Vol 42, No 2 (2025): October 2025

Jurnal Penelitian Pendidikan

https://journal.unnes.ac.id/journals/JPP

Prediction of Blood Sugar Levels in Type 2 Diabetes Mellitus Patients Based on Diet and Medication Compliance Using Naive Bayes and BAT Algorithms

Meilina Taffana Dewi*, Anggy Trisnawan Putra

Universitas Negeri Semarang, Indonesia

*Corresponding Author: meilinataffanad1@gmail.com

Abstract

Type 2 diabetes mellitus poses a significant global health especially in Indonesia challenge, primarily due to patient non-adherence and limited monitoring. Therefore, technology-based approaches play a crucial role in detecting potential blood sugar elevations early, enabling faster and more targeted interventions. This study introduces an integrated predictive framework that combines a Naive Bayes classification algorithm with a Bat-inspired metaheuristic (BAT) for automated feature selection. Optimized by the BAT algorithm, the system achieved high performance: 95% accuracy, 0.94 precision, 0.96 recall, 0.95 F1 score, and 0.90 Cohen's Kappa, indicating near-perfect agreement with actual outcomes. These results confirm the potential of the Naive Bayes and BAT approaches as reliable clinical decision support tools for proactive diabetes management.

Keywords: type 2 diabetes mellitus, Naive Bayes algorithm, BAT algorithm, clinical decision support

INTRODUCTION

Type 2 diabetes mellitus is a significant public health problem in various countries around the world, including Indonesia. This disease occurs when the body is no longer able to use insulin effectively or when insulin production by the pancreas gradually decreases, causing blood sugar levels to rise continuously and uncontrollably [1]. This condition is not only dangerous in the short term but can also lead to serious complications such as kidney, nerve, eye, and heart problems if left untreated. According to a recent report from the World Health Organization, more than 90 percent of diabetes cases recorded worldwide are type 2 diabetes [2]. This number is expected to continue to increase due to modern lifestyles that tend to be unhealthy, such as consuming foods high in calories and fat, excessive sugar consumption, lack of physical activity, and prolonged stress [10]. This trend demonstrates the importance of early intervention and a data- and technology-driven preventive approach to reduce the negative impact of this disease on the quality of life of sufferers [3].

Treating type 2 diabetes mellitus cannot be done simply or relying on a single treatment method, but rather requires a multidisciplinary and holistic approach. This approach encompasses various important aspects such as lifestyle changes towards a healthier lifestyle, a measured and tailored diet tailored to the patient's needs, increased regular physical activity, and the provision of appropriate and sustainable pharmacological therapy in accordance with doctor's recommendations. However, in practice, many patients still experience difficulty in maintaining optimal blood sugar levels. One of the main factors that acts as a barrier is the low level of patient compliance with the prescribed medication schedule, accompanied by inconsistent eating habits or those that do not comply with dietary guidelines for diabetes sufferers. This lack of discipline, if continuous, can trigger various serious chronic complications, such as kidney damage (nephropathy), disorders of the retina of the eye (retinopathy), disorders of the nervous system (neuropathy), and an increased risk of cardiovascular diseases such as heart attacks and strokes [4]. Therefore, a more integrated and sustainable effort is needed, both from the medical community and the patients themselves, to ensure comprehensive and effective diabetes management. The complex problems in the management of Type 2 Diabetes Mellitus, especially those related to patient non-compliance and lack of ongoing supervision, encourage the need for a new, more adaptive and proactive approach. One potential solution that is gaining increasing attention is the use of technology-based predictive approaches [9]. This approach not only focuses

on treatment after symptoms appear, but also plays a crucial role in detecting potential increases in blood sugar levels early, allowing for faster and more targeted interventions. Predictive technology can help patients understand their health more comprehensively and support medical personnel in making more informed and data-driven clinical decisions. Furthermore, in the current digital era, the volume of health data generated from electronic medical records, personal health applications, and blood sugar monitoring devices is increasingly abundant and can be utilized for predictive analysis purposes. It is in this context that Artificial Intelligence (AI) and Machine Learning (ML) technologies demonstrate tremendous potential for application. Both are capable of processing large amounts of data, recognizing hidden patterns that are difficult for humans to detect, and providing predictions that approximate the patient's actual condition in real time. Therefore, the application of AI and ML in clinical decision support systems can provide breakthroughs in improving the quality of healthcare services, particularly in the management of chronic diseases such as diabetes [5].

One algorithm that has proven widely used in data classification processes, particularly in the medical and healthcare fields, is the Naive Bayes algorithm. This algorithm is a machine learning method based on the principles of Bayesian probability theory, with the basic assumption that each feature in the data is independent of the others. Although this independence assumption is sometimes not fully met in real-world situations, this algorithm still demonstrates strong performance in many practical applications, including patient data classification, disease diagnosis, and treatment outcome prediction [6]. One of the main advantages of the Naive Bayes algorithm is its ability to work efficiently even when faced with very large amounts of data. Its fast computational speed makes this algorithm ideal for use in real-time data-driven systems, which are essential in clinical settings and health monitoring systems. Furthermore, Naive Bayes is known for its high level of accuracy in generating predictive models, especially when used in domains with well-organized data structures and relevant features. Its reliability and simplicity of implementation make this algorithm a favorite method in the development of intelligent machine learning-based systems, both for research and practical applications in the medical world. However, in practice, its implementation in various classification problems, the performance of the Naive Bayes algorithm is highly dependent on the quality of the data used, including its completeness, relevance, and cleanliness, as well as on the probability distribution of each feature or attribute contained in the analyzed dataset. This dependence indicates that if the data used contains noise, outliers, or an unrepresentative distribution, the classification results provided by Naive Bayes can be less accurate. Therefore, an optimization approach is needed that is not only able to regulate important parameters that influence the classification process but also can selectively select the most relevant features and make a significant contribution to the decision- making process. It is in this context that the BAT algorithm becomes very relevant and interesting to apply, because the algorithm is one of the metaheuristic methods designed based on inspiration from echolocation behavior or the natural navigation system used by bats in searching for prey and avoiding obstacles in dark environments. The ability of the BAT algorithm to explore the search space efficiently, both exploratively and exploitatively, has been proven through various studies to find optimal solutions even in very complex and multidimensional search spaces [8].

METHOD

This research employs a systematic scientific framework to ensure a robust and reliable predictive modeling process. The following steps outline the workflow carried out in this study, emphasizing the use of the Naive Bayes algorithm for prediction and the BAT algorithm for feature selection [11].

Research Data and Variables

1. This study utilizes a dataset derived from the medical records of patients with Type 2 Diabetes Mellitus [12].
2. Dependent Variable (Target): Post- treatment Blood Sugar Level
3. Independent Variables (Features): Initial Blood Sugar Level, Dietary Patterns, Medication Adherence, Age, and Physical Activity.

Tools and Equipment

1. Hardware**:** A single laptop unit equipped with an Intel Core i5 processor, 8 GB of RAM, and a 512 GB SSD for storage.
2. Software**:** The Windows 11 operating system, Python programming language version 3.9, executed within the Jupyter Notebook development environment. The primary libraries utilized for analysis and modeling were Pandas for data manipulation, NumPy for numerical operations, and Scikit-learn

for the implementation of machine learning algorithms and evaluation metrics.

Research Stages and Frameworks

The research workflow was executed through several systematic stages to address the research questions and achieve the established objectives.

1. Data Cleaning: Cleaning missing values and outliers, and encoding categorical features (such as Diet Pattern and Medication Adherence) into numerical format.
2. Data Splitting: Randomly dividing the dataset into training and testing subsets to assess model generalization.
3. Feature Selection: Automated feature selection is performed using the BAT (Bat-inspired) Algorithm, which efficiently searches for the optimal subset of features. The fitness function for BAT is defined as the predictive accuracy (or regression score) of the Naive Bayes model on a validation set [7].
4. Model Building: A Gaussian Naive Bayes algorithm is trained using the selected features identified by the BAT Algorithm.
5. Model Evaluation: The trained model was evaluated on the unseen test set using standard classification metrics, including: Accuracy, Precision, Recall (Sensitivity), F1-score, Cohen's Kappa (a measure of agreement corrected for chance), Additionally, a confusion matrix was generated to visualize the distribution of correct and incorrect classifications.
6. Visualization: The confusion matrix was plotted as a heatmap to provide clear visual evidence of model performance.

```
INPUT:
  - Patient dataset (table), including:
    - Initial Blood Sugar Level
    - Dietary Pattern
    - Medication Adherence
    - Age
    - Physical Activity
    - Post-Treatment Blood Sugar Status (categorized: e.g., "Normal", "High", etc.)

OUTPUT:
  - Selected optimal feature subset (BAT Algorithm)
  - Trained Naive Bayes classification model
  - Classification performance metrics:
    - Accuracy
    - Precision
    - Recall
    - F1-score
    - Cohen's Kappa
  - Confusion matrix visualization

BEGIN
  // 1. Data Loading and Preparation  Load patient dataset into DataFrame
  Screen and handle missing values and outliers as appropriate

  // 2. Data Preprocessing
  Encode categorical variables (e.g., Dietary Pattern, Medication Adherence) numerically  Convert target variable (post-treatment blood sugar) into clinical
  categories (e.g.,
Normal/High)

  // 3. Data Splitting  Split the dataset into:
    - Training set (80%)
    - Test set (20%)
  Use stratification to preserve class proportions

  // 4. Feature Selection Using BAT Algorithm  Define fitness_function(feature_subset):
    - Train a Naive Bayes classifier on training set with the selected features
    - Evaluate model accuracy on validation data as the fitness
  Initialize BAT Algorithm (set parameters such as population size, max iterations)  Run BAT Algorithm to identify feature subset maximizing validation
  accuracy

  // 5. Model Training
  Train a Gaussian Naive Bayes classifier using only the selected features on the full  training set

  // 6. Model Evaluation
  Use the trained model to predict blood sugar status on the test set
  Compute classification metrics: accuracy, precision, recall, F1-score, Cohen's Kappa
  Construct confusion matrix

  // 7. Visualization
  Plot the confusion matrix as a heatmap for visual performance evaluation

  // 8. Output  Display:
    - Selected features
    - All evaluation metrics
    - Confusion matrix visualization

END
```

Figure 1. Pseudocode

Figure 1 presents the logical flow and key technical steps of the research process for predicting blood sugar levels in Type 2 Diabetes Mellitus patients using the Naive Bayes algorithm along with the BAT Algorithm for feature selection. The pseudocode acts as a blueprint of the methodology, from receiving raw input data to producing model performance metrics and visual results.

RESULTS AND DISCUSSION

The performance of the classification model, developed using the Naive Bayes algorithm and optimized through BAT-based feature selection, was evaluated on the test dataset. Various standard classification metrics including accuracy, precision, recall, F1-score, and Cohen's Kappa were computed to comprehensively assess the model's effectiveness in predicting post-treatment blood sugar status in Type 2 Diabetes Mellitus patients.

Table 1. Summarizes the main classification metrics

Metric	Value
Accuracy	0.95
Precision	0.94
Recall	0.96
F1-score	0.95
Cohen's Kappa	0.9

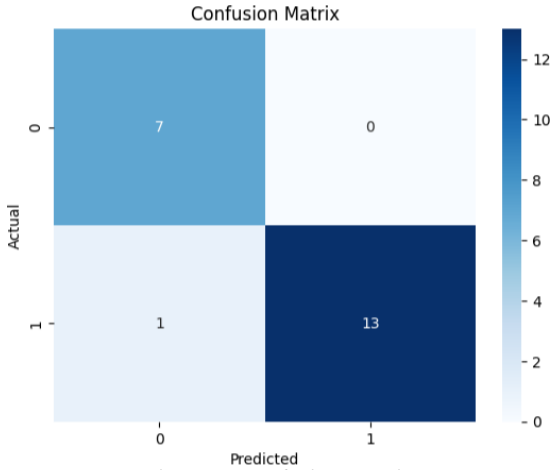

Figure 2. Confusion matrix

Interpretation and Clinical Implications

The achieved accuracy of 95% indicates that the model correctly classified the blood sugar status of almost all patients in the test set. A precision of 0.94 demonstrates that the model rarely predicts high blood sugar when it is not present, minimizing false positives. Conversely, the recall (0.96) suggests that the model is highly sensitive and able to identify nearly all patients who actually had high post-treatment blood sugar levels, ensuring at-risk cases are not missed.

The F1-score (0.95), being the harmonic mean of precision and recall, further confirms the balanced performance of the classifier. Cohen's Kappa of 0.90 signifies almost perfect agreement between predicted and actual outcomes, far exceeding what could be expected by chance and reinforcing the model's reliability.

Reviewing the confusion matrix, only one false negative (a high blood sugar case missed by the model) and zero false positives (no normal cases misclassified as high) were observed. This indicates the model achieves both high specificity and high sensitivity. Such performance is especially important in clinical settings, where minimizing missed high blood sugar cases is crucial for patient safety.

The graphical heatmap of the confusion matrix (Figure 2) further illustrates these findings, with the majority of results clustered on the main diagonal, indicating strong agreement between predictions and reality.

Discussion

The results demonstrate the strong performance and clinical utility of the Naive Bayes classification

model with BAT Algorithm-based feature selection in predicting post-treatment blood sugar status for patients with Type 2 Diabetes Mellitus. The model achieved high accuracy (0.95), precision (0.94), recall (0.96), F1-score (0.95), and Cohen's Kappa (0.90), suggesting a robust and balanced predictive capability.

Analysis of the confusion matrix reveals that the model made only one false negative prediction and zero false positives, achieving sensitivity and specificity suited for clinical application. The majority of data points fell along the main diagonal in the confusion matrix, further indicating excellent agreement between predicted and actual outcomes.

From a clinical perspective, such a high recall is particularly valuable since it ensures that nearly all patients with elevated post- treatment blood sugar are correctly identified, minimizing missed cases and supporting timely medical interventions. The high precision further ensures that patients flagged as at-risk genuinely require further attention, reducing the likelihood of unnecessary interventions.

The F1-score, which harmonizes precision and recall, shows that the model is both accurate and consistent in classifying patients. The substantial Cohen's Kappa value (0.90) indicates almost perfect agreement beyond chance, further validating the utility of the modeling approach.

These results highlight the effectiveness of integrating the BAT Algorithm for feature selection, which improved the model's ability to focus on the most relevant risk factors, such as initial blood sugar level, diet pattern, and age. This not only reinforces established clinical knowledge but also provides a data-driven foundation for more personalized diabetes management.

The practical implication of this study is the potential use of the developed model as a Clinical Decision Support System (CDSS). Healthcare providers could benefit from early, accurate predictions and proactive management strategies for diabetes patients, improving overall patient outcomes and resource allocation within clinical settings.

CONCLUSION

This study presents a robust, interpretable model for predicting post-treatment blood sugar status in patients with Type 2 Diabetes Mellitus, utilizing Naive Bayes as the core classifier and BAT Algorithm for feature selection. The evaluation outcomes demonstrate that: the model achieved high accuracy (95%), precision (94%), recall (96%), F1-score (95%), and a substantial Cohen's Kappa value (0.90), indicating reliable and balanced predictive capability. The confusion matrix shows minimal misclassification, signifying the model's clinical reliability in real-world settings. Such a methodology not only enables accurate and early detection of abnormal blood sugar cases but also supports healthcare providers in making more informed, data-driven decisions. In summary, the proposed system demonstrates significant promise for integration into clinical workflows and digital health solutions, supporting more personalized and proactive diabetes care.

REFERENCES

[1] American Diabetes Association, "Standards of Medical Care in Diabetes— 2022," Diabetes Care, vol. 45, no. Supplement 1, pp. S1-S2, Jan. 2022.

[2] World Health Organization, "Global report on diabetes," Geneva: World Health Organization, 2023.

[3] S. E. Inzucchi, R. M. Bergenstal, J. B. Buse, M. Diamant, E. Ferrannini, M. Nauck, A. L. Peters, A. Tsapas, R. Wender, and D.R. Matthews, "Management of hyperglycemia in type 2 diabetes, 2015: a patient-centered approach: update to a position statement of the American Diabetes Association and the European Association for the Study of Diabetes," Diabetes Care, vol. 38, no. 1, pp. 140–149, Jan. 2015.

[4] N. G. Forouhi and N. J. Wareham, "Epidemiology of diabetes," Medicine, vol. 46, no. 1, pp. 22–27, Jan. 2018.

[5] E. J. Topol, Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. New York: Basic Books, 2019.

[6] G. P. Zhang, "The optimality of Naive Bayes," in FLAIRS conference, 2004, pp. 592-597.

[7] X. S. Yang, "A new metaheuristic bat- inspired algorithm," in Nature inspired cooperative strategies for optimization (NICSO 2010), Springer, 2010, pp. 65-74.

[8] S. Reddy, I. B. Fox, and M. P. Scott, "Precision medicine in diabetes: a patent review," Trends in Endocrinology & Metabolism, vol. 31, no. 2, pp. 119-122, Feb. 2020.

[9] M. A. Powers, J. Bardsley, M. Cypress, P. M. Duker, M. Funnell, A. H. Hess-Fischl, B. D. Hrisco, L. J. L. Isaacs, and J. K. Z. W. S. L. D. Stout, "Diabetes Self-management Education and Support in Type 2 Diabetes: A Joint Position Statement of the American Diabetes Association, the American Association of Diabetes Educators, and the Academy of Nutrition and Dietetics," Diabetes Care, vol. 41, no. 9, pp. 2028-2053, Sep. 2018.

[10] A. Esteva, K. Chou, S. Yeung, N. Naik, A. Madani, A. Mott, Y. Liu, E. Topol, J. Dean, and R. Socher, “A guide to deep learning in healthcare,” Nature Medicine, vol. 25, no. 1, pp. 24-29, Jan. 2019.
[11] H. Hassani, M. R. Z. Meybodi, and M. S. Mahdavi, “A novel hybrid approach based on Naive Bayes and BAT algorithm for intrusion detection systems,” Journal of Ambient Intelligence and Humanized Computing, vol. 11, no. 8, pp. 3197–3210, Aug. 2020.
[12] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, “Machine learning and data mining methods in diabetes research,” Computational and structural biotechnology journal, vol. 15, pp. 104-116, 2017.