Research Article • DOI: 10.2478/jsiot-2025-0003 JSIoT • 2025(1) • 2025 • 23-30

Journal of Smart Internet of Things

Machine learning-based classification of DNA sequences for diabetes mellitus type prediction

Albegli Ahmed Hasan Ahmed†, Kusum Yadav

College of Computer Science and Engineering. University of Ha'il. Kingdom of Saudi Arabia

Abstract

A machine learning (ML) algorithm is used to classify DNA sequences and predict diabetes risk using the results of this study. Researchers use the INS Insulin Dataset to explore multiple preprocessing strategies such as k-mer representations, ordinal encodings, oversamplings, and min-max normalizations of DNA sequences from diabetic and non-diabetic subjects. The performance of the model was enhanced by using feature selection techniques such as F-regressors and Mutual Information. A study based on accuracy, precision, recall, and F1-score values has been done on four bioinformatics classifiers, including Random Forest, Gaussian Naive Bayes, and Support Vector Machines (SVM). Results demonstrated that Random Forest achieved the highest accuracy (0.89 with F-regressor), followed by SVM and Decision Tree, while Gaussian Naïve Bayes showed moderate performance. The findings highlight the effectiveness of machine learning in uncovering genetic patterns associated with diabetes and emphasize the potential of DNA-based predictive modeling in precision medicine. This work contributes to advancing computational genomics and provides a foundation for early diagnosis and personalized treatment strategies for diabetes mellitus

Keywords

Machine Learning • DNA Sequence Classification • Diabetes Mellitus Prediction • Genomic Data Analysis • Random Forest / Support Vector Machine (SVM).

Received 21 March 2025; Accepted 11 May 2025; Published 15 June 2025

Introduction

There is no better metabolic disorder than diabetes mellitus (DM), characterized by chronic hyperglycemia that occurs as a result of failures within either the secretion process or the action pathway of insulin. International Diabetes Federation reports that diabetes continues to increase in prevalence, and millions of people are at risk of developing heart disease, kidney disease, and neuropathy as a result. Diabetes can be diagnosed early and treated effectively with personalized strategies if it is detected and predicted early.

A recent technological advancement has enabled researchers to gather unprecedented amounts of genomic data that can reveal the genetic basis of complex diseases, including diabetes. Genetic variants are known to increase susceptibility to Type 1 and Type 2 diabetes, and DNA sequences contain essential biological information. Genomics datasets, however, pose significant challenges for traditional statistical methods of analysis because of their vastness and complexity.

DNA sequence data can be analyzed using machine learning to identify hidden patterns and predictive biomarkers, which is a powerful tool for handling large-scale biological data. A ML model can benefit from classification algorithms in order to distinguish between genomic profiles associated with diabetes risk and those without it. By integrating machine learning with genomic data, you can improve prediction accuracy as well as tailor your healthcare decisions based on your genetic background.

A great deal of progress has been made in understanding complex human diseases through GWAS [1]. It is the primary objective of these studies to develop a precise model of the complex structure of gene regulators associated with disease. This genetic disease develops when the pancreas gland fails to produce enough insulin hormone or the insulin hormone cannot be properly used by the body, causing diabetes to develop and last indefinitely. Diabetes Mellitus is an autoimmune disease caused by glucose not being used by the body and raising blood sugar levels. A person with type 2 diabetes (T2D) produces insufficient insulin or fails to respond to insulin in a timely manner [2]. There is a much greater incidence of type 2 diabetes (T2DM) than type 1 diabetes (T1D), which is caused by insulin resistance. In T1D, the body produces no insulin; genes and environmental factors contribute to T2D [3]. There are

†*Corresponding author: Albegli Ahmed Hasan Ahmed*

Email: y.kusum@uoh.edu.sa

Open Access. © 2025 Albegli Ahmed Hasan Ahmed et al., published by Sciendo

This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 License.

some patients with diabetes who are misdiagnosed, and diabetes can actually take either of three forms: 1A, 1B, or 2B. The world is also full of people who have diabetes caused by genetic mutations [4]. It is unnecessary for many of these people to be treated with insulin rather than low-dose medications.

T2D disease is a common condition, and studies are being conducted on how to correctly diagnose it. There are few studies regarding the diagnosis of diseases based on DNA sequences. DNA sequences are converted into digital signals and then analyzed to identify diabetes with spectral images. Using a deep learning approach, this paper identifies the genetic basis (and risk mechanisms) of diabetes-related diseases. In this manner, new drug targets for diabetes-related diseases may be identified.

In addition to its multifaceted nature, diabetes mellitus also has numerous comorbidities, requiring extensive treatment options for all individuals affected by the disorder [5]. Statistical models based on linear statistics used to determine the onset and progression of diabetes mellitus before machine learning algorithms were introduced [6]. By using machine learning, these previously published metadata sets were refined to identify vulnerable groups in need of clinical intervention as well as define biomarkers used for defining pathology [7]. It is possible to predict diabetes severity more accurately and assess diabetes severity more accurately by using HbA1c combined with biomarkers such as 8-hydroxy-2-deoxyguanosine (8-OHdG) and other metabolites.

A machine learning model was used in the current study in order to integrate cardiac physiological, biochemical, genomic, and epigenetic biomarker data in order to determine diabetes type 2. Using machine-learning algorithms, 50 patients were classified as diabetic, mitochondrial function was analyzed, and methylation status was investigated. This study emphasizes the use of novel biomarkers to enhance existing diagnostic standards and to provide more precise methods to identify populations that may be at risk of type 2 diabetes, such as those with prediabetes, and determine whether or not they are developing the condition. As we used machine-learning algorithms to analyze physiological, biochemical, and molecular data, in order to evaluate whether it was possible to predict health outcomes using the best predictive features alone or in conjunction with HbA1c, we looked for features that had the highest predictive accuracy. For the purpose of determining which biomarkers perform best overall, we compared models that did not attain 50% predictive accuracy to models that did in the absence of HbA1c.

Using machine learning, heart imaging diagnostics, such as echocardiography and computed tomography angiography, have been evaluated in order to assess cardiovascular health and outcomes [8]. Research into cardiovascular diseases is expected to grow exponentially as machine-learning applications become more prevalent [9]. Although machine-learning models based on image data are becoming increasingly popular, little is known about their predicative power on heart genomics, epigenetics, proteomics, and metabolomics. A decade ago, datasets were accumulating and compartmentalized, but recent advances have enabled hierarchical predictive algorithms to be integrated with biological processes through metadata, deep sequencing, and omics-based approaches. In order to provide feedback to patients and the general population suffering from disease, machine-learning will be essential. Patients will have access to their personal “omics” profiles, which will enhance the health practices of care providers.

There were 285 million people globally predicted to suffer from diabetes in 2010, a metabolic disorder characterized by high blood sugar levels caused by inadequate insulin production or release. As the disease continues to grow at its present rate, it is projected that by 2030, there will be 552 million diabetes cases. By 2040, one in ten people is expected to have diabetes [10]. As a result, it is of great importance to research diabetes identification and treatment in an effective and timely manner. Using genomic patterns to diagnose diabetes will result in a greater degree of accuracy and precision as well as ensuring better habits are followed that will prevent diabetes from developing. It is possible to slow or postpone the development of future diseases by identifying them very effectively and enjoying better health overall if they are identified early. Using machine learning techniques, one can screen future illnesses and diagnose abnormalities. For example, by analyzing current medical conditions and prior illnesses, forward prediction techniques can anticipate diabetes in real time [11].

Type 2 diabetes, or T2D, is a metabolic condition where the body either doesn’t produce enough insulin or doesn’t use insulin properly, leading to high blood sugar levels. A person’s lifestyle, such as what they eat, how much they exercise, and other daily habits, can greatly affect whether their child develops this disease. T2D can also lead to shorter life expectancy and a lower quality of life. Medications and lifestyle modifications may be used to control illness. To avoid life-threatening consequences, it is crucial that T2D is diagnosed and treated early. Research studies on medical diagnoses have demonstrated impressive accuracy in predicting illness and forecasting the future. Gene combinations are often the main cause of most diseases. To find these specific gene sequences, scientists compare the DNA of healthy people with that of people who have the disease. DNA is important for how cells develop and is passed down from one generation to the next. It is made up of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T). These bases

arrange themselves in a specific order to form the unique genetic code found in nearly every human cell, as mentioned in [12]. The molecular composition analysis of human genes is extensively used to forecast diseases linked to ancestral lineage. Genomics research can aid in modifying individual lifestyles, thereby reducing the likelihood of future diseases. DNA analysis can help predict diseases resulting from DNA mutations. Biomedical engineering has acknowledged a vast gene dataset that can contribute to predicting various health conditions. Using Neural Networks helps identify gene patterns that are harmful to human cells and likely to cause disease. When combined with other methods, this approach can predict illnesses accurately while keeping the processing time reasonable.

Machine Learning is a rapidly growing technology for tackling unavoidable issues across different fields. It employs supervised, semi-supervised, or weakly supervised methods to analyze data from sources such as medical records and wearables to predict illnesses. However, in these methods, early detection of sickness remains limited, and patients cannot significantly alter their lifestyle in a short time to prevent the illness. The polygenic scores method is one of the most widely used strategies for early illness prediction. This method has already been carefully checked and tested before it was used in clinical trials, and it's also used for screening diseases [13]. Recent studies and genetic analysis can help people change their habits and lower the chances of getting conditions like heart attacks, cardiovascular diseases, cancer, and Alzheimer's. Creating polygenic risk scores involves two main steps: discovery and validation. In the discovery stage, risks are found by using statistical tests like linear or logistic regression. The validation stage then confirms these findings and gathers information about specific genetic changes called Single Nucleotide Polymorphisms (SNPs).

Related works

Research on predicting diabetes risk using genetic data combines statistical genomics, feature engineering for sequence data, and modern machine learning and deep learning techniques. Early studies primarily relied on genome-wide association study (GWAS) results and combinations of single-nucleotide polymorphisms (SNPs) as features for classical classifiers and risk scores. For instance, efforts to uncover genotype-phenotype links and apply machine learning models to genotype data showed that supervised methods can identify SNP combinations linked to Type 2 Diabetes (T2D) and enhance risk stratification beyond single-variant analyses.

Recent comparative studies have started to evaluate classical machine learning algorithms (such as SVM, Random Forest, and gradient-boosted trees) against deep learning models using DNA-sequence-derived features for diabetes prediction. Research published between 2024 and 2025 indicates that when DNA sequences are properly encoded—such as through k-mers, TF-IDF on k-mers, or learned embeddings—both well-tuned gradient-boosted trees and deep learning models can be competitive. However, the advantages of deep learning models often depend on having larger labeled datasets and effective regularization. A comprehensive 2025 PLOS study examining various ML and deep learning pipelines for DNA-sequence-based diabetes classification emphasizes both the potential and the challenges related to reproducibility in this emerging area [14].

Classical machine learning models and feature selection in genomics

A support vector machine (SVM) model was developed as soon as genomic classification began to be studied [15, 16]. Microarray data can be effectively selected using SVMs for gene selection, particularly cancer classification. An extension of the approach to genomic sequence classification demonstrated that the model was capable of identifying genetic patterns specific to each class. A linear and kernel-based classifier can be applied to biological data based on this foundational research.

Natural language processing techniques and data imbalance solutions

It is recommended that natural language processing (NLP) methods be applied to genomic sequences for the purpose of extracting structured features. As a symbolic language, DNA sequences are encoded. A combination of deep learning and NLP has been used to classify regulatory DNA elements, outperforming traditional statistical descriptors. There is a significant problem of class imbalance in genomic studies. In order to increase the representation of minorities, synthetic minority sampling techniques (SMOTE) were proposed [20]. Several applications of it have been successful, resulting in improved model sensitivity and reduced bias substantially [21]. Beyond simple classification, there is increasing interest in integrating genetic sequence data with other molecular layers such as gene expression and methylation, along with clinical measures and explainability tools. Multi-omics machine learning frameworks have enhanced Type 2 diabetes prediction and identified potential biomarkers, such as methylation or expression signals, that aid in interpreting model outcomes. Simultaneously, progress in variant effect prediction, like large models that assess missense variants, offers complementary tools to identify potentially pathogenic sequence changes that can be utilized by downstream diabetes prediction models. These advancements indicate a trend towards hybrid pipelines that

merge variant-level pathogenicity scoring, sequence-level representation learning, and clinical data integration [22].

Diabetes mellitus (DM) is a condition that affects how the body processes blood sugar, leading to high levels over time. It also impacts how the body handles fats, carbs, and proteins. Type 2 diabetes mellitus (T2DM) is connected to problems with lipids and lipoproteins in the blood. This includes lower levels of high-density lipoprotein (HDL) cholesterol, higher numbers of small, dense low-density lipoprotein (LDL) particles, and increased triglyceride (TG) levels. These issues can happen even if LDL cholesterol is in the normal range. These changes are part of a condition called insulin resistance syndrome, which plays a big role in the development of T2DM. People with diabetes often have different lipid-related risks, such as higher total cholesterol (TC), LDL, and triglycerides, as well as lower HDL levels [23].

An impaired insulin production or reduced insulin sensitivity is commonly observed in diabetes mellitus, which is a chronic metabolic disorder with elevated blood sugar levels. By 2045, there will be 693 million adults affected by this disease, making it one of the world's fastest-growing diseases. In order to detect and intervene early when diabetic complications arise, predictive modeling is necessary, particularly in regard to kidney disease, retinal disease, and neuropathy. Diabetics who suffer from these complications are at an increased risk of death, vision loss, kidney failure, and a reduced quality of life. Despite their importance, clinical risk factors and glycemic control fail to predict vascular complications reliably. Currently, genetic biomarkers and machine learning techniques can be used to assess diabetes risk and complications.

There is a great deal going on in the public health system when it comes to diabetes mellitus (DM) and its high prevalence, as well as the complex interaction of genetic factors and environmental factors that contribute to its development. It is estimated that there will be 537 million adults worldwide affected by type 2 diabetes mellitus (T2DM) by 2021, out of which over 90% will have type 2 diabetes mellitus (T2DM). Patients suffering from this disease have a wide range of phenotypes due to its complex causes, including their age at onset, complications they experience, and the effectiveness of their management techniques. It is estimated that 69% of T2DM is inherited genetically in individuals between 35 and 60 years of age, in spite of the fact that environmental and lifestyle factors are known risk factors [24].

The use of machine learning techniques in genomic diagnostics is widespread, especially the use of algorithms that analyze time series and deep learning. As a result of these approaches, it is now possible to analyze functional DNA sequences, which are key to understanding gene function and regulation. Recently, machine learning has been successfully applied to a variety of biomedical challenges such as genome annotation, variant calling, variant classification, and genotype-to-phenotype prediction [25]. This review will highlight promising integrations and solutions from recent advancements in state-of-the-art.

Proposed methodology

Diabetes can be diagnosed using machine learning techniques with the proposed system. Initially, the data undergoes preprocessing, followed by the application of four distinct classification methods, with the results being compared. Figure 1 depicts the proposed system in a schematic form, offering further detail.

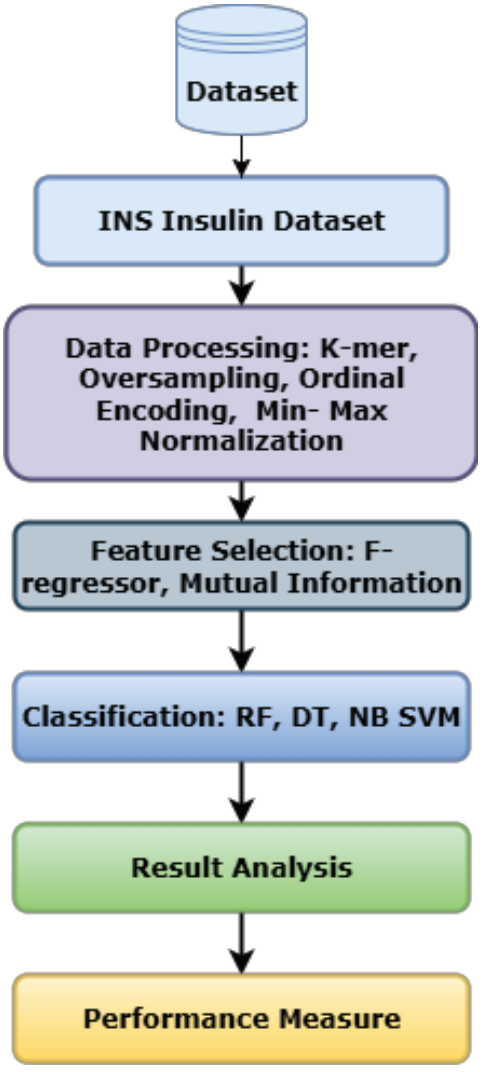

Figure 1. Proposed System.

Dataset

INS Insulin Dataset has been made available by the National Agency for Health, a renowned research institute in health and medicine. A genetic variation in insulin was found in the diabetes mellitus dataset. In addition to DNA sequences and clinical data, the INS Insulin Dataset includes additional diabetic patients with T1D and T2D. Ten,199 sequences are diabetes-related, while 4,371 are non-diabetic [26].

Data preprocessing

Machine learning requires effective preprocessing of DNA data in order to correctly identify diabetes mellitus types. A one-shot encoding approach, K-mer representations, minimum-maximum normalizations, oversampling, and converting lists to text are investigated as ways. Described in this paper are the different techniques for diabetes mellitus disease study, as well as their significance. Machine learning requires the correction of missing values in diabetes mellitus study DNA data before it can be analyzed. Among the methods employed in diabetes mellitus genetic pattern analysis and interpretation are deletion, imputation, and multiple imputations. Using machine learning models for diabetic mellitus research improves reliability and contributes to major discoveries.

K-mer

It is difficult to analyze long, complicated DNA sequences. By overlapping k-length sequences with different patterns, by using the k-mer method, this problem can be overcome. Through the reduction of dimensionality and the collection of patterns and motifs specific to diabetics, in this algorithm, diabetic features are identified using machine learning. AGA, GAT, and AGAT are composed of four monomers, three dimers, two tetramers, and one tetramer of AGAT. Six-k-mer sequences are derived from the DNA sequence of the molecules.

Oversampling

It is common for genetic research to use datasets that underrepresent one class. Increasing minority class samples balances classes. As a result of random duplication or synthetic minority oversampling, minorities have sufficient training data to avoid bias when using random sampling methodologies. The algorithm's ability to handle imbalanced scenarios is enhanced by this strategy.

Ordinal Encoding

Machine learning preprocessors commonly use ordinal encoding. Diabetes research often uses it to process DNA. By converting categorical data into numerical representations, including diabetes types, categorical data can be maintained while also being transformed into numerical representations. Unique integers are used in ordinal encoding to preserve variable order. Computer algorithms can then handle input features more easily, facilitating classification exploration.

Min- Max Normalization

It is common to normalize feature values using min-max normalization, which rescales them between 0 and 1.

$$X\_norm = (X - X\_min) / (X\_max - X\_min) \quad (1)$$

There are four main values in this dataset - Xmin, Xmax, Xmin norm and Xmin norm.

Feature selection

When categorizing diabetes mellitus from DNA sequences, machine learning must select features. Feature selection using F-regressors and Mutual Information is compared in an academic study. Genes associated with diabetes are found using this approach. The principles, approaches, and their usefulness in this investigation are presented.

F-regressor

Based on ANOVA F-values, F-regressors select features. The feature is correlated to the target variable individually, and the F-value is calculated to identify features that are strongly associated with the type of diabetes mellitus. A higher value indicates a stronger association, in determining whether diabetics are discriminated against, it is the most important feature.

Mutual Information

It evaluates statistical dependence based on mutual information. The target variable information is evaluated in feature selection. It is capable of detecting linear and non-linear correlations in DNA sequences with complex dependencies. It is more useful and preferable to analyze feature-target variables based on Mutual Information values. Set entropy is calculated with the help of equations (3) and (4), based on the values of H(X) and H(Y).

$$(x) = - \sum i \in x\ (x)\ log\ (x) \quad (3)$$

$$(y) = - \sum j \in y\ (y)(y) \quad (4)$$

We can calculate mutual data as follows:

$$(x, y) = (x) + (y) - (x, y) \quad (5)$$

Random Forest

By combining DT algorithms, RF predicts or classifies the value of variables [27]. As a result, multiple regression trees

are constructed and the results are averaging based on input vectors *x* and evidential feature values. A RF regression predictor based on such trees $\{T(x)\}_1^K$ is as follows:

$$\int_{rf}^{K} (x) = \frac{1}{K}\sum_{k=1}^{K} T(x)$$

GaussianNB

Gaussian distributions are used in GNB's probabilistic ML classification algorithms. A Gaussian Naive Bayes model predicts an outcome based on each feature. All components are combined to predict the probability of categorizing a group dependent variable. Assuming Gaussian feature probabilities, equation (6) is as follows:

$$p(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\ exp\ \left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right) \quad (6)$$

Support Vector Machine

For classification, supervised learning is an effective method. The goal of SVMs is to create into categories as much as possible. A SVM can handle separable data with linear and nonlinear components using kernel functions. This equation is used to calculate RBF kernels:

$$k(\times i + x_j) = exp\ exp\ (\gamma\|x_i + x_j\|^2) \quad (7)$$

A learnable constraint, gamma (γ), denotes the RBF kernel in this situation. Due to its robust generalization and ability to deal with high-dimensional datasets, Researchers studying diabetes mellitus often use SVMs to classify DNA sequence patterns and correlations.

Decision Tree

There is an algorithm called DT that is used in data mining [28]. A classification algorithm constructs models using inductive learning processes involving reclassified data sets. Attributes are used to define data items. In decision trees, attributes are mapped to categories based on their values, and its values are used to classify data items. These attributes are used to partition data items, and the process is repeated recursively once a subset of data items has been partitioned. There are no differences between data items within a subset, DTs separate data based on attributes at each node. There are several edges on each node, and each edge is labeled according to its parent attribute. Even the leaves are categorized according to their decision values. An important part of decision trees is the use of statistical classifiers to classify data. In a recursive selection process, each data point is classified according to the classes that differentiate the target application from others. Assume that *X* is the features of the data point and *Y* is the class; the decision is based on the ratio between *X* and *Y* as follows:

$$RATIO\ (X \mid Y) = \frac{H(X) - H(X \mid Y)}{H(X)} \quad (5)$$

Result and discussion

An analysis of DNA genes, a large dataset, and the presentation of features all influence diabetes classifiers. In order to evaluate their performance, actual and predicted labels are compared in terms of accuracy, precision, recall, and F1-score [29]. Dataset properties and research goals should be considered when designing algorithms, and redundancy should be avoided through preprocessing. As a part of this study, we used f-regressors and mutual information as a method for selecting features for analyzing diabetic DNA. In order to evaluate the model accuracy, we assessed 75 features and 5–10 K-mer values, with an 80/20 split used for train-test comparisons. Based on the results shown in Figure 2, we can compare the performance of the models among different methods of selecting features.

Random Forest demonstrated strong accuracy with 0.89 using the F-regressor and 0.88 with the mutual information method, even though the precision, recall, and F1-Score were competitive. Gaussian Naive Bayes performed less effectively than Random Forest, as seen in Figure 3, showing only decent performance during F-regressor trials and overall average results with acceptable precision across test cases. This method is suitable for many single hypothesis problems but less effective in scenarios with more dimensions. A significant improvement was made in the classification algorithms by using SVMs and Decision Trees. Mutual information and the F-regressor were used to evaluate each model. SVMs based on mutual information and F-regressor had accuracy of 0.83 and 0.84, respectively. It was found that SVMs, like Random Forests, improved reliability and F1-scores as features were selected. Using the mutual information approach, there was good precision and recall for Decision Trees, however, they were less accurate than Random Forests and SVM classifier algorithms. It is more reliable and easier to achieve academic success when multiple classifiers are combined in deep neural networks or ensembles.

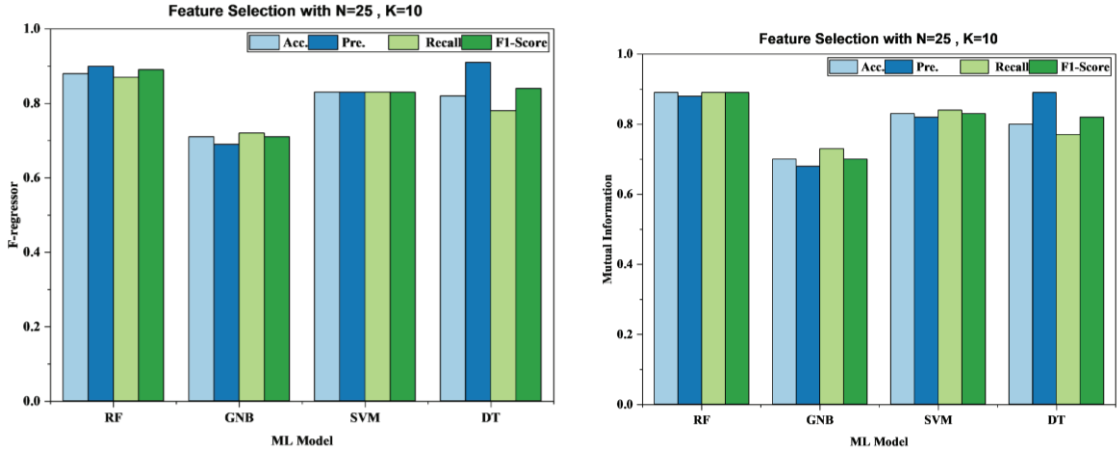

Figure 2. Performance analysis of ML model with N=75 and K=10.

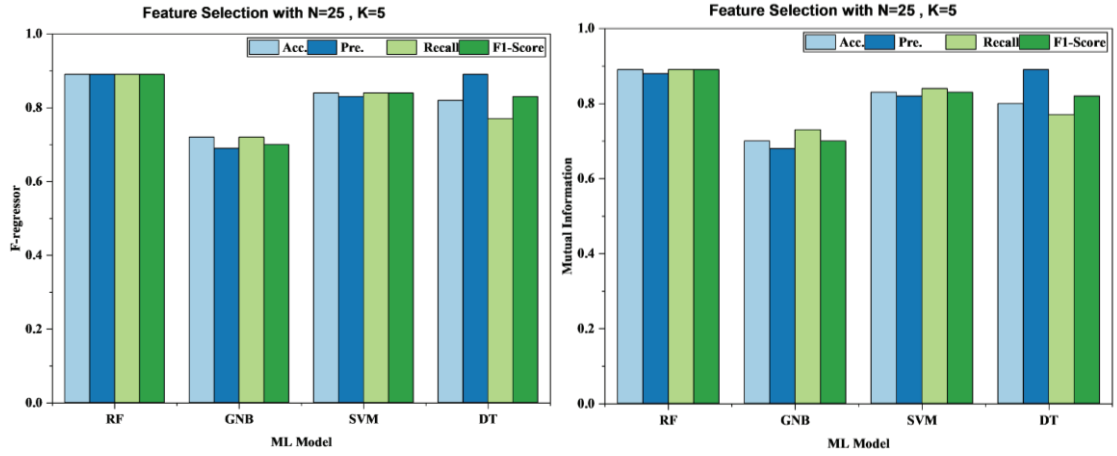

Figure 3. Performance analysis of ML model with N=75 and K=5.

Conclusion

The study demonstrates that machine learning models, when combined with effective feature representation and selection techniques, can accurately classify DNA sequences for predicting diabetes mellitus types. Among the tested models, Random Forest and SVM consistently outperformed other classifiers, showing strong predictive power and robustness across evaluation metrics. The integration of genomic data with machine learning approaches highlights the feasibility of early and precise diabetes detection, which is crucial for effective intervention and personalized healthcare strategies. Despite limitations such as data imbalance and model dependency on preprocessing, this research establishes a framework for using genomic information in clinical diagnostics. Future work may extend these findings by incorporating multi-omics data, deep learning architectures, and larger datasets to further enhance predictive accuracy and clinical applicability.

Funding Statement: The authors received no specific funding for this study

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] D. S. W. Ho, W. Schierding, M. Wake, R. Saffery, and J. O'Sullivan, "Machine Learning SNP Based Prediction for Precision Medicine," *Front. Genet.*, vol. 10, p. 267, Mar. 2019, doi: 10.3389/fgene.2019.00267.

[2] R. A. DeFronzo *et al.*, "Type 2 diabetes mellitus," *Nat Rev Dis Primers*, vol. 1, no. 1, p. 15019, Jul. 2015, doi: 10.1038/nrdp.2015.19.

[3] A. Mahajan *et al.*, "Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps," *Nat Genet*, vol. 50, no. 11, pp. 1505–1513, Nov. 2018, doi: 10.1038/s41588-018-0241-6.

[4] J. W. Kleinberger and T. I. Pollin, "Personalized medicine in diabetes mellitus: current opportunities and future prospects," *Annals of the New York Academy of Sciences*, vol. 1346, no. 1, pp. 45–56, Jun. 2015, doi: 10.1111/nyas.12757.

[5] E. Capobianco, "Systems and precision medicine approaches to diabetes heterogeneity: a Big Data perspective," *Clinical & Translational Med*, vol. 6, no. 1, p. e23, Dec. 2017, doi: 10.1186/s40169-017-0155-4.

[6] M. Massi-Benedetti, "Changing targets in the treatment of type 2 diabetes," *Current Medical Research and Opinion*, vol. 22, no. sup2, pp. S5–S13, Aug. 2006, doi: 10.1185/030079906X112714.

[7] 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, doi: 10.1016/j.csbj.2016.12.005.

[8] S. J. Al'Aref *et al.*, "Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging," *European Heart Journal*, vol. 40, no. 24, pp. 1975–1986, Jun. 2019, doi: 10.1093/eurheartj/ehy404.

[9] K. Shameer, K. W. Johnson, B. S. Glicksberg, J. T. Dudley, and P. P. Sengupta, "Machine learning in cardiovascular medicine: are we there yet?," *Heart*, vol. 104, no. 14, pp. 1156–1164, Jul. 2018, doi: 10.1136/heartjnl-2017-311198.

[10] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting Diabetes Mellitus With Machine Learning Techniques," *Front. Genet.*, vol. 9, p. 515, Nov. 2018, doi: 10.3389/fgene.2018.00515.

[11] H. M. Deberneh and I. Kim, "Prediction of Type 2 Diabetes Based on Machine Learning Algorithm," *IJERPH*, vol. 18, no. 6, p. 3317, Mar. 2021, doi: 10.3390/ijerph18063317.

[12] A. Arshad and Y. D. Khan, "DNA Computing A Survey," in *2019 International Conference on Innovative Computing (ICIC)*, Lahore, Pakistan: IEEE, Nov. 2019, pp. 1–5. doi: 10.1109/ICIC48496.2019.8966707.

[13] H.-C. So and P. C. Sham, "Exploring the predictive power of polygenic scores derived from genome-wide association studies: a study of 10 complex traits," *Bioinformatics*, vol. 33, no. 6, pp. 886–892, Mar. 2017, doi: 10.1093/bioinformatics/btw745.

[14] S. A. Salloum, K. M. Alomari, and A. Salloum, "DNA sequence classification for diabetes mellitus using NuSVC and XGBoost: A comparative," *PLoS One*, vol. 20, no. 7, p. e0328253, Jul. 2025, doi: 10.1371/journal.pone.0328253.

[15] Y. Jiang *et al.*, "Immunomarker support vector machine classifier for prediction of gastric cancer survival and adjuvant chemotherapeutic benefit," *Clinical Cancer Research*, vol. 24, no. 22, pp. 5574–5584, 2018.

[16] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification using Support Vector Machines," *Machine Learning*, vol. 46, no. 1–3, pp. 389–422, Jan. 2002, doi: 10.1023/A:1012487302797.

[17] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA: ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[18] B. López, F. Torrent-Fontbona, R. Viñas, and J. M. Fernández-Real, "Single Nucleotide Polymorphism relevance learning with Random Forests for Type 2 diabetes risk prediction," *Artificial Intelligence in Medicine*, vol. 85, pp. 43–49, Apr. 2018, doi: 10.1016/j.artmed.2017.09.005.

[19] Y.-J. Huang, C. Chen, and H.-C. Yang, "AI-driven Integration of Multimodal Imaging Pixel Data and Genome-wide Genotype Data Enhances Precision Health for Type 2 Diabetes: Insights from a Large-scale Biobank Study," Jul. 26, 2024. doi: 10.1101/2024.07.25.24310650.

[20] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *jair*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[21] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," *Information Sciences*, vol. 250, pp. 113–141, Nov. 2013, doi: 10.1016/j.ins.2013.07.007.

[22] T. Rönn, A. Perfilyev, N. Oskolkov, and C. Ling, "Predicting type 2 diabetes via machine learning integration of multiple omics from human pancreatic islets," *Sci Rep*, vol. 14, no. 1, p. 14637, Jun. 2024, doi: 10.1038/s41598-024-64846-3.

[23] R. M. Krauss, "Lipids and Lipoproteins in Patients With Type 2 Diabetes," *Diabetes Care*, vol. 27, no. 6, pp. 1496–1504, Jun. 2004, doi: 10.2337/diacare.27.6.1496.

[24] for the Botnia Study Group *et al.*, "Heritability and familiality of type 2 diabetes and related quantitative traits in the Botnia Study," *Diabetologia*, vol. 54, no. 11, pp. 2811–2819, Nov. 2011, doi: 10.1007/s00125-011-2267-5.

[25] "Accurate Genomic Prediction of Human Height," *Genetics*, vol. 214, no. 1, pp. 231–231, Jan. 2020, doi: 10.1534/genetics.119.302946.

[26] "National Library of Medicine." [Online]. Available: https://www.ncbi.nlm.nih.gov

[27] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[28] B. Ingre, A. Yadav, and A. K. Soni, "Decision Tree Based Intrusion Detection System for NSL-KDD Dataset," in *Information and Communication Technology for Intelligent Systems (ICTIS 2017) - Volume 2*, vol. 84, S. C. Satapathy and A. Joshi, Eds., in Smart Innovation, Systems and Technologies, vol. 84. , Cham: Springer International Publishing, 2018, pp. 207–218. doi: 10.1007/978-3-319-63645-0_23.

[29] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A Fast Learning Algorithm for Deep Belief Nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006, doi: 10.1162/neco.2006.18.7.1527.