Al-Bahir Journal for Engineering and Pure Sciences
BJEPS

Volume 3 | Issue 2 Article 3

Predictive Diabetes Mellitus From DNA Sequences Using Deep Learning

Lena abed ALraheim Hamza
College for Women, University of Babylon, Babylon, Iraq.

Hussein Attya Lafta
Hussein Attya Lafta works at College of Information Technology, University of Babylon, Babylon, Iraq

Sura Zaki Al-Rashid
College of Information Technology, University of Babylon, Babylon, Iraq.

ORIGINAL STUDY

Predictive Diabetes Mellitus from DNA Sequences Using Deep Learning

Lena abed A. Hamza [a,*], Hussein A. Lafta [b], Sura Z. Al_Rashid [b]

[a] College for Women, University of Babylon, Babylon, Iraq
[b] College of Information Technology, University of Babylon, Babylon, Iraq

Abstract

Diabetes is a chronic metabolic disorder characterized by elevated blood sugar levels. It manifests in different forms, with type 1 and type 2 being the most prevalent. Type 1 diabetes results from the autoimmune destruction of insulin-producing cells, whereas type 2 diabetes primarily stems from insulin resistance.

Despite advancements in treatment, accurate detection and prediction of diabetes remain challenging. Early diagnosis is crucial for effective management and prevention of complications. Another obstacle lies in interpreting vast amounts of health data, including DNA sequencing, which poses difficulties for healthcare professionals in identifying relevant patterns and associations.

Artificial intelligence (AI) holds promise in healthcare by developing and training deep learning algorithms to analyze health data and DNA sequences. The research paper focuses on applying both Convolutional Neural Networks (CNNs) algorithm, in addition to Long Short-Term Memory (LSTM) algorithm for predicting types of diabetes based on DNA sequencing. The study aims to leverage the power of CNN and LSTM, known for their proficiency in analyzing image and sequence data, to accurately classify diabetes types.

The experimental results of the proposed CNN-LSTM model showcased remarkable performance, achieving a recorded accuracy of 100% on a labeled dataset that included DNA sequencing and corresponding diabetes types. The model's evaluation encompassed several metrics, including accuracy, recall, precision, and the F1 score.

Keywords: DNA sequencing, Genetic data analysis, Long short-term memory (LSTM), Kmer, Convolutional Neural Networks (CNNs), Deep learning

1. Introduction

Diabetes, also known as diabetes mellitus, is a group of metabolic disorders characterized by high blood sugar levels (elevated amount of glucose in the blood) due to disorder in insulin production, inactivity of insulin, or both of the previous reasons. The insulin is a hormone produced by beta cells in the pancreas and regulates the level of sugar amount in the blood. The prevalence of diabetes is increasing worldwide, leading to significant public health implications [1].

The diagnosis and management of diabetes have been defined by the World Health Organization (WHO) since 1965. The diagnostic criteria for diabetes include fasting plasma glucose concentration and, in some cases, random plasma glucose concentration. The minimum fasting plasma glucose concentration for diagnosis is 126 mg/dL (7 mmol/L), and the minimum random plasma glucose concentration is 200 mg/dL (11.1 mmol/L) in the presence of typical symptoms. These criteria are used for routine diagnosis and epidemiological studies of diabetes [1,2].

Diabetes can be classified into two main types: Type 1 diabetes (T1DM) and Type 2 diabetes (T2DM) [1]. Type 1 diabetes manifest when the immune system in human body mistakenly attacks the beta cells in the pancreas that produce insulin, resulting in little or no insulin production. It is

Received 12 July 2023; revised 26 July 2023; accepted 28 July 2023.
Available online 30 August 2023

* Corresponding author.
E-mail addresses: lena.alrahiem.gsci112@student.uobabylon.edu.iq (L.A. Hamza), wsci.husein.attia@uobabylon.edu.iq (H.A. Lafta), sura_os@itnet.uobabylon.edu.iq (S.Z. Al_Rashid).

https://doi.org/10.55810/2313-0083.1042

usually diagnosed in childhood or early adolescence and requires insulin replacement therapy [2]. On the other hand, Type 2 diabetes occurs when the body cells resisting the consumption of insulin, resulting in glucose accumulation in the blood. Some of the factors contributing to T2D are physical inactivity, obesity, and unhealthy practices contribute to the development of T2D. Initially, it can be controlled by altering the daily lifestyle, oral medications, or insulin therapy if needed [3].

Diabetes is a chronic condition, and when poorly controlled, it can lead to various complications affecting the structure and function of body tissues. It can cause symptoms such as excessive thirst (polydipsia), increased urination (polyuria), unexplained weight loss, and increased hunger [4]. Severe cases of diabetes, particularly those with high blood sugar levels and accompanying ketone acidosis (increased ketone levels), can lead to serious complications and, in severe cases, may result in death if left untreated [4].

In recent years, DNA sequencing has played a crucial role in disease prediction and understanding the underlying mechanisms of human diseases [5]. Genetic factors are known to contribute to the development of diabetes, and mutations in specific genes, such as the insulin gene, can disrupt insulin production and lead to the onset of diabetes [6]. Analyzing gene expressions and DNA sequences can provide valuable insights into disease susceptibility and potential therapeutic targets.

AI, particularly deep learning, in recent years, contributed positively to various areas, including clinical medicine and genomics research. AI systems have the ability to analyze large health datasets, interpret patterns, and produce a prognosis to aid in disease detection and management. The classification of DM into four types, as illustrated in Fig. 1 [7], is endorsed by the World Health Organization (WHO).

In clinical medicine, AI techniques have been applied to tasks such as electrocardiogram analysis, radiological image interpretation, and natural language processing for health record analysis [8]. In genomics research, AI can assist in analyzing genetic data and identifying disease-related patterns and associations.

Several studies have explored the potential of AI techniques, particularly deep learning methods, in diabetes detection and prediction. Collaborative computing-based approaches, machine learning algorithms, and bioinformatics methods have been utilized to classify gene expressions, predict and diagnose diabetes, and analyze genetic data associated with the disease. In study [9], a collaborative computing-based approach was used to classify gene expressions for Type 2 Diabetes. The study employed the K-Nearest Neighbour (KNN) classifier to differentiate between control samples and insulin-exposed samples, achieving a test classification accuracy of 100%. Another study [10] aimed to predict diabetes using machine learning

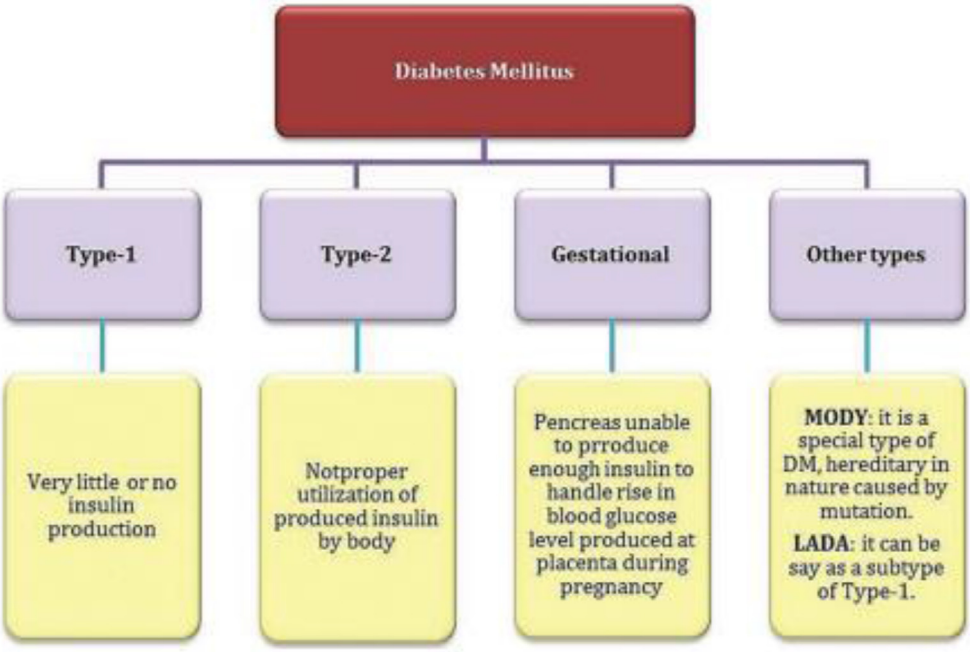

Fig. 1. Illustrates the classification of diabetes mellitus according to the (WHO) [7].

algorithms, including Linear Regression and Support Vector Machine (SVM). The Linear Regression model demonstrated a moderate determination coefficient, while SVM achieved a testing accuracy of 0.8290960451977402. In the proposed work [11], decision tree, random forest, and neural network algorithms were utilized to predict and diagnose diabetes mellitus. The results obtained from Random Forest algorithm is the highest value. With accuracy value of 0.8084 when considering all attributes. Study [12] proposed a new algorithm for identifying Type 2 Diabetes (T2D) risk variables using a combination of techniques, including CNN methods (ResNet and VGG19) and classifiers like SVM and k-NN. The developed algorithm achieved an accuracy rate of 99.09%.Research [13] focused on bioinformatics systems and the analysis of genetic data, exploring alignment-free methods to assess similarity between DNA sequences. The study utilized a dataset of 860 genomes and experimented with different word sizes (k values) for k-mer analysis. In study [14], authors aimed to predict DNA binding sites specific to transcription factors (TFs) using the SVM algorithm. The identification of TF binding sites contributes to understanding gene regulation. However, specific details about the dataset and findings were not mentioned in the provided abstract.

These studies collectively highlight the potential of AI techniques, particularly deep learning, in diabetes detection and prediction using various approaches and datasets. The Table 1 below summarizes the key findings from these relevant studies.

In previous studies, researchers have primarily focused on utilizing healthcare indicators for prediction or classification tasks related to diabetes. However, there has been a limited emphasis on DNA analysis in this context. Therefore, our research aims to address this gap by proposing a deep learning model based on CNNs for detecting diabetes using DNA sequencing data. To provide an overview of the existing literature, Table 1 summarizes the key findings from relevant studies related to predicting diabetes mellitus.

Table 1. Summary of the previous researches on diabetes mellitus prediction

Ref	models	Accuracy rate %
[9] (2022)	KNN	100
[10] (2021)	LR-SVC	70–83
[11] (2018)	RF-ANN	80.84
[12] (2021)	CNN-RNN	99.9 CNN- SVM
[13] (2019)	alignment-based and alignment-free methods	–
[14] (2019)	CNN-SVC	–

In this work, a more sophisticated model will be proposed with higher and better results obtained. The proposed model and the related data preprocessing procedure will be explained properly.

2. Materials and methods

Machine learning has revolutionized the automatic detection of crucial patterns in data. Recent years in research, showed that machine learning has become widely used and effective approach in various tasks that involve discovering patterns in big amount of data [15].

Deep learning, a subfield of the wider field of Machine Learning. Deep Learning relies on the Neural Network to learn from data and produce future prognosis. It takes inspiration from the structure and functioning of the human brain, aiming to replicate the process of learning and extracting meaningful patterns from vast amounts of data. In deep learning, artificial neural networks consist of multiple interconnected layers of nodes called neurons. Each neuron receives input from the previous layer, performs computations, and passes the output to the next layer. Deep learning models typically have numerous hidden layers, enabling the extraction of hierarchical representations of the input data.

One significant advantage of deep learning is its capacity to autonomously learn hierarchical representations directly from raw data, eliminating the need for manual feature engineering. By training on extensive datasets, deep learning models excel at capturing intricate patterns and relationships. Consequently, they have proven highly effective in some jobs. These including signal and image processing and recognition, speech recognition, the process of human natural languages, and recommendation systems.

Several well-known deep learning (DL) architectures have emerged, each designed for specific data types and tasks. CNNs are widely used for image and video analysis, while LSTM models excel in processing sequential data. Transformer models have gained prominence in natural language processing tasks. The application of deep learning has brought about remarkable advancements in various domains, including computer vision, speech recognition, and healthcare [16–18].

2.1. Convolutional Neural Networks (CNNs)

CNN, or Convolutional Neural Network, is a powerful and widely used model for large-scale neural networks. It draws inspiration from the visual mechanism observed in living organisms,

making it particularly effective in tasks involving image and pattern recognition [19].

The strength of CNN lies in its ability to extract higher-level abstract features by analyzing artificial neurons across the input matrix. It focuses on identifying translation-invariant patterns at each position through the computation of locally weighted sums. By doing so, CNN can generate the expected output values [20].

CNN's architecture typically consists of several key components. It starts with an input layer, followed by a convolutional layer. The contents of convolutional layer are number of filters. These filters will be applied to the input signal, extracting important features. The Rectified Linear Unit (ReLU) is an activation function. Which apply the non-lineraty by into the model. Subsequently, a pooling layer operates on each feature map from the convolutional layer, reducing the dimensionality and capturing the most salient features. Finally, the output layer is a fully connected neural network that produces the final output based on an activation function [22].

Through its multiple layers, a CNN can capture the intrinsic features of raw datasets, representing different levels of abstraction [21]. This hierarchical representation allows CNNs to excel at tasks such as image classification, object detection, and image segmentation. Fig. 2 provides an illustration of the standard CNN structure. Fig. 2 provides an illustration of the standard CNN architecture.

CNNs have found extensive application in various image processing tasks, leveraging their multilayered structure to effectively identify image features that enhance the classification process. Furthermore, CNNs have emerged as an appealing approach for text classification, particularly for character-based analysis. Consequently, CNNs have been utilized in the analysis of DNA sequences to detect and characterize crucial elements such as promoters and binding sites [23].

2.2. Long Short-Term Memory (LSTMs)

LSTM, which stands for Long Short-Term Memory, is a specific type of recurrent neural network (RNN) that addresses the challenge of learning long-term dependencies in sequential data. It overcomes the limitations of traditional RNNs by incorporating three gates. Which are input, forget, and output gates. In addition to memory cell.

The memory cell in LSTM allows it to retain and propagate information over longer sequences, enabling the model to capture and remember relevant context from the past. The input gate controls the flow of new information into the memory cell, while the forget gate determines which information should be discarded. The output gate regulates the output based on the current input and the information stored in the memory cell.

LSTM has proven to be effective in various tasks such as speech recognition and time series prediction, where long-term dependencies play a crucial role. The ability of LSTM to capture and retain important information over extended sequences makes it well-suited for handling sequential data.

Researchers have explored different variations and extensions of LSTM to further enhance its performance in specific domains. These modifications aim to improve memory retention, increase model capacity, and address other challenges encountered in practical applications.

Overall, LSTM is robust approach for processing sequential data, and it has demonstrated impressive results in various fields. Its default structure consists of a memory cell and three gating mechanisms, as depicted in Fig. 3 [24].

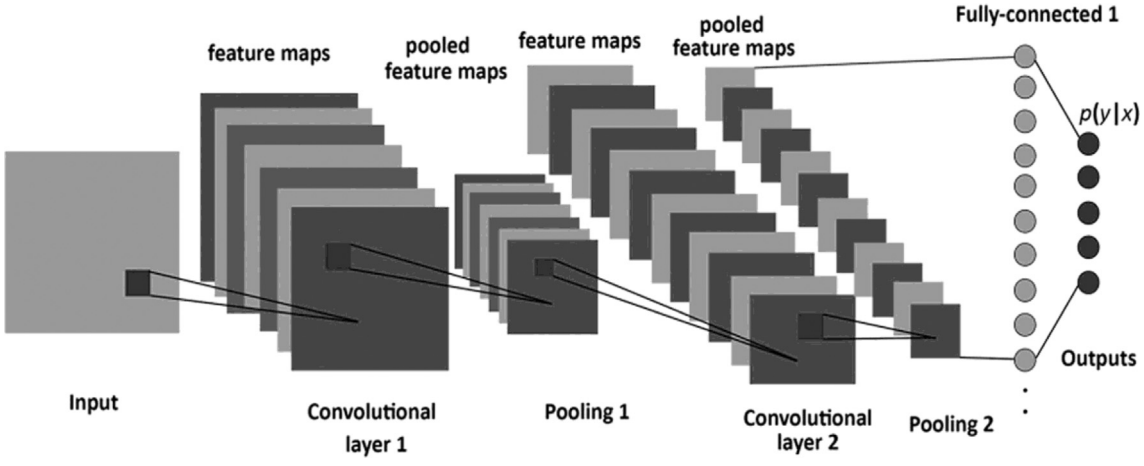

Fig. 2. CNN example.

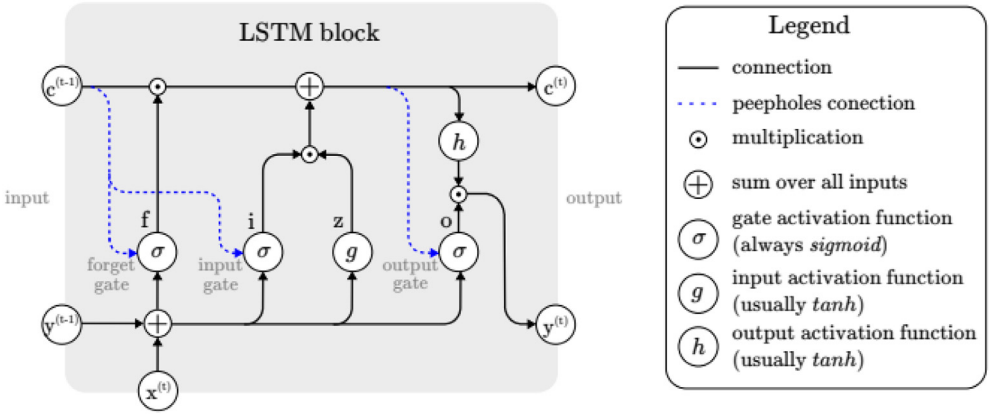

Fig. 3. The default architecture of the LSTM.

2.3. Dataset description

We utilize the dataset representing the INS (Insulin) gene, which can be found at the following: [https://www.ncbi.nlm.nih.gov/gene?Db=gene&Cmd=DetailsSearch&Term=3630] (https://www.ncbi.nlm.nih.gov/gene?Db = gene&Cmd = DetailsSearch&Term = 3630). The INS gene, symbolized by INS, is a well-studied gene in humans. It has the official full name of Insulin and a gene ID of 3630. The gene is located on chromosome 11p15.5. INS is a protein-coding gene, meaning it encodes a protein. Its primary function is to produce insulin, a hormone that plays a crucial role in regulating blood sugar (glucose) levels in the body. Insulin facilitates the uptake of glucose into cells and regulates its utilization, storage, and production in various tissues, including the liver, muscles, and adipose (fat) tissue. This hormone is essential for maintaining glucose homeostasis.

Genetic variations or mutations in the INS gene can lead to various forms of diabetes. Type 1 diabetes mellitus (T1DM) is one of these conditions, characterized by insufficient production of insulin by the pancreas. Mutations in INS can also cause other rare forms of diabetes, such as permanent neonatal diabetes mellitus (PNDM) and maturity-onset diabetes of the young (MODY). These disorders are associated with disruptions in insulin function and glucose regulation. Fig. 4 shown The default structure of the INS gene [25].

This dataset is selected due to its public available, and easy to interpret. Fig. 5 shows a sample of the used datasets.

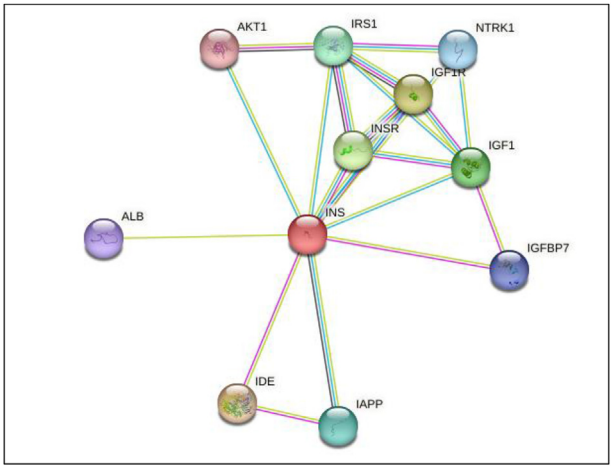

Fig. 4. The default structure of the INS gene [25].

2.4. Proposed methodology

Predicting diabetes using DNA sequence data involves several stages, including preprocessing, K-mer representation, oversampling, encoding, normalization, model training, and testing. This is illustrated in the block diagram representing the proposed model in Fig. 6.

1. The first stage, preprocessing, focuses on preparing the DNA dataset for analysis. This step typically involves cleaning the data, handling missing values, and ensuring data consistency. It may also include removing irrelevant features or filtering out noise to enhance the quality of the dataset.
2. Next, the K-mer representation stage is applied, where DNA sequences are divided into overlapping subsequences of length K. This technique captures the local composition and order of DNA bases and can be a powerful tool for DNA sequence analysis. By representing the DNA sequences in this way, the model can identify important patterns and motifs that contribute to the prediction of diabetes.

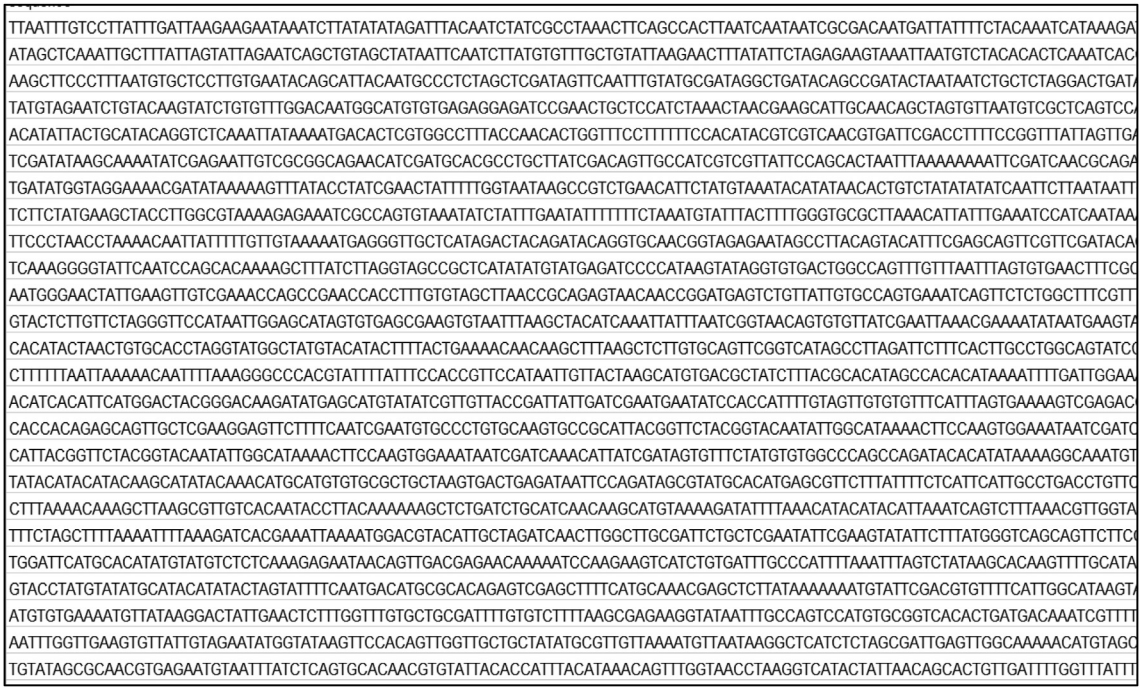

Fig. 5. sample of the used dataset.

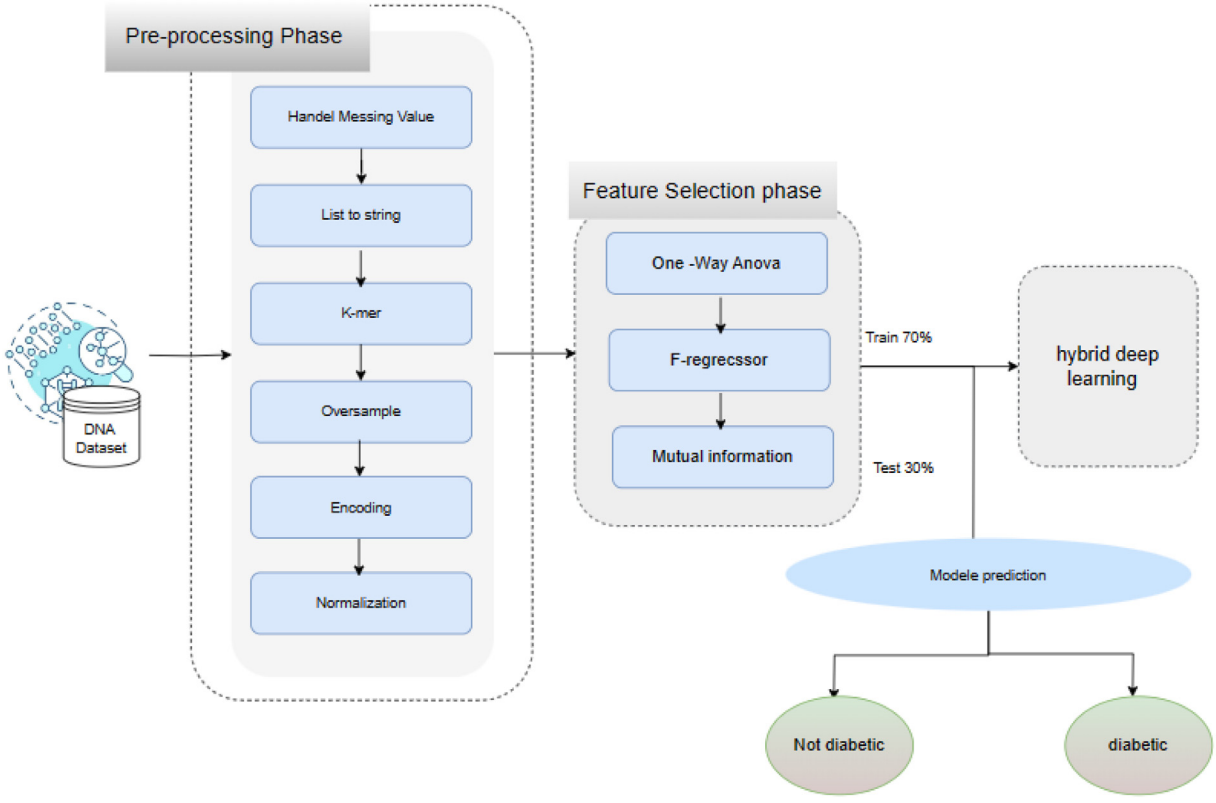

Fig. 6. Proposed System diagram.

3. To address class imbalance issues in the dataset, oversampling techniques can be employed. Oversampling involves generating synthetic samples of the minority class to balance the distribution and improve the model's ability to learn from both classes effectively. This step helps prevent bias towards the majority class, leading to more accurate predictions.
4. Ordinal encoding the DNA sequences is crucial for transforming the raw genetic information into a format that deep learning models can understand. Various encoding schemes can be utilized, such as one-hot encoding or embedding techniques, to represent the DNA sequences as numerical inputs. This encoding allows the model to capture the underlying patterns and relationships within the DNA sequences.
5. Normalization is another essential stage in the pipeline, which aims to standardize the features' scale and range. This step ensures that different features have a similar influence on the model's training process. Normalization prevents certain features from dominating others, which can result in biased predictions and hinder the model's performance. the min–max normalization method is applied to all genes for normalizing the values to avoid high values that might affect the calculation of the results, according to the Equation (1):

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

All the numeric values of the genes in the DNA dataset that serve as input to the feature selection method were normalized to be in the range zero and one.

6. Feature selection: The sequential feature selection method was applied in a study on the detection of diabetes using DNA sequencing. Three feature selection methods, namely "One-way ANOVA," "F-regressor," and "Mutual Information," were employed.

In the first step, the "One-way ANOVA" method was used to identify variables that exhibit statistical significance in distinguishing between different groups of diabetic patients and healthy individuals. The probability value was utilized to determine variables that show significant statistical differences between the different groups.

The equation for calculating the Mutual Information between a feature X and a target variable Y is as equation (2):

$$MI(X, Y) = \Sigma\Sigma\ p(x, y) * \log(p(x, y) / (p(x) * p(y))) \quad (2)$$

Where:

p (x, y) is the joint probability mass function of X and Y.

p(x) and p(y) are the marginal probability mass functions of X and Y, respectively.

Next, in the subsequent step, the "F-regressor" method was employed to identify variables that are highly correlated with the disease. This method relies on estimating the strength of the relationship between independent variables and the dependent variable using regression analysis. Variables with the highest correlation strength with the disease were selected to form a subset of the data.

The equation for calculating the Fisher Ratio for a feature is as in equation (3):

$$FR = (mean1 - mean2)\hat{}2 / (var1 + var2) \quad (3)$$

Where:

mean1&2 are the means of the feature in two different classes.

var1 &2 are the variances of the feature in two different classes.

In the final step, the "Mutual Information" method was utilized to select variables that carry important and informative insights for distinguishing between diabetic patients and healthy individuals. This method measures the mutual dependence between variables to determine the extent of their shared influence.

The equation for calculating the F-value for a feature using one-way ANOVA is as in equation (4):

$$F = (MSB / MSW) \quad (4)$$

Where:

Mean Square Between (MSB) is the mean square of the variance between the groups.

Mean Square Within (MSW) is the mean square of the variance within the groups.

By employing the sequential feature selection methods, a subset of the data containing the most important variables for diabetes detection using DNA sequencing was obtained. This selected subset was then used as input for deep learning models (specifically, a CNN-LSTM model) to achieve improved accuracy in discriminating between diabetic patients and healthy individuals. The classification models were trained on 70% of the subset as training data and then used to predict the

classification of patients and healthy individuals in the remaining 30% of the data.

E. Hybrid deep learning layers

The proposed approach illustrated in Fig. 6 combines LSTM and CNN architectures. It consists of several layers, including Conv1D, MaxPooling1D, LSTM, and Dense layers. It has a total of 19,874 trainable parameters. The model follows a sequential connectivity, where each layer transforms the input data and passes it to the next layer. The model starts with Conv1D layers for feature extraction, followed by MaxPooling1D layers for downsampling. LSTM layers are then used to capture temporal dependencies. Another Conv1D layer is applied, followed by MaxPooling1D and LSTM layers. Finally, a Flatten layer reshapes the data, and a Dense layer with two output units is used for classification purposes. The model's architecture enables efficient processing of sequential data and learning meaningful representations for prediction tasks. Fig. 7 shows the CNN- LSTM architectures.

Layer (type)	Output Shape	Param #
conv1d_1 (Conv1D)	(None, 26, 16)	96
max_pooling1d_1 (MaxPooling1	(None, 26, 16)	0
conv1d_2 (Conv1D)	(None, 22, 32)	2592
max_pooling1d_2 (MaxPooling1	(None, 22, 32)	0
lstm_1 (LSTM)	(None, 22, 32)	8320
max_pooling1d_3 (MaxPooling1	(None, 22, 32)	0
conv1d_3 (Conv1D)	(None, 18, 32)	5152
max_pooling1d_4 (MaxPooling1	(None, 18, 32)	0
lstm_2 (LSTM)	(None, 18, 16)	3136
max_pooling1d_5 (MaxPooling1	(None, 18, 16)	0
flatten_1 (Flatten)	(None, 288)	0
dense_1 (Dense)	(None, 2)	578

Total params: 19,874
Trainable params: 19,874
Non-trainable params: 0

Fig. 7. The CNN- LSTM architectures.

3. Results and discussion

The experimental tests were performed on the INS gene dataset, and the dataset was divided into two groups. The first group comprised 50% of the features and involved 6 batches of kmer, while the second group represented 20% of the features and also consisted of 6 batches of kmer. Table 2 shows the number of the selected features with respect to k-mer sizes. The performance of the trained model, CNN-LSTM, was evaluated using performance metrics, namely accuracy, recall, and F1 score. Notably, most metrics achieved a remarkable accuracy of 100%. The experimental results, including the performance of the proposed model, are presented in Tables 3–14 of the research. These findings highlight the effectiveness of the CNN-LSTM

Table 2. Number of the selected features with respect to k-mer sizes.

K-mer size	Number of words	Number of features 20%	Number of features 50%	Number of features 75%
3	299	60	149	224
4	298	60	149	223
5	297	60	148	222
6	296	59	148	222
7	295	59	147	221
8	294	59	147	220

Table 3. Feature selection with $N = 50$ and $K = 3$.

Feature method	Accuracy	Precision	Recall	F1-Score
One-way ANOVA	1.00	1.00	1.00	1.00
FR	1.00	1.00	1.00	1.00
MI	1.00	1.00	1.00	1.00

Table 4. Feature selection with $N = 50$ and $K = 4$.

Feature method	Accuracy	Precision	Recall	F1-Score
One-way ANOVA	1.00	1.00	1.00	1.00
FR	1.00	1.00	1.00	1.00
MI	1.00	1.00	1.00	1.00

Table 5. Feature selection with $N = 50$ and $K = 5$.

Feature method	Accuracy	Precision	Recall	F1-Score
One-way ANOVA	1.00	1.00	1.00	1.00
FR	1.00	1.00	1.00	1.00
MI	1.00	1.00	1.00	1.00

Table 6. Feature selection with $N = 50$ and $K = 6$.

Feature method	Accuracy	Precision	Recall	F1-Score
One-way ANOVA	1.00	1.00	1.00	1.00
FR	1.00	1.00	1.00	1.00
MI	1.00	1.00	1.00	1.00

Table 7. Feature selection with N = 50 and K = 7.

Feature method	Accuracy	Precision	Recall	F1-Score
One-way ANOVA	1.00	1.00	1.00	1.00
FR	1.00	1.00	1.00	1.00
MI	1.00	1.00	1.00	1.00

Table 8. Feature selection with N = 50 and K = 8.

Feature method	Accuracy	Precision	Recall	F1-Score
One-way ANOVA	1.00	1.00	1.00	1.00
FR	1.00	1.00	1.00	1.00
MI	1.00	1.00	1.00	1.00

Table 9. Feature selection with N = 20 and K = 3.

Feature method	Accuracy	Precision	Recall	F1-Score
One-way ANOVA	0.99	0.99	0.99	0.99
FR	0.99	0.99	0.99	0.99
MI	0.94	0.94	0.94	0.94

Table 10. Feature selection with N = 20 and K = 4.

Feature method	Accuracy	Precision	Recall	F1-Score
One-way ANOVA	0.98	0.98	0.98	0.98
FR	0.96	0.96	0.96	0.96
MI	0.97	0.97	0.97	0.97

Table 11. Feature selection with N = 20 and K = 5.

Feature method	Accuracy	Precision	Recall	F1-Score
One-way ANOVA	0.96	0.96	0.96	0.96
FR	0.92	0.92	0.92	0.92
MI	0.99	0.99	0.99	0.99

Table 12. Feature selection with N = 20 and K = 6.

Feature method	Accuracy	Precision	Recall	F1-Score
One-way ANOVA	0.97	0.97	0.97	0.97
FR	1.00	1.00	1.00	1.00
MI	0.98	0.98	0.98	0.98

Table 13. Feature selection with N = 20 and K = 7.

Feature method	Accuracy	Precision	Recall	F1-Score
One-way ANOVA	0.99	0.99	0.99	0.99
FR	1.00	1.00	1.00	1.00
MI	0.91	0.91	0.91	0.91

Table 14. Feature selection with N = 20 and K = 8.

Feature method	Accuracy	Precision	Recall	F1-Score
One-way ANOVA	1.00	1.00	1.00	1.00
FR	0.99	0.99	0.99	0.99
MI	0.99	0.99	0.99	0.99

model in accurately predicting and analyzing the INS gene dataset.

The tables above present the results of feature selection experiments with different values of N (number of features) and K (kmer value). Each table shows accuracy, precision, recall, and F1 score for three feature selection methods: one-way ANOVA, FR, and MI.

1. The feature selection methods achieved high performance when 50% of N (number of features) were selected. The accuracy, precision, recall, and F1 score were consistently recorded as 1.00 for each feature selection method.
2. The performance varied for different values of K (kmer value) by 20%. The table indicates that the path 8 achieved the highest performance in terms of accuracy, precision, recall, and F1 score for all three feature selection methods: one-way ANOVA, FR, and MI. However, the recall and F1 score decreased to 94% and 88%, respectively, for the MI method at K = 7, as shown in Table 12.

When comparing and discussing the results from the above tables based on the 50% and 20% feature ratios, along with the K values, the following main points can be noted:

- The results indicate that the feature selection methods (one-way ANOVA, FR, MI) are highly effective in selecting features from the INS gene dataset.
- The 50% feature ratio achieved higher accuracy in classifying types of diabetes with high precision.
- The performance of K (kmer value) varied at the 20% feature ratio, and its optimal path was observed at K = 8.

Overall, the experimental results shows that CNN-LSTM model is effective in accurate prediction and analysis of the INS gene dataset. The model achieves high accuracy, recall, and F1 score, and different feature selection methods show similar performance.

4. Conclusion

The experimental results presented in this paper shows how the CNN-LSTM model is effective in accurately predicting and analyzing the INS gene dataset. The model achieved a remarkable accuracy of 100% across all feature selection methods and performed well in both of the metrics of recall and F1 score.

The feature selection methods (one-way ANOVA, FR, MI) proved to be highly effective in selecting features from the INS gene dataset. When 50% of the features were selected, the classification accuracy and precision were notably high. This indicates that the selected features were able to successfully classify different types of diabetes.

Additionally, the performance of the model varied with different values of K (kmer value) at the 20% feature ratio. The optimal performance was observed at K = 8, where the model gained higher values of accuracy, precision, recall, and F1 score metrics.

Overall, the hybrid LSTM-CNN model has demonstrated excellent results in predicting diabetes using DNA sequence data. It achieved a remarkable accuracy rate and perfect scores in precision, recall, and F1 score, highlighting the effectiveness of deep learning approaches in addressing complex biological problems.

5. Future works

Moving forward, there are several important avenues for future research. Firstly, incorporating CNNs into the research can further enhance the automation, detection, and prediction of diabetes. CNNs excel in analyzing complex patterns in large datasets, making them well-suited for diabetes detection and prediction can lead to improved accuracy and efficiency.

Furthermore, future research can explore the possibility of applying the proposed model in other healthcare and medical domains. Predicting other diseases or enhancing the diagnosis of complex conditions could benefit from similar deep learning approaches.

References

[1] Forouhi NG, Wareham NJ. Epidemiology of diabetes. Medicine Jan 2019;47(1):22–7. https://doi.org/10.1016/j.mpmed.2018.10.004.

[2] Morris AP. Progress in defining the genetic contribution to type 2 diabetes susceptibility. Curr Opin Genet Dev Jun. 2018;50:41–51. https://doi.org/10.1016/j.gde.2018.02.003.

[3] Seo J-W, Lee K-J. Post-translational Modifications and Their Biological Functions: Proteomic Analysis and Systematic Approaches. BMB Rep Jan 2004;37(1):35–44. https://doi.org/10.5483/bmbrep.2004.37.1.035.

[4] Baslé E, Joubert N, Pucheault M. Protein Chemical Modification on Endogenous Amino Acids. Chem Biol Mar. 2010;17(3):213–27. https://doi.org/10.1016/j.chembiol.2010.02.008.

[5] Kinsner W. Towards cognitive analysis of DNA. In: Proc. 9th IEEE int. Conf. On cognitive informatics, Beijing, China; 2010. https://doi.org/10.1109/coginf.2010.5599728.

[6] Nguyen NG, et al. DNA Sequence Classification by Convolutional Neural Network. J Biomed Sci Eng 2016;9(5):280–6. https://doi.org/10.4236/jbise.2016.95021.

[7] El-Attar NE, Moustafa BM, Awad WA. Deep learning model to detect diabetes mellitus based on DNA sequence. Intell Autom Soft Comput 2022;31(1):325–38. https://doi.org/10.32604/iasc.2022.019970.

[8] Shrestha A, Mahmood A. Review of Deep Learning Algorithms and Architectures. IEEE Access 2019;7:53040–65. https://doi.org/10.1109/access.2019.2912200.

[9] Al Rashid SZ. Collaborative Computing-Based K-Nearest Neighbour Algorithm and Mutual Information to Classify Gene Expressions for Type 2 Diabetes. Int J e-Collab 2022;18(2):12. https://doi.org/10.4018/IJeC.304044.

[10] Rajeswari SVKR, VijayakumarPonnusamy. Prediction of Diabetes Mellitus Using Machine Learning Algorithm. In: Annals of the Romanian society for cell biology; 2021. p. 5655–62. Retrieved from, https://www.annalsofrscb.ro/index.php/journal/article/view/6653.

[11] Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. Predicting Diabetes Mellitus With Machine Learning Techniques. Front Genet Nov 2018;9. https://doi.org/10.3389/fgene.2018.00515.

[12] Das B. A deep learning model for identification of diabetes type 2 based on nucleotide signals. Neural Comput Appl Mar 2022;34(15):12587–99. https://doi.org/10.1007/s00521-022-07121-8.

[13] Shanan NAA, Lafta HA, Alrashid SZ. Using alignment-free methods as preprocessing stage to classification whole genomes. Research Paper. Babylon, Iraq: Computer Department, Science College for Women, University of Babylon; 2021. https://doi.org/10.22075/IJNAA.2021.5281.

[14] Aziz FA, Al-Rashid SZ. Prediction of DNA Binding Sites Bound to Specific Transcription Factors by the SVM Algorithm. Int J Sci 2022;63(11):37. https://doi.org/10.24996/ijs.2022.63.11.37.

[15] Shalev-Shwartz Shai, Ben-David Shai. Understanding machine learning : from theory to algorithms. Delhi: Cambridge University Press; 2015.

[16] Mathew A, Amudha P, Sivakumari S. Deep Learning Techniques: An Overview. In: Advances in intelligent systems and computing; May 2020. p. 599–608. https://doi.org/10.1007/978-981-15-3383-9_54.

[17] Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, Santamaría J, Fadhel MA, Al-Amidie M, Farhan L. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data 2021;8(1):53. https://doi.org/10.1186/s40537-021-00421-4.

[18] Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge, Massachusetts: The Mit Press; 2016.

[19] Khan S, Rahmani H, Shah A, Bennamoun M. Convolutional neural network. Jan 2018. p. 43–68. https://doi.org/10.1007/978-3-031-01821-3_4.

[20] Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, et al. A primer on deep learning in genomics. Nat Genet 2019;51(1):12–8. https://doi.org/10.1038/s41588-018-0295-5.

[21] Nonis F, Barbiero P, Cirrincione G, Olivetti EC, Marcolin F, Vezzetti E. Understanding Abstraction in Deep CNN: An Application on Facial Emotion Recognition. 2021. p. 281–90. https://doi.org/10.1007/978-981-15-5093-5_26.

[22] El Attar NE, Hassan MK, Alghamdi OA, Awad WA. Deep learning model for classification and bioactivity prediction of essential oil producing plants from Egypt. Sci Rep 2020;10(1):1–10. https://doi.org/10.1038/s41598-020-78449-1.

[23] Mohammadpoor M, Sheikhi M. A deep learning algorithm to detect coronavirus (COVID-19) disease using CT images. Peer J Comp Sci 2021;3:1–12. https://doi.org/10.1007/s00330-021-07715-1.

[24] Van Houdt G, Mosquera C, Nápoles G. A review on the long short-term memory model. Artif Intell Rev 2020;53:5929–55. https://doi.org/10.1007/s10462-020-09819-9.

[25] The INS gene and its putative association with human ageing. https://genomics.senescence.info/genes/entry.php?hgnc=INS (accessed Jul. 1, 2023).