Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 12, 2025

A COMPARATIVE MACHINE LEARNING APPROACH TO DIABETES PREDICTION: INTEGRATING BEST FIRST SEARCH FEATURE SELECTION WITH SVM AND NAÏVE BAYES CLASSIFIERS

Toufeeq Ur Rehman[1], Junaid Iqbal[2], Malak Roman[*3], Abdullah[4], Zakir Ahmad[5], Bashir Ahmad[6], Muhammad Rafiq[7]

[1,2,4,5,6,7] *BS-Computer Science, Department of Computer Science, University of Chitral, KP-Pakistan.*
[*3] *Lecturer, Department of Computer Science, University of Chitral, KP-Pakistan*

[*3]malak_5116@uoch.edu.pk

DOI: https://doi.org/10.5281/zenodo.17854483

Keywords
Machine Learning, Artificial Intelligence, Diabetes Diseases, Best First Search, Support Vector Machine, Naïve Bayes Classifiers.

Article History
Received: 13 October 2025
Accepted: 23 November 2025
Published: 08 December 2025

Corresponding Author: *
Malak Roman

Abstract
Data mining is the process of using machine learning, statistical, and other database systems techniques to extract useful, hidden patterns from large data sets. It is beneficial for decision-making and prediction tasks across various domains. In the healthcare domain, data mining has been recognized as crucial for analyzing patient data to enhance the precision of patient diagnosis, provide best practices in treatment, and improve the ability to predict disease. Recent advances in Artificial Intelligence (AI) and Machine Learning (ML) have shown remarkable potential to transform diabetic data by enabling predictive, personalized, and data-driven healthcare systems. This study presents a comprehensive comparative analysis of multiple AI-driven models for the robust early prediction and clinical evaluation of Diabetes. By utilizing the Best First Search algorithm for feature selection in conjunction with Support Vector Machine and Naïve Bayes classifiers to identify the most significant attributes in the dataset, our study seeks to enhance predictive performance. With an accuracy of 91.34%, our analysis showed that the SVM (Support Vector Machine) with Best First Search performed better than the other classifiers, while the Naive Bayes classifier with Best-First Search produced a reasonable result, with an accuracy of 87.88%. To improve the diabetes prediction set for early diagnosis and health management, the results of this study demonstrate the effective application of combining feature selection with classification.

INTRODUCTION

Machine learning applies to multiple algorithms that learn from previous data to make predictions. It uses computer algorithms to improve its task performance with each iteration of the task, by optimizing mathematical models with training data [1]. The field of data mining is an emerging multidisciplinary area of computer science, which combines the traditional principles of machine learning, database management systems, artificial intelligence, and statistical computation [2]. Data mining is a subset of artificial intelligence that utilizes large datasets to uncover meaningful information hidden in unrecognized patterns. It is increasingly employed in healthcare for clinical diagnostics and disease prediction. [3]. ML methods are broadly applicable to the fields of image recognition, natural language processing,

and predictive analytics [4]. The main objective of data mining is to obtain relevant, high-value information from the data sets and to provide it to the users in an understandable and correlated form [5]. That type of information could benefit many kinds of decision-making systems in any organization [6].

Data mining is the process of identifying patterns, correlations, and trends in large datasets to transform raw data into meaningful information. It is a crucial part of data analytics, helping businesses make informed forecasts and decisions by revealing hidden relationships through statistical and machine learning techniques [7]. It has additional consequences in healthcare facilities. It smooths the progress of health organizations to precisely as well as pigeonhole the inefficiencies, also to reduce large expenditures [8]. Data mining is useful in the health sector for numerous numbers of vital applications, supporting a variety of important applications, some of which are precise diagnoses, effective risk management, clinical decision support systems (CDSS), and innovation in medical research [9]. Adding technology to day-to-day life has become crucial, significantly streamlining routine tasks. IoT enables M2M communication through innovative information-sharing methods and supports P2M connectivity. It manages information queries and carries out instructions on a variety of hardware devices with different functionalities. The emergence of existing technologies has fundamentally transformed human existence [10]. With the assistance of data mining, recognizing risk factors and associating indications with historical information will allow early action to enhance clinical outcomes and reduce healthcare costs. For example, predictive models can flag patients who may be at risk for chronic diseases based on lifestyle, genetic, demographic, and bio specimen data. [6]

DIABETES DISEASES:

Diabetes is a lifelong illness affecting millions worldwide. It involves high blood glucose levels. And is an important cause of death. It may result in critical downstream effects such as heart disease, neuropathy, and kidney failure. Demographic factors like population growth and population aging, along with way of life factors like urbanization, poor diet, and decreased physical activity, will steadily increase the global burden of diabetes and its associated problems [11]. Diabetes is a long-term condition caused by evaluated blood glucose levels due to limited production of insulin or poor insulin response. This condition is estimated to affect 853 million adults globally in 2025, with estimates of 1.31 billion adults by 2050 [12]. This condition leads to approximately 1.6 million deaths annually due to complications. Insulin, a hormone secreted by the pancreas, regulates glucose uptake into cells for energy. If left unmanaged, diabetes is a critical complication that is more likely to occur, such as cardiac failure, nerve damage, and long-lasting vision loss [13].

The three major types of diabetes include:

Type 1 Diabetes. Here pancreatic beta cells are destroyed by the immune system in this autoimmune condition, resulting in little to no insulin production and requiring lifelong insulin treatment. [14].

Type 2 Diabetes: It represents over 90% of all diabetes cases, is mostly affected by insulin defiance, regularly related to fatness and a lack of physical activity. Management usually concerns lifestyle changes, medical treatment such as metformin, and, in some cases, insulin therapy.

Gestational Diabetes involves hyperglycemia during pregnancy, typically diagnosed between 24–28 weeks, increasing risks of maternal and fetal complications, such as preeclampsia and future Type 2 diabetes [15].

LITERATURE REVIEW

Utilizing advanced technologies, numerous research studies have explored how machine learning and deep learning are used in medical areas, since they can significantly increase the accuracy of diabetic disease prediction This section shows studies that made use of publicly available datasets, including data from Kaggle, local hospitals, and labs, are shown in this section.

Rastogi, R., & Bansal, M. (2023) [16] established a framework to identify and prevent diabetes using the Naïve Bayes, Random Forest, Support Vector Machine, and Logistic Regression classifiers. They used a Kaggle dataset, statistically pre-processed, and then fitted each classifier, finally evaluating their respective performance based on accuracy. The Logistic Regression achieved the highest performance with an accurate score of 82.46%. Fadnavis, R., et al. (2021) [17] A framework was built to predict heart disease using the 14-attribute-Cleveland dataset, and employed the Naïve Bayes classifier, achieving 85.25% accuracy. You can see from their work the strength of machine learning algorithms in helping to analyze complex medical datasets to identify risk factors.

Sisodia, D., et al. (2018) [18] proposed a framework for predicting diabetes by applying data mining, supervised learning methods. The researchers applied Naïve Bayes, Support Vector Machine, and Decision Tree, then examined the performance of each algorithm. Their study reported that Naïve Bayes achieved 76.3% accuracy, Support Vector Machine achieved 65.1%, and the Decision Tree model achieved 73.82%. Vijayarani, S., & Dhayanand, S. (2015) [19] developed a framework that predicts kidney disease with data mining classification methods, implementing and comparing two machine learning models: Naïve Bayes and Support Vector Machine (SVM). Researchers measured the models with different metrics and observed 70.96% accuracy for Naïve Bayes and 76.32% accuracy for SVM.

Zriqat, I. A., et al. (2016) [20]. They conducted a broad study that used data mining classification methods to predict heart disease. The researchers trained and tested Decision Tree, Random Forest, SVM, and Naïve Bayes on the Cleveland heart disease dataset. SVM achieved the maximum accuracy of 76.57%. Singh, M. S., et al. (2024) [21] recommended a framework for predicting congestive heart failure (CHF) and They applied multiple algorithms—DT, NB, KNN, SVM, RF, LR, and DNN—and used a C4.5-based preprocessing method for feature selection. and KNN for missing data imputation, after comparing the results of the algorithms, it was found that the Support Vector Machine model achieved the top accuracy of 85.13%. Similarly, Sakib, S., et al. (2021) [22] investigated the efficiency of machine learning methods for diabetes using the PIMA Indian Diabetes Dataset. They applied various algorithms, including Decision Tree (DT), Logistic Regression (LR), XGBoost, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Random Forest (RF). A top accuracy of 80.21% was reached by the SVM algorithm, according to their analysis. Roman, M. et al. (2022) [8] proposed a machine learning framework for predicting cardiovascular disease, employing K-Nearest Neighbors (KNN) and Fuzzy KNN classifier with an emphasis on feature selection during preprocessing. They explored three preprocessing methods: simple preprocessing, symmetric uncertainty, and a genetic search algorithm. The Fuzzy KNN model achieved accuracies of 65% with simple preprocessing, 74% with symmetric uncertainty, and 75% with the genetic search algorithm. In contrast, the standard KNN model demonstrated higher performance, yielding 90% accuracy with simple preprocessing, 93% with the genetic search algorithm, and 95% with symmetric uncertainty.

Kangra, K., & Singh, J. (2023) [23] They applied GA-based feature selection in their study, boosting the performance of machine learning algorithms for diabetes prediction. They applied these algorithms directly to the diabetes datasets and compared model performance with and without GA in the preprocessing stage. The results demonstrated that using GA for feature selection notably enhanced the accuracy of the predictive models. Zeinalnezhad, M., & Shishehchi, S. (2024) [24] propose a hybrid method integrating a data mining classifier with a meta-heuristic method to forecast diabetes in patients. Datasets are obtained from the UC Irvine Machine Learning Repository, and Researchers evaluated Genetic algorithm along with SVM, Random Forest, and Neural Network to find the most effective predictive model for diabetes. The results indicate that the Support

Vector Machine (SVM) obtained an accuracy of 72%, whereas both the Genetic Algorithm–optimized SVM (GA-SVM) and Random Forest (RF) achieved 74%. The Neural Network (NN) reached 70% accuracy. GA-SVM and RF showed equal, higher performance among all the models."

El-Sofany, H., et al. (2024) [25] developed a machine learning-based approach, integrated into a mobile application, for predicting diabetes. To identify the method that generates the most accurate diabetes prediction. The experiment includes ten machine learning (ML) classification approaches, such as logistic regression, random forest, KNN, decision tree, bagging, AdaBoost, XGBoost, voting, SVM, and Naive Bayes. They trained their models using the PIMA and private datasets and evaluated their performance. The highest accuracy, 83.1%, was attained with the XGB algorithm when combined with the SMOTE method. Guleria, P., et al. (2024) [26] proposed a framework combining Shapley Additive Explanations (SHAP) and Data Science as a Service with ML to predict diabetes. They utilized the PIMA dataset for training models and evaluated their performance. The results were as follows: Decision Tree (68.8%), SVM (76.6%), KNN (70.7%), Naïve Bayes (75.4%), AdaBoost (75.9%), Bagged Tree (76.0%), and Neural Network (77.9%)

Alkalifah, B., et al. (2025) [27] investigated various machine learning-based regression techniques to predict fluctuations in diabetes levels. They applied Neighborhood Component Analysis for feature reduction and evaluated their models' performance. The models achieved the following accuracies: Linear Regression (LR) 78.33%, Linear Regression with Stochastic Gradient Descent (LRSGD) 78.33%, Support Vector Machine (SVM) 80%, Gaussian Process Regression (GPR) 89.59%, Boosted Decision Tree Ensemble (BSTE) 92.04%, Bagged Decision Tree (BDT) 92.58%, Stepwise Regression (SW) 77.96%, and Artificial Neural Network (ANN) 79.46%. Modak, S.K.S., Jha, V.K. (2024) [28] developed a disease prediction system for diabetes using different machine learning methods, aiming to enhance early discovering and interference for diabetes. They sourced datasets from Kaggle and applied Principal Component Analysis (PCA) for feature selection. Their models accomplished as following accuracies are XGBoost (0.94), LightGBM (0.947), CatBoost (0.954), AdaBoost (0.946), Bagging (0.909), Random Forest (0.928), Support Vector Machine (SVM) (0.81), Naïve Bayes (NB) (0.783), and Logistic Regression (LR) (0.787).

Bigdeli, S. K., et al. (2025) [29] utilize machine learning algorithms to predict the possibility of gestational diabetes mellitus (GDM) in pregnant women at the Vali-e-Asr Hospital's fertility health center in Tehran, Iran. Their models, based on data from 2020–2022, achieved the following accuracies: Decision Tree (DT) 82%, Multilayer Perceptron (MLP) 74%, K-Nearest Neighbors (KNN) 77%, Naïve Bayes (NB) 72%, Random Forest (RF) 94%, and XGBoost 88%. Ram, A., & Vishwakarma, H. (2021) [30] utilized machine learning and data mining techniques to predict diabetes disease, using the PIMA dataset for their study. Their results showed that Logistic Regression using all nine features achieved an accuracy of 84.7%, while Logistic Regression with five selected features yielded a slightly higher accuracy of 85%. Identified the possible features to increase accuracy using the recursive feature removal technique.

Khadragy, S., et al. (2022) [31] developed a predictive system for diabetes in the United Arab Emirates (UAE) with data collected from top hospitals in Dubai, UAE, consisting of 2856 patient records. They found patterns and risk factors for diabetes by applying K-means clustering, dividing the dataset into four clusters determined by patient characteristics. The paper was not similar to the other studies mentioned earlier in that this research was primarily classification-based, with no direct accuracy metrics or prediction of diabetes, but instead with clustering and looking at relationships between medical characteristics. Kumar, M., et al. (2014) [32] showed that feature selection using the Best First Search (BFS) algorithm improves diabetes prediction. The BFS algorithm selected 4 out of 9 attributes to train an ant colony neural network (ACO-NN). Their results proved that

BFS improves feature selection and, therefore, the accuracy of the machine learning classifier is also improved. For example, the ACO-NN gets 90% accuracy using the BFS algorithm.

Ahad, A., et al. [33] suggested the model of estimating the stroke disease using fuzzy K-Nearest Neighbor (F-KNN) and Artificial Neural Network (ANN) classifier. They use Chi-Squared and Heuristic-based search algorithm (BFS) data mining techniques to support attribute evaluations and proper pre-processing. Using the BFS algorithm enabled the authors to have only 4 attributes of 12 taken and obtained 97% accuracy with the ANN classifier and 96% accuracy with the F-KNN. Karthikeyan, T., & Thangaraju, P. [34] developed a model to predict Hepatitis disease using the Naïve Bayes algorithm, along with a feature selection mechanism which used the Best First Search (BFS) algorithm to select 10 of the 19 attributes. They found that the Naïve Bayes model performed at 84% accuracy without feature selection and then improved to 88% accuracy once BFS was implemented. Similarly Roman, M., et al. [35] examined two machine learning classifiers, Decision Table and Random Forest, to resolve how to use student data to predict academic success. They used the Best First Search (BFS) algorithm, which is a heuristic-based feature selection approach applicable to both algorithms. Their findings specified that the Decision Table had slightly better performance than Random Forest, getting a root mean squared error of 1.92. Their study highlights the efficiency of BFS in enhancing feature selection and improving predictive accuracy in educational data mining.

RESEARCH METHODOLOGY:

The preliminary phase of this research work involves the collection of data related to diabetes, analyzing, and validating. The Best First Search algorithm is used to refine and select significant features, enhancing the model's capability to recognize the most important attributes. Support Vector Machine and naïve Bayes classification algorithms are trained and tested. Figure 1 demonstrates our research methodology.

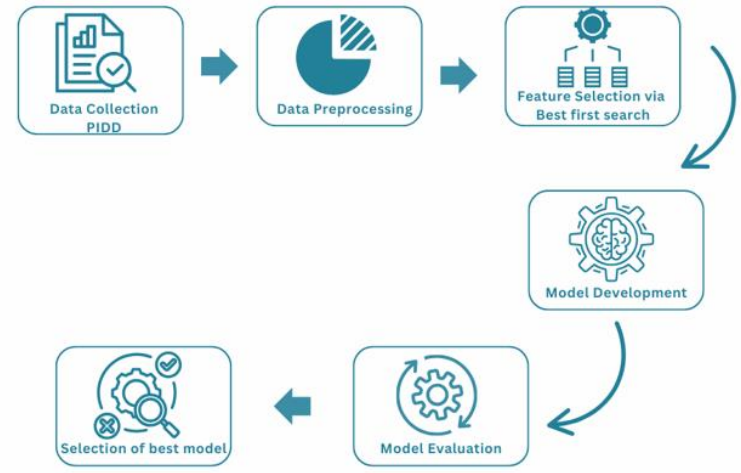

Figure 1: Research Methodology

DATA COLLECTION:

Diabetes patient dataset obtained from the publicly available Pima Indian Diabetes repositories, a commonly used for medical, data mining, and machine learning research. The datasets contain 768 attributes with 9 features as discussed in Table 1.

Table 1: PIMA Diabetes Dataset Description

S.No	Attributes	Description
1.	Pregnancies	Number of times pregnant
2.	Glucose	Plasma glucose concentration 2 hours in an oral glucose tolerance test
3.	Blood Pressure	Diastolic blood pressure (mm Hg)
4.	Skin Thickness	Triceps skin fold thickness (mm)
5.	Insulin	2-Hour serum insulin (mu U/ml)
6.	BMI	Body mass index (weight in kg/(height in m)^2)
7.	DiabetesPedigreeFunction	Diabetes pedigree function
8.	Age	Age in years
9.	Outcome	Class variable (0 or 1) 268 of 768 are 1, the others are 0

PREPROCESSING AND FEATURE SELECTION:

Feature selection helps machine learning classifiers achieve better performance by pinpointing the most relevant health indicators for diabetes. Feature selection is a method that reduces dimensionality, eliminates redundant attributes, and improves model prediction [36].

BEST FIRST SEARCH ALGORITHM

The Best-First Search technique determines the next course of action by applying a unique rule known as an evaluation function. There are two main types: There are two primary types: A* Search and Greedy BFS. Greedy BFS employs a heuristic function to determine the most promising option at each step [37]. In BFS, the classifiers prioritize node expansion based on their heuristic scores. It maintains two lists: a closed list for nodes already processed and an open list for nodes generated but still unexamined. At each iteration, the algorithm selects from the open list the node with the highest heuristic priority (e.g., lowest estimated cost to the goal) for expansion. Once expanded, the node is moved to the closed list, and its successor nodes are generated and added to the open list for further evaluation. This process repeats until the goal is reached or the open list is exhausted. The optimal attribute selection procedure using Best First Search is summarized in Table 2. The mathematical formulation of the BFS algorithm is presented as shown [35].

$$f(n) = g(n) + h(n) \qquad \text{equ. 1}$$

- f(n) This value estimates the full cost of the route that travels from node n to the goal.
- g(n) This value shows the total expense spent moving from the initial node to node n.
- h(n) This value uses a heuristic to estimate the cost of going from node n to the target.

Table 2: Optimal Attribute Selected by BFS.

S.No	Attribute
1	Pregnancies
2	Glucose
3	Skin Thickness
4	BMI
5	DiabetesPedigreeFunction

The Best First Search (BFS) algorithm helps diabetic prediction systems become more efficient and accurate when data mining is applied. It is particularly helpful in selecting optimal features, optimizing model structures, and minimizing computational overhead in classification and prediction tasks.

BUILDING CLASSIFICATION MODELS: SUPPORT VECTOR MACHINE AND NAÏVE BAYES

Two machine learning algorithms are used for classification on the preprocessed datasets after attributes were selected via the Best-First Search algorithm. The Support Vector Machine (SVM) and Naïve Bayes (NB) algorithms were trained using 70% of the dataset, while the remaining 30% was used for testing.

SUPPORT VECTOR MACHINE (SVM):

Support Vector Machine (SVM) is a fundamental machine learning algorithm characterized by its ability to select flexible kernel functions and its robustness against noisy data.SVM is useful for solving classification and regression problems as well, leveraging support vectors to define decision boundaries that maximize the margin between classes, thereby improving generalization performance. [38]. The decision boundary in SVM is optimized by maximizing the margin to the closest data points from each class. In the context of diabetes classification, this approach ensures a robust separation between diabetic and non-diabetic instances. The decision boundary's location and direction are primarily determined by the support vectors, those critical data points closest to the margin, which play a major role in defining the classifier's effectiveness [39].

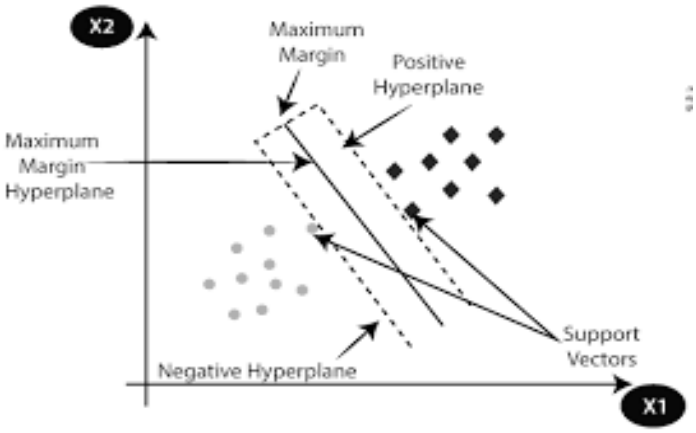

Figure 2: Support Vector Machine Hyperplanes

SVM works through constructing a hyperplane in the feature space that maximally separates distinct data classes, ensuring clear and optimal class boundaries even in high-dimensional environments. This line is called a hyperplane. The goal is to make the gap between the groups as wide as possible. Both discrete and continuous data types can be managed using SVM. It creates a model by placing the training data points in a multi-dimensional space and then sorts each point into its correct group [40]. Linear SVM separates data into two groups using a straight line or hyperplane, while non-linear SVM uses kernel functions to handle data that cannot be separated linearly.

The decision function is utilized to determine the hyperplane:

$$f(x) = w^t x + b \qquad \text{equ. 2}$$

- $x \in R^n$ is the standardized explanatory variable vector,
- $w \in R^n$ is the learning coefficient vector,
- $b \in R$ is the phrase of bias

To meet the optimization conditions, the margin will be maximized while ensuring the classes are correctly separated, resulting in the following primal problem [41].

$\min_{w,b} \frac{1}{2}||w||^2$

Subject to $y_i(w^T x_i + b) \geq 1, \forall_i$

where $y_i \in \{-1, +1\}$ is the target class.

The Support Vector Machine (SVM) algorithm plays a major part in diabetes disease prediction using data mining by providing a highly accurate, robust, and scalable classification approach for medical datasets. It is widely applied to differentiate between diabetic and non-diabetic individuals based on physiological and biochemical indicators.

NAÏVE BAYES

The Naïve Bayes algorithm is, type of classifier that uses Bayes' theorem to make predictions based on probabilities, to estimate class membership likelihoods by analyzing the occurrence and co-occurrence of feature values in a dataset. It operates under the simplifying assumption of conditional independence among features given the class label, which rarely holds in real-world data; therefore, the term "naive." Despite this assumption, Naive Bayes demonstrates rapid learning and effective performance across various supervised classification tasks due to its efficiency in parameter estimation and scalability with limited training data [42]. NB determines that the classification of one characteristic within a class does not influence the classification of another characteristic. This probabilistic relationship is mathematically expressed in the following equation 2.

$$p(c|x) = \frac{[p(x|c)\, p(c)]}{p(x)} \qquad \text{equ. 3}$$

Where,

- **P(c)** It shows the posterior probability for class c, representing the initial belief about the class distribution before observing the data.
- **P(x|c)** possibility that represents the chance of seeing the feature vector x for a specific class c.
- **P(c|x)** represents the posterior probability of class c after observing x, reflecting the updated belief.
- **P(x)** is the proof or overall likelihood of feature vector x, which remains constant for every class. [43]

The Naive Bayes algorithm is a probabilistic technique in data mining for diabetes prediction due to its simplicity, efficiency, and solid performance, even with high-dimensional datasets typical of medical records. This algorithm quickly processes and interprets clinical data, making it valuable for early intervention and diagnosis.

RESULTS AND DISCUSSIONS

In healthcare, especially when performing early detection of diabetes, the accuracy of a diagnostic model has serious implications, not only for the technical measure of accuracy itself. But for influencing the health outcomes of individuals. Mistaking a diabetic patient for a non-diabetic, or a FN (false negative), could mean not diagnosing someone who requires immediate treatment, missing the opportunity to intervene to help prevent complications, and therefore posing greater risks to the patient's health. On the other hand, a non-diabetic patient mistakenly identified as diabetic, or FP (false positive), may undergo further unnecessary procedures that can cause undue concern. Considering consequences in cases where high stakes are involved.

In this research work, we implemented two different machine learning classifiers, i.e., support vector machine (SVM) and Naïve Bayes (NB), using different evaluation metrics like accuracy, precision, recall(sensitivity), false positive rate (FPR), and False negative rate (FNR). These metrics were calculated using the following formulas [44].

$$\text{Acucuracy} = \frac{\text{no of correct predicton}}{\text{Total no of prediction}} \qquad \text{(equ.4)}$$

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive+False Positive}} \qquad \text{(equ. 5)}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{False Positive+False Negative}} \qquad \text{(equ.6)}$$

$$\text{F1 Score} = \frac{\text{Precision X Recall}}{\text{Precision+Recall}} \qquad \text{(equ. 7)}$$

The Support Vector Machine classifier showed good performance, achieving 91/34% accuracy, while the Naïve Bayes classifier achieved an

accuracy of 87.88% also with feature selection as presented in Table 3.

Table 3: Classifier results comparison with best first search algorithm

MODEL	ACCURACY	PRECISION	RECALL	F1 MEASURE
Support Vector Machine + Best First Search Algorithm	91.34%	93%	93%	93%
Naive Bayes+ Best first search algorithm	87.88%	91%	91%	91%

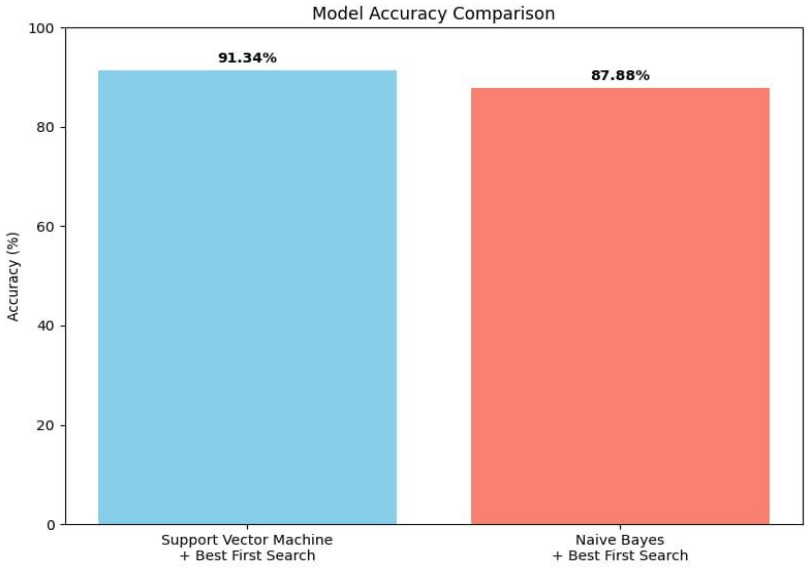

Figure 3: Model's Accuracy Comparison

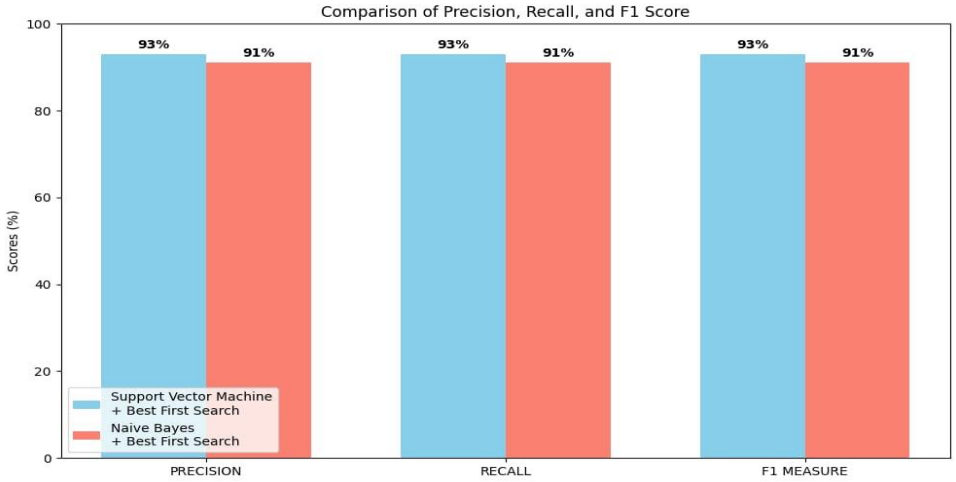

Figure 4: Visual representation of Precision, Recall, and F1 Score

These results show with combination of Best First Search feature selection with SVM provides superior performance on the dataset compared to Naïve Bayes. These results are consistent with previous research on the Pima dataset, where SVM models are consistently found in the top

tier of performance. In prior work, reported accuracy was usually between 75% and 85% accuracy for SVM models without advanced feature selection, indicating that the use of the Best First Search in conjunction with the feature selection approach has enhanced model performance.

DISCUSSION

The results show that feature selection improves classification accuracy for predicting diabetes. The models achieved high accuracy and increased computational efficiency by reducing the dataset from 8 to 5 attributes. The SVM model achieved 91.34% accuracy, performing better than the Naïve Bayes algorithm, which means that utilizing the BFS feature selection method demonstrates the impact that dimensionality reduction can have on predictive modeling. The selected indicators of patients are the most relevant factors associated with predicting diabetes. It delivers insights that support clinical decision-making and future research.

While these results are encouraging, there are limitations. Because of the small sample size, the findings might not apply to larger populations. Future work should include investigating hybrid feature selection methods, using multiple datasets for validation, and an ensemble model to construct a more generalizable and robust system for predicting the risk of diabetes.

REFERENCES

Rahman, M. A. U., Roman, M., Ahmad, S., Jan, M. S., & Nawab, H. U. (2021). Optimizing Collaboration: Insights into Inter-Team Coordination and Self-Management in Distributed Agile Software Development. Webology (ISSN: 1735-188X), 18(5).

Ahmed, N., Ameer, R., Khan, A., Nooruddin, & Saba Gull. (2024). Data mining in Healthcare: An overview of applications, techniques, and challenges. In IJAIMS International Journal of Artificial Intelligence and Mathematical Sciences (Vol. Volume3, Issue 01).

Kolling, M.L.; Furstenau, L.B.; Sott, M.K.; Rabaioli, B.; Ulmi, P.H.; Bragazzi, N.L.; Tedesco, L.P.C. Data Mining in Healthcare: Applying Strategic Intelligence Techniques to Depict 25 Years of Research Development. Int. J. Environ. Res. Public Health 2021, 18, 3099. https://doi.org/10.3390/ijerph18063099

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347-1358.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37-37.

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications, and research directions. SN computer science, 2(3), 160.

Hussain, F., Shams, S. S., Roman, M., Anwar, M., Shah, F. A., Meraj, A. B. E., ... & Uddin, M. A. (2025). Machine Learning in Healthcare: Predicting Chronic Kidney Disease Through Feature-Driven Heuristic Models. Frontier in Medical and Health Research, 3(7), 318-328.

Roman, M., Nawab, H. U., Ahmad, S., & Khan, I. A. (2022). K-Nearest Neighbor and Fuzzy K-Nearest Neighbor Algorithm Performance Analysis for Heart Disease Classification. Webology (ISSN: 1735-188X), 19(1).

Chakraborty, S., & Mali, K. (2020). An overview of biomedical image analysis from the deep learning perspective. Applications of advanced machine intelligence in computer vision and object recognition: emerging research and opportunities, 197-218.

Anwar, M., Rahman, T., & Roman, M. (2025). Voice-Activated Smart Environments: Deep Learning Approach for Pashto Speech Command Processing. Spectrum of Engineering Sciences, 3(5), 541-550.

Tayyab, S.L., Seher, W., Hussain, K., Murtaza, I. (2024). Diabetes: A Global Health Concern and Potential Strategies to Reduce Its Prevalence. In: Rezaei, N. (eds) Integrated Science for Sustainable Development Goal 3. Integrated Science, vol 25. Springer, Cham.

International Diabetes Federation. (2025). IDF Diabetes Atlas (10th ed.). IDF Press. https://www.idf.org/

Lin, X., Xu, Y., Pan, X., Xu, J., Ding, Y., Sun, X., ... & Shan, P. F. (2020). Global, regional, and national burden and trend of diabetes in 195 countries and territories: an analysis from 1990 to 2025. Scientific Reports, 10(1), 1-11.

Hossain, M. J., Al-Mamun, M., & Islam, M. R. (2024). Diabetes mellitus, the fastest growing global public health concern: Early detection should be focused. Health Science Reports, 7(3), e2004.

Sweeting, A., Wong, J., Murphy, H. R., & Ross, G. P. (2022). A clinical update on gestational diabetes mellitus. Endocrine reviews, 43(5), 763-793.

Rastogi, R., & Bansal, M. (2023). Diabetes prediction model using data mining techniques. Measurement: Sensors, 25, 100605.

Fadnavis, R., Dhore, K., Gupta, D., Waghmare, J., & Kosankar, D. (2021, May). Heart disease prediction using data mining. In Journal of physics: conference series (Vol. 1913, No. 1, p. 012099). IOP Publishing.

Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. Procedia computer science, 132, 1578-1585.

Vijayarani, S., & Dhayanand, S. (2015). Data mining classification algorithms for kidney disease prediction. Int J Cybernetics Inform, 4(4), 13-25.

Zriqat, I. A., Altamimi, A. M., & Azzeh, M. (2017). A comparative study for predicting heart diseases using data mining classification methods. arXiv preprint arXiv:1704.02799.

Singh, M. S., Thongam, K., Choudhary, P., & Bhagat, P. K. (2024). An integrated machine learning approach for congestive heart failure prediction. Diagnostics, 14(7), 736.

Sakib, S., Yasmin, N., Tasawar, I. K., Aziz, A., Siddique, M. A. B., & Khan, M. M. R. (2021, September). Performance analysis of machine learning approaches in diabetes prediction. In 2021 IEEE 9th Region 10 Humanitarian Technology Conference (R10-HTC) (pp. 1-6). IEEE.

Kangra, K., & Singh, J. (2024). A genetic algorithm-based feature selection approach for diabetes prediction. International Journal of Artificial Intelligence (IJ-AI), 13(2), 1489-1498.

Zeinalnezhad, M., & Shishehchi, S. (2024). An integrated data mining algorithms and meta-heuristic technique to predict the readmission risk of diabetic patients. Healthcare Analytics, 5, 100292.

El-Sofany, H., El-Seoud, S. A., Karam, O. H., Abd El-Latif, Y. M., & Taj-Eddin, I. A. (2024). A proposed technique using machine learning for the prediction of diabetes disease through a mobile app. International Journal of Intelligent Systems, 2024(1), 6688934.

Guleria, P., Srinivasu, P. N., & Hassaballah, M. (2024). Diabetes prediction using Shapley additive explanations and DSaaS over machine learning classifiers: a novel healthcare paradigm. Multimedia Tools and Applications, 83(14), 40677-40712.

Alkalifah, B., Shaheen, M. T., Alotibi, J., Alsubait, T., & Alhakami, H. (2025). Evaluation of machine learning-based regression techniques for prediction of diabetes levels fluctuations. Heliyon, 11(1).

Modak, S. K. S., & Jha, V. K. (2024). Diabetes prediction model using machine learning techniques. Multimedia Tools and Applications, 83(13), 38523-38549.

Bigdeli, S. K., Ghazisaedi, M., Ayyoubzadeh, S. M., Hantoushzadeh, S., & Ahmadi, M. (2025). Predicting Gestational Diabetes Mellitus in the first trimester using machine learning algorithms: a cross-sectional study at a hospital fertility health center in Iran. BMC Medical Informatics and Decision Making, 25(1), 3.

Ram, A., & Vishwakarma, H. (2021). Diabetes prediction using machine learning and data mining methods. In IOP Conference Series: Materials Science and Engineering (Vol. 1116, No. 1, p. 012135). IOP Publishing.

Khadragy, S., Elshaeer, M., Mouzaek, T., Shammass, D., Shwedeh, F., Aburayya, A., & Aljasmi, S. (2022). Predicting diabetes in United Arab Emirates healthcare: artificial intelligence and data mining case study. South East. Eur. J. Public Heal, 5.

Kumar, M., Sharma, A., & Agarwal, S. (2014). Clinical decision support system for diabetes disease diagnosis using optimized neural network. In 2014 Students Conference on Engineering and Systems (pp. 1-6). IEEE.

Ahad, A., Puspitasari, I., Zheng, J., Ullah, S., Ullah, F., Bakhsh, S. T., & Pires, I. M. (2025). Machine Learning Stroke Prediction in Smart Healthcare: Integrating Fuzzy K-Nearest Neighbor and Artificial Neural Networks with Feature Selection Techniques. Computers, Materials & Continua, 82(3).

Karthikeyan, T., & Thangaraju, P. (2015). Best first and greedy search based CFS-Naïve Bayes classification algorithms for hepatitis diagnosis. Biosciences and Biotechnology Research Asia, 12(1), 983-990.

Roman, M., Ullah, A., Ullah, M. A., Hussain, F., Shams, S. S., Bint-e-Meraj, A., & Ali, S. (2025). Predicting Academic Success: A Machine Learning Approach Using Decision Tables and Random Forests Algorithms. Spectrum of Engineering Sciences, 3(5), 205-213.

Shah, F. A., Meraj, A. B. E., Roman, M., Anwar, M., Zaib, A., Shams, S. S., & Hussain, F. (2025). Enhancing Weather Forecasting Accuracy: A Machine Learning Approach Using Genetic Algorithm and Random Forest. Global Research Journal of Natural Science and Technology.

Huang, S., & Zhou, J. (2025). An enhanced stability evaluation system for entry-type excavations: Utilizing a hybrid bagging-SVM model, GP and kriging techniques. Journal of Rock Mechanics and Geotechnical Engineering, 17(4), 2360-2373.

Bommala, H., Krishna, K. V., Supriya, A., Biradar, R. K., Mayabrahma, B., Ushasree, D., & Kotov, E. V. (2024). Fine-Tunining the Future: Optimizing svm hyper-parameters or enhanced diabetes prediction. In MATEC Web of Conferences (Vol. 392, p. 01082). EDP Sciences.

Patil, R., & Tamane, S. (2018). A comparative analysis on the evaluation of classification algorithms in the prediction of diabetes. International Journal of Electrical and Computer Engineering, 8(5), 3966-3975.

Patil, T. R., & Sherekar, S. S. (2013). Performance analysis of Naive Bayes and J48 classification algorithm for data classification. International journal of computer science and applications, 6(2), 256-261.

Ullah, M. A., Ullah, A., Roman, M., Anwar, M., Siddiqi, M. U. M., Jaffar, M. H., & Ali, J. (2025). A Comparative Study of Machine Learning Algorithms for Cardiovascular Risk Prediction: Support Vector Machine, Gradient Boosting, And Rotation Forest. Spectrum of Engineering Sciences, 481-491.

Bayes, T. (1968). Naive bayes classifier. Article Sources and Contributors, 1-9.

Konan, A. P., Coulibaly, A., Saha, K. B., & Oumtanaga, S. (2025). Diabetes diagnosis using machine learning: A SVM-based approach. Open Journal of Applied Sciences, 15, 1695–1705. https://doi.org/10.4236/ojapps.2025.156116

Roman, M., Naz, I., Luqman, M. A., Ali, J., Jan, M. S., & Nawab, H. U. (2024). Stroke Disease Prediction Using K-Nearest Neighbor and Decision Tree Algorithms with Machine Learning Pre-Processing Techniques. Migration Letters, 21(S4), 2015-2027.