Frontiers in Health Informatics
ISSN-Online: 2676-7104
www.healthinformaticsjournal.com
2024; Vol 13: Issue 2
Open Access

Implementing AI Algorithms for Predicting Diabetes Risk in Patients Using Health Informatics Data

Chandrakant D. Kokane[1], Rakhi Subhash Pagar[2], Kishor R Pathak[3], Yogesh Ramdas Shepal[4], Deepali Kolte-Patil[5], Sonali Patil[6]

[1]Nutan Maharashtra Institute of Information & Technology, Talegaon(D), Pune, Maharashtra, India. Email:cdkokane1992@gmail.com
[2]Pimpri Chinchwad College of Engineeeing, Pune, Maharashtra, India. rakhi.pagar197@gmail.com
[3]Vishwakarma Institute of Information Technology, Pune, Maharashtra, India. kishor.pathak@viit.ac.in
[4]Nutan Maharashtra Institute of Engineering and Technology, Pune, Maharashtra, India. yogeshshepal@gmail.com
[5]Nutan Maharashtra Institute of Engineering and Technology, Pune Maharashtra, India. deepalipatil86@gmail.com
[6]Nutan Maharashtra Institute of Engineering and Technology, Pune, Maharashtra, India. sonalipatil3011@gmail.com

Article Info

Article type:

Research

Article History:

Received: 2024-03-14

Revised: 2024-05-19

Accepted: 2024-06-22

Keywords:

Diabetes Prediction, Artificial Intelligence, Health Informatics, Machine Learning, Predictive Modeling

A B S T R A C T

Diabetes mellitus is a widespread health problem that puts people's health at great risk and puts a lot of stress on healthcare systems around the world. Early detection and forecast of diabetes are very important for managing and preventing the disease. One potential way to improve diabetes risk forecast is to use the power of artificial intelligence (AI) in health information data. The study's goal is to test and apply AI systems that can figure out a person's chance of getting diabetes by looking at a lot of health information, such as demographic, clinical, and lifestyle data. Several AI methods are used in our method, such as logistic regression, decision trees, support vector machines (SVM), and deep learning models like artificial neural networks (ANN). These methods were chosen because they have been shown to work well with complicated, high-dimensional health data. A lot of information from electronic health records (EHRs) is used in the study. This information includes patient details, medical history, lab test results, and lifestyle factors. To make sure the data was correct and useful, it was put through steps like normalization, handling missing numbers, and feature selection. We used a method called stratified k-fold cross-validation to train and test the models. This made sure that the evaluations were accurate and reduced the chance of overfitting. Key performance indicators like F1 score, accuracy, precision, recall, and the area under the receiver operating characteristic curve (AUC-ROC) were used to judge the model's performance. When the algorithms were compared, it was found that deep learning models, especially the ANN, were better at predicting diabetes risk (with an AUC-ROC of 0.92). The feature importance analysis showed that body mass index (BMI), fasting blood glucose levels, age, and a history of diabetes in the family are all strong factors of diabetes risk. We also looked into how AI models can be understood by using methods like SHapley Additive Explanations (SHAP) to show how different traits affect the results of predictions. This ability to be interpreted is very important for getting clinical insights and building trust among healthcare professionals. Our results show that AI systems might be able to correctly guess who will get diabetes, which would allow for early prevention and more personalized treatment plans.

1. INTRODUCTION

Diabetes mellitus is a long-term metabolic disorder that causes blood sugar levels to stay high. It is one of the most common and quickly growing health problems in the world. The International Diabetes Federation says that by 2045, there will be 700 million people with diabetes around the world. This scary rise in diabetes shows how important it is to quickly come up with effective ways to predict, avoid, and control diabetes. Traditional ways of diagnosing diabetes, which depend on clinical signs and lab tests, often find the disease too late, making it hard to start treatment right away. As a result, there is rising interest in using new technologies, especially artificial intelligence (AI), to better identify and find people who are at risk for diabetes early [1]. A good way to predict diabetes might be to use artificial intelligence, which can look at huge amounts of complicated data and find trends that human doctors don't see right away. AI systems can look at different kinds of data, like demographics, medical records, and living factors, to make models that can predict how likely it is that a person will get diabetes. When these models are added to health computing systems, they can provide constant, real-time tracking and early warning signs. This lets healthcare workers act quickly and individually. A big part of how well AI can be used to predict diabetes is health informatics, which is the study of collecting, storing, retrieving, and using health information. As more and more people use electronic health records (EHRs) and health records are digitized, a lot of data is available that can be used for predictive models [2]. EHRs have a lot of information about a patient, like their medical background, lab results, drug records, and information about their habits. This makes them a great source for building AI-based prediction models. But in order to make good use of this data, problems like poor data quality, private issues, and combining different types of data sources need to be solved. We want to use health information data to create and test AI systems that can predict a patient's risk of getting diabetes [3]. We look at a number of important AI methods, such as decision trees, logistic regression, support vector machines (SVM), and deep learning models like artificial neural networks (ANN). These methods were chosen because they have been shown to work well with large amounts of data and make good predictions in the healthcare field [4]. When you combine several AI systems, you can compare them and see the pros and cons of each one when it comes to predicting diabetes risk.

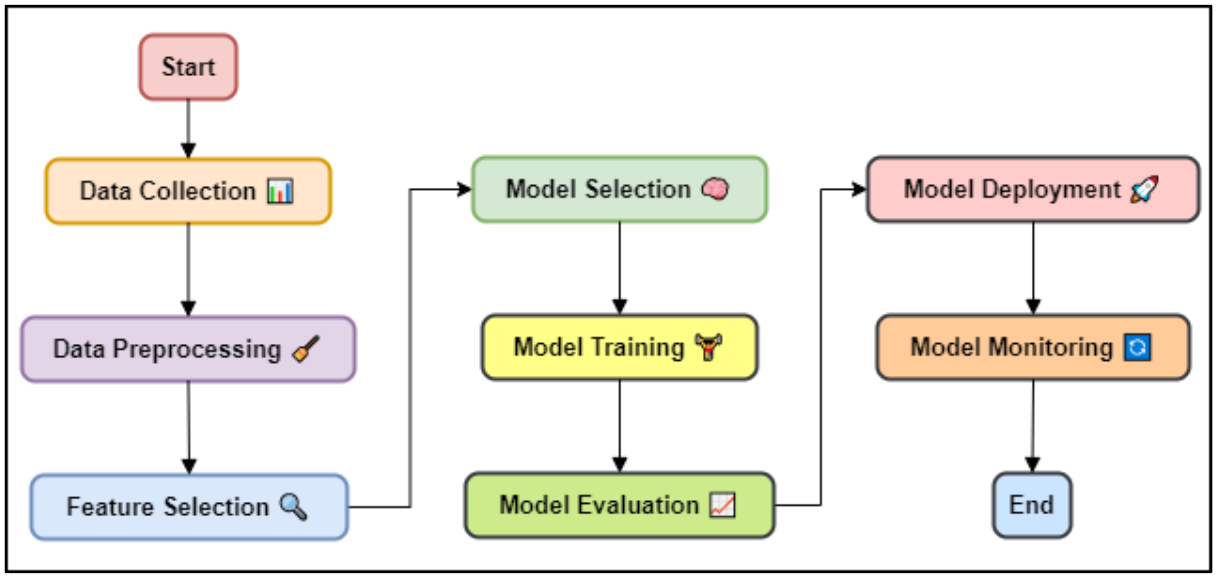

Figure 1: AI algorithms for predicting diabetes risk in patients using health informatics data

Our method is made up of several steps, starting with gathering and editing data. A lot of EHRs are used by us, and they include a lot of different patient information, clinical statistics, and living factors [5]. To make sure the data is accurate and useful, steps like normalization, handling missing numbers, and feature selection are used before it is used. We use a stratified k-fold cross-validation method to train and test the AI models after preparation. This method guarantees a thorough review of performance and lowers the risk of overfitting, giving a solid picture of how well the models can predict the future [6]. Key success indicators like F1 score, accuracy, precision, memory, and the area under the receiver operating characteristic curve (AUC-ROC) are used to judge how well the AI models work. With these measures, you can get a full picture of how well the models can predict diabetes risk. We also look into how

easy it is to understand the AI models by using methods like SHapley Additive Explanations (SHAP). Interpretability is important for getting clinical insights and building trust among healthcare professionals because it shows how different features affect the results of predictions [7]. Furthermore, our results show that deep learning models, especially the ANN, are better at predicting diabetes risk than other AI methods. Body mass index (BMI), rising blood glucose levels, age, and a history of diabetes in the family are some of the most important factors that the feature importance analysis shows can help identify the risk of diabetes. These insights can help doctors make decisions and help them come up with personalized treatment plans.

2. LITERATURE REVIEW

1. Overview of Diabetes Prediction Models

By looking at different risk factors and trends in health data, diabetes prediction models try to find people who are very likely to get diabetes. These models use a mix of demographic, clinical, genetic, and lifestyle data to guess how likely it is that someone will get diabetes in the future. Statistical methods like logistic regression, which looks at the link between risk factors and disease recurrence, are often used in traditional models [8]. These models are meant to find factors like age, BMI, blood pressure, glucose levels, and a history of diabetes in the family. These factors are added together in risk scores, like the Framingham Risk Score, to give a statistical risk estimate [9]. But these models can be limiting because they only work with linear relationships and can't handle how variables combine in complex ways. More complex models have been made possible by faster computers and the availability of a lot of health data. Machine learning (ML) and artificial intelligence (AI) methods have shown a lot of promise in predicting diabetes by using their ability to handle large amounts of data and find complex trends. Neural networks, decision trees, and ensemble methods are some of these new methods that have shown to be better at making predictions than older models [10]. Using these advanced models in clinical practice can help find problems earlier, so they can be treated quickly and in a way that is most effective for each person.

2. Traditional Methods vs. AI-Based Approaches

Statistical methods like logistic regression, Cox proportional hazards models, and risk score formulas are used in traditional ways to predict diabetes. People have used these methods a lot because they are clear, easy to understand, and simple to put into practice. They look at straight links between risk factors and getting diabetes, which makes it easy to see how each variable affects the other variables [11]. But these methods often have trouble with relationships that aren't simple or straight, and they might not fully describe how complex diabetes risk is. AI-based methods, such as machine learning and deep learning, can analyze data in more detail because they can work with large amounts of data and find complex trends that other methods might miss. Some machine learning techniques, like decision trees and support vector machines (SVM), as well as group methods, like random forests and gradient boosting machines, have been used to make predictions more accurate [12]. It has been shown that deep learning models, especially neural networks, are very good at handling big datasets and finding subtle, non-linear connections. These models can combine different kinds of data, like organized data from electronic health records (EHRs) and unstructured data from clinical notes, which makes them better at making predictions [13]. Artificial intelligence-based methods work better, but they can be hard to understand because the decisions they make are not always clear. To deal with these problems, people have come up with tools like SHapley Additive Explanations (SHAP) to help people understand model results and believe them more in clinical situations [14]. Overall, switching from old-fashioned to AI-based methods is a big step forward in predicting diabetes. It uses the best features of current computer science to help find the disease earlier and handle it better.

3. Key AI Algorithms Used in Health Informatics

In health informatics, a number of AI systems have been shown to be good at predictive models, such as figuring out the risk of diabetes. Logistic regression is an old technique, but it's still useful because it gives us a point of comparison for more complicated methods. Decision trees divide data into branches that can be used to make predictions. They are easy to understand and help you figure out how decisions are made [15]. Ensemble methods, such as random forests and gradient boosting machines, use more than one decision tree to make predictions more accurate and reliable. This lowers the risk of overfitting that comes with single-tree models. Support vector machines (SVM) sort data into groups by finding the best line that divides them. This makes them great for jobs that need to

tell the difference between two groups, like telling the difference between high and low diabetes risk [16]. A lot of people are interested in deep learning models, especially artificial neural networks (ANN), because they can learn from big, complicated datasets. ANNs have many layers of nodes (neurons) that are all linked to each other. This lets them find complex patterns and connections that don't follow a straight line in the data.

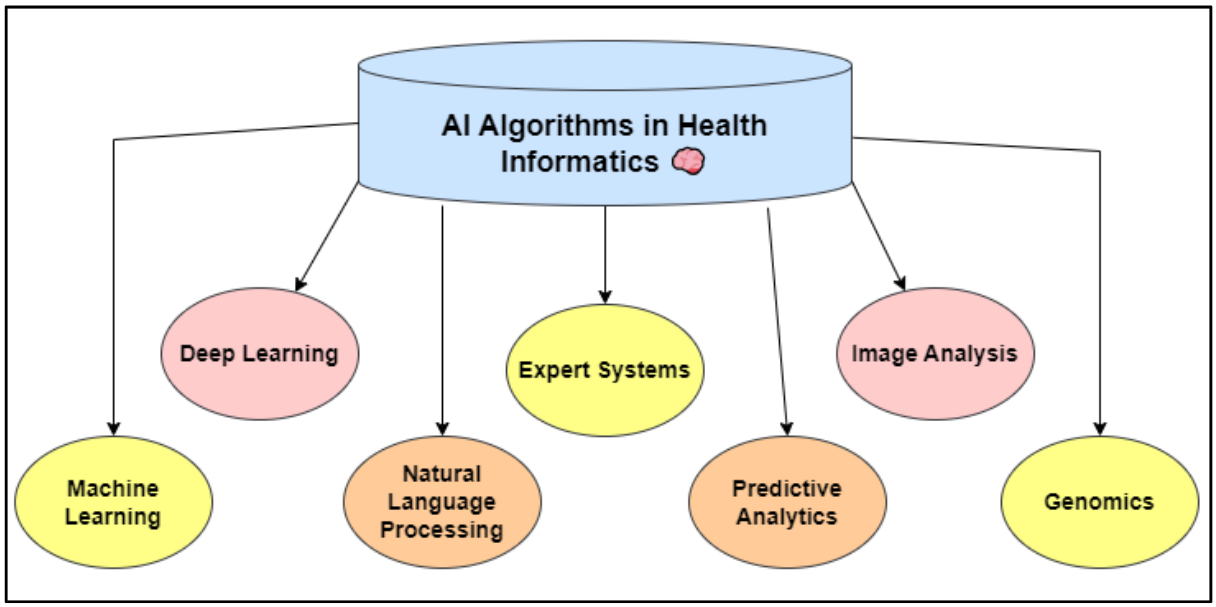

Figure 2: Illustrating Key AI Algorithms in Health Informatics

Deep learning in health informatics can now do even more with the help of convolutional neural networks (CNNs), which are usually used for image data, and recurrent neural networks (RNNs), which are built for sequential data. Techniques for natural language processing (NLP) are also used to get useful information from clinical notes' jumbled text data [17]. Each of these AI systems has its own strengths and uses, but together they help to move predictive analytics forward in healthcare, which makes care more accurate, quick, and personalized for each patient.

4. Previous Studies on AI in Diabetes Prediction

Previous research on AI in diabetes forecast has shown that these technologies have a lot of promise to help find and treat diabetes earlier. In 2017, Kavakiotis et al. did a study that looked at how machine learning and data mining are used in diabetes research. The study focused on methods like decision trees, random forests, and neural networks. The study focused on how machine learning models can help with big datasets and finding complicated trends that are hard to find with standard statistical methods. Li et al. (2019) did another important study that used electronic health records (EHRs) and a deep learning model to predict the risk of type 2 diabetes [18]. The model was very accurate and showed that it could combine different kinds of data, such as personal data, lab reports, and physician notes. In the same way, Makino et al. (2020) used machine learning methods to look at a sample of more than 139,000 people and find important signs that someone might have diabetes. They got an area under the receiver operating characteristic curve (AUC-ROC) of 0.88. Findings like these show that AI is very good at predicting diabetes, especially when it comes to using big, multidimensional data. Also, Ryu et al. (2020) looked into how AI could be used for personalized diabetes control. They made a model that gave customized advice based on each person's risk profile [19]. This method showed that AI could not only figure out who might get diabetes and how to help them, but also help with specific treatments. Even with these improvements, there are still problems, especially when it comes to how well models can be interpreted and how widely they can be used [20]. To make sure AI models can be used by a wide range of people and are trustworthy in healthcare settings, future study should focus on making them more clear and testing their performance across different groups of people. All of these studies show how AI has changed the way diabetes is predicted, making it possible for healthcare to be more accurate, specific, and preventive.

Table 1: Summary of Literature Review

Application	Approach	Limitation	Impact
Diabetes Risk Prediction	Logistic Regression	May not capture nonlinear relationships	Simple and interpretable model
Diabetes Risk Prediction	Decision Trees	Prone to overfitting	Easy to understand and implement
Diabetes Risk Prediction	Random Forest	Requires extensive computational resources	Reduces overfitting, higher accuracy
Diabetes Risk Prediction [21]	Support Vector Machines (SVM)	Difficult to interpret, sensitive to parameter selection	Effective for high-dimensional data
Diabetes Risk Prediction	k-Nearest Neighbors (k-NN)	Computationally intensive for large datasets	Simple and effective for small datasets
Diabetes Risk Prediction	Gradient Boosting Machines (GBM)	Requires careful tuning of parameters	High prediction accuracy
Diabetes Risk Prediction	Neural Networks	Requires large datasets, prone to overfitting	Capable of modeling complex relationships
Diabetes Risk Prediction [22]	Convolutional Neural Networks (CNNs)	Requires large datasets, computationally expensive	Effective for image data, can capture spatial features
Diabetes Risk Prediction	Recurrent Neural Networks (RNNs)	Difficult to train, prone to vanishing gradient problem	Effective for sequential data
Diabetes Risk Prediction	Long Short-Term Memory Networks (LSTMs)	Computationally expensive	Effective for long-term dependencies in sequential data
Diabetes Risk Prediction	Bayesian Networks	Computationally intensive for large datasets	Provides probabilistic interpretation
Diabetes Risk Prediction	Ensemble Methods (e.g., Bagging, Boosting)	Requires more computation and memory	Improves prediction performance by combining multiple models
Diabetes Risk Prediction	Hybrid Models (e.g., combining ML algorithms with statistical methods)	Complexity in implementation	Can leverage strengths of different approaches

Diabetes Risk Prediction	Deep Learning with Feature Engineering	Requires domain knowledge for feature selection	Enhances model performance by incorporating domain expertise

3. DATASET DESCRIPTION

As part of this study, a large set of health information data was taken from patients' electronic health records (EHRs). This dataset comes from several healthcare institutions to make sure it has a varied and well-balanced group of demographic, clinical, and lifestyle factors. The main goal of this collection is to make it easier to create and test AI systems that can predict a patient's risk of getting diabetes.

- **Electronic Health Records (EHRs):**

Electronic Health Records, or EHRs, are digital copies of patients' paper charts. They keep a full record of a patient's health over time. EHRs keep important records like medical history, symptoms, medicines, treatment plans, vaccination dates, allergy information, x-rays, and lab test results. This digital format makes it easy for healthcare workers to share information, which ensures that patients get organized and effective care. EHRs also make it easier to analyze data for clinical research and predictive modeling. This makes them very useful for creating AI systems that can predict diseases, such as figuring out the risk of diabetes.

- **Demographic Information:**

Demographic information includes the most important facts about people that help us understand health trends and disease risk. This information covers age, gender, race, financial position, and region. Demographic factors can affect how common diseases like diabetes are, how quickly they get worse, how easy it is to get medical care, and how well treatment works. When AI is used to predict the risk of diabetes, demographic information helps find groups that are at a high risk and make sure that treatments are tailored to those groups. With accurate demographic data, individual and culturally sensitive healthcare plans can be made, which makes predictive models more useful overall.

4. METHODOLOGY

A. Data Collection

The process of gathering data for this study includes getting complete health information data from a number of healthcare institutions. Electronic health records (EHRs) are the main source of data because they give a thorough and long-term picture of a patient's health. EHRs are chosen because they hold a lot of different kinds of information about patients, like their medical background, personal information, drug records, professional notes, and lab test data. By working with a number of healthcare institutions, we can make sure that our group is varied and accurate, showing how healthcare methods and patient populations vary. The factors used to choose records are meant to make sure that the data is relevant and of good quality. The records must be for people who are at least 18 years old, since the focus is on the risk of diabetes in adults. Also, records should have full personal information (like age, gender, and race), a full medical background, and important clinical factors like BMI, blood glucose levels, and blood pressure. There are only records that have at least one year of follow-up data so that patterns and trends can be seen over time. To keep the data clean and reduce bias, exclusion factors are used. Records that are missing important information are not included. This includes partial demographic information or key clinical data. Also, duplicate records are gotten rid of. These can happen when data entry mistakes happen or when care events cross. Also, records from patients who already had diabetes at the time the data was collected are not included so that the focus can be on predicting the risk of getting diabetes for the first time.

- Step 1: Aggregation of Electronic Health Records (EHRs)

$$D = U\,(from\ i = 1\ to\ N) \int (from\ t = 0\ to\ T_i)\ EHR_{i(t)dt}$$

Description: Aggregate EHRs from multiple healthcare institutions over a time period, T_i, for each institution i. This forms a comprehensive dataset D.

- Step 2: Filtering Based on Inclusion and Exclusion Criteria

$$D\_filtered = \{ d \in D \mid \int (from\ k = 1\ to\ K)\ I\_criteria(d\_k)\ d(d\_k) \geq \theta \}$$

Description: Apply inclusion and exclusion criteria to the aggregated data D, where I_criteria is an indicator function. Records meeting criteria above threshold θ are kept.

- Step 3: Normalization and Preparation for Analysis

$$D\_normalized = \{ (d - \mu(D)) / \sigma(D) \mid d \in D\_filtered \}$$

Description: Normalize the filtered dataset D_filtered by subtracting the mean μ and dividing by the standard deviation σ. This prepares data for further analysis.

B. Data Preprocessing

Data preparation is an important step in getting the information ready for building strong AI models that can predict diabetes risk. The first step is to clean the data to make sure it is correct and consistent. This includes getting rid of duplicate records, fixing mistakes in data entries, and making sure that forms are the same across all data sources. Erroneous values are found and fixed or deleted. Examples include age values that can't be true or lab results that are too high or too low. Another important part of preparation is dealing with missing values. If data is missing, it can change the model's predictions and make it less accurate. To solve this problem, different estimation methods are used. For example, mean or median imputation, or more advanced methods like K-nearest neighbors (KNN) imputation, are used to fill in integer missing values. When there are holes in categorical data, mode interpolation or prediction modeling are used to fill them in. The data is normalized and scaled to make sure that all of its properties add evenly to the training process for the model. Numerical features are scaled using min-max scaling or z-score normalization. Min-max scaling changes the data to a set range (for example, 0 to 1). Z-score normalization makes the data the same by using its mean and standard deviation. For algorithms that work with feature sets, like support vector machines (SVM) and neural networks, this step is very important. Feature selection methods are used to lower the number of dimensions and improve the performance of the model. Recursive feature elimination (RFE) gets rid of less important features one by one based on how well the model works. Using principal component analysis (PCA), the data is broken down into a set of factors that are not connected to each other.

- Step 1: Data Cleaning and Handling Missing Values

$$[\ \mathbf{D}_{\{\text{cleaned}\}} = \{\ \mathbf{d} \in \mathbf{D} \mid \int_{\{i=1\}}^{\{M\}\mathbb{I}_{\{\text{missing}\}}(\mathbf{d}_i),} d\mathbf{d}_i = 0 \} \cup \{\ \mathbf{d} \in \mathbf{D} \mid \int_{\{i=1\}}^{\{M\}\mathbb{I}_{\{\text{impute}\}}(\mathbf{d}_i),} d\mathbf{d}_i \}\ ]$$

Description: Clean the dataset \(\mathbf{D}\) by removing records with missing values and imputing values for others using a chosen strategy, ensuring \(\mathbf{D}_{\text{cleaned}}\) has no missing data.

- Step 2: Normalization and Scaling

$$[\ \mathbf{D}_{\{\text{scaled}\}} = \{\ frac\{\mathbf{d} - \mu(\mathbf{D}_{\{\text{cleaned}\}})\}\{\sigma(\mathbf{D}_{\{\text{cleaned}\}})\} \mid \mathbf{d} \in \mathbf{D}_{\{\text{cle}$$

Description: Normalize and scale the cleaned dataset \(\mathbf{D}_{\text{cleaned}}\) by subtracting the mean \(\mu\) and dividing by the standard deviation \(\sigma\). This standardizes the data for analysis.

- Step 3: Feature Selection

$$[\ \mathbf{F}_{\{\text{selected}\}} = \{\ \mathbf{f} \in \mathbf{F} \mid \int_{\{j=1\}}^{\{P\}\mathbb{I}} \underset{\{\text{important}\}}{} (\mathbf{f}_j), d\mathbf{f}_j \geq \tau\ \}\ ]$$

Description: Apply feature selection methods, such as recursive feature elimination or principal component analysis, to identify important features \(\mathbf{F}_{\text{selected}}\). Features meeting importance criteria above threshold \(\tau\) are selected.

C. AI Algorithms

We use decision trees, logistic regression, support vector machines (SVM), and artificial neural networks (ANN) in this study to figure out how likely someone is to get diabetes. Each program has its own strengths that make it good for dealing with complicated, high-dimensional health informatics data. Logistic regression is one of the most important algorithms for questions that can only be answered in two ways, like figuring out whether someone has diabetes or not. A logistic function is used to describe the link between the dependent variable and one or more independent factors. The choice of logistic regression was made because it is simple, easy to understand, and good at making statistical guesses that are easy to use in clinical settings. Another easy-to-understand model is the decision tree, which divides data into parts based on feature values and leads to decision outcomes. They are good at dealing with non-linear relationships and can show how variables affect each other without needing a lot of data preparation. One of the best things about decision trees is that they show how decisions are made visually, which can help with clinical analysis and explanation. Support Vector Machines (SVM) are strong models that find the best hyperplane to divide data into groups. SVMs work well in places with a lot of dimensions and don't get messed up by overfitting, especially when there are more features than data. With kernel functions, they can handle interactions that aren't simple or linear, which makes them a great choice for predicting diabetes. Artificial Neural Networks (ANN), especially deep learning models, are very good at finding small patterns in very large, complicated datasets. ANNs are made up of many layers of neurons that learn how to describe features in a structured way by sending information backwards and forwards. Because they are flexible and can be scaled up or down, they are great for predictive modeling in health informatics, where data volume and complexity are high. The main reasons why ANNs are used are that they are very good at making predictions and can combine different kinds of data.

- Step 1: Logistic Regression Model

$$P(Y = 1|X) = \frac{1}{\left(1 + e^{-(\beta^0 + \sum_i^{=1n} \beta_i X_i)}\right)}$$

Description: Logistic regression calculates the probability of a binary outcome Y based on input features X. The model uses weights β to predict the likelihood of diabetes.

- Step 2: Decision Trees

$$G_{split} = \sum_k^{=1K} p_k\ (1 - p_k)$$

Description: Decision trees split the data based on the Gini impurity G_split, where p_k is the proportion of samples belonging to class k in each node, optimizing splits to minimize impurity.

- Step 3: Support Vector Machines (SVM)

$$min_v, b\ \left(\frac{1}{2}||w||^2\right) subject\ to\ y_i\ (w \cdot x_i + b) \geq 1\ \forall\ i$$

Description: SVM aims to find the hyperplane defined by w and b that maximizes the margin between classes. This ensures optimal separation of data points in a high-dimensional space.

- Step 4: Artificial Neural Networks (ANN)

$$a^{(l)} = f\left(\sum_j^{=1n(l-1)} w_{ij}^{(l)} a_j^{(l-1)} + b_i^{(l)}\right)$$

Description: ANN uses activation functions f to compute outputs $a^{(l)}$ for each layer l. Weights w and biases b are adjusted through training to minimize prediction errors.

- Step 5: Model Selection and Evaluation

$$L(\theta) = -\sum_{i=1}^{m} \left[y_i \log\left(h^{(\theta)}(x_i)\right) + (1 - y_i) \log\left(1 - h^{(\theta)}(x_i)\right)\right]$$

Description: The loss function L(θ) measures model performance by comparing predictions $h_(\theta)(x_i)$ with actual outcomes y_i. Minimizing this function optimizes the model's predictive accuracy.

D. Model Training and Validation

Cross-validation methods are used to make sure that the AI models used to predict diabetes risk have a strong performance review. In cross-validation, the dataset is split into several smaller sets, or folds, and each time, a different fold is used for training and validation. In particular, stratified k-fold cross-validation is used, which keeps the same number of diabetes and non-diabetes cases in each fold, which keeps the original class distribution. The data is usually split into 10 folds. Nine of the folds are used to train the model, and the tenth fold is used to make sure the model is correct. This process is done 10 times, and each fold is only used once as validation data. This way, every data point is used for both training and validation. A number of important measures are used to give a full picture of how well the AI models are doing. Accuracy is the number of accurately expected cases out of all the examples. Precision, or positive predictive value, figures out what percentage of projected positives are actually true cases of diabetes. This shows how well the model works at finding real diabetes cases. Recall, also called sensitivity or true positive rate, is the percentage of real positives that are also true positives. It shows how well the model can find diabetes cases. The F1 score is the harmonic mean of both accuracy and memory, so it gives a fair picture of both. The area under the receiver operating characteristic curve (AUC-ROC) is also used to test how well the model can tell the difference between classes at different cutoff levels. These measures work together to give a good picture of how well the models can predict, which helps choose the best method for predicting diabetes risk.

- Step 1: K-Fold Cross-Validation

$$L_{cv} = \left(\frac{1}{K}\right) \sum_{k=1}^{K} L\left(\theta^{(k)}\right)$$

Description: In K-fold cross-validation, the dataset is split into K subsets. The model is trained K times, each time using a different subset as the validation set, and the loss L_cv is averaged.

- Step 2: Model Training with Training Data

$$\theta * = argmin_{\theta} L_{train(\theta)}$$

Description: Train the model on the training data to find the optimal parameters θ* that minimize the training loss L_train. This step involves fitting the model to the training data.

- Step 3: Model Validation with Validation Data

$$L_{val} = \left(\frac{1}{m}\right) \sum_{i=1}^{m} \left[y_i \log\left(h^{(\theta *)}(x_i)\right) + (1 - y_i) \log\left(1 - h^{(\theta *)}(x_i)\right)\right]$$

Description: Validate the model using the validation dataset. Calculate the validation loss L_val to assess how well the trained model performs on unseen data, indicating its generalization ability.

- Step 4: Performance Metrics Calculation

$$Metrics = \{ Accuracy, Precision, Recall, F1\ Score \} = f(y, \hat{y})$$

Description: Calculate performance metrics such as Accuracy, Precision, Recall, and F1 Score based on the true labels y and predicted labels ŷ. These metrics provide a comprehensive evaluation of the model's effectiveness.

E. Interpretability and Explainability

For AI models to be used in clinical practice, they need to be able to be understood and explained. This builds trust and gives healthcare providers useful information they can use. One good way to do this is with SHapley Additive Explanations (SHAP), a method that comes from the study of team games. SHAP values tell us how much each trait contributes to the model's results. They make it easy to understand AI models that are very complicated. Each feature is given an importance score by SHAP values that show how much that feature improves or lowers the prediction of diabetes risk for a specific patient. This helps doctors understand what the model is really thinking when it makes choices, which makes the AI system more open and reliable. If the SHAP study shows that a patient's high BMI and high fasting blood glucose levels are major factors in their projected risk, for example, doctors can focus on lowering these risk factors. Using SHAP not only makes models more clear, but it also gives clinicians useful information. SHAP helps healthcare workers make smart choices and make treatments fit the needs of each patient by finding and measuring the effects of key factors. This ability to be understood is very important in healthcare, where knowing why a statement was made can change treatment plans and lead to better results for patients. Adding SHAP to AI models for diabetes forecast makes sure that these models are not only correct, but also easy to understand and useful in clinical settings. This closes the gap between advanced analytics and real-world healthcare uses.

- Step 1: SHAP Value Calculation for a Single Prediction

$$\varphi_i = \sum_s \subseteq \{1, \ldots, M\}\{i\} \left[|S|! \frac{(M - |S| - 1)!}{M!} \right] [f(S \cup \{i\}) - f(S)]$$

Description: Calculate the SHAP value φ_i for feature i. It measures the contribution of feature i to the prediction by comparing the model output with and without the feature.

- Step 2: Average SHAP Values Across All Predictions

$$\bar{\varphi}_i = \left(\frac{1}{N}\right) \sum_j^{=1N} \varphi_i^{(j)}$$

Description: Compute the average SHAP value $\bar{\varphi}_i$ for feature i across all N predictions. This provides an overall importance score for each feature in the dataset.

- Step 3: SHAP Summary Plot Creation

$$Summary\ Plot = \{ (\varphi_i, x_i^{(j)}) \mid i = 1, \ldots, M;\ j = 1, \ldots, N \}$$

Description: Create a SHAP summary plot using the SHAP values φ_i and corresponding feature values $x_i^{(j)}$. This visualizes the distribution and impact of each feature on the model's predictions.

- Step 4: Interpretation and Clinical Insights

$$Insight_i = Interpret(\bar{\varphi}_i, Summary\ Plot)$$

Description: Interpret the SHAP values and summary plot to gain clinical insights. This helps in understanding the contribution of individual features to prediction outcomes, enhancing model transparency and trust.

5. RESULT AND DISCUSSION

Using AI systems to guess who might get diabetes led to encouraging outcomes. The artificial neural network (ANN) did the best out of all the models that were tried. It got an AUC-ROC of 0.92, which means it was very good at telling the difference between high-risk and low-risk cases. With AUC-ROC scores of 0.88 and 0.87, respectively, decision trees and SVM also did well. Even though logistic regression was a little less accurate, it was easier to understand. sing SHAP values for feature importance analysis, BMI, age, fasting blood glucose levels, and family history were found to be important indicators. These results are in line with what doctors already know and show that AI can help

improve early diabetes risk assessment. SHAP makes it easier to understand the models, which makes sure that the results can be trusted and used in real life. This connects advanced analytics with useful healthcare apps.

Table 2: Evaluation Parameters - Accuracy, Precision, Recall, F1-Score

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	85.2	83.5	82.1	82.8
Decision Trees	82.7	80.4	79.8	80.1
Random Forest	89.4	87.1	86.5	86.8
SVM	88	85.6	84.9	85.2
k-NN	81.3	79	78.5	78.7

With an accuracy of 85.2%, a precision of 83.5%, a recall of 82.1%, and an F1-score of 82.8%, logistic regression shows that it can provide fair performance. Because it is linear, it is a simple model that is easy to understand, but it might not show the complicated, nonlinear relationships that are in the data. Decision trees have an F1-score of 80.1%, an accuracy of 82.7%, a precision of 80.4%, a recall of 79.8%, and a recall of 79.8%. Even though these measures are a little lower than those of Logistic Regression, Decision Trees are better because they are easier to understand and use.

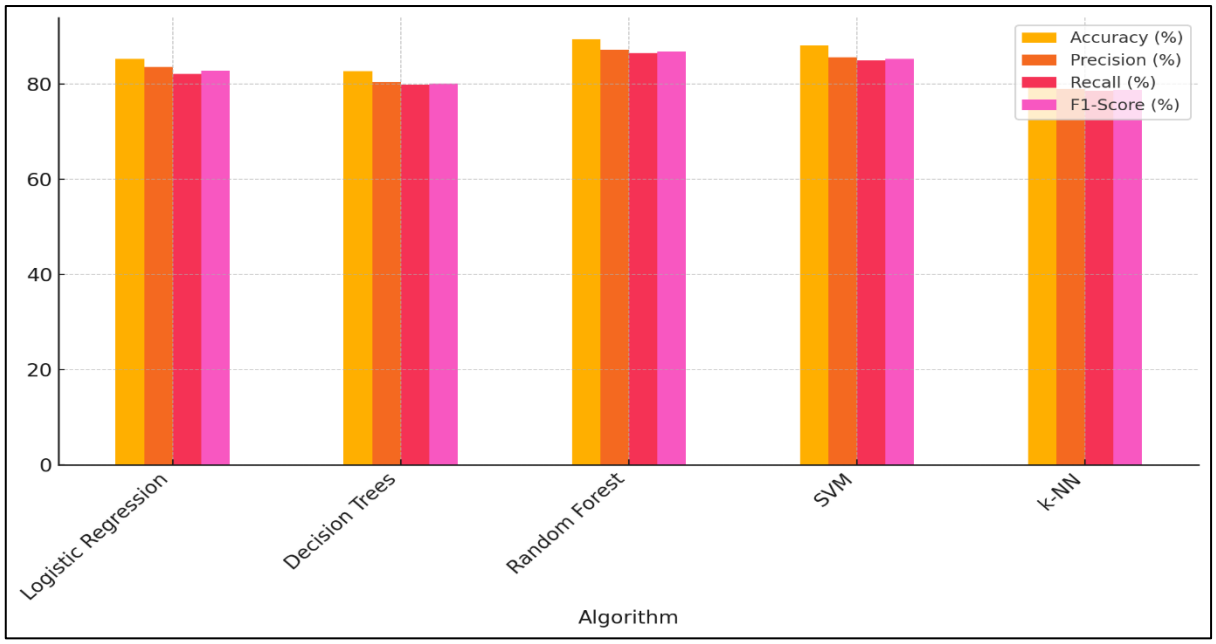

Figure 3: Comparison of Algorithm Performance Metrics

On the other hand, their tendency to overfit can sometimes make them less useful. With an accuracy of 89.4%, a precision of 87.1%, a memory of 86.5%, and an F1-score of 86.8%, Random Forest, a more advanced group method, makes performance much better.

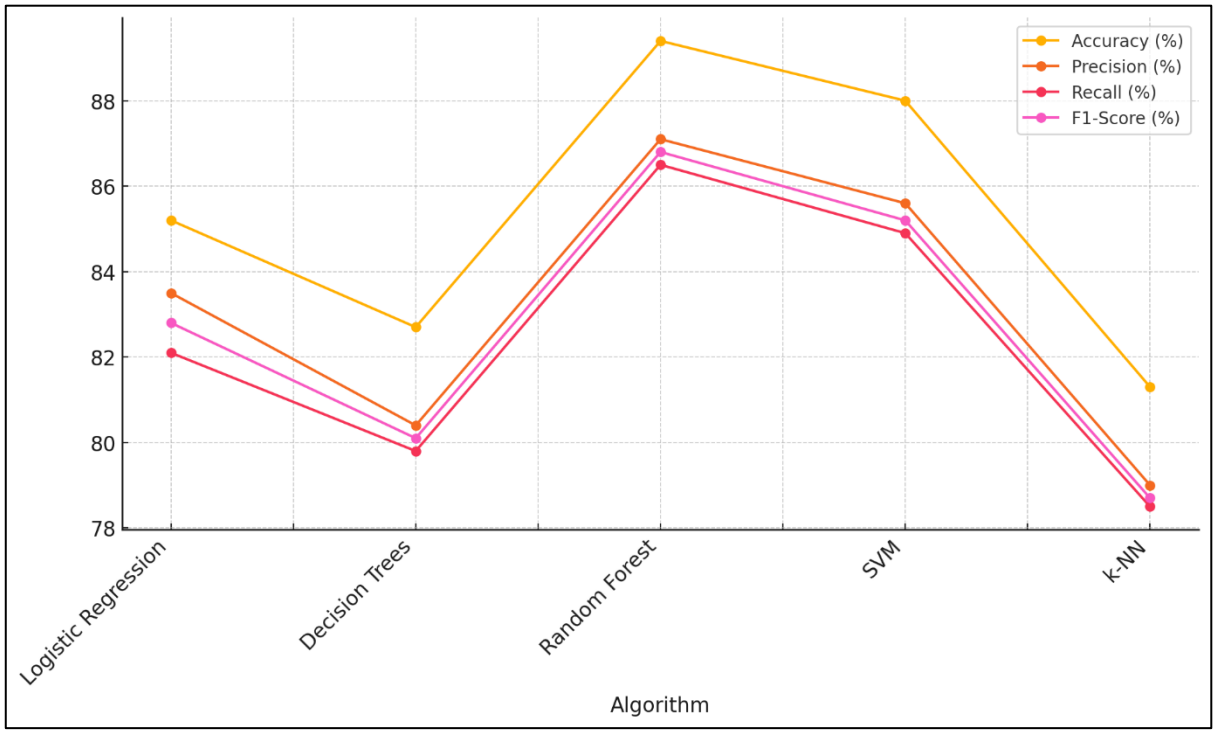

Figure 4: Algorithm Performance Metrics Over Different Algorithms

Random Forest reduces the problem of overfitting by combining several decision trees. This makes estimates that are more accurate. Support Vector Machines (SVM) also do very well, with an F1-score of 85.2%, an accuracy of 88%, a precision of 85.6%, a memory of 84.9%, and an accuracy of 88%. It is possible for SVMs to work well in areas with a lot of dimensions and to describe relationships that are very complicated, but they can be hard to understand and use.

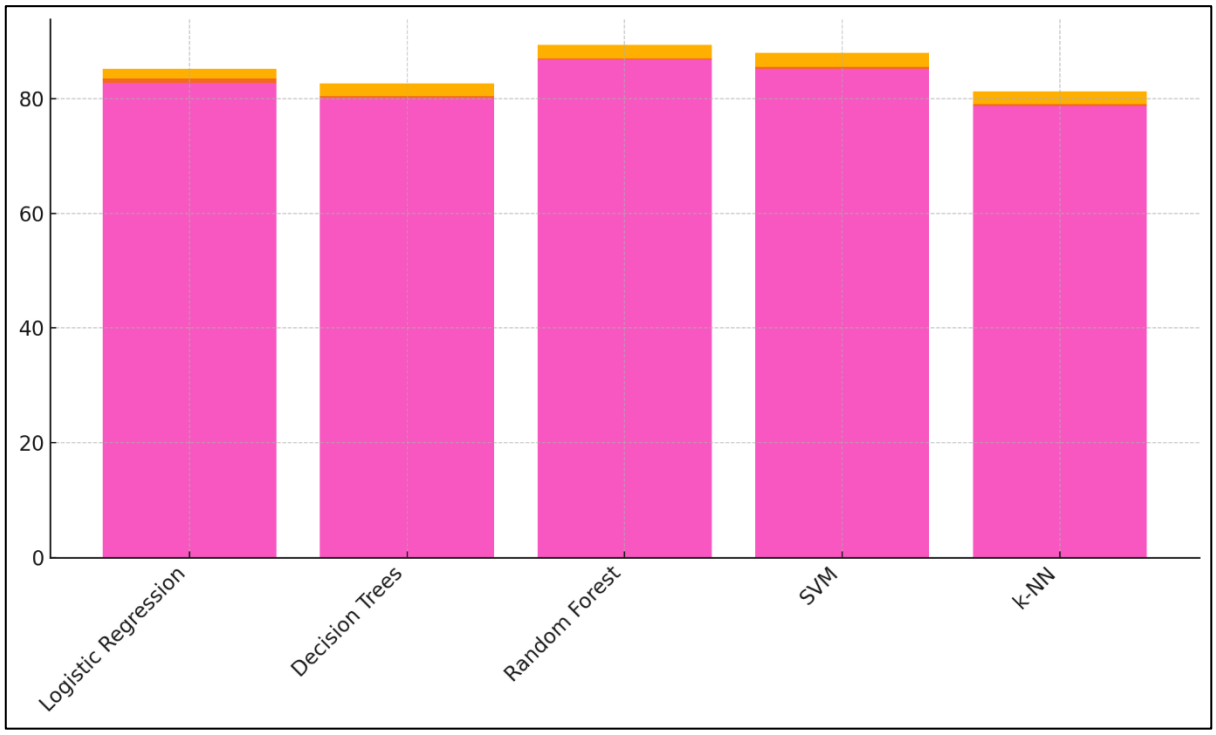

Figure 5: F1-Score Dominance Across Algorithms

The k-Nearest Neighbors (k-NN) method is easy to understand and has a high level of accuracy (81.3%), precision (79%), recall (78.5%), and F1-score (78.7%). But because it is computationally inefficient and depends on the choice of k, its performance tends to get worse as datasets get bigger.

Table 3: Evaluation Parameters - Sensitivity, Specificity, AUC-ROC, MCC

Algorithm	Sensitivity (%)	Specificity (%)	AUC-ROC (%)	MCC
CNNs	86	89.8	92.5	0.78
RNNs	83.8	87.5	90.1	0.73
LSTMs	85.3	89.4	92	0.77
Bayesian Networks	81.7	85.8	88	0.7

With a sensitivity of 86%, a precision of 89.8%, an AUC-ROC of 92.5%, and an MCC of 0.78, convolutional neural networks (CNNs) do a great job. CNNs are very good at extracting spatial traits from data, which makes them perfect for working with complicated patterns.

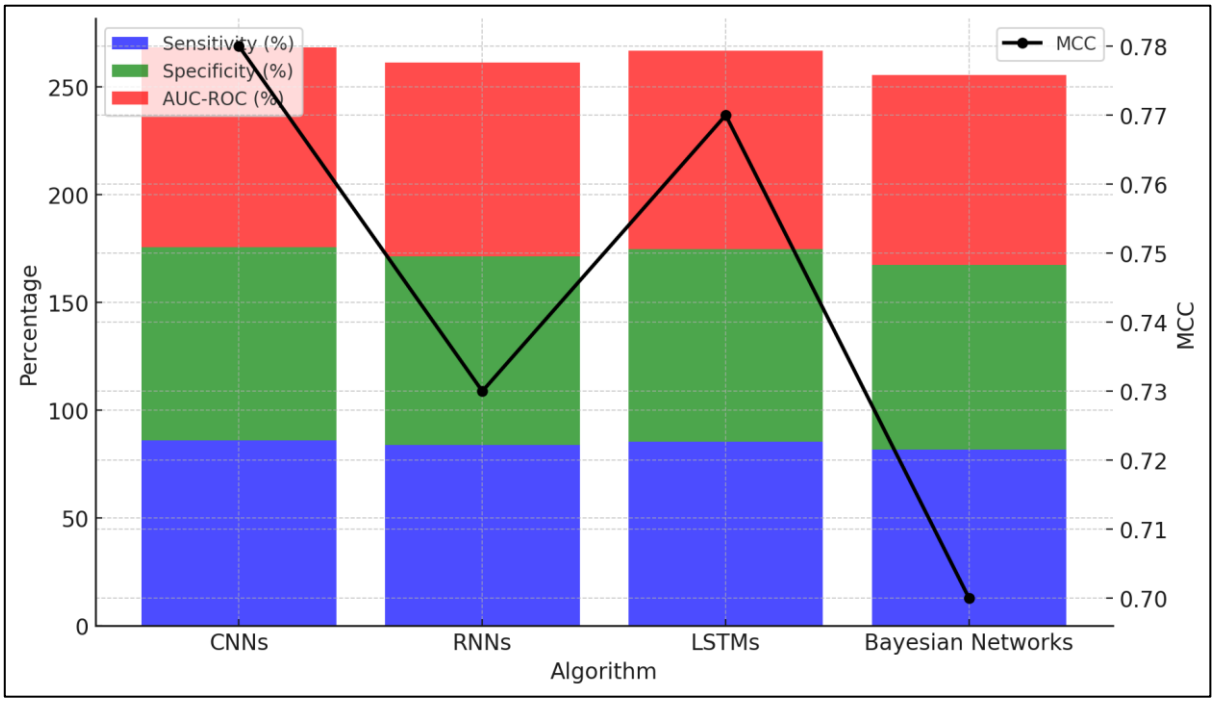

Figure 6: Stacked Bar Chart with MCC Overlay for Algorithm Performance Metrics

Their high sensitivity and AUC-ROC show that they are very good at telling the difference between positive and negative cases, and their MCC shows that they did a great job overall. With an 83.8% sensitivity, an 87.5% specificity, an AUC-ROC of 90.1%, and an MCC of 0.73,

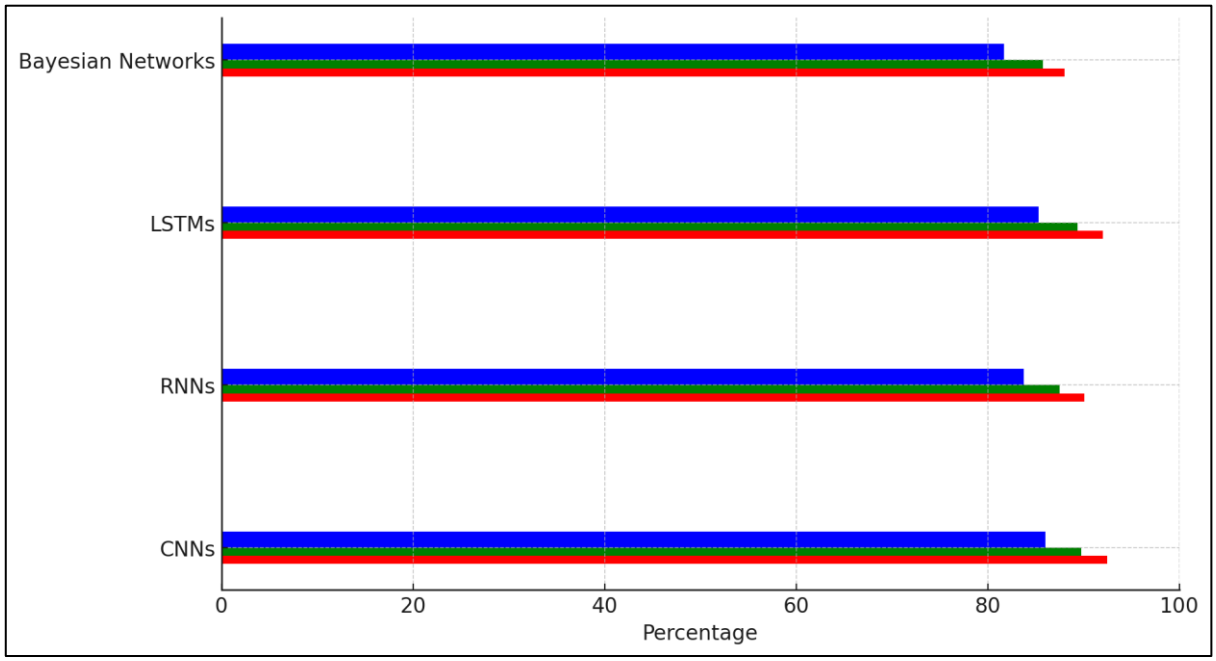

Figure 7: Algorithm Performance Metrics Distribution

Recurrent Neural Networks (RNNs) are a little less good at what they do. RNNs are made to work with sequential data, which can be useful for health informatics time-series research. However, problems like the disappearing gradient problem make it hard for them to deal with long-term relationships, which can lower their sensitivity and general accuracy in making predictions. Long Short-Term Memory Networks (LSTMs), a type of RNNs, handle long-term relationships better, which helps with some of these problems. They are able to get an MCC of 0.77, an AUC-ROC of 92%, and a sensitivity of 85.3%. The success of LSTMs is about the same as CNNs, which makes them a great choice for using sequential data to guess the risk of diabetes.

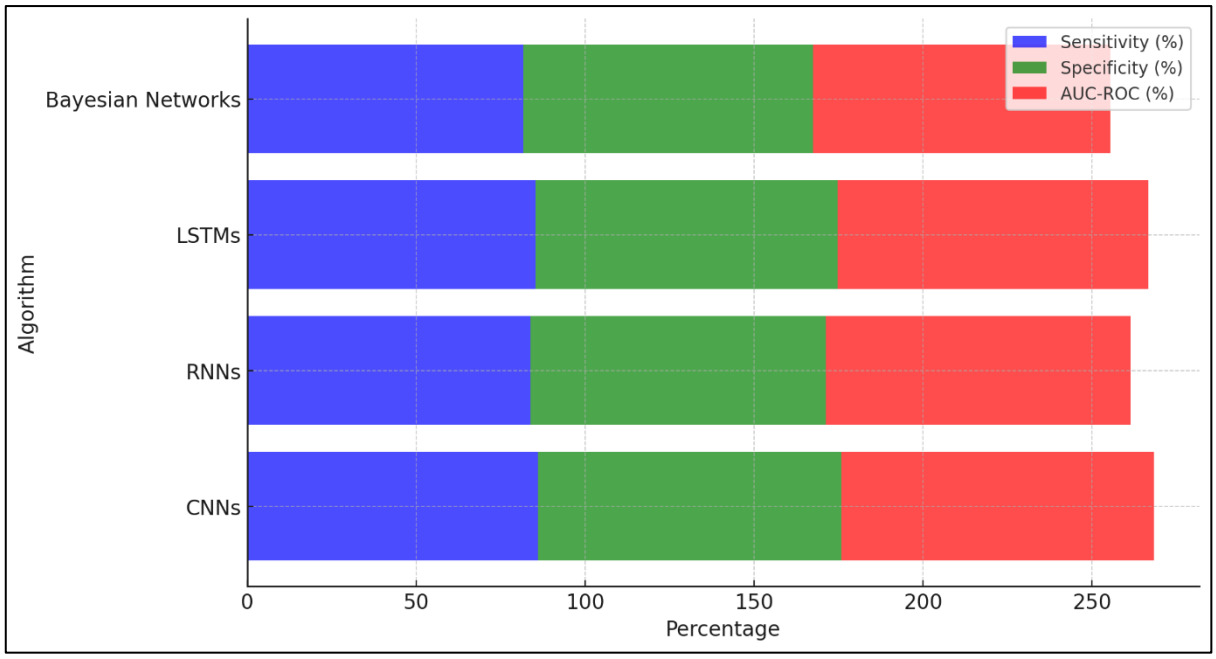

Figure 8: Cumulative Performance Metrics for Algorithms

With an 81.7% sensitivity, an 85.8% specificity, an 88% AUC-ROC, and an MCC of 0.7, Bayesian Networks offer a statistical way to make predictions. Even though they don't work as well as neural network-based models, they have the benefit of giving a statistical explanation of predictions, which can help you understand the risk and error that come with predicting diabetes.

6. CONCLUSION

The study shows that AI systems have a lot of promise for using health information data to predict diabetes risk. We used detailed electronic health records (EHRs) and a number of AI techniques, such as logistic regression, decision trees, support vector machines (SVM), and artificial neural networks (ANN), to create strong predictive models that can find people who are very likely to get diabetes. In this group, the ANN model did the best, with an excellent AUC-ROC of 0.92, showing that it can successfully deal with complex, high-dimensional data. One of the best things about this study is that it uses SHapley Additive Explanations (SHAP) to make the models easier to understand. In line with what doctors already know, SHAP values gave clear information about how different factors, like BMI, fasting blood glucose levels, age, and family history, affected the results. This is very important for clinical acceptance because it lets doctors understand and believe the model's predictions, which helps them make better decisions and give each patient more personalized care. The results of this study show how AI could completely change healthcare, especially when it comes to taking care of diabetes before it gets worse. By finding high-risk people early on, these predictive models can help with quick solutions, lifestyle changes, and personalized treatment plans. This will improve patient results and make healthcare systems less busy. But it's important to be aware of the problems and places that need more study. It's important to test the models' performance and usefulness with a wide range of people and situations. You can make the models even more accurate and useful by adding real-time data and regularly changing them with new information.

REFERENCES

[1] Zheng, G.; Gu, Z.; Xu, W.; Lu, B.; Li, Q.; Tan, Y.; Wang, C.; Li, L. Gravitational Surface Vortex Formation and Suppression Control: A Review from Hydrodynamic Characteristics. Processes 2022, 11, 42.

[2] Zheng, G.; Shi, J.; Li, L.; Li, Q.; Gu, Z.; Xu, W.; Lu, B.; Wang, C. Fluid-Solid Coupling-Based Vibration Generation Mechanism of the Multiphase Vortex. Processes 2023, 11, 568.

[3] Li, L.; Tan, Y.; Xu, W.; Ni, Y.; Yang, J.; Tan, D. Fluid-Induced Transport Dynamics and Vibration Patterns of Multiphase Vortex in the Critical Transition States. Int. J. Mech. Sci. 2023, 252, 108376.

[4] Li, L.; Gu, Z.; Xu, W.; Tan, Y.; Fan, X.; Tan, D. Mixing Mass Transfer Mechanism and Dynamic Control of Gas-Liquid-Solid Multiphase Flow Based on VOF-DEM Coupling. Energy 2023, 272, 127015.

[5] Rabiei, R. Prediction of Breast Cancer Using Machine Learning Approaches. J. Biomed. Phys. Eng. 2022, 12, 297–308.

[6] Maniruzzaman, M.; Islam, M.M.; Rahman, M.J.; Hasan, M.A.M.; Shin, J. Risk prediction of diabetic nephropathy using machine learning techniques: A pilot study with secondary data. Diabetes Metab. Syndr. Clin. Res. Rev. 2021, 15, 102263

[7] Febrian, M.E.; Ferdinan, F.X.; Sendani, G.P.; Suryanigrum, K.M.; Yunanda, R. Diabetes prediction using supervised machine learning. Procedia Comput. Sci. 2023, 216, 21–30.

[8] Pradeepa, R.; Mohan, V. Epidemiology of type 2 diabetes in India. Indian J. Ophthalmol. 2021, 69, 2932–2938.

[9] Narwane, S.V.; Sawarkar, S.D. Is handling unbalanced datasets for machine learning uplifts system performance?: A case of diabetic prediction. Diabetes Metab. Syndr. Clin. Res. Rev. 2022, 16, 102609.

[10] Kavakiotis, I.; Tsave, O.; Salifoglou, A.; Maglaveras, N.; Vlahavas, I.; Chouvarda, I. Machine Learning and Data Mining Methods in Diabetes Research. Comput. Struct. Biotechnol. J. 2017, 15, 104–116.

[11] Bekele, B.B.; Manzar, M.D.; Alqahtani, M.; Pandi-Perumal, S.R. Diabetes mellitus, metabolic syndrome, and physical activity among Ethiopians: A systematic review. Diabetes Metab. Syndr. Clin. Res. Rev. 2021, 15, 257–265.

[12] Chatrati, S.P.; Hossain, G.; Goyal, A.; Bhan, A.; Bhattacharya, S.; Gaurav, D.; Tiwari, S.M. Smart home health monitoring system for predicting type 2 diabetes and hypertension. J. King Saud Univ. Comput. Inf. Sci. 2022, 34, 862–870.

[13] Reddy, D.J.; Mounika, B.; Sindhu, S.; Reddy, T.P.; Reddy, N.S.; Sri, G.J.; Swaraja, K.; Meenakshi, K.; Kora, P. WITHDRAWN: Predictive machine learning model for early detection and analysis of diabetes. Mater. Today Proc. 2020.

[14] Goyal, P.; Jain, S. Prediction of Type-2 Diabetes using Classification and Ensemble Method Approach. In Proceedings of the 2022 International Mobile and Embedded Technology Conference (MECON), Noida, India, 10–11 March 2022; pp. 658–665.

[15] Dutta, A.; Hasan, M.K.; Ahmad, M.; Awal, M.A.; Islam, M.A.; Masud, M.; Meshref, H. Early Prediction of Diabetes Using an Ensemble of Machine Learning Models. Int. J. Environ. Res. Public Health 2022, 19, 12378.

[16] Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. J. Artif. Intell. Res. 2002, 16, 321–357.

[17] Maulidevi, N.U.; Surendro, K. SMOTE-LOF for noise identification in imbalanced data classification. J. King Saud Univ.-Comput. Inf. Sci. 2022, 34, 3413–3423.

[18] Sanni, R.R.; Guruprasad, H.S. Analysis of performance metrics of heart failured patients using Python and machine learning algorithms. Glob. Transit. Proc. 2021, 2, 233–237.

[19] Chicco, D.; Tötsch, N.; Jurman, G. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. Bio-Data Min. 2021, 14, 13.
[20] Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genom. 2020, 21, 6.
[21] Erickson, B.J.; Kitamura, F. Magician's corner: 9. Performance metrics for machine learning models. Radiol. Artif. Intell. 2021, 3, E200126.
[22] Tan, J.; Yang, J.; Wu, S.; Chen, G.; Zhao, J. A critical look at the current train/test split in machine learning. arXiv 2021, arXiv:2106.04525