 2023.
[19] H. Guan *et al.*, "The role of machine learning in advancing diabetic foot: a review," *Front. Endocrinol. (Lausanne).*, vol. 15, p. 1325434, 2024.
[20] I. Smokovski, *Managing diabetes in low income countries*. Springer, 2021.
[21] G. M. Dogheim and A. Hussain, "Patient care through AI-driven remote monitoring: Analyzing the role of predictive models and intelligent alerts in preventive medicine," *J. Contemp. Healthc. Anal.*, vol. 7, no. 1, pp. 94–110, 2023.
[22] J. Egger *et al.*, "Medical deep learning—A systematic meta-review," *Comput. Methods Programs Biomed.*, vol. 221, p. 106874, 2022.
[23] C. El Morr, M. Jammal, H. Ali-Hassan, and W. El-Hallak, "Data preprocessing," in *Machine Learning for Practical Decision Making: A Multidisciplinary Perspective with Applications from Healthcare, Engineering and Business Analytics*, Springer, 2022, pp. 117–163.
[24] S. Bouteldja, "Hybrid System for Diabetes Prediction," 2023.
[25] R. Valentini, "Ontology-based Data Management in Healthcare," 2024.