JURNAL INFOTEL
VOL. 16, NO. 2, MAY 2024, PP. 353–368. DOI:10.20895/INFOTEL.V16I2.1117

RESEARCH ARTICLE

In-Depth Exploration and Comparison of Machine Learning Performances for Early-Stage Diabetes Risk Prediction

Nor Kumalasari Caecar Pratiwi

Department of Electrical Engineering, Telkom University, Bandung 40257, Indonesia

Corresponding email: caecarnkcp@telkomuniversity.ac.id

Received: December 29, 2023; Revised: February 28, 2024; Accepted: March 14, 2024.

Abstract: Diabetes mellitus is distinguished by an inability of the human system to produce insulin on an ongoing basis, as well as by the inefficient utilization of the insulin hormone, resulting in an elevated level of blood glucose. Global diabetes rates have nearly doubled since 1980, reaching 9.3% among adults. Alarmingly, of the 463 million individuals with diabetes, 50.1% are unaware of their condition. Indonesia ranks seventh globally with 10.7 million diabetes cases. In 2019, it was fifth globally for adults (20–79 years) with undiagnosed diabetes. This silent epidemic demands urgent attention and comprehensive strategies for early detection and management. In recent years, researchers have increasingly studied machine learning for early diabetes recognition. In this study, we aim to predict early-stage diabetes risk by utilizing 16 health condition features. We explore 12 distinct machine learning algorithms, applying a hyperparameter grid to tune each algorithm. This involves systematically testing combinations of hyperparameters to identify the optimal settings for achieving the most accurate and reliable predictive models. The results indicate that the LightGBM algorithm achieved the highest accuracy of 0.9692. By contrast, the logistic regression and naïve Bayes algorithms demonstrated the lowest performance, each with an accuracy of 0.8923. The implications of these results underline the capability of employing machine learning algorithms to precisely and effectively detect individuals susceptible to diabetes, enabling the implementation of individualized healthcare approaches.

Keywords: diabetes mellitus, LightGBM, logistic regression, machine learning, naïve Bayes, prediction

 ISSN: 2085-3688; e-ISSN: 2460-0997

1 Introduction

As per the International Diabetes Federation (IDF) definition, diabetes mellitus is characterized by a deficiency in long-term insulin production within the body, also marked by the ineffective utilization of the insulin hormone, leading to an elevation in blood glucose levels [1]. Diabetes mellitus represents a significant public health concern, impacting approximately 424.9 million individuals globally, one-third of those affected are aged 65 years or older [2]. By 2035, the larger Asia Pacific Region expects a 30–40% increase in diabetes prevalence [3]. The alarming surge in diabetes prevalence on a global scale reveals a stark reality, since 1980, the rates have almost doubled, soaring from 4.7% to 9.3% among adults [4]. Shockingly, within the staggering count of 463 million individuals grappling with diabetes, an unsettling 50.1% remain oblivious to their condition [5]. Indonesia takes the seventh spot in the global ranking, bearing the weight of 10.7 million cases of diabetes mellitus [6]. If not properly managed, diabetes can lead to various complications affecting different organs, potentially contributing to morbidity and mortality. This stark figure emphasizes the urgent need for strategic healthcare efforts and comprehensive measures to tackle the escalating diabetes's prevalence.

Machine learning, allows machines to access and learn from data automatically [7]. Several prior researchers have undertaken diabetes detection utilizing image inputs. Harahap developed a system for classifying foot ulcers in diabetic patients using convolutional neural network (CNN) [8]. Study [9], digital pathology and machine learning models were utilized to analyze human pancreata images. Aslan and Sabanci [10] transform numerical data into images, emphasizing feature importance to improve CNN models for early diabetes diagnosis. The result demonstrates the images' robustness in early diabetes diagnosis. The study [11] introduces a CNN to predict diabetes based on retinal images, to classify into diabetic or nondiabetic classes. In both studies [12] and [13], there were investigations into the early detection of diabetes with panoramic tongue imaging.

Diabetes prediction models can indeed be developed using raw data that includes various features related to health conditions, lifestyle, age, and more. Figure 1 shown a general representation of data collection for diabetes prediction [14]. Using raw data for diabetes prediction has the advantage of incorporating a holistic view of an individual's health and lifestyle, and it may reduce computational complexity.

Febrian conducted research employing the k-nearest neighbors (KNN) and naïve Bayes algorithms for predicting diabetes based on various health attributes [15]. In addition to health attributes, lifestyle factors can also serve as features for the prediction process, as research conducted by Mujumdar and Vaidehi [16]. Tasin *et al.* [17] predicted diabetes by utilizing multiple features including glucose levels, insulin, age, blood pressure, and body mass index (BMI), then the XGBclassifier exhibited the most optimal performance, with accuracy of 81%. Qin *et al.* [18] conducted a comparative analysis of five distinct machine-learning for predicting diabetes. The researchers utilized lifestyle statistics obtained from NHANES database. Ahmed and Li [19] introduced a diabetes predictive model for enhanced categorization, incorporating external variables influencing diabetes. Ganie *et al.* [20] experimented with five boosting model with PIMA dataset to predict diabetes. Gradient boosting achieving an impressive accuracy rate of 92.85%. Employing the Smote-Tomek Link and random forest algorithms, study [21] deals with the issue of dataset imbalance. Eight attributes were used as inputs in this study, and one trait was used as output.

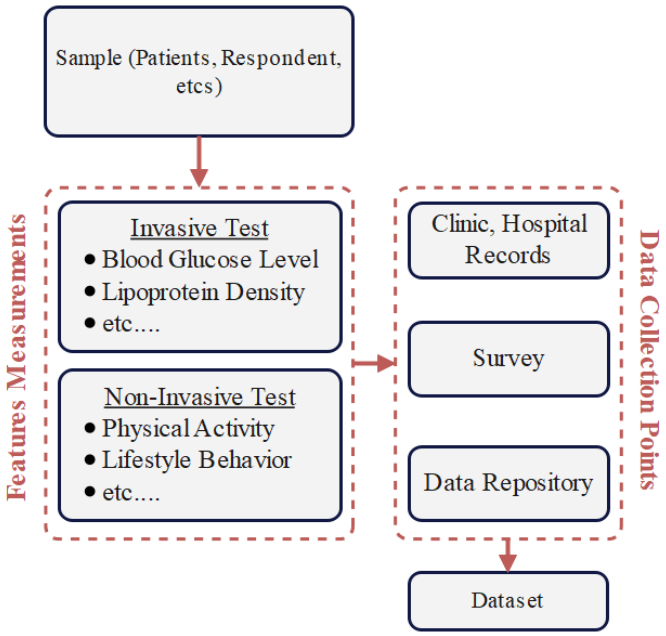

Figure 1: Data recording process for the diabetes predictions.

The review suggests that machine learning holds potential in forecasting diabetes at an early stage. It also noted the importance of incorporating raw data to gain a thorough understanding of an individual's health and lifestyle, while also streamlining computational complexity. Therefore, the study systematically investigates and compares the performance of twelve distinct machine learning algorithms for the early-stage prediction of diabetes risk, utilizing a dataset comprising sixteen health condition features. This study aims to identify the most accurate and reliable predictive models among the tested algorithms, with a focus on optimizing hyperparameters through a rigorous hyperparameter grid search. By achieving a comprehensive understanding of the strengths and limitations of each algorithm in predicting early-stage diabetes risk, this research seeks to contribute to the development of effective and individualized healthcare strategies for diabetes prevention and management.

2 Research Method

The objective of this study was conduct a comprehensive evaluation of twelve machine learning models in forecasting the risk of diabetes in individuals. For each algorithm, we employed a hyperparameter grid—a structured approach defining potential values for tuning during the hyperparameter optimization process. This tuning process carefully try out different conditions of hyperparameter to find the best machine learning model. Figure 2 delivers a detailed overview of the research approach.

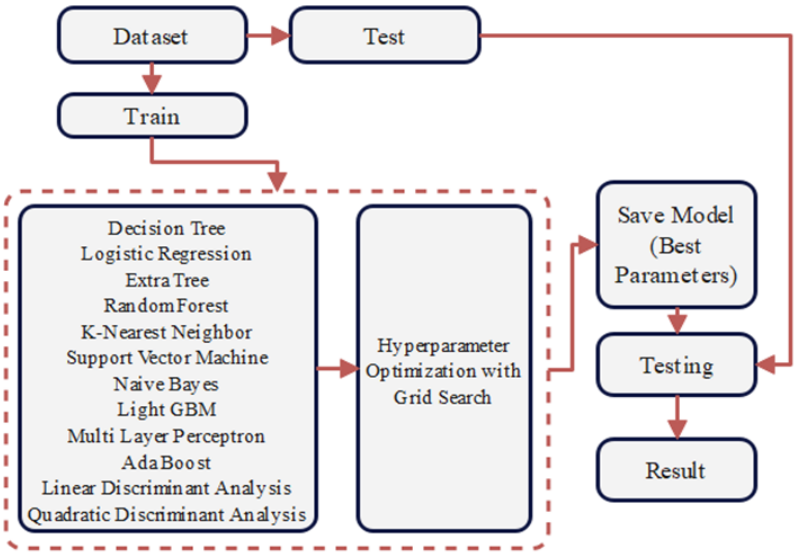

Figure 2: Research methodology for diabetes prediction.

2.1 Dataset

This study utilizes public data accessible on Kaggle [22]. The set of data consists of vital sign and symptom information pertaining to those who are at risk of getting diabetes or who exhibit early symptoms of the disease. Table 1 offers a detailed description of the attributes presented in the dataset. The dataset comprises information from 520 individuals, with 320 identified as positive for diabetes and the remaining as negative.

The selection of health condition features for diabetes risk prediction is crucial. each feature should have a strong rationale backed by either statistical analysis or relevant literature demonstrating its association with diabetes. Age is a well-established risk factor for diabetes. As individuals age, the risk of insulin resistance and impaired glucose tolerance increases [23, 24]. Gender differences exist in the prevalence and risk factors of diabetes. For instance, women with a history of gestational diabetes are at higher risk of developing type 2 diabetes later in life. Additionally, hormonal differences between males and females can influence insulin sensitivity and glucose metabolism [25–27].

Excessive urination is a classic symptom of diabetes, particularly when blood sugar levels are elevated. It occurs due to the kidneys attempting to remove excess glucose from the blood by excreting it in urine. Polyuria is a hallmark feature of diabetes and is strongly associated with the condition [28]. Polydipsia, intense thirst is closely linked with polyuria in diabetes. When individuals experience excessive urination, they become dehydrated, leading to increased thirst as the body attempts to compensate for fluid loss [29].

Unexplained weight loss can be an early sign of diabetes [30], it occurs due to the body's inability to properly utilize glucose for energy, leading to the breakdown of fat and muscle tissue for fuel. General weakness or fatigue, polyphagia, visual blurring, , muscle stiffness , irritability, delayed healing, partial paresis, genital thrush, alopecia, obesity, and itching are nonspecific symptom that can occur in individuals with diabetes [31, 32]. Each of these

Table 1: Dataset attributes for early-stage diabetes risk prediction

Attributes	Annotation
Age	Age distribution of the respondents (range at 16 – 65)
Sex	Gender of the participants (female or male)
Polyuria	Presence of excessive urination (yes or no)
Polydipsia	Intense thirst (yes or no)
Sudden Weight Loss	Significant weight decline (yes or no)
Weakness	Weakness in general (yes or no)
Polyphagia	Excessive hunger (yes or no)
Genital Thrush	Existence of Genital Thrush: Indicates whether there is the presence of fungal infection in the genital area (yes or no)
Visual Blurring	Refers to vision impairment or haziness (yes or no)
Itching	Indicates the presence of skin irritation or scratching (yes or no)
Irritability	Refers to the manifestation of irritability (yes or no)
Delayed Healing	Indicates a slower-than-normal healing process of wounds (yes or no)
Partial Paresis	Refers to a condition where there is a partial reduction in voluntary motor control (yes or no)
Muscle Stiffness	Indicates the existence of stiffness or inflexibility in muscles (yes or no)
Alopecia	Refers to the occurrence of hair loss or thinning of hair (yes or no)
Obesity	Indicates the presence of excess body weight or obesity (yes or no)
Class	The diabetes classification (Negative or Positive)

features contributes to the overall risk profile for diabetes and has been selected based on its documented association with the condition.

2.2 Machine Learning Method

Machine learning (ML) is a broad term encompassing a variety of algorithms that can make smart predictions using a set of data, allow machines to acquire knowledge without programming, and provide automatic data access and improved experience as the machine learns [7], [33, 34]. In healthcare industry, explainable machine learning models empower healthcare professionals to make informed, data-driven decisions, enabling personalized interventions and contributing to an enhanced quality of healthcare services [35].

A decision tree is a step-by-step process for determining the result of a function, denoted as $f(x)$. It involves conducting tests on the given input of x, outcome for each test guides sequentially until the function $f(x)$ is accurately determined [36]. A Decision tree simplifies decision-making by breaking down complex choices into simpler steps, offering more solutions to problems efficiently. It also uncovers relationships between input and target variables, making it effective for decision-making [37].

Logistic regression is a predictive model designed to assess the association between a categorical dependent variable (target), typically with nominal or ordinal scale, and an independent variable (predictor) that is categorical with interval or ratio scale [38]. Logistic regression is a widely employed statistical technique that facilitates the multivariate investigation of a binary dependent variable [39].

The extra tree, or extremely randomized tree, is an ensemble technique based on trees used in both supervised classification and regression scenarios. By employing randomization in the selection of cut-points for numerical input features, the goal is to have the

optimal cut-point account for a significant portion of the induced tree's variance [39]. In order to construct a robust model, both random forest and extra trees employ multiple decision trees; their primary distinction is in their feature selection methods [40].

In random forest, the best features randomly picked for each decision point. Meanwhile, extra trees use random features and random values for each decision point. The KNN algorithm is predominantly utilized for classification tasks, the concept is predicated on a parameter k that is variable and signifies the amount of 'nearest neighbors' [41]. The KNN works by determining the adjacent or neighboring points or neighbors from a training input in response to a specified data. This is done by finding the nearest distances to the data point. Once we identify the k nearest data points, we figure out the most common class by using a majority voting rule.

The core concept of support vector machines (SVM) revolves around decision functions hyperplanes, that adeptly differentiate between positive and negative data by maximizing the margins [42]. This optimization process tries to make a big gap between the nearest positive example and the hyperplane and make a big gap between the nearest negative example and the hyperplane. Naïve Bayes uses a simple probability-based classification using Bayes' theorem, assumes that whether a specific feature is there or not in a class is unrelated to the presence or absence of any other feature [43].

LightGBM is a current adaptation of the gradient boosting algorithm that enhances its capacity for scaling without compromising the algorithm's inherent effectiveness [44]. Operating in the feed-forward direction, the multi-layer perceptron (MLP) is an artificial neural network comprised of a minimum of three node layers: input, hidden, and output [45]. To facilitate training proses, MLP utilizes the supervised learning technique known as backpropagation.

Adaptive Boosting (AdaBoost) is an iterative algorithm in which a feeble classifier is modified in every iteration till the classification error rate achieves a predetermined level of minimization [46]. If the classifier accurately classifies, there should be a concurrent reduction in both the sample's weight and the probability of its selection. A prevalent classification technique, quadratic discriminant analysis (QDA) and linear discriminant analysis (LDA) both seek to distinguish data points by locating straight lines (in LDA) or curved surfaces (in QDA) [47]. LDA is simple and robust against increasing dimensionality, but its effectiveness relies on the assumption of equal covariances. On the other hand, QDA allows for different data variabilities but comes with a higher flexibility cost as it estimates more parameters, demanding a larger sample size.

2.3 Performance Parameters

To assess the efficacy of classification models, various performance metrics are employed, including accuracy, precision, recall, and F1-score. The overall correctness is measured by accuracy, which computes the ratio of true predictions to the total number of instances. Precision evaluates the capacity of the model to accurately detect positive instances. The recall function computes the capacity of the model to capture every positive instance. To achieve an equilibrium between recall and precision, the F1-score is computed using the harmonic mean of the metrics. These metrics aid in the assessment and comprehension of the performance of a classification model across various dimensions.

3 Results

In this section, exhaustive yet succinct study results for the twelve employed algorithms are presented. As delineated in the preceding section, this investigation employs a grid search parameter that autonomously seeks the optimal hyperparameter combination from a set of multiple hyperparameters. Explanation of the grid hyperparameters assigned to each algorithm is provided in Table 2. The table outlines the hyperparameter grid combinations employed for fine-tuning each machine learning predictor in the analysis.

Table 2: Hyperparameter grid combination for each predictor

Predictor	**Hyperparameter Grid Combination**
Decision Tree	Criterion (gini or entropy), Splitter (best or random), max_depth (10,20,30), min_samples_split (2,5,10), and min_samples_leaf (2 or 4)
Logistic Regression	Penalty (l1 or l2), C value (0.001, 0.01, 0.1, 1, 10, 100), solver (liblinear or saga)
Extra Tree	n_estimators (50, 100, 150), max_depth (None, 10, 20), min_samples_split (2, 5, 10), and min_samples_leaf (1, 2, 4)
Random Forest	n_estimators (100, 200, 300), criterion (gini, entropy), max_depth (10, 20, 30), min_samples_split (2, 5, 10), min_samples_leaf (2, 4), max_features (sqrt, log2)
K-Nearest Neighbor	k value (range from 1 - 21), Distance (Manhattan or Euclidean)
Support Vector Machine	Regularization parameter (0.1, 1, 10, 100), kernel (linear, rbf, poly), gamma (scale, auto, 0.1, 1), degree (2, 3, 4)
LightGBM	num_leaves (31, 63, 127), learning_rate (0.05, 0.1, 0.2), n_estimators (50, 100, 200), subsample (0.8, 0.9, 1.0), colsample_bytree (0.8, 0.9, 1.0)
Multi-Layer Perceptron	Size of Hidden Layer ((50,), (100,), (50, 50)), activation (relu, tanh, logistic), solver/ optimizer (adam, sgd), alpha (0.0001, 0.001, 0.01), learning_rate (constant, invscaling, adaptive)
AdaBoost	n_estimators (50, 100, 200), learning_rate (0.05, 0.1, 0.2)

For decision tree, parameters criterion refers to the metric used to measure the quality of a split at each node of the tree. The two commonly used criteria are gini impurity and entropy. The splitter decides how the tree makes splits at each step. When set to 'best,' it looks at all features and picks the one that gives the most information gain. When set to 'random,' it randomly picks a group of features and chooses the best split from that group. The 'max_depth' setting determines how deep the decision tree can go. A deeper tree can understand more intricate patterns in the training data, but it might also lead to overfitting. The 'min_samples_split' decides smallest samples needed to split a node, and 'min_samples_leaf' sets the minimum samples required in a leaf node. For logistic regression explores hyperparameters such as penalty, C value, and solver type. The 'penalty' determines the type of regularization applied to the model.

Regularization is a method employed to avoid overfitting by including a penalty term in the loss function. The 'C' hyperparameter is the inverse of the regularization strength. Smaller values of C lead to stronger regularization, encouraging simpler models with smaller coefficients. The 'solver' selects the optimization algorithm during the training of the logistic regression model. Liblinear is suitable for smaller to medium-sized datasets, while 'saga' is a good option for larger datasets. While Extra Trees and Random Forests are both tree-based ensemble methods and share some similarities, they do have differences in

how they introduce randomness during the training process. The KNN algorithm involves two key hyperparameters, k represents the number of nearest neighbors. Picking the correct value is essential. A small k may make the model sensitive to noise, while a large k might result in oversmoothing and the loss of crucial patterns in the data. The distance metric is crucial in determining how we measure the space between data points. Manhattan distance calculates the sum of absolute differences in coordinates, while Euclidean distance measures the straight-line distance between two points.

SVM algorithm involves several hyperparameters that play a crucial role in shaping the model. The regularization parameter, C, determines how much emphasis is given to reducing both training and testing errors in a balanced manner. A smaller encourages a simpler decision boundary, while a larger allows for a more complex decision boundary that may fit the training data more closely. SVMs have the capability to employ various kernel functions to transform input data into a higher-dimensional space. Linear kernel suitable for linearly separable data., radial basis function (RBF) suitable for non-linear data, commonly used when the decision boundary is complex and not easily linear, and polynomial for polynomial decision boundaries. Gamma determines the reach or influence of a single training example in a model. A low gamma means a far reach, and a high gamma means a narrow reach. The degree is specific to the polynomial kernel and represents the degree of the polynomial used to find the decision boundary. Higher degrees allow the model to fit more complex curves.

LightGBM designed for distributed and efficient training. The 'num_leaves' parameter decides the most leaves a tree can have. A higher value lets the model understand more complicated patterns but could result in overfitting. The learning rate decides how much each tree contributes to the final prediction. A lower learning rate needs more trees for the model to come together, but it often leads to better generalization. N_estimators refer to number of boosting rounds or trees to be built. Adding more trees can improve the accuracy of the model, but it also demands more computational resources. Subsample controls the fraction of samples used for training each tree. The MLP is a form of artificial neural network that includes several layers of neurons. Hidden_layer_sizes indicate the quantity of neurons in each hidden layer. The activation function determines the output of each neuron in the network. The solver used to update the weights during training process, two common choices are stochastic gradient descent (SGD) and adaptive moment estimation (Adam). Alpha represents the L2 regularization term. The crucial hyperparameter for AdaBoost, an ensemble learning method that combines predictions from multiple weak learners (usually decision trees) to form a strong learner, is n_estimators, defines the count of weak learners (base models) to be trained.

The F1-score, accuracy, precision, and recall for each predictor algorithm are detailed in Table 3. The findings of the present study line up with what has been demonstrated in prior investigations, machine learning algorithms still an accurate early prediction of diabetes based on large dataset, in this case 12 attributes of health problems and lifestyle. In this study, LightGBM outperformed other algorithms. LightGBM is known for its efficiency in handling large datasets due to its gradient-based approach and leaf-wise tree growth strategy. Our dataset have contained a large number of features or instances, making LightGBM a suitable choice for efficient computation and modeling. The dataset exhibited class imbalance, with one class (negative cases of diabetes) being significantly less frequent than the other.

LightGBM's ability to handle imbalanced data through techniques like gradient-based learning and bagging may have contributed to its superior performance. Diabetes risk prediction often involves complex, non-linear relationships between predictor variables and the target outcome. LightGBM's capability to model non-linear relationships efficiently, especially with high-dimensional data, might have provided it with an advantage over other algorithms like logistic regression or decision trees. In conclusion, LightGBM's efficiency in handling large datasets, robustness to class imbalance, and have an ability to model non-linear could collectively explain its superior performance in our study.

Table 3: Performance for each machine learning predictors

Machine Learning Predictor	Accuracy	Precision	Recall	F1-score
LightGBM	0.9692	0.9708	0.9692	0.9690
Extra Trees	0.9615	0.9639	0.9615	0.9612
K-Nearest Neighbor	0.9615	0.9622	0.9615	0.9613
Ada Boost	0.9615	0.9622	0.9615	0.9613
Decision Tree	0.9538	0.9551	0.9538	0.9535
Support Vector Machine	0.9538	0.9572	0.9538	0.9533
Quadratic Discriminant Analysis	0.9462	0.9506	0.9462	0.9454
Random Forest	0.9385	0.9443	0.9385	0.9375
Multi-Layer Perceptron	0.9385	0.9414	0.9385	0.9378
Linear Discriminant Analysis	0.9077	0.9103	0.9077	0.9081
Logistic Regression	0.8923	0.8962	0.8923	0.8906
Naïve Bayes	0.8923	0.8993	0.8923	0.8900

The study rigorously explores and evaluates twelve distinct machine learning algorithms, providing a comprehensive overview of their performance in predicting early-stage diabetes risk. Utilizing a grid search parameter approach allows for the systematic optimization of hyperparameters, enhancing the precision and reliability of the predictive models. Detailed explanations of the hyperparameter grid combinations for each algorithm are provided, facilitating transparency and reproducibility of the study's methodology. The inclusion of diverse algorithms and hyperparameter settings enables a thorough comparison of predictive capabilities, offering valuable insights into the strengths and weaknesses of each approach. While the study examines a wide range of machine learning algorithms, it may not encompass all possible methodologies or variations, potentially limiting the generalizability of the findings. The performance of the algorithms is assessed using a single dataset, which may not fully represent the diversity of populations or healthcare settings, potentially affecting the generalizability of the results. Additional validation studies using external datasets or real-world clinical data are needed to confirm the robustness and applicability of the predictive models in practical healthcare settings.

A comparison of the accuracy, precision, recall, and F1-score of each predictor algorithm is depicted in Figure 3 to Figure 6, respectively. The degree of performance parameter achieved by each algorithm does not vary substantially.

4 Discussion

Based on the findings outlined in the preceding section, it is noticeable that in general, each machine learning algorithm employed can generate timely prognostications regarding a

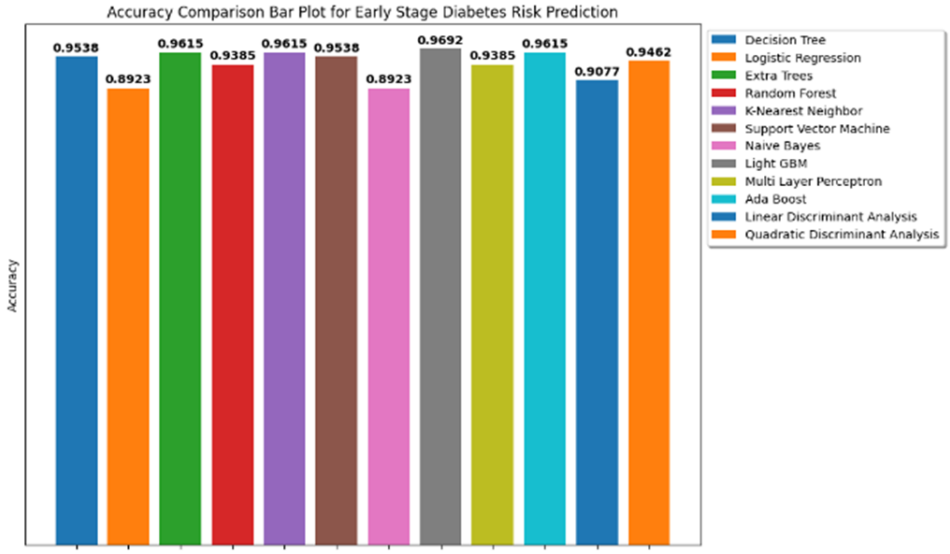

Figure 3: Accuracy for each Predictors.

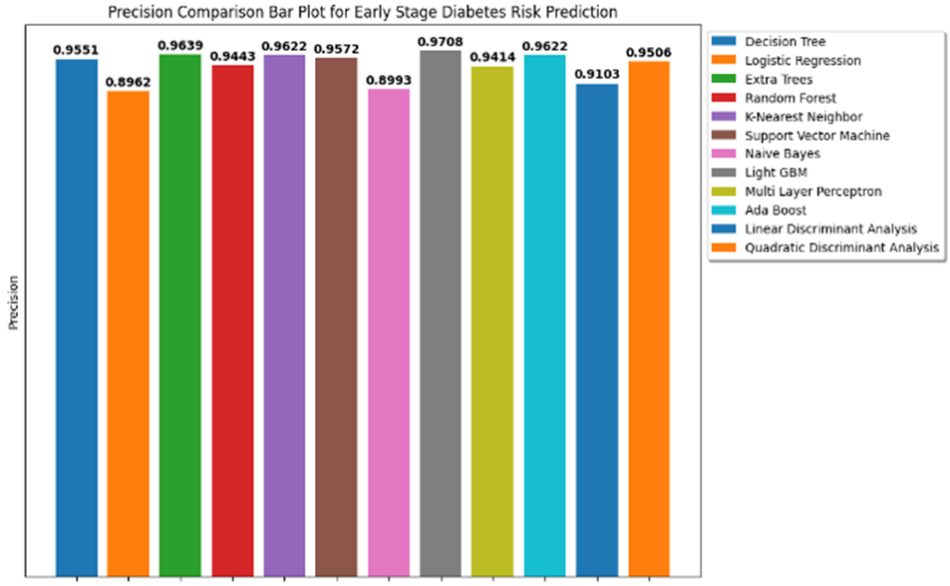

Figure 4: Precision for each predictors.

possibility of getting diabetes. The results presented in the table indicate that the Light-GBM algorithm achieved the highest performance with an accuracy of 0.9692. By contrast, the logistic regression and naïve Bayes algorithms demonstrated the lowest performance, each with an accuracy of 0.8923. The success of a machine learning algorithm in predicting the risk of diabetes depends on various factors, including the quality of the data, the algo-

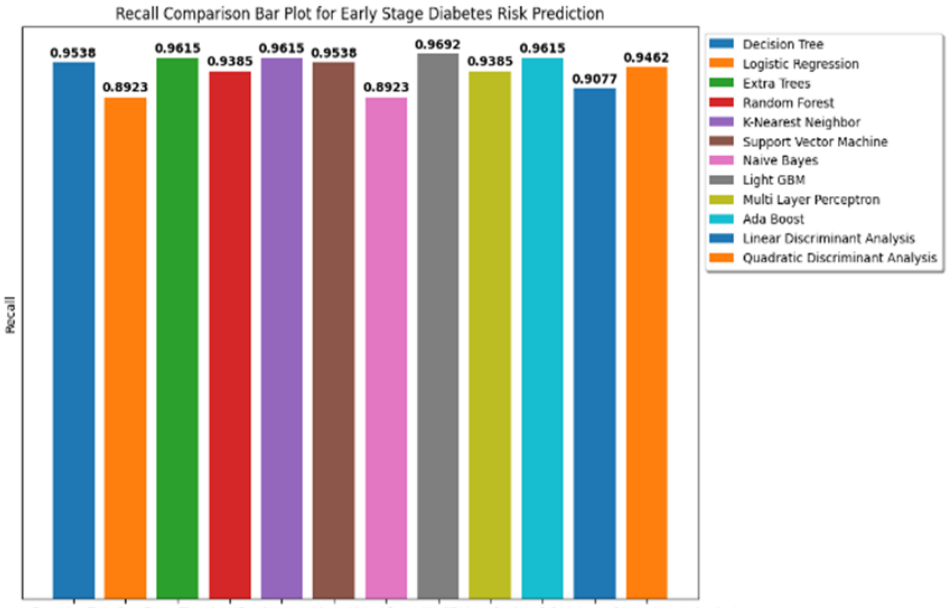

Figure 5: Recall for each predictors.

rithm used, and the hyperparameter-tuning. High-quality, well-curated data is crucial for effective predictions. Fine-tune the hyperparameters of the chosen algorithm to optimize its performance. The used of grid search or randomized search to find the optimal set of hyperparameters.

Table 4: Analysis regarding previous research

Method	Dataset	Result
Proposed Model (LightGBM)	520 recors with 16 attributes (features)	Acc = 0.97
Naïve Bayes [15]	Pima Indians Diabetes (with 8 independent variables / features)	Acc = 0.76
Logistic Regression [16]	800 records and 10 attributes	Acc = 0.96
XGBclassifier [17]	Private dataset of female patients in Bangladesh (203 records,6 attributes)	Acc = 0.81
CATBoost [18]	NHANES dataset (124,821 records with 18 diabetes-relevant factors)	Acc = 0.82
Decision Tree [19]	MCH dataset	Acc = 0.99
Gradient boosting [20]	Pima Indians Diabetes (with 8 independent variables / features)	Acc = 0.92

Table 4 provides the result comparison regarding previous similar research. Note that each study used different datasets, optimization techniques, and simulation configurations for assessment, potentially making direct comparisons inconclusive. The proposed model, utilizing LightGBM with 520 records and 16 features, stands out with an impressive accuracy of 0.97. In comparison, naïve Bayes on the Pima Indians Diabetes dataset achieves an accuracy of 0.76, while logistic regression, using 800 records and 10 attributes, performs

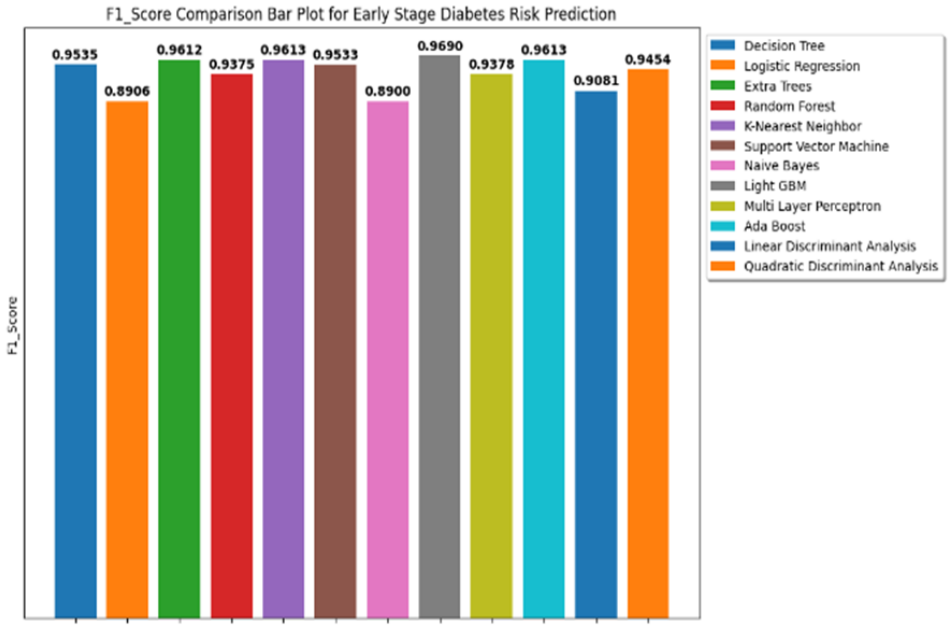

Figure 6: F1-score for each predictors.

well with an accuracy of 0.96. The XGBclassifier on a private dataset of female patients in Bangladesh (203 records, 6 attributes) achieves an accuracy of 0.81. CATBoost, applied to the NHANES dataset with 124,821 records and 18 diabetes-relevant factors, reaches an accuracy of 0.82. Decision tree, implemented on the MCH dataset, shows remarkable accuracy at 0.99. Lastly, gradient boosting on the Pima Indians Diabetes dataset achieves an accuracy of 0.92. The results highlight the efficacy of the proposed LightGBM model and demonstrate the diverse performance levels across different machine learning algorithms and datasets in predicting early-stage diabetes risk.

5 Conclusion

In conclusion, this study addresses the urgent need for effective strategies in combating the rising prevalence of diabetes mellitus. By leveraging machine learning algorithms, particularly the LightGBM model, the research achieves a remarkable accuracy of 0.9692 in predicting early-stage diabetes risk. This underscores the potential of machine learning in providing precise and timely prognostications for individuals susceptible to diabetes. The novelty of this research lies in its comprehensive evaluation of twelve distinct machine learning algorithms, coupled with a systematic hyperparameter optimization process. Through this approach, the study not only identifies the most accurate predictive model but also sheds light on the strengths and limitations of each algorithm. This contributes to the development of individualized healthcare strategies tailored to early diabetes detection and prevention. Moreover, the comparison with previous research highlights the superiority of the proposed LightGBM model in terms of accuracy, reaffirming its effectiveness in predicting

diabetes risk. However, it's important to acknowledge the limitations of the study, such as the reliance on a single dataset and the potential lack of generalizability across diverse populations and healthcare settings. For future research, it is recommended to validate the predictive models using external datasets or real-world clinical data to ensure robustness and applicability in practical healthcare settings. Additionally, exploring ensemble methods or hybrid models combining machine learning algorithms could further enhance prediction accuracy and reliability. Overall, this study paves the way for continued advancements in leveraging machine learning for early diabetes detection and personalized healthcare interventions.

References

[1] S. I. Oktora and D. Butar Butar, "Determinants of Diabetes Mellitus Prevalence in Indonesia," *Jurnal Kesehatan Masyarakat*, vol. 18, pp. 266–273, Nov. 2022.

[2] E. J. Kim, K. H. Ha, D. J. Kim, and Y. H. Choi, "Diabetes and the Risk of Infection: A National Cohort Study," *Diabetes & Metabolism Journal*, vol. 43, no. 6, p. 804, 2019.

[3] A. R. Cholil, D. Lindarto, T. G. D. Pemayun, W. Wisnu, P. Kumala, and H. H. S. Puteri, "DiabCare Asia 2012: diabetes management, control, and complications in patients with type 2 diabetes in Indonesia," *Medical Journal of Indonesia*, vol. 28, pp. 47–56, May 2019.

[4] C.-H. Jung, J. W. Son, S. Kang, W. J. Kim, H.-S. Kim, H. S. Kim, M. Seo, H.-J. Shin, S.-S. Lee, S. J. Jeong, Y. Cho, S. J. Han, H. M. Jang, M. Rho, S. Lee, M. Koo, B. Yoo, J.-W. Moon, H. Y. Lee, J.-S. Yun, S. Y. Kim, S. R. Kim, I.-K. Jeong, J.-O. Mok, and K. H. Yoon, "Diabetes Fact Sheets in Korea, 2020: An Appraisal of Current Status," *Diabetes & Metabolism Journal*, vol. 45, pp. 1–10, Jan. 2021.

[5] P. Saeedi, I. Petersohn, P. Salpea, B. Malanda, S. Karuranga, N. Unwin, S. Colagiuri, L. Guariguata, A. A. Motala, K. Ogurtsova, J. E. Shaw, D. Bright, and R. Williams, "Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition," *Diabetes Research and Clinical Practice*, vol. 157, p. 107843, Nov. 2019.

[6] A. Z. Safitri, R. N. Fajariyah, and E. Astutik, "Risk Factors of Diabetes Mellitus in Urban Communities in Indonesia (IFLS 5)," *Jurnal Berkala Epidemiologi*, vol. 9, p. 184, May 2021.

[7] M. M. Taye, "Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions," *Computers*, vol. 12, p. 91, Apr. 2023.

[8] M. Harahap, S. K. Anjelli, W. A. M. Sinaga, R. Alward, J. F. W. Manawan, and A. M. Husein, "Classification of diabetic foot ulcer using convolutional neural network (CNN) in diabetic patients," *JURNAL INFOTEL*, vol. 14, pp. 196–202, Aug. 2022.

[9] X. Tang, I. Kusmartseva, S. Kulkarni, A. Posgai, S. Speier, D. A. Schatz, M. J. Haller, M. Campbell-Thompson, C. H. Wasserfall, B. O. Roep, J. S. Kaddis, and M. A. Atkinson, "Image-Based Machine Learning Algorithms for Disease Characterization in the

Human Type 1 Diabetes Pancreas," *The American Journal of Pathology*, vol. 191, pp. 454–462, Mar. 2021.

[10] M. F. Aslan and K. Sabanci, "A Novel Proposal for Deep Learning-Based Diabetes Prediction: Converting Clinical Data to Image Data," *Diagnostics*, vol. 13, p. 796, Feb. 2023.

[11] M. Ragab, A. S. A.-M. AL-Ghamdi, B. Fakieh, H. Choudhry, R. F. Mansour, and D. Koundal, "Prediction of Diabetes through Retinal Images Using Deep Neural Network," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–6, June 2022.

[12] S. Balasubramaniyan, V. Jeyakumar, and D. S. Nachimuthu, "Panoramic tongue imaging and deep convolutional machine learning model for diabetes diagnosis in humans," *Scientific Reports*, vol. 12, p. 186, Jan. 2022.

[13] B. R, M. K T, J. D, and R. K. C, "Diabetes Mellitus Diagnosis based on Tongue Images using Machine Learning," in *2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS)*, (Coimbatore, India), pp. 1614–1618, IEEE, Mar. 2023.

[14] B. F. Wee, S. Sivakumar, K. H. Lim, W. K. Wong, and F. H. Juwono, "Diabetes detection based on machine learning and deep learning approaches," *Multimedia Tools and Applications*, vol. 83, pp. 24153–24185, Aug. 2023.

[15] M. E. Febrian, F. X. Ferdinan, G. P. Sendani, K. M. Suryanigrum, and R. Yunanda, "Diabetes prediction using supervised machine learning," *Procedia Computer Science*, vol. 216, pp. 21–30, 2023.

[16] A. Mujumdar and V. Vaidehi, "Diabetes Prediction using Machine Learning Algorithms," *Procedia Computer Science*, vol. 165, pp. 292–299, 2019.

[17] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, "Diabetes prediction using machine learning and explainable AI techniques," *Healthcare Technology Letters*, vol. 10, pp. 1–10, Feb. 2023.

[18] Y. Qin, J. Wu, W. Xiao, K. Wang, A. Huang, B. Liu, J. Yu, C. Li, F. Yu, and Z. Ren, "Machine Learning Models for Data-Driven Prediction of Diabetes by Lifestyle Type," *International Journal of Environmental Research and Public Health*, vol. 19, p. 15027, Nov. 2022.

[19] U. Ahmed and C. Li, "Machine Learning for Diabetes Prediction," in *2021 International Conference on Information and Communication Technology Convergence (ICTC)*, (Jeju Island, Korea, Republic of), pp. 16–19, IEEE, Oct. 2021.

[20] S. M. Ganie, P. K. D. Pramanik, M. Bashir Malik, S. Mallik, and H. Qin, "An ensemble learning approach for diabetes prediction using boosting techniques," *Frontiers in Genetics*, vol. 14, p. 1252159, Oct. 2023.

[21] H. Hairani, A. Anggrawan, and D. Priyanto, "Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link," *JOIV : International Journal on Informatics Visualization*, vol. 7, p. 258, Feb. 2023.

[22] K. , "Early stage diabetes risk prediction - Anticipating Diabetes Onset: Early Stage Risk Prediction Dataset," 2023.

[23] L.-Y. Huang, C.-H. Liu, F.-Y. Chen, C.-H. Kuo, P. Pitrone, and J.-S. Liu, "Aging Affects Insulin Resistance, Insulin Secretion, and Glucose Effectiveness in Subjects with Normal Blood Glucose and Body Weight," *Diagnostics*, vol. 13, p. 2158, June 2023.

[24] M. R. Refaie, N. A. Sayed-Ahmed, A. M. Bakr, M. Y. Abdel Aziz, M. H. El Kannishi, and S. S. Abdel-Gawad, "Aging is an Inevitable Risk Factor for Insulin Resistance," *Journal of Taibah University Medical Sciences*, vol. 1, no. 1, pp. 30–41, 2006.

[25] N. Sattar, "Gender aspects in type 2 diabetes mellitus and cardiometabolic risk," *Best Practice & Research Clinical Endocrinology & Metabolism*, vol. 27, pp. 501–507, Aug. 2013.

[26] P. Estoppey, C. Clair, D. Auderset, and J. J. Puder, "Sex differences in type 2 diabetes," *Cardiovascular Medicine*, May 2023.

[27] A. Kautzky-Willer, J. Harreiter, and G. Pacini, "Sex and Gender Differences in Risk, Pathophysiology and Complications of Type 2 Diabetes Mellitus," *Endocrine Reviews*, vol. 37, pp. 278–316, June 2016.

[28] D. Care and S. S. Suppl, "2. Classification and Diagnosis of Diabetes: *Standards of Medical Care in Diabetes—2022*," *Diabetes Care*, vol. 45, pp. S17–S38, Jan. 2022.

[29] R. Gundamaraju and R. Vemuri, "Pathophysiology of Greedy Colon and Diabetes: Role of Atropine in worsening of Diabetes," *Euroasian Journal of Hepato-Gastroenterology*, vol. 4, pp. 51–54, Jan. 2014.

[30] D. O. F. Diabetes, "Diagnosis and Classification of Diabetes Mellitus," *Diabetes Care*, vol. 32, pp. S62–S67, Jan. 2009.

[31] J. K. Jensen, "Risk Prediction: Are We There Yet?," *Circulation*, vol. 134, pp. 1441–1443, Nov. 2016.

[32] B. A. C. Permana, R. Ahmad, H. Bahtiar, A. Sudianto, and I. Gunawan, "Classification of diabetes disease using decision tree algorithm (C4.5)," *Journal of Physics: Conference Series*, vol. 1869, p. 012082, Apr. 2021.

[33] J. A. Nichols, H. W. Herbert Chan, and M. A. B. Baker, "Machine learning: applications of artificial intelligence to imaging and diagnosis," *Biophysical Reviews*, vol. 11, pp. 111–118, Feb. 2019.

[34] P. Larrañaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J. A. Lozano, R. Armañanzas, G. Santafé, A. Pérez, and V. Robles, "Machine learning in bioinformatics," *Briefings in Bioinformatics*, vol. 7, pp. 86–112, Mar. 2006.

[35] G. Stiglic, P. Kocbek, N. Fijacko, M. Zitnik, K. Verbert, and L. Cilar, "Interpretability of machine learning-based prediction models in healthcare," *WIREs Data Mining and Knowledge Discovery*, vol. 10, p. e1379, Sept. 2020.

[36] H. Blockeel, L. Devos, B. Frénay, G. Nanfack, and S. Nijssen, "Decision trees: from efficient prediction to responsible AI," *Frontiers in Artificial Intelligence*, vol. 6, p. 1124553, July 2023.

[37] T. M. M. Keumala, M. Melinda, and S. Syahrial, "Decision tree method to classify the electroencephalography-based emotion data," *JURNAL INFOTEL*, vol. 14, pp. 37–49, Feb. 2022.

[38] T. Ciu and R. S. Oetama, "Logistic Regression Prediction Model for Cardiovascular Disease," *IJNMT (International Journal of New Media Technology)*, vol. 7, pp. 33–38, July 2020.

[39] M. E. Shipe, S. A. Deppen, F. Farjah, and E. L. Grogan, "Developing prediction models for clinical use using logistic regression: an overview," *Journal of Thoracic Disease*, vol. 11, pp. S574–S584, Mar. 2019.

[40] M. Ghazwani and M. Y. Begum, "Computational intelligence modeling of hyoscine drug solubility and solvent density in supercritical processing: gradient boosting, extra trees, and random forest models," *Scientific Reports*, vol. 13, p. 10046, June 2023.

[41] S. Uddin, I. Haque, H. Lu, M. A. Moni, and E. Gide, "Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction," *Scientific Reports*, vol. 12, p. 6256, Apr. 2022.

[42] A. S. Abobakr Yahya, A. N. Ahmed, F. Binti Othman, R. K. Ibrahim, H. A. Afan, A. El-Shafie, C. M. Fai, M. S. Hossain, M. Ehteram, and A. Elshafie, "Water Quality Prediction Model Based Support Vector Machine Model for Ungauged River Catchment under Dual Scenarios," *Water*, vol. 11, p. 1231, June 2019.

[43] A. Meiriza, E. Lestari, P. Putra, A. Monaputri, and D. A. Lestari, "Prediction Graduate Student Use Naive Bayes Classifier," in *Proceedings of the Sriwijaya International Conference on Information Technology and Its Applications (SICONIAN 2019)*, (Palembang, Indonesia), Atlantis Press, 2020.

[44] J. Zhang, D. Mucs, U. Norinder, and F. Svensson, "LightGBM: An Effective and Scalable Algorithm for Prediction of Chemical Toxicity–Application to the Tox21 and Mutagenicity Data Sets," *Journal of Chemical Information and Modeling*, vol. 59, pp. 4150–4158, Oct. 2019.

[45] M. Nahiduzzaman, M. J. Nayeem, M. T. Ahmed, and M. S. U. Zaman, "Prediction of Heart Disease Using Multi-Layer Perceptron Neural Network and Support Vector Machine," in *2019 4th International Conference on Electrical Information and Communication Technology (EICT)*, (Khulna, Bangladesh), pp. 1–6, IEEE, Dec. 2019.

[46] J.-K. Tsai and C.-H. Hung, "Improving AdaBoost Classifier to Predict Enterprise Performance after COVID-19," *Mathematics*, vol. 9, p. 2215, Sept. 2021.

[47] R. Wu and N. Hao, "Quadratic discriminant analysis by projection," *Journal of Multivariate Analysis*, vol. 190, p. 104987, July 2022.