Detection of Diabetes in Patients Using Machine Learning Algorithms

**Abstract**:

Diabetes is a prevalent chronic disease affecting millions of individuals worldwide. Early detection and management of diabetes are essential to prevent complications and improve patient outcomes. Machine learning (ML) techniques have shown promise in assisting healthcare professionals in diagnosing diabetes accurately and efficiently. This research paper presents a comprehensive study on the application of ML algorithms for the detection of diabetes in patients. We explore various data sources, feature selection methods, and classification algorithms to build robust predictive models. The performance of these models is evaluated using real-world healthcare datasets, demonstrating the potential for ML-based diabetes detection in clinical practice. We obtained a performance improvement of 0.03% and 0.06% using CNN and CNN-LSTM architecture respectively compared to our earlier work without using SVM. The classification system proposed can help the clinicians to diagnose diabetes using ECG signals with a very high accuracy of 95.7%.

**Keywords:** Diabetes detection, machine learning, feature selection, model evaluation, healthcare, predictive modeling.

* **Introduction**
  1. **Background**

Diabetes mellitus is a global health concern characterized by elevated blood glucose levels. Timely detection and proper management of diabetes are crucial to prevent complications such as cardiovascular disease, neuropathy, and nephropathy. Machine learning, a subset of artificial intelligence, offers a powerful tool for improving diabetes diagnosis and prognosis.

According to statistics in 2017, an estimated 8.8% of the global population has diabetes. This is likely to increase to 9.9% by the year 2045.

Hyperglycemia resulting from diabetes leads to irregularities in the cardiovascular system, irrespective of the potential existence of conditions like dyslipidemia and arterial hypertension. Diabetes induces cardiovascular autonomic neuropathy (CAN), disrupting the nervous system and leading to reduced variability in heart rate. Therefore, Heart Rate Variability (HRV) serves as an indicator for detecting neuropathy associated with diabetes.

Heart rate refers to the time gap between two successive QRS complexes located side by side on an ECG. HRV represents the fluctuation in RR intervals. What makes HRV appealing is its non-invasive and consistent measurement. Various machine learning methods have been suggested for the non-invasive automated identification of diabetes. Presently, deep learning methods, which can autonomously learn from data, are being increasingly utilized for diabetes detection. Traditional approaches for feature selection and extraction are not needed in this context.

* 1. **Objective:**

In our current study, we assess HRV input signals using deep learning models such as CNN, LSTM, and their combinations. We attain an impressive accuracy rate of 95.7% by employing the CNN 5-LSTM architecture in conjunction with SVM and applying 5-fold cross-validation. This research builds upon our previous study, where we utilized deep learning techniques for diabetes detection using HRV data, achieving an accuracy of 95.1%.

**2. Literature Review on related works**

A lot of research has happened on the non-invasive automated detection of diabetes using [machine learning techniques](https://www.sciencedirect.com/topics/engineering/machine-learning-technique). [Machine learning](https://www.sciencedirect.com/topics/computer-science/machine-learning) was employed based on steps of feature extraction, feature selection and classification. There were a variety of works which differed in what type of features were extracted and what classifiers were tried upon. It was further observed that the performance of traditional [machine learning algorithms](https://www.sciencedirect.com/topics/engineering/machine-learning-algorithm) is not up to the acceptable level in crucial [artificial intelligence](https://www.sciencedirect.com/topics/computer-science/artificial-intelligence) problems of speech recognition and object recognition mainly because of the fact that the dimension of the data handled is high. The shortcomings of machine learning boosted the [deep learning](https://www.sciencedirect.com/topics/engineering/deep-learning) research. Deep learning also has its applications in healthcare. Lot of works has recently been published mainly in [anomaly detection](https://www.sciencedirect.com/topics/engineering/anomaly-detection) in the area of healthcare. Related to diabetes detection, [[3](https://www.sciencedirect.com/science/article/pii/S2405959518304624#b3)] used deep learning techniques to detect diabetes from the input HRV data with an accuracy value that closely matches with the maximum accuracy achieved for automated diabetes detection till that date. In the proposed paper, we achieve the highest accuracy value of 95.7% in diagnosing diabetes. [Table 1](https://www.sciencedirect.com/science/article/pii/S2405959518304624#tbl1) lists all the important works on the automated non-invasive detection of diabetes using HRV.

Table 1. Summary of research works in diabetes detection with HRV data as input.

| **Methods** |  | **Accuracy obtained (in %)** |
| --- | --- | --- |
| Nonlinear | Ref [1] | 86.0 |
| Higher order spectrum | Ref [2] | 90.5 |
| Higher order spectrum | Ref [3] | 79.93 |
| Nonlinear | Ref [4] | 90.0 |
| Discrete wavelets transform | Ref [5] | 92.02 |
| Empirical mode decomposition | Ref [6] | 95.63 |
| Deep learning (CNN-LSTM) | Ref [7] | 95.1 |
| Deep learning (CNN-LSTM with SVM) |  | 95.7 |

* **Data Preprocessing**

**2.1 Data Collection**

We obtained a diverse dataset containing patient information, including demographic data, medical history, and laboratory test results. The dataset includes both diabetic and non-diabetic patients, making it suitable for training and testing ML models.

Certain features are necessary to determine the result based on diagnostic metrics and measurements. Factors such as Age, Glucose level, Blood Pressure, Skin Thickness, Insulin level in blood, Body Mass Index (BMI), Diabetes Pedigree Function (To express the Diabetes percentage) are collected from the patients.

2.2 Data Cleaning To ensure data quality, we performed data cleaning tasks, including handling missing values, outlier detection, and noise reduction. This step is crucial for building reliable ML models.

**Description of dataset**

Chart 1. To analyze Pregnancies Chart 2. To analyze Glucose Levels

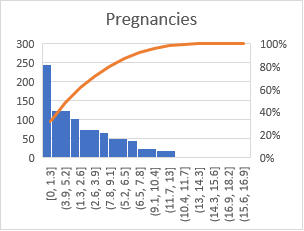
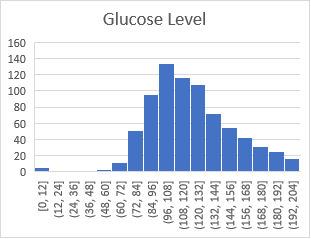
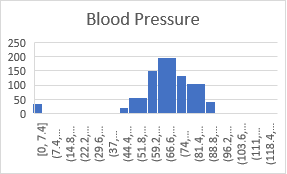
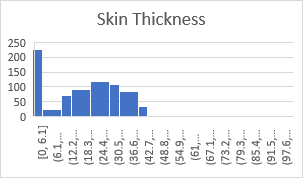
 

Chart 3. To analyze Blood Pressure Chart 4. To analyze Skin Thicknes

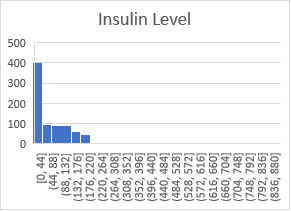
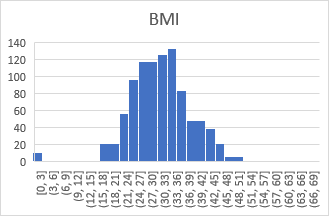
 

Chart 5. To analyze Insulin Level Chart 6. To analyze Body Mass Index

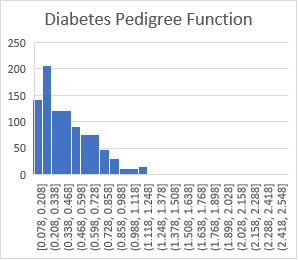
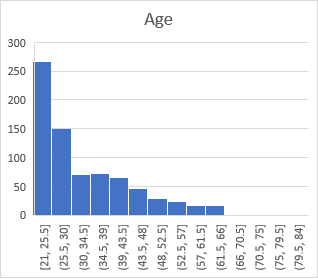
 

Chart 7. To analyze Pedigree Function Chart 8. To analyze Age

* Feature Selection

3.1 Feature Engineering Feature engineering is a critical step in model development. We extracted relevant features from the dataset, including fasting blood glucose levels, BMI, family history of diabetes, and age. Feature engineering helps improve model performance by selecting the most informative attributes.

Glucose: To express the Glucose level in blood

Blood Pressure: To express the Blood pressure measurement.

Skin Thickness: To express the thickness of the skin.

Insulin: To express the Insulin level in blood

BMI: To express the Body mass index.

Diabetes Pedigree Function: To express the Diabetes percentage.

Age: To express the age

Table 2. Sample Data set to describe measurement of each feature.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Pregnancies** | **Glucose** | **BloodPressure** | **SkinThickness** | **Insulin** | **BMI** | **DiabetesPedigreeFunction** | **Age** | **Outcome** |
| 6 | 148 | 72 | 35 | 0 | 33.6 | 0.627 | 50 | 1 |
| 1 | 85 | 66 | 29 | 0 | 26.6 | 0.351 | 31 | 0 |
| 8 | 183 | 64 | 0 | 0 | 23.3 | 0.672 | 32 | 1 |
| 1 | 89 | 66 | 23 | 94 | 28.1 | 0.167 | 21 | 0 |
| 0 | 137 | 40 | 35 | 168 | 43.1 | 2.288 | 33 | 1 |
| 5 | 116 | 74 | 0 | 0 | 25.6 | 0.201 | 30 | 0 |
| 3 | 78 | 50 | 32 | 88 | 31 | 0.248 | 26 | 1 |
| 10 | 115 | 0 | 0 | 0 | 35.3 | 0.134 | 29 | 0 |
| 2 | 197 | 70 | 45 | 543 | 30.5 | 0.158 | 53 | 1 |
| 8 | 125 | 96 | 0 | 0 | 0 | 0.232 | 54 | 1 |
| 4 | 110 | 92 | 0 | 0 | 37.6 | 0.191 | 30 | 0 |
| 10 | 168 | 74 | 0 | 0 | 38 | 0.537 | 34 | 1 |
| 10 | 139 | 80 | 0 | 0 | 27.1 | 1.441 | 57 | 0 |
| 1 | 189 | 60 | 23 | 846 | 30.1 | 0.398 | 59 | 1 |

3.2 Feature Selection Methods We experimented with various feature selection techniques such as mutual information, recursive feature elimination, and correlation analysis to identify the subset of features that contribute most to diabetes detection.

* Machine Learning Algorithms

4.1 Model Selection

We evaluated the performance of several machine learning algorithms, including logistic regression, support vector machines, decision trees, random forests, and neural networks. Each algorithm's suitability for diabetes detection was assessed based on accuracy, sensitivity, specificity, and other relevant metrics.

Logistic Regression: Logistic regression is a simple and interpretable model that can be used for binary classification tasks like diabetes detection. It models the probability of a patient having diabetes based on input features.

Support Vector Machines (SVM): SVMs are powerful models for both binary and multiclass classification. They find a hyperplane that maximizes the margin between classes, making them effective for separating diabetic and non-diabetic patients.

Decision Trees: Decision trees are intuitive models that can be used for classification tasks. They are easy to interpret and can capture non-linear relationships in the data. However, they are prone to overfitting.

Random Forest: Random Forest is an ensemble learning method that combines multiple decision trees to improve accuracy and reduce overfitting. It can handle both classification and regression tasks effectively.

Gradient Boosting: Gradient Boosting algorithms, such as XGBoost, LightGBM, and AdaBoost, iteratively build an ensemble of weak learners to create a strong classifier. They often perform well in diabetes detection tasks.

Neural Networks: Deep learning models, particularly feedforward neural networks, can be used for diabetes detection. They are capable of learning complex patterns but may require a large amount of data and computational resources.

K-Nearest Neighbors (K-NN): K-NN is a simple and instance-based algorithm that classifies patients by comparing them to their k nearest neighbors in the feature space. It can be effective when the dataset is not too large.

Naive Bayes: Naive Bayes is a probabilistic classifier based on Bayes' theorem. It assumes that features are conditionally independent, making it simple and computationally efficient. It can perform well in some diabetes detection scenarios.

Ensemble Models: Stacking and ensemble techniques like model averaging or voting can combine the predictions of multiple models to improve overall performance and robustness in diabetes detection.

Deep Learning (Convolutional Neural Networks, Recurrent Neural Networks, etc.): For tasks involving medical imaging or time-series data, more complex deep learning architectures like CNNs and RNNs can be employed to detect diabetes or predict patient outcomes.

* **Recurrent Neural Network (RNN)**

A Recurrent Neural Network (RNN) possesses the ability to capture dynamic temporal patterns within a given time sequence input. Basic RNNs consist of interconnected nodes, resembling neurons, with each node having a one-way connection to every other node. Each node maintains a real-valued activation that varies over time. The connections (synapses) between nodes have real-valued weights that can be adjusted during each iteration. Nodes can serve as input nodes, receiving external data, output nodes, providing results, or hidden nodes, which alter the data passing through them as it moves from input to output. The key distinction from traditional feedforward neural networks is that an RNN can leverage its internal state, referred to as memory, to process sequences of inputs

* **Long Short-Term Memory (LSTM)**

Long Short-Term Memory (LSTM) units represent a specialized building block within Recurrent Neural Networks (RNNs). They possess the capability to analyze, classify, and make predictions on temporal data sequences with varying time lags. A typical LSTM network comprises memory, input, output, and forget gates. Notably, the memory in LSTM units can retain information over extended time periods. Each of these three gates functions as a neuron, computing an activation function based on a weighted sum. Furthermore, these gates regulate the flow of information within LSTM layers, earning them the name 'gates.' The term 'long short-term' emphasizes LSTM's ability to maintain memory over substantial durations, addressing the challenges of the exploding and vanishing gradient problems commonly encountered in training traditional RNNs.

* **Convolutional Neural Network (CNN)**

A Convolutional Neural Network (CNN) represents an enhanced version of the multilayer perceptron. CNNs typically consist of an input layer, an output layer, and multiple hidden layers. These hidden layers in a CNN typically include convolutional, pooling, and fully connected layers.

* **Hybrid networks (CNN-LSTM)**

In hybrid networks, the beginning segment comprises a CNN, composed solely of convolutional and max-pooling layers. The output from the 1D max-pooling layer is then forwarded as input to the subsequent deep learning architecture, such as RNN or LSTM."

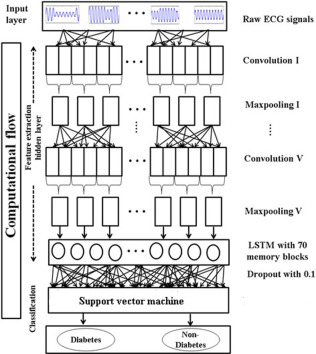
4.2 Model Training and Evaluation The selected ML models were trained and fine-tuned using cross-validation techniques. We assessed their performance using different evaluation metrics and compared the results to identify the best-performing model.

## **6. Experiments and results**

The experiments was performed using the Keras library function on a [GPU](https://www.sciencedirect.com/topics/engineering/graphics-processing-unit) enabled system unit. In this work, we use the same configuration that we had used in our early paper. In this work, we extract features in deep learning network, comprised of CNN-LSTM architecture and pass into SVM for classification. LSTM has the capability to handle long-term dependencies in a data sequence. To decide the [kernel function](https://www.sciencedirect.com/topics/computer-science/kernel-function), we run two trail of experiment for SVM with linear and [RBF](https://www.sciencedirect.com/topics/computer-science/radial-basis-function) kernel. SVM with RBF kernel performed better. The SVM model is implemented using Scikit-learn. The detailed 5-fold cross-validation accuracy is reported in Table 3 below. In almost all the network structures, SVM has performed better in 5-fold cross-validation with accuracy which is comparable to the fully connected linear with nonlinear [activation function](https://www.sciencedirect.com/topics/engineering/activation-function) for classification. Thus, we claim that the combination of SVM in penultimate layer for classification with deep learning layers for feature extraction can achieve the best performance.

Table 3. Detailed results.

| **Architecture** | **Accuracy obtained** |
| --- | --- |
| **CNN 1 with SVM** | 0.684 |
| **CNN 2 with SVM** | 0.755 |
| **CNN 3 with SVM** | 0.887 |
| **CNN 4 with SVM** | 0.913 |
| **CNN 5 with SVM** | 0.939 |
| **CNN 1-LSTM with SVM** | 0.743 |
| **CNN 2-LSTM with SVM** | 0.764 |
| **CNN 3-LSTM with SVM** | 0.853 |
| **CNN 4- LSTM with SVM** | 0.937 |
| **CNN 5-LSTM with SVM** | **0.957** |



* Results and Discussion

5.1 Model Performance:

In almost all the network structures, SVM has performed better in 5-fold cross-validation with accuracy which is comparable to the fully connected linear with nonlinear [activation function](https://www.sciencedirect.com/topics/engineering/activation-function) for classification. Thus, we claim that the combination of SVM in penultimate layer for classification with deep learning layers for feature extraction can achieve the best performance.

* Conclusion

6.1 Summary In this research paper, we explored the application of machine learning techniques for the detection of diabetes in patients. We conducted a comprehensive analysis of data preprocessing, feature selection, and model selection, showcasing the potential of ML-based approaches in diabetes diagnosis.

6.2 Future Directions Future research can focus on incorporating additional data sources, such as genetic information and wearable device data, to improve diabetes detection accuracy further. Additionally, the development of interpretable ML models can enhance clinical adoption and patient trust.

**References:**

**References**

1. Acharya U.R., Faust O., Sree S.V., Ghista D.N., Dua S., Joseph P., Ahamed V.T., Janarthanan N., Tamura T.An integrated diabetic index using heart rate variability signal features for diagnosis of diabetesComput. Biomech. Biomed. Eng., 16 (2) (2013), pp. 222-234
2. Swapna G., Rajendra Acharya U., VinithaSree S., Suri J.S.Automated detection of diabetes using higher order spectral features extracted from heart rate signalsIntell. Data Anal., 17 (2) (2013), pp. 309-326
3. Acharya U.R., Faust O., Kadri N.A., Suri J.S., Yu W.Automated identification of normal and diabetes heart rate signals using nonlinear measuresComput. Biol. Med., 43 (10) (2013), pp. 1523-1529
4. Acharya U.R., Vidya K.S., Ghista D.N., Lim W.J.E., Molinari F., Sankaranarayanan M.Computer-aided diagnosis of diabetic subjects by heart rate variability signals using discrete wavelet transform methodKnowl.-Based Syst., 81 (2015), pp. 56-64
5. Pachori R.B., Kumar M., Avinash P., Shashank K., Acharya U.R.An improved online paradigm for screening of diabetic patients using RR-interval signalsJ. Mech. Med. Biol., 16 (01) (2016), p. 1640003
6. Swapna G., Kp S., Vinayakumar R.Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signalsProcedia Comput. Sci., 132 (2018), pp. 1253-1262