E-ISSN: 3123-2507
DOI: https://doi.org/10.64803/cessmuds.vxnx.xxx | Page: 753-761
Arnita, Syahputra, Marpaung, Setia Sinaga, Junaidi, and Ramadhani

Critical Evaluation of Machine Learning Models for Glucose Prediction: A Case Study Of Overfitting and Small Data

Arnita[1*], Hermawan Syahputra[2], Faridawaty Marpaung[2], Marlina Setia Sinaga[2], Agus Junaidi[3], Fanny Ramadhani[2]

[1,2]Department of Mathematics, Faculty of Mathematics and Natural Sciences, Medan State University, Jl. William Iskandar Ps. V, 20221, North Sumatera, Indonesia
[3]Department of electrical Engineering, Faculty of Engineering, Universitas Negeri Medan, Jl. William Iskandar Ps. V, 20221, North Sumatera, Indonesia
E-mail: [1]*arnita@unimed.ac.id, [2]hsyahputra@unimed.ac.id, [2]faridawaty@unimed.ac.id, [2]marlinasetiasin@unimed.ac.id, [3]agusjunaidi@unimed.ac.id, [2]Fannyr@unimed.ac.id

Abstract

Abstract: Glucose level prediction is an important area of research, but the application of machine learning models often faces challenges such as overfitting and limited data. This study aims to evaluate the performance of seven regression models—Linear Regression, Lasso, SVR, Decision Tree, Random Forest, Gradient Boosting, and K-Nearest Neighbors in predicting glucose levels on a small dataset that has undergone preprocessing and dimensionality reduction using Principal Component Analysis (PCA). The evaluation was conducted using the R-squared, Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) metrics. The results showed that most models suffered from severe overfitting, where good performance on the training data was not followed by performance on the test data, as reflected by negative R-squared values. Although Random Forest produced relatively lower errors (MSE = 0.009; RMSE = 0.095), the overall performance remained inadequate due to data limitations. These findings confirm that small sample sizes and overfitting simultaneously hamper the generalizability of models and emphasize the need for more robust approaches in biomedical applications with limited data.

Keywords: *Decision Support System, SAW, WP, Integrity Zone.*

I. INTRODUCTION

Blood sugar prediction is a crucial issue in diabetes prevention and management. Accurate prediction models can help support medical decision-making and prevent serious complications resulting from fluctuating blood sugar levels. Therefore, the development of reliable models is increasingly urgent, as the number of diabetes sufferers worldwide increases (K. Liu et al., 2023). This clinical relevance makes research related to blood sugar prediction a significant contributor to healthcare systems.

The application of machine learning (ML) offers significant opportunities because it can capture complex patterns in medical and physiological data. Several studies have shown that ML algorithms can utilize sensor data such as continuous glucose monitoring (CGM), physical activity, and even dietary patterns to generate more accurate predictions (van Doorn et al., 2021). With this potential, ML is not only a predictive tool but also plays a role in supporting more adaptive and personalized health monitoring systems. However, fundamental challenges often limit the performance of predictive models. Small sample sizes and limited data representation often make it difficult for models to generalize to broader populations. Furthermore, the problem of overfitting is a major obstacle, where models appear excellent on training data but fail when tested on new data or real-world conditions (Sung, 2024). The combination of these two factors worsens model performance and reduces reliability.

The implications of these limitations are crucial because inaccurate models can pose risks to medical decision-making. Several studies emphasize that patient safety must be a priority, so researchers should not ignore data limitations or the risk of overfitting in research (Emerson et al., 2023). Therefore, research in the field of blood glucose prediction must focus not only on achieving high accuracy but also on ensuring the model's robustness in real-world conditions and its continued use in clinical practice. In this study, in addition to addressing specific challenges of data and overfitting, we also used seven different machine learning methods for comparison and to bolster the robustness of the glucose prediction model. These methods are linear regression, Lasso, support vector regression (SVR), decision tree, random forest, gradient boosting, and K-nearest neighbors (KNN). These methods were chosen because each has advantages that can address various aspects of blood glucose data, such as speed, interpretability, or the ability to handle nonlinearity.

Linear regression and lasso regression were chosen because they are simple and interpretive, excellent as baselines, and can work with limited data. Lasso Regression, with its L1 regularization, has the ability to select important features and reduce

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

overfitting by removing irrelevant features, which is especially useful if the dataset has many variables but a lmited sample size (Soleh & Aunuddin, 2013) . SVR was chosen because of its ability to handle nonlinear data with the kernel trick and is often more robust to outliers than simple linear regression (Bagaskara et al., 2024). Decision Tree and Random Forest were included because of their ability to handle interactions between features, nonlinearity, and data variability (Kinasih et al., 2024). Random Forest is good because it combines many decision trees, making it relatively stable and robust to noise, and can also provide the best estimate (Hameed & Kleinberg, 2020). Gradient boosting was chosen because of its ability to improve accuracy through ensemble boosting and its ability to iteratively correct the weaknesses of weak learners, although it may require careful hyperparameter tuning [9](Belyadi & Haghighat, 2021). KNN is used because it is simple and intuitive: if some data points are very "similar" in features, KNN can be very effective, especially on data where anomalies or small changes arise from similar examples (Shi et al., 2022).

Relevant research conducted by (Hameed & Kleinberg, 2020) explored blood glucose level prediction by comparing various machine learning algorithms ranging from simple models such as linear regression and ARMA to more complex models such as random forest, support vector regression (SVR), and recurrent neural networks (RNN/LSTM). This comparison was conducted on two different data types: the OhioT1DM dataset, a controlled study with more uniform clinical conditions, and Patient Generated Health Data (PGHD) from OpenAPS, which reflects real-world data with variations in patient behavior and lifestyle. The results showed that algorithm performance is not always consistent between the two data sources; models that perform very well on small, structured datasets do not necessarily maintain their superiority on larger, more heterogeneous real-world data. Studies have shown that models such as Random Forest and SVR perform well in blood glucose prediction and diabetes management. For example, in the study "Machine Learning Models for Blood Glucose Level Prediction: Systematic Review and Network Meta-Analysis," neural networks generally performed best for long prediction horizons, but random forest and SVR were often strong choices in many conditions (K. Liu et al., 2023). Also, a comparative study of BG prediction methods by (Hameed & Kleinberg, 2020) shows the superiority of baseline models such as linear regression and random forest in more heterogeneous real situations.

Based on the above background, this study aims to evaluate and compare the performance of seven machine learning algorithms Linear Regression, Lasso, Support Vector Regression (SVR), Decision Tree, Random Forest, Gradient Boosting, and K-Nearest Neighbors (KNN) in predicting non-invasive blood glucose levels based on physiological multisensor data. By analyzing the advantages of each method, this study is expected to contribute to identifying the most accurate and reliable prediction model to support a smart technology-based health monitoring system.

II. RESEARCH METHODOLOGY

The research methods contain explanations in the form of paragraphs about the research design or descriptions of the experimental settings, data sources, data collection techniques, and data analysis conducted by the researcher. This guide will explain how to write the headings. If there are more than one heading, use the second level headings as below.

A. Data

A passive infrared (PIR) mobility sensor, a non-contact infrared temperature sensor, blood sugar level examination data obtained with a pulse oximeter and heart rate sensor, and elderly measurement findings obtained with invasive instruments comprise the research data utilized.51 individuals having both normal sample criteria and high blood sugar were included in the sample.After that, this data was separated into two sections: 40 for training data and 11 for testing. According to the guidelines of supervised learning methods, this section aims to evaluate the model's capacity for generalization (Goodfellow et al., n.d.).Thirteen factors were employed in this investigation.These variables, which include body temperature, BPM, SpO2, PPG_Red, PPR_RR, Nir61, Nir680, Nir730, Nir760, Nir810, Nir860, and Glucose_manual, are the outcomes of blood sugar level measurement extraction generated by each sensor.

B. Preprocessing

The initial stage after data collection is preprocessing, where the data is checked for completeness and distribution. Preprocessing is then continued with data normalization using min-max standardization and dimensionality reduction using Principal Component Analysis (PCA) to address non-normal data distribution and reduce outliers. Min-max standardization can be written as the following formula:

$$Standardize = \frac{(x - x_{min})}{(x_{max} - x_{min})} \quad (1)$$

PCA was chosen because it is able to simplify the number of features while maintaining the main variability, thereby reducing the risk of overfitting. The steps of the PCA process are (Jollife & Cadima, 2016):

1. Form a Covariance/Correlation Matrix that describes the relationship between variables.
2. Calculate the Eigen Values and Eigen Vectors which show the amount of variance that can be explained by each principal component.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

3. Sort the principal components from largest to smallest eigenvalues. PC1 (the first principal component) explains the largest variance in the data, followed by PC2, PC3, etc.
4. Forming Principal Components calculated by multiplying the standardized data by the selected eigenvectors.
5. Selecting the relevant number of components is determined based on the percentage of variance explained. Generally, components are chosen that explain a total of ≥70–90% of the variance..
6. Interpretation of results is based on the loading matrix value, which states the relationship between the original variables and the PC. Visualization can be in the form of a scree plot, biplot, or scatter plot of PC1 vs. PC2.

C. Regression modeling

Seven regression algorithms were used to build the glucose prediction model, namely linear regression, Lasso regression, support vector regression (SVR), decision tree, random forest, gradient boosting, and K-nearest neighbors (KNN). The selection of these algorithms was based on the differences in their respective characteristics; for example, linear regression is simple but interpretive, tree-based methods are robust to non-linear data, and ensemble methods such as random forest and gradient boosting are effective in increasing accuracy (Alanazi, 2022) .

a. Linear Regression

Linear regression is a supervised learning method that models the relationship between independent variables (e.g., physiological signals: PPG, GSR, skin temperature, activity, diet) and dependent variables (e.g., blood sugar levels). This model attempts to find the best straight line (line of best fit) to predict the target value. The mathematical model (Jollife & Cadima, 2016):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \in \quad (2)$$

Where
Y = dependent variable,
X_i= independent variable, β_i = regression coefficient,
$\in$ = error (residual).

The steps in regression modeling are (Amalia Dewayanti & Utami, n.d.; Ayu & Srinadi, 2014):

1. Building a linear regression model.
2. Estimating regression coefficients using least squares.
3. Model evaluation using evaluation metrics: MSE and RMSE.
4. Interpretation and implementation
5. Using the model for prediction.

b. Lasso Regression

Lasso regression is an extension of linear regression that adds L1 regularization to the loss function. Its goal is to reduce model complexity by shrinking the regression coefficients and performing automatic feature selection, as some coefficients can be forced to zero. Lasso objective gsi (Soleh & Aunuddin, 2013) :

$$Minimize\ \frac{1}{2n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p}|\beta_j| \quad (3)$$

Where
y_i = target,
$\hat{y}_i$ = prediction result,
β_j = regression coefficient,
λ = regularization parameter (L1 penalty).

Steps to build a Lasso regression model:

1. Build the Lasso Model.
2. Determine the regularization parameter λ.
3. Calculate the regression coefficients using least squares and the L1 penalty.
4. Select parameters (tuning) by cross-validating to find the best λ.
5. Evaluate the odel using MSE and RMSE metrics.
6. Interpret the Results

c. Support Vector Regression (SVR)

Support Vector Regression (SVR) is a development of the Support Vector Machine (SVM) used for regression problems. Its working principle is to find the best regression function that can predict data with a certain error tolerance (ϵ) while maintaining a maximum margin. SVR does not attempt to minimize the error of each data point but rather ensures that most of the data falls within the tolerance margin (ϵ-tube). SVR has several advantages, including the ability to capture non-linear relationships using kernels, robustness against outliers, suitability for small- to medium-sized data with high dimensions, and good generalizability because SVR focuses on support vectors (Drucker et al., 1997).Its main functions:

$$Mf(x) = w.\phi(x) + b \quad (4)$$

With the provision of

$$|y_i - f(x_i)| \leq \epsilon \quad (5)$$

Where

$\phi(x)$ = mapping function to high-dimensional space using kernel),
ϵ = error tolerance margin,
C = penalty parameter to control the trade-off between margin and error.

SVR Implementation:
(1) Data Preparation and Preprocessing
(2) Kernel Selection
(3) Parameter Determination
(4) SVR Model Training
(5) Model Evaluation
(6) Interpretation and Implementation.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

d. Decision Tree

Decision Tree (DT) is a supervised learning method used for classification and regression. This model is in the form of a tree structure with nodes (tests on features), branches (decision results), and leaves (output/class or predicted values). The concept is similar to the human decision-making process: breaking down problems into sequential questions. The advantages of decision trees include being easy to understand and interpret, not requiring much data pre-processing, being able to handle both numerical and categorical data, being fast and efficient for medium-sized datasets, being able to capture interactions between variables naturally, and being able to be used as the basis for ensemble methods (Breiman, 2001a).

Steps of the Decision Tree algorithm (Quinlan, 1986):

1. Attribute Selection: The algorithm selects the best features for splitting using measures such as Information Gain (based on Entropy), Gini Index, and Variance Reduction (for regression).
2. Node Creation
3. Recursion (Recursive Partitioning)
4. The splitting process continues until all data is classified or a stopping criterion (e.g., maximum tree depth) is met.
5. Stopping Criteria: If there are no more features to split and all samples in a node are homogeneous, the maximum tree depth is reached.
6. Pruning: To reduce overfitting, the tree is pruned by removing insignificant branches.
7. Prediction: The finished tree is used to predict new data.

e. Random Forest

Several decision trees are combined in Random Forest, a bagging-based ensemble learning technique, to increase prediction accuracy and lower the chance of overfitting.Only a random subset of features is taken into consideration at each split, and each tree is trained on a random portion of the data (by bootstrap sampling). Regression utilizing the average of the trees' predictions and majority vote of the trees is used to classify the final forecast (Breiman, 2001b). Among the benefits of Random Forest are its high accuracy, low overfitting, flexibility, robustness to missing data, ability to handle huge data sets with numerous features and samples, ability to directly evaluate feature relevance, and ability to function well with imbalanced data (Genuer et al., 2017).

Random Forest Steps (Breiman, 2001a):

1. Take bootstrap samples from the original dataset to create a subset of the training data for each tree.
2. Build a decision tree for each data subset.
3. At each node, select a random subset of features.
4. Determine the best split based on that subset.
5. Continue building the tree until a certain depth is reached or the tree cannot be further split.
6. Repeat the process for a certain number of trees (e.g., 100, 500, or 1000).
7. Combine the tree predictions by averaging the predicted values (regression).

f. Gradient Boosting

Gradient Boosting (GB) is a boosting-based ensemble learning algorithm that combines many weak models (usually shallow decision trees) sequentially. Unlike bagging (e.g., Random Forest), which builds trees in parallel, boosting works sequentially: each new model is built to correct the errors of the previous model. The main idea of GB is to minimize the loss function using the gradient descent method (Friedman, 2001).

Gradient Boosting Steps (Trevor Hastie, Robert Tibshirani, 2009).

1. Initialize the initial model with a simple prediction, for example, the target mean (for regression).
2. Calculate the residual (error), which is the difference between the actual value and the model prediction.
3. Fit a weak decision tree to the residual to learn the error pattern.
4. Update the model by adding the new tree prediction, multiplied by the learning rate (η).

$$F_m(x) = F_{m-1}(x) + \eta . h_m(x) \quad (6)$$

Where $h_m(x)$ is the weak model at the mth iteration.

5. Repeat the process until the number of iterations is reached or the error does not decrease significantly.
6. The final output is a combination of all weak models.

g. K-Nearest Neighbors (KNN)

K-Nearest Neighbors Regression (KNN Regression) is a non-parametric method used to predict target values by calculating the average (or a certain weighting function) of the target value at the K nearest neighbors of a data point based on the distance in the feature space. This algorithm does not build an explicit model like linear regression but rather stores all training data and makes predictions when new data comes in (lazy learning) (Altman, 1992). The advantages of KNN Regression include being simple and easy to understand, not requiring data distribution assumptions, being flexible, being able to capture local patterns in the data, and performing well for small to medium datasets with moderate dimensions (Trevor Hastie, Robert Tibshirani, 2009).

KNN Regression Steps (Cover & Hart, 1952):

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

1. Determine the parameter K (the number of neighbors to consider).
2. Calculate the distance between the new data and all training data (e.g., Euclidean, Manhattan, or Minkowski distance).
3. Identify the K nearest neighbors based on the smallest distance.
4. Predict the target value by calculating the average (or distance-weighted average) of the target values of these K neighbors.

$$\hat{y}(x) = \frac{1}{K}\sum_{i=1}^{K} y_i \quad (7)$$

Or if it is weighted:

$$\hat{y} = \frac{\sum_{i=1}^{K} \frac{1}{d(x,x_i)} y_i}{\sum_{i=1}^{K} \frac{1}{d(x,x_i)}} \quad (8)$$

Where $d(x,x_i)$ is the distance between the test point and its i-th neighbor.

1. Repeat the process for each new data point you want to predict.

D. Model Evaluation

Model performance was evaluated using the Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) metrics. These two metrics were chosen because they are able to measure prediction accuracy and detect potential overfitting (Chicco et al., 2021). Smaller MSE and RMSE values indicate better model performance in approximating actual glucose values.

III. RESULTS AND DISCUSSION

A. Data preprocessing

The data analysis process begins with normalizing physiological sensor data (body temperature, BPM, SpO2, PPG, and NIR signals) to align the feature scales so that all variables have a similar value range before being processed by the machine learning algorithm. After all features have been normalized, the next stage is data visualization using boxplots. The purpose of visualization using boxplots is to show data distribution, identify outliers by seeing whether there are values that are much higher or lower than the majority of the data that can be immediately seen, assess symmetry or skewness, and help data exploration (McGill et al., 1978). The boxplot generated from the research data is shown in Figure 1.

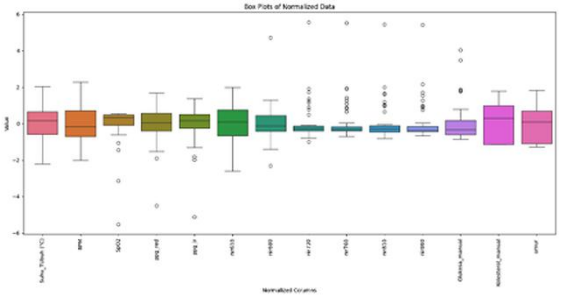


Figure 1. Boxplot of data after normalization.

A Figure 1 shows that the data distribution after normalization is relatively balanced, although there are still a number of outliers in certain PPG and NIR variables. The presence of these outliers can be interpreted as natural variations in physiological signals that are sensitive to body and environmental conditions, but do not disrupt the overall stability of the model. Therefore, the analysis can proceed to the next stage, dimensionality reduction.

B. Dimensionality reduction with Principal Component Analysis (PCA).

Principal Component Analysis (PCA) is performed to reduce the dimensionality of data. This dimensionality reduction plays a crucial role in improving the quality of regression modeling, especially when dealing with small data sets. One of its main benefits is addressing the problem of multicollinearity in sensor data, where predictor variables, particularly physiological variables, are highly correlated with each other, leading to instability in the estimation of regression coefficients. By transforming the original variables into a set of orthogonal principal components, PCA can generate new, correlation-free predictors, enabling the blood glucose prediction model to learn more stably and accurately (Jollife & Cadima, 2016). The results of dimensionality reduction on the research data are shown in Figure 2.

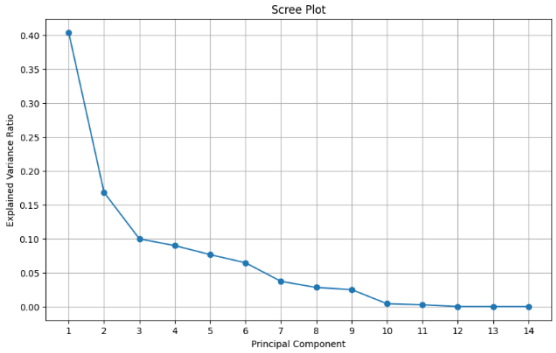


Figure 2. Principle Component Analysis (PCA) screen plot

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

The scree plot indicates that the first and second principal components account for the greatest share of the variation, contributing around 50% and 80%, respectively, when examined up to the sixth principal component. This suggests that six main components can capture the majority of the data's critical information, allowing the principal components to increase the model's efficiency without sacrificing any important information.

C. Predictive modeling

During the modeling phase, the performance of several regression algorithms was compared, including Random Forest Regressor, Lasso, Gradient Boosting Regressor, Linear Regression, Support Vector Regressor (SVR), Decision Tree Regressor, and K-Nearest Neighbors (KNN). A summary of the model goodness-of-fit evaluation results is shown in Table 1 below.

Table 1. Recapitulation of model goodness evaluation results.

ALGORITMA	MSE	RMSE
Random Forest Regressor	0.009	0.095
Lasso	0.012	0.110
Gradient Boosting Regressor	0.013	0.112
Linear Regression	0.015	0.122
SVR	0.022	0.147
Decision Tree Regressor	0.031	0.176
K-Nearest Neighbors Regressor	0.027	0.166

The evaluation metrics used to evaluate model goodness-of-fit in this study are Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) are widely used because they provide a quantitative overview of the difference between predicted and actual values. MSE is calculated as the average of the squared errors, thus imposing a greater penalty on large prediction errors (outliers). This is important to encourage the model to produce more consistent and accurate predictions. On the other hand, RMSE is the root of MSE, so it has the same units as the target variable. This makes RMSE easier to interpret in practical contexts (Chicco et al., 2021).

Based on the evaluation results shown in Table 1, it can be seen that Random Forest Regressor produces the best performance with MSE = 0.009 and RMSE = 0.095, lower than other algorithms. The superior performance of Random Forest is due to its ability to handle non-linear data, be robust against outliers, and overcome overfitting through ensemble learning (Breiman, 2001a; Ao et al., 2019). These results are in line with recent studies that demonstrate the effectiveness of Random Forest in predictions based on physiological and biomedical data. For example, (Q. Liu & Liu, 2017) showed that Random Forest is more accurate than linear regression for health biomarker prediction. In addition, (Woldaregay et al., 2019) reported that Random Forest is superior in blood glucose prediction compared to traditional regression models.

Meanwhile, linear algorithms such as Lasso and linear regression produced fairly good accuracy but still had higher errors, indicating limitations in capturing non-linear relationships between variables. Gradient boosting performed close to Lasso, although it tended to require more complex parameter tuning. Meanwhile, SVR, Decision Tree, and KNN had higher errors, likely influenced by sensitivity to noise and limited generalization to large-scale data. Thus, it can be concluded that ensemble learning-based approaches, particularly Random Forest, are more suitable for use in the context of predictions based on physiological multi-sensor data. However, these results need further evaluation with broader cross-validation tests and application to real-time data to test the model's robustness under dynamic field conditions.

The machine learning regression models used in this study demonstrated varying performance in non-invasive blood glucose prediction. Linear regression and SVR tended to be less stable due to their limitations in accommodating nonlinear relationships and sensitivity to multicollinearity and noise in physiological signals. The KNN model was also not optimal for high-dimensional data, while ANN required significantly larger data sets to generalize effectively. In contrast, ensemble models such as gradient boosting and especially random forest demonstrated significantly superior accuracy. Random Forest emerged as the best model due to its ability to manage complex physiological signals through a bagging mechanism, its robustness to noise, and its ability to address multicollinearity between sensors without requiring intensive hyperparameter tuning. The addition of PCA also improved the performance of certain models by reducing dimensionality and minimizing feature redundancy, particularly in models sensitive to correlations between variables such as linear regression, SVR, and ANN.

However, all models still faced challenges due to the highly variable physiological characteristics between individuals and the quality of sensor signals, which are susceptible to interference from movement or environmental conditions. The relatively limited dataset size also limited the ability of complex models to consistently generalize physiological patterns (Laha et al., 2022). Furthermore, even after model optimization, physiological changes may not always directly reflect blood glucose dynamics, potentially leading to prediction errors. These findings indicate that although Random Forest offers the most stable and accurate performance, non-invasive glucose

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

prediction approaches still require improved sensor data quality, better noise management, and dataset expansion for machine learning models to achieve more reliable performance under real-world conditions.

This study provides important insights that multimodal physiological data such as PPG, NIR, body temperature, and SpO_2 contain complex non-linear patterns, which are difficult to capture by traditional linear regression models. By applying ensemble learning algorithms, specifically Random Forest, this study successfully demonstrated better prediction performance with a low error rate. These results confirm that utilizing a combination of physiological signals can improve the accuracy of prediction systems while opening up opportunities for the development of non-invasive wearable and IoT-based health systems. The practical impact is quite significant, as it enables more affordable, painless real-time glucose monitoring and has the potential to reduce long-term medical costs through early detection of complications.

However, this study is not without limitations. The relatively small sample size may limit the model's generalizability to a broader population. While Random Forest is capable of capturing complex (non-linear) data patterns well, this algorithm is difficult to explain simply to users (medical professionals). The study also hasn't fully compared its performance with current deep learning algorithms, which are potentially superior in handling complex data.

IV. CONCLUSION

Based on the research results, the data normalization process successfully aligned the scales between physiological sensor variables, and the boxplot showed a relatively balanced data distribution despite the presence of natural outliers in the PPG and NIR signals. The dimensionality reduction stage using PCA was able to summarize more than 80% of the data variation with only six principal components, thereby increasing modeling efficiency while reducing the risk of multicollinearity. At the modeling stage, a comparison of the performance of the regression algorithms showed that Random Forest Regressor provided the best performance with MSE = 0.009 and RMSE = 0.095, superior to Lasso, Gradient Boosting, Linear Regression, SVR, Decision Tree, and KNN. This superiority demonstrates the effectiveness of the ensemble learning approach in handling non-linear data and is robust to outliers.

Practically, this study demonstrates that utilizing multimodal physiological data can improve the accuracy of non-invasive glucose level prediction, potentially applicable to IoT-based wearable systems for real-time monitoring. These findings support digital transformation efforts in healthcare, particularly in the early detection of diabetes complications and reducing reliance on invasive methods. However, the study still has limitations, particularly related to the small sample size, the limited interpretability of the Random Forest model for medical practitioners, and the lack of in-depth comparisons with current deep learning algorithms. Therefore, further research with larger datasets, long-term clinical validation, and exploration of deep learning models are needed to improve the system's generalizability and reliability in real-world conditions.

V. RECOMMENDATIONS

Further development can focus on completing and validating the assessment data for all faculty alternatives, particularly the raw criterion values used for FIP, and on rechecking the numerical consistency between the decision matrix, normalization, preference values, and final ranking. Subsequent work can also improve the DSS interface and provide systematic testing of the implemented calculation module so that the displayed rankings can be traced directly to the underlying assessment data

ACKNOWLEDGMENTS

We would like to thank Universitas Negeri Medan for funding this study. Especially the rector, LPPM, the dean of FMIPA Universitas Negeri Medan, and the department head who have provided support and facilities for this series of studies.

FUNDING INFORMATION

The authors would like to express their sincere gratitude to Universitas Negeri Medan for providing financial support for this research through the IPTEKS Research Grant Program administered by the Institute for Research and Community Service (LPPM), Universitas Negeri Medan, under Research Contract No. 0001/UN33.8/PPKM/PPI/2025. This financial support has contributed significantly to the successful completion of this research.

REFERENCES

Alanazi, A. (2022). Using machine learning for healthcare challenges and opportunities. In *Informatics in Medicine Unlocked* (Vol. 30). Elsevier Ltd. https://doi.org/10.1016/j.imu.2022.100924

Altman, N. S. (1992). An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. In *Source: The American Statistician* (Vol. 46, Number 3).

Amalia Dewayanti, A., & Utami, H. (n.d.). *ESTIMASI ROBUST PADA MODEL REGRESI UNTUK MENANGANI OUTLIER DAN HETEROSKEDASTISITAS (ROBUST ESTIMATION IN REGRESSION MODEL*

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

FOR HANDLING OUTLIER ANDHETEROSCEDASTICITY). *03*(01), 2021.

Ayu, G., & Srinadi, M. (2014). *Pengaruh Outlier Terhadap Estimator Parameter Regresi dan Metode Regresi Robust*. https://www.researchgate.net/publication/285592493

Bagaskara, K., Perdana, H., Aprizkiyandari, S., & September, D. (2024). PENERAPAN SUPPORT VECTOR REGRESSION KERNEL LINIER DALAM MERAMALKAN HARGA CPO INDONESIA Informasi Artikel ABSTRAK Sejarah Artikel. *Equator: Journal of Mathematical and Statistical Sciences (EJMSS)*, *3*(2).

Belyadi, H., & Haghighat, A. (2021). Supervised learning. In *Machine Learning Guide for Oil and Gas Using Python* (pp. 169–295). Elsevier. https://doi.org/10.1016/B978-0-12-821929-4.00004-4

Breiman, L. (2001a). RANDOM FORESTS. *Random Forest*, *7*(6), 1–33. https://doi.org/10.14569/ijacsa.2016.070603

Breiman, L. (2001b). Random Forests. In *Machine Learning* (Vol. 45). https://doi.org/https://doi.org/10.1023/A:1010933404324

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, *7*, 1–24. https://doi.org/10.7717/PEERJ-CS.623

Cover, T. M., & Hart, P. E. (1952). Nearest Neighbor Pattern Classification. *IEEE TRANSACTIONS ON INFORMATION THEORY*, *24*(1), 335–342.

Emerson, H., McConville, R., & Guy, M. (2023). *The Safety Challenges of Deep Learning in Real-World Type 1 Diabetes Management*. http://arxiv.org/abs/2310.14743

Friedman, J. H. (2001). I1999 REITZ LECTURE GREEDY FUNCTION APPROXIMATION: A GRADIENT BOOSTING MACHINE. *The Annals of Statistics*, *29*(5), 1189–1232.

Genuer, R., Poggi, J. M., Tuleau-Malot, C., & Villa-Vialaneix, N. (2017). Random Forests for Big Data. *Big Data Research*, *9*, 28–46. https://doi.org/10.1016/j.bdr.2017.07.003

Hameed, H., & Kleinberg, S. (2020). Comparing Machine Learning Techniques for Blood Glucose Forecasting Using Free-living and Patient Generated Data. In *Proceedings of Machine Learning Research* (Vol. 85). https://github.com/health-

Jollife, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. In *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* (Vol. 374, Number 2065). Royal Society of London. https://doi.org/10.1098/rsta.2015.0202

Kinasih, A. N. S., Handayani, A. N., Ardiansah, J. T., & Damanhuri, N. S. (2024). Comparative analysis of decision tree and random forest classifiers for structured data classification in machine learning. *Science in Information Technology Letters*, *5*(2), 13–24. https://doi.org/10.31763/sitech.v5i2.1746

Laha, S., Rajput, A., Laha, S. S., & Jadhav, R. (2022). A Concise and Systematic Review on Non-Invasive Glucose Monitoring for Potential Diabetes Management. In *Biosensors* (Vol. 12, Number 11). MDPI. https://doi.org/10.3390/bios12110965

Liu, K., Li, L., Ma, Y., Jiang, J., Liu, Z., Ye, Z., Liu, S., Pu, C., Chen, C., & Yi, W. (2023). Machine Learning Models for Blood Glucose Level Prediction in Patients With Diabetes Mellitus: Systematic Review and Network Meta-Analysis. *JMIR Medical Informatics*, *11*(1). https://doi.org/10.2196/47833

Liu, Q., & Liu, M. (2017). Energy-efficient clustering algorithm based on game theory for wireless sensor networks. *International Journal of Distributed Sensor Networks*, *13*(11). https://doi.org/10.1177/1550147717743701

McGill, R., Tukey, J. W., & Larsen, W. A. (1978). Variations of box plots. *American Statistician*, *32*(1), 12–16. https://doi.org/10.1080/00031305.1978.10479236

Quinlan, J. R. (1986). Induction of Decision Trees. In *Machine Learning* (Vol. 1).

Shi, Y., Yang, K., Yang, Z., & Zhou, Y. (2022). Primer on artificial intelligence. In *Mobile Edge Artificial Intelligence* (pp. 7–36). Elsevier. https://doi.org/10.1016/b978-0-12-823817-2.00011-5

Soleh, A. M., & Aunuddin. (2013). Lasso: An Alternative Solution for Selection and Shrinkage Linear Regression Models. *FSK : Indonesian Journal of Statistics*, *18*(1), 21–27.

Sung, J. J. Y. (2024). Artificial intelligence and the future of medicine. In *Artificial Intelligence in Medicine: From Ethical, Social, and Legal Perspectives* (pp. 1–12). Elsevier. https://doi.org/10.1016/B978-0-323-95068-8.00001-7

Trevor Hastie, Robert Tibshirani, J. F. (2009). The Elements of Statistical learning. In *Spring Series in Statistic*.

van Doorn, W. P. T. M., Foreman, Y. D., Schaper, N. C., Savelberg, H. H. C. M., Koster, A., van der Kallen, C. J. H., Wesselius, A., Schram, M. T., Henry, R. M. A., Dagnelie, P. C., de Galan, B. E., Bekers, O., Stehouwer, C. D. A., Meex, S. J. R., & Brouwers, M. C. G. J.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

(2021). Machine learning-based glucose prediction with use of continuous glucose and physical activity monitoring data: The Maastricht Study. *PLoS ONE*, *16*(6 June). https://doi.org/10.1371/journal.pone.0253125

Woldaregay, A. Z., Årsand, E., Walderhaug, S., Albers, D., Mamykina, L., Botsis, T., & Hartvigsen, G. (2019). Data-driven modeling and prediction of blood glucose dynamics: Machine learning applications in type 1 diabetes. In *Artificial Intelligence in Medicine* (Vol. 98, pp. 109–134). Elsevier B.V. https://doi.org/10.1016/j.artmed.2019.07.007

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.