International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048
Volume 9, Issue 7, pp 71-80, July-2025
https://doi.org/10.47001/IRJIET/2025.907008

DiREcT AI: Development and Validation of a Machine Learning Tool for Diabetes Complications Risk Education in South Indian Patients

[1]Debdeep Saha, [2]James Devasia, [3]Jayaprakash Sahoo, [4]Subitha Lakshminarayanan

[1]Intern, Jawaharlal Institute of Postgraduate Medical Education and Research, Pondicherry, India
[2,4]Department of Preventive and Social Medicine, Jawaharlal Institute of Postgraduate Medical Education and Research, Pondicherry, India
[3]Department of Endocrinology, Jawaharlal Institute of Postgraduate Medical Education and Research, Pondicherry, India
Corresponding Author's E-mail: subitha.l@gmail.com

Abstract

Background and Aims: **Diabetes mellitus, a group of metabolic diseases requiring multifactorial risk reduction and continuous medical care, poses significant challenges for complications prevention beyond glycemic control. This study addresses the contemporary emphasis on Artificial Intelligence (AI) and machine learning (ML) to develop algorithms capable of learning patterns and decision rules from data. Despite the existence of risk scores for complications, their limitations in accurately estimating both types of diabetes complications underscore the need for predictive models based on local data applicable in bedside and clinic settings.**

Methods: **In the Endocrinology OPD setting of a tertiary care hospital, we have utilized four Neural Network-based algorithms (Random Forest, Decision Tree, K Neighbour [KNN], and Artificial Neural Networks [ANN]) to predict complications in type 2 diabetes patients.**

Results: **A Deep Neural Network model integrating these algorithms achieved optimal results, particularly with the ANN GRU model exhibiting a sensitivity of 89%, specificity of 97%, F1 score of 0.93, and AUC ROC of 0.98.**

Conclusion: **This study outlines the successful development and validation of a machine learning-based model for predicting adverse outcomes associated with diverse diabetes complications, underscoring the potential of machine learning in individual risk predictions and offering a practical application for patient education, facilitating behavior change for risk reduction and overall wellness.**

Keywords: DiREcT AI, Machine Learning, Tool for Diabetes, Risk Education, South Indian Patients, Artificial Neural Networks, ANN, KNN, Artificial Intelligence, AI.

I. INTRODUCTION

Diabetes is a group of metabolic diseases associated with hyperglycemic changes due to insulin secretion defects, insulin resistance or both. It requires multifactorial risk-reduction in addition to glycemic control and continuous medical care (1). Hyperglycemia may lead to complications, including both macrovascular (coronary artery disease, peripheral arterial disease, and stroke) and microvascular complications (diabetic neuropathy, nephropathy, and retinopathy) (2).

According to the International Diabetes Federation (IDF) the prevalence of diabetes is expected to rise from 537 million (20-79 years) in 2021 to 783 million 2045, a world where 1 out of 8 individuals may fall prey to this disease and its consequences (3). The burden of diabetes has been on the rise in India, with the number of patients being 65 million in 2016 as compared to 26 million in 1990. Diabetes has the highest contribution to the increase in health loss since 1990 with respect to major non-communicable diseases (4). Moreover, research on diabetes complications is advancing, revealing a multifaceted interplay of various causative factors. A comprehensive strategy is essential for promptly identifying these complications (5). Preventing long term complications is crucial, and education about self-management and support are the key factors in reducing the risk of long-term microvascular and macrovascular complications. Late detection of complications associated with diabetes can cause increased morbidity and financial burden in patients, especially in developing countries like India.

Though some risk scores for development of complications in diabetes have been developed, they misestimate both type of complications (6). There is a need to predict future complications based on local data that can be used in the bedside and in clinics. There is a gap in knowledge regarding prediction of diabetes related complications,

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 7, pp 71-80, July-2025

https://doi.org/10.47001/IRJIET/2025.907008

specifically in the Indian context. With the current three-pronged situation comprising of rising burden of diabetes in India, increased usefulness of Artificial Intelligence (AI) in medicine, and advent of precision prognostics; there is a great impetus in terms of prioritizing patient treatment to develop such models for risk scores (7, 8). Among the various models tested in recent times, most machine learning models (among which neural networks were the most frequently utilized) reported a positive mean relative Area Under Curve (AUC) when compared to non-machine learning methods, with random forest models showing the higher accuracy in a substantial number of studies (9,10). A recent study was aimed at improving the degree of perception about developing diabetes related complications. An educational tool was developed and validated through literature review and Delphi technique. The DiREct tool developed categorized 11 risk factors into low, moderate and high (11).

Recently, a great emphasis has been put to the Artificial intelligence and machine learning, aimed at developing algorithms which are able to learn patterns and decision rules from data (12). Such attempts combined with technological and biomedical advances enable early detection and diagnosis of diabetes (13). In a study by Fiarni et al, Data Mining algorithm was employed to predict complicated diabetic disease in Indonesia. The overall accuracy of the model was found to be 68% (14). Machine learning-based models have been developed using administrative health data to predict adverse outcomes associated with diabetes (15).

Development of such accurate, user-friendly tools will help to provide a better risk perception among health-care professionals and patients. There will be better compliance to lifestyle modification among diabetic patients who know they are at an increased risk for development of diabetic complications. In this context, this study was done to develop a machine learning model for predicting development of diabetes related complications based on risk factors, in patients with type 2 diabetes mellitus attending diabetes clinic in a tertiary care hospital in a developing country setting.

II. MATERIAL AND METHODS

Study design and setting: This was a descriptive study conducted in the Endocrinology OPD setting of a tertiary care hospital. Jawaharlal Institute of Postgraduate Medical Education and Research (JIPMER) is a tertiary care hospital situated in Puducherry. Diabetes clinics are conducted twice weekly in both departments of Medicine and Endocrinology. Around 3000 patients with diabetes are on follow up and investigations are carried out to assess for complications. Data collection was completed from August – October 2021.

Study Participants: Patients more than 18 years of age and on treatment for type 2 Diabetes; receiving treatment for a minimum of three months were included in the study.

Sample size and sampling: It was proposed to enrol 1000 patients in this studyas we had selected 12 variables to be studied and big data is required for machine learning algorithms. However, we could complete an enrolment of 537 patients. Universal sampling was followed.

Data collection: Patients with Diabetes were contacted in their clinics and written informed consent was obtained. Data was collected based on a pre tested proforma through patient interviews and chart reviews.

- Disease details: Age of onset, Family history, Duration of diabetes
- Lifestyle factors: Physical activity, Dietary pattern, Medication adherence, Smoking status
- Anthropometry & Biochemical parameters: Body mass index, Glycaemic control (Fasting blood sugar), Systolic BP (mmHg)

Standardised definitions were used for compliance to lifestyle factors and obtained through patient interviews. Presence of macrovascular and microvascular complications like Coronary heart disease, diabetic foot, retinopathy, nephropathy and neuropathy were recorded from case sheets. Electronic data capture was done using Epidata version 3.1 for mobile phone. The patients were categorized based on the level of risk development of diabetes related complications.

RISK STRATIFICATION FOR DEVELOPMENT OF DIABETES RELATED COMPLICATIONS (DiReCT)

	LOW RISK	MODERATE RISK	HIGH RISK
1. Glycaemic control (Fasting)	<100	101-125	≥ 126
2. Systolic BP (mmHg)	120 – 139	140- 159	>160

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 9, Issue 7, pp 71-80, July-2025
https://doi.org/10.47001/IRJIET/2025.907008

3. Obesity	18.5-22.9	23-24.9	25-29.9
4. Smoking	Neversmoker	Ex-smoker	Current smoker
5. Physical activity	≥150 min/week	100-149 min/week	<100 min/week
6. Dietary pattern	More vegetables& Proteins+ less carbohydrate	Less proteins vegetables+ more carbohydrate	Only carbohydrate
7.Medication adherence	High adherence	Medium adherence	Low adherence
8. Duration of diabetes	<5 years	5-11 years	>10 years
9. Family history	1st degree relative* - no DM complication and HTN	1st degree relative- H/O either one	legree relative- H/O both
10. Age of onset	≥ 60 years	41-60 years	20- 40 years

Figure 1: Risk stratification for development of diabetes related complications based on DiReCT tool

Ethical Considerations: This study received ethical approval from the Institutional Ethics Committee for Observational Studies of Jawaharlal Institute of Postgraduate Medical Education and Research (JIPMER). Written informed consent was taken before enrolment into study. Confidentiality of data collected was ensured. The work adhered to the guidelines and regulations outlined by the Indian Council of Medical Research (ICMR), ICMR-AI ethics guidelines, and the International Committee on Harmonization of Good Clinical Practice (ICH-GCP). We also addressed the risk in AI applications according to WHO guidance for Ethics and Governance of Artificial Intelligence for Health.

Model to predict the Diabetes related complications

In this study, we employed four algorithms - Random Forest, Decision Tree, K Neighbour, and Artificial Neural Networks - for predicting diabetes complications. These models were developed using the TensorFlow framework (version 2.7.0, developed by the Google Brain Team, available at https://tensorflow.org), and Keras (version 2.7.0, accessible at https://keras.io). Python (version 3.7, from the Python Software Foundation, https://python.org) served as the programming language for implementing these deep learning frameworks. The computational work was performed on a desktop computer configured with an Intel i9-9820X CPU @3.30 GHz, 64GB of RAM, and dual NVIDIA GeForce RTX 2080Ti GPUs each with 11GB of memory.

Random Forest Classifier (RFC)

Random forest or random decision tree is the supervised machine learning algorithm widely used in classification problem. In this work we employed scikit learn model selection tool "GridsearchCV" for finding the optimal parameters for Random Forest Classifier. The optimal parameters were then obtained from the grid search to train the classifier. Number of estimators were 100, max depth of the tree was 90 were obtained from model selection tool.

Decision Tree Classifier (DTC)

Decision Tree Classifier is a non-parametric supervised machine learning algorithm that use set of rules to make decision and widely used in classification. The Optimal parameters for Decision Tree classifier, mainly depth were obtained from GridsearchCV method, with tree depth obtained was 80. The classifier trained with optimal parameters for prediction of diabetics complications

KNN Classifier (KNN)

The K-nearest neighbors algorithm is a non-parametric supervised machine learning algorithm used in classification. The algorithm use proximity to make classifications. GridsearchCV used to get the optimal number of neighbours for KNN algorithm. The search result obtained the optimal number of neighbours was 3.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 7, pp 71-80, July-2025

https://doi.org/10.47001/IRJIET/2025.907008

Artificial Neural Networks (ANN) classifier

Artificial Neural Networks are comprised of interconnected nodes, containing an input layer, with one or more hidden layers and an output layer. In this work we developed an ANN classifier with two Dense layer and RELU as activation layer. The final classifier layer with single node dense layer with sigmoid activation. The model were compiled using binary cross entropy (BCE) loss.

Pseudocode for RFC, DTC, and KNN Implementation

1. Load the training data

Load dataset with 537 instances (complications: 17.5% positive) X = features y = target (complications)

2. Prepare data

Scaling

Scale X using StandardScaler

Missing Value Treatment

Impute missing values using mean/median strategy

Apply SMOTE to handle class imbalance

Apply SMOTE to balance dataset (result: 443 positive, 443 negative)

Split data into training and validation sets (80% train, 20% test)

Split X, y into X_train, X_test, y_train, y_test (test_size=0.2)

3. Find optimal parameters using GridSearchCV

For Random Forest Classifier (RFC)

Define param_grid_RFC = {n_estimators: [50, 100, 200], max_depth: [50, 90, 100]} Run GridSearchCV with RFC on X_train, y_train using param_grid_RFC Get best_params_RFC (n_estimators = 100, max_depth = 90)

For Decision Tree Classifier (DTC)

Define param_grid_DTC = {max_depth: [50, 80, 100]} Run GridSearchCV with DTC on X_train, y_train using param_grid_DTC Get best_params_DTC (max_depth = 80)

For K-Nearest Neighbors (KNN)

Define param_grid_KNN = {n_neighbors: [3, 5, 7]} Run GridSearchCV with KNN on X_train, y_train using param_grid_KNN Get best_params_KNN (n_neighbors = 3)

4. Fit the models with optimal parameters

RFC

Initialize RFC with n_estimators = 100, max_depth = 90 Fit RFC on X_train, y_train

DTC

Initialize DTC with max_depth = 80 Fit DTC on X_train, y_train

KNN

Initialize KNN with n_neighbors = 3 Fit KNN on X_train, y_train

5. Predict class value for new data

For RFC

Predict y_pred_RFC = RFC.predict(X_test)

For DTC

Predict y_pred_DTC = DTC.predict(X_test)

For KNN

Predict y_pred_KNN = KNN.predict(X_test)

Dataset

The dataset featured a binary predictor variable named 'complications,' with the incidence of positive cases being 17.5% out of a total of 537 instances, indicating a significant class imbalance. To address this imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was utilized. SMOTE selects a random instance from the minority class and identifies its k-nearest neighbors. New instances are then synthesized along the line segments connecting the selected instance to its neighbors. By applying SMOTE, the dataset was rebalanced to consist of 443 instances each of positive and negative 'complications,' thereby mitigating the initial imbalance.

ANN model training

For the training of the ANN model (Fig. 2), the Train/Test split approach was employed, allocating 80% of the

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 7, pp 71-80, July-2025

https://doi.org/10.47001/IRJIET/2025.907008

data for training and 20% for testing purposes. The data underwent normalization (pre-processing) utilizing the StandardScaler to ensure adherence to a Standard Normal Distribution. The model underwent training for a total of 500 epochs with batch processing set at increments of ten. The input layer was configured to accommodate ten nodes, corresponding to the ten independent variables incorporated within the model.

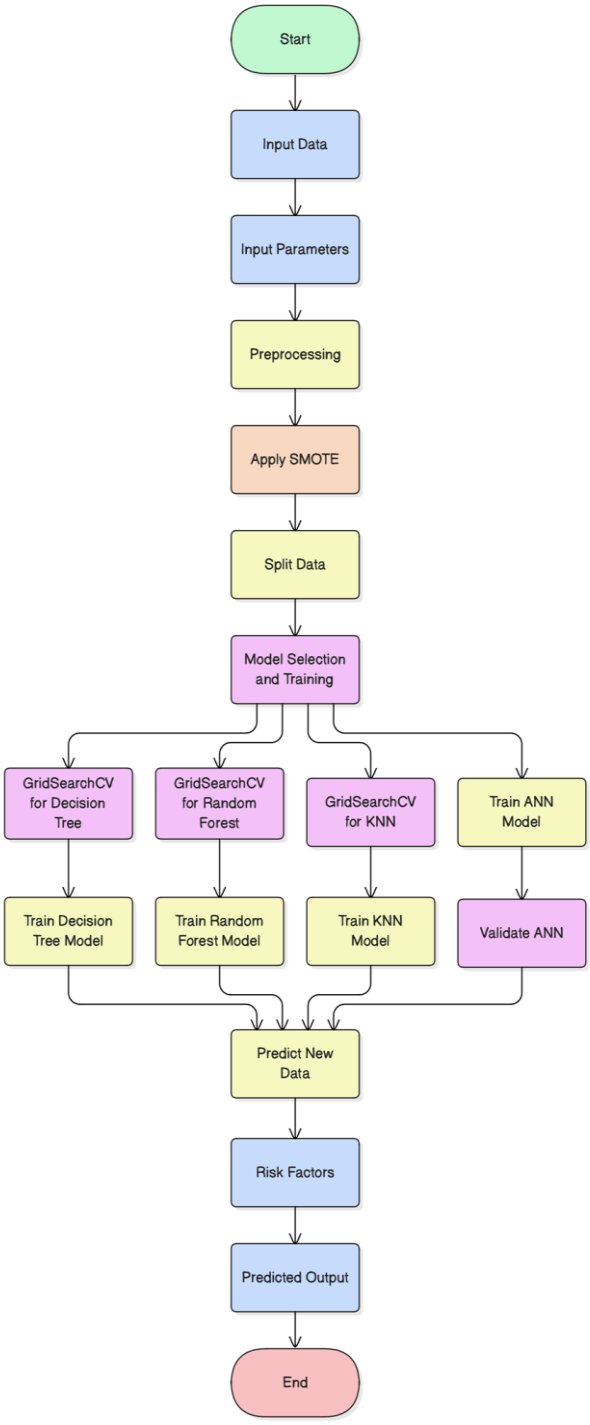

Figure 2: Flow chart for Machine Learning model – DiREcT AI

Statistical analysis

Data was digitally recorded using Epidata 3.1 optimized for mobile devices. Statistical evaluations were conducted utilizing SPSS version 19. Quantitative variables were described using mean and standard deviation, while qualitative variables were represented as percentages.

The efficacy of the Machine Learning Model was gauged through the following statistical measures: True Positives (TP) and True Negatives (TN) denote correct classifications, whereas False Positives (FP) and False Negatives (FN) denote misclassifications.

1. Accuracy quantifies the overall correct predictions, calculated as (TP + TN) / (Total Predictions).
2. Precision or Positive Predictive Value (PPV) assesses the proportion of actual positives among positive predictions, computed as TP / (TP + FP).
3. Specificity measures the proportion of actual negatives correctly identified, calculated as TN / (Actual Negatives).
4. Sensitivity reflects the ability to identify actual positives, computed as TP / (Actual Positives).
5. F1 Score provides a harmonic mean of Precision and Sensitivity, calculated as 2 * ((Precision * Sensitivity) / (Precision + Sensitivity)).
6. AUC of the ROC curve serves as a comparative metric for model performance across different thresholds.

III. RESULTS

The study's socio-demographic composition is a representation of the population visiting endocrine OPDs for treatment of type-2 DM in South India (table 1). Age distribution reveals 4.8% in the 18-30 group, 66.8% aged 31-60, and 28.3% aged 61 or above. Gender distribution shows a balanced mix, with 41% males and 59% females. Glycemic control data indicates 12.3% with fasting glucose levels of 100 or below, 20.6% in the 101-125 range, and 67% with levels of 126 or above. Systolic blood pressure categorization includes 74.5% below 140 mmHg, 19.7% at 140-159, and 5.6% at 160 or above. In terms of obesity, 18.7% have a BMI below 23, 16.9% fall in the 23-24.9 range, and 64.3% have a BMI of 25 or above.

Additionally, the study encompasses lifestyle factors such as smoking status, physical activity, dietary patterns, medication adherence, duration of diabetes, and family history, contributing to a comprehensive socio-demographic profile. This multifaceted data provides a foundation for training the models with potential risk factors and health behaviours within the study population.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 7, pp 71-80, July-2025

https://doi.org/10.47001/IRJIET/2025.907008

Model Performance

Among the four models tested, the ANN surpassed the others across all performance indicators. Both the Decision Tree and KNN classifiers demonstrated comparable levels of sensitivity, specificity, accuracy, precision, and F1 score, all of which were modest when contrasted with the superior outcomes of the Random Forest and ANN classifiers.

In this comparative analysis (Table. 3) of machine learning algorithms, the data reveals distinct performance characteristics. The Random Forest algorithm demonstrates superior proficiency with an AUC of 0.97, indicating excellent class separation ability. It also shows high sensitivity (0.87) and specificity (0.93), suggesting effective identification of true positives and negatives, with an overall accuracy of 0.90. The Decision Tree and KNN algorithms exhibit moderate performance; both have an accuracy of 0.73 and similar sensitivity and specificity, but their AUC values are lower (0.80 for Decision Tree, 0.89 for KNN), indicating less effective class discrimination. The ANN algorithm outperforms others with the highest AUC of 0.98, excellent sensitivity (0.89), and near-perfect specificity (0.97), leading to an impressive accuracy of 0.93.

The ROC curve analysis (Fig 3) presents the Random Forest and ANN algorithms as top performers with AUCs nearly at 0.975, indicating excellent classification capabilities. The Decision Tree has a lower AUC of 0.8015, showing weaker discriminative power, while KNN is better at 0.8865 but still trails the leading two algorithms.

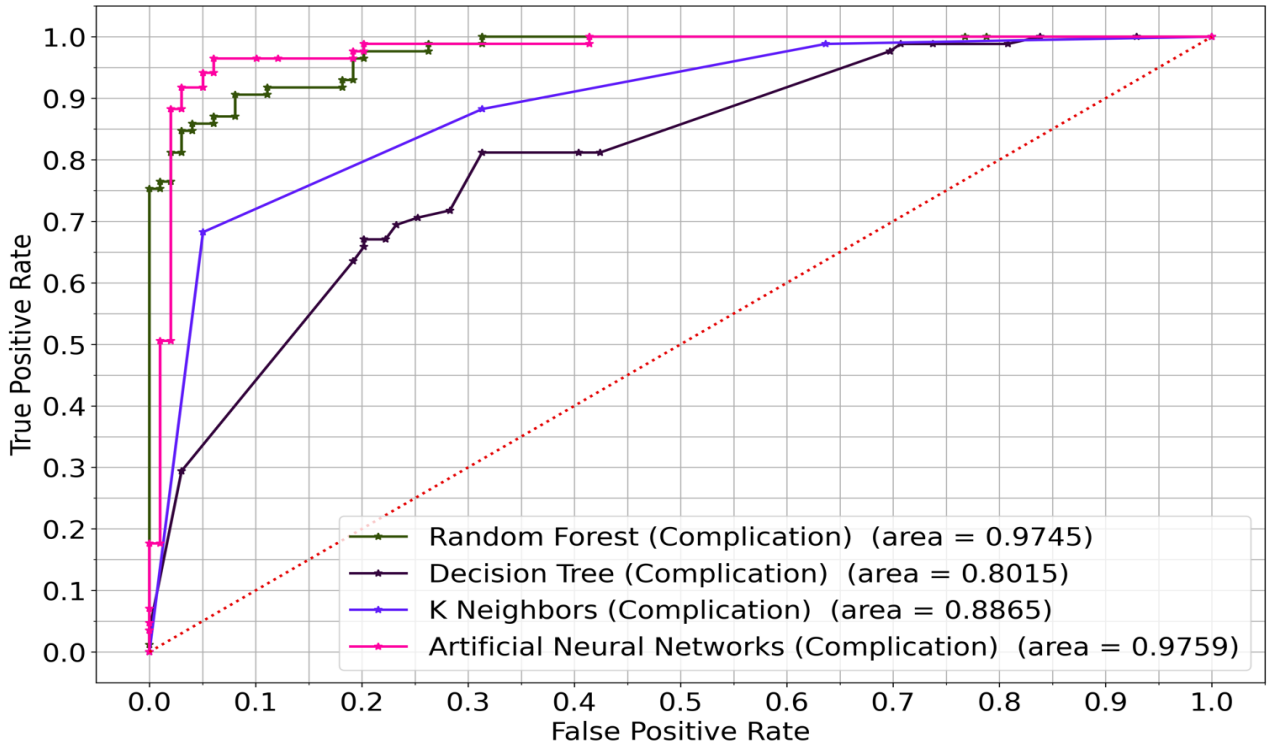

Figure 3: ROC of the DiREcT AI model in predicting 'Any complication'

IV. DISCUSSION

This study aims to develop and validate a deep learning model for risk prediction for development of complications in patients with diabetes. Our model can be used to predict the risk of developing diabetes-related complications with accuracy of 75% and a test AUC of 0.84. Further, web-app based on the trained DNN model can be developed for personal prediction of the probability of developing any complication following diabetes as a percentage. The entry parameters for such personalized predictions can be obtained from patient records and through routine interview.

A large number of studies have reported on the prediction model of diabetic complications in various socio-demographic populations. In a study by Fiarni et al, Data Mining algorithm was employed to predict complicated diabetic disease in Indonesia. The overall accuracy of the model was found to be 68%. Retinopathy is commonly associated with female patients having hypertensive crisis. Diabetes for a duration more than 4 years may contribute significantly to incidence of nephropathy. BMI over 25 and female gender are often the key risk factors for neuropathy (14).

ISSN (online): 2581-3048
Volume 9, Issue 7, pp 71-80, July-2025
https://doi.org/10.47001/IRJIET/2025.907008

In a study which employed advanced machine learning algorithms, including Recurrent neural network (RNN), long short-term memory (LSTM) and RNN gated recurrent unit (GRU) to predict diabetic complications, random forest and multiplayer perception traditional models were compared with the models designed. RNN GRU gave the highest prediction accuracy (between 73%-83%), as compared to 66%-76% accuracy of traditional models (15).

Studies done in different parts of the world may show varying results. This might be partially attributable to the variable epidemiology owing to the unique genetic and environmental factors prevalent in the target population. A study in conducted in Sudan found logistic regression model to perform better than random forest and KNN in their study population, with the highest recall score of 81% (as compared to 62% and 57% respectively) and F1 score of 75% (16).

A recent study from Canada has developed a machine learning-based model which uses administrative health data to predict adverse outcomes associated with diabetes. This model was able to predict the risk of adverse outcomes over a period of three years due to microvascular and macrovascular complications. It had strong discrimination (average test

AUC = 77.7, range 77.7–77.9). This highlighted the potential of AI and electronic health data by including 700 features from multiple diverse data sources of cohort of 1,567,636 patients (17).

Early prediction of diabetes related complications in diabetes clinics is important for risk stratification and management. Use of such algorithms can be facilitated through desktop application and performed at the screening counters in these clinics. Early assessment of risk of diabetic complications will increase the willingness of lifestyle modifications (like physical activities, diet control and continuous blood sugar monitoring) in patients with higher risk, thereby reducing morbidity associated with diabetic complications.

This is one of the first studies which focuses on development of AI based model for risk prediction among patients with diabetes using database developed from our setting. The model is developed and validated on the south Indian population which is more specific and powerful to apply in our setting than existing models developed from US or other populations. Further the risk factors utilized in this model has been derived from studies conducted in the local setting, and developed as a risk prediction tool. There are several limitations to the current study in addition to those associated with retrospective data collection. The models, in order to have robust performance and more accurate prediction possibilities need to have larger sample size. Much of AI based studies in this topic have employed data mining techniques for using EHR data, but these are major constraints in a developing country setting (18). The model also needs to explore further on individual types of complications listed eg. nephropathy, neuropahy, etc. Thirldy, since this model is trained with a dataset of almost exclusively South Indian patients, it's performance may vary when tested on other populations. The inclusion of a South Indian population, predominant among patients attending outpatient departments in our region, presents a dual facet wherein it may be construed as both a limitation and a strength, particularly in terms of enhancing specificity for this demographic. This characteristic has the potential to bolster precision prognostics tailored to the unique genetic and environmental factors prevalent in the South Indian populace (19). Fourthly, the seamless day-to-day usability and widespread acceptance of AI algorithms face challenges, similar to other AI-related technologies, mainly due to the intricacies of their algorithms (20). Lastly, as with all data-driven machine learning models the accuracy of input parameters is of paramount importance for optimal outcome. Since some of these parameters are based on patient recall rather than concrete data management systems, it may have a component of recall bias affecting the overall performance.

The escalating integration of technology and artificial intelligence (AI) in healthcare has propelled the application of machine learning principles beyond the prediction of diabetes risk and its complications. These principles now extend to encompass continuous monitoring of symptoms and biomarkers, as well as facilitating self-management of diabetes. Moreover, the evolving landscape of AI has increasingly contributed to aiding clinical decision-making processes. This paradigm shift has the potential to yield a comprehensive enhancement in the comprehension of the disease and its associated risk factors. Furthermore, it facilitates improved glycemic control, fostering a holistic approach. Empowered by AI-driven risk predictions, both physicians and patients are equipped to partake in a collaborative and targeted optimization of healthcare strategies tailored to the unique needs of each individual (21).

V. CONCLUSION

In conclusion, we outline the development and validation of a model based on machine learning which aids in predicting a range of diabetes complications. We have tried to demonstrate the potential of machine learning for individual risk predictions that can be used as a simple application for patient education aimed at behaviour change for risk reduction and overall wellness.

ISSN (online): 2581-3048
Volume 9, Issue 7, pp 71-80, July-2025
https://doi.org/10.47001/IRJIET/2025.907008

REFERENCES

[1] American Diabetes Association. Standards of Medical Care in Diabetes—2020. Diabetes Care 2020;43(Suppl. 1):S1–S212

[2] Chawla A, Chawla R, Jaggi S. Microvasular and macrovascular complications in diabetes mellitus: Distinct or continuum? Indian J Endocrinol Metab. 2016 JulAug;20(4):546-51.

[3] International Diabetes Federation. IDF Diabetes Atlas, 10th edn. Brussels, Belgium: 2021. Available at: https://www.diabetesatlas.org

[4] India State-Level Disease Burden Initiative Diabetes Collaborators. The increasing burden of diabetes and variations among the states of India: the Global Burden of Disease Study 1990– 2016. Lancet Glob Health 2018; 6: e1352–62.

[5] Papatheodorou K, Banach M, Edmonds M, Papanas N, Papazoglou D. Complications of diabetes. Journal of diabetes research. 2015 Jul 12;2015.

[6] Basu S, Sussman JB, Berkowitz SA, Hayward RA, Bertoni AG, Correa A et al. Validation of Risk Equations for Complications of Type 2 Diabetes (RECODe) Using Individual Participant Data From Diverse Longitudinal Cohorts in the U.S. Diabetes Care 2018;41:586– 595

[7] Haug CJ, Drazen JM. Artificial Intelligence and Machine Learning in Clinical Medicine, 2023. N Engl J Med. 2023;388(13):1201-1208.

[8] Schiborn, C., Schulze, M.B. Precision prognostics for the development of complications in diabetes. Diabetologia 65, 1867–1882 (2022). https://doi.org/10.1007/s00125-022-05731-4

[9] Tan KR, Seng JJB, Kwan YH, Chen YJ, Zainudin SB, Loh DHF, Liu N, Low LL. Evaluation of Machine Learning Methods Developed for Prediction of Diabetes Complications: A Systematic Review. J Diabetes Sci Technol. 2023 Mar;17(2):474-489. doi: 10.1177/19322968211056917. Epub 2021 Nov 3. PMID: 34727783; PMCID: PMC10012374.

[10] Uddin, S., Khan, A., Hossain, M. et al. Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform DecisMak 19, 281 (2019). https://doi.org/10.1186/s12911-019-1004-8

[11] Anjali, C., Olickal, J.J., Arikrishnan, K., Zunatha Banu, A., Sahoo, J., Kar, S.S., & Lakshminarayanan, S. (2021). Development and testing of Diabetes Complications Risk Educational Tool (DiREcT) for improving risk perception among patients with diabetes mellitus: a mixed method study. International Journal of Diabetes in Developing Countries, 41, 504 - 510.

[12] Dagliati A, Marini S, Sacchi L, Cogni G, Teliti M, Tibollo V, De Cata P, Chiovato L, Bellazzi R. Machine Learning Methods to Predict Diabetes Complications. J Diabetes Sci Technol. 2018 Mar;12(2):295-302.

[13] Chaki J, Ganesh ST, Cidham SK, Theertan SA. Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review. Journal of King Saud University-Computer and Information Sciences. 2022 Jun 1;34(6):3204-25.

[14] Fiarni C, Sipayung E M and Maemunah S 2019 Analysis and Prediction of Diabetes Complication Disease Using Data Mining Algorithms Procedia Computer ence 161 449-457

[15] Ljubic B, Hai AA, Stanojevic M, Diaz W, Polimac D, Pavlovski M, Obradovic Z. Predicting complications of diabetes mellitus using advanced machine learning algorithms. J Am Med Inform Assoc. 2020 Jul 1;27(9):1343-1351. doi: 10.1093/jamia/ocaa120. PMID: 32869093; PMCID: PMC7647294.

[16] Abaker AA, Saeed FA. A comparative analysis of machine learning algorithms to build a predictive model for detecting diabetes complications. Informatica. 2021 Mar 15;45(1).

[17] Ravaut, M., Sadeghi, H., Leung, K.K. et al. Predicting adverse outcomes due to diabetes complications with machine learning using administrative health data. Npj Digit. Med. 4, 24 (2021). https://doi.org/10.1038/s41746-021-00394-8.

[18] Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal. 2017 Jan 1;15:104-16.

[19] Chung WK, Erion K, Florez JC et al (2020) Precision medicine in diabetes: a Consensus Report from the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD). Diabetologia 63(9):1671–1693. https://doi.org/10.1007/s00125-020-05181-w

[20] Basu K, Sinha R, Ong A, Basu T. Artificial Intelligence: How is It Changing Medical Sciences and Its Future? Indian J Dermatol. 2020 Sep-Oct;65(5):365-370. doi: 10.4103/ijd.IJD_421_20. PMID: 33165420; PMCID: PMC7640807.

[21] Ellahham S. Artificial intelligence: the future for diabetes care. Am J Med 2020 Aug;133(8):895-900.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 7, pp 71-80, July-2025

https://doi.org/10.47001/IRJIET/2025.907008

APPENDIX

Table 1: Socio-demographic and Clinical characteristics of patients with Type 2 Diabetes mellitus, attending Tertiary care centre, Puducherry 2021 (N= 537)

Variables	Category	Frequency	Percentage (%)
Age	18 – 30	29	4.8
	31 – 60	399	66.8
	> 61	169	28.3
Gender	Male	243	41.0
	Female	354	59.0
Glycemic control (Fasting)	<=100	74	12.3
	101 - 125	123	20.6
	>=126	400	67.0
Systolic BP (mmHg)	<140	445	74.5
	140 - 159	118	19.7
	>=160	34	5.6
Obesity (BMI)	<23	112	18.7
	23 – 24.9	101	16.9
	>=25	384	64.3
Smoking status	Never smoker	538	90.1
	Ex-smoker	34	5.7
	Current smoker	25	4.2
Physical activity (min / week)	>150	206	34.5
	100 - 149	37	6.2
	<100	354	59.2
Dietary pattern	More vegetables & proteins	522	97.2
	Only carbohydrates	15	2.79
Medication adherence	High adherence	508	85.1
	Medium adherence	45	7.5
	Low adherence	44	7.3
Duration of diabetes (in years)	< 5	244	45.4
	5 - 10	167	31.1
	10 <	126	23.4
Family history (1st degree relative)	No Diabetes Mellitus and Hypertension complication	369	68.7
	History of either one	136	25.3
	History of both	32	5.9
Age of onset (in years)	>60	47	8.7

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 7, pp 71-80, July-2025

https://doi.org/10.47001/IRJIET/2025.907008

	41 - 60	285	53.0
	20 - 40	205	38.1

Table 2: Frequency of Diabetes-Related Complications in Patients with Type 2 Diabetes Mellitus at a Tertiary Care Center, Puducherry, 2021 (N=537)

S. No	Complications	Frequency	Percentage (%)
1	Retinopathy	19	3.5
2.	Nephropathy	17	3.1
3	Neuropathy	35	6.5
4	Diabetic foot	10	1.8
5	Cardiovascular diseases	44	8.1
6	Stroke	4	0.7
7	Any of the above complications	**94**	**17.5**

Footnote: Data were collected from 537 patients with Type 2 Diabetes Mellitus at a tertiary care center in Puducherry in 2021. Complications were diagnosed based on clinical records following standard diagnostic criteria. The row for "Presence of Any Complication" is bolded to highlight the overall prevalence.

Table 3: Performance Metrics of AI Models for Predicting Complications in Patients with Diabetes

Algorithm	AUC [95% CI]	Sensitivity [95% CI]	Specificity [95% CI]	Accuracy	Precision	F1 Score
Random Forest	0.97 [0.95 - 1.00]	0.87 [0.80 - 0.94]	0.93 [0.88 - 0.98]	0.90	0.91	0.89
Decision Tree	0.80 [0.74 - 0.87]	0.67 [0.57 - 0.77]	0.78 [0.70 - 0.86]	0.73	0.72	0.69
KNN	0.89 [0.84 - 0.94]	0.67 [0.57 - 0.77]	0.78 [0.70 - 0.86]	0.73	0.72	0.69
ANN	**0.98 [0.95 - 1.00]**	**0.89 [0.83 - 0.96]**	**0.97 [0.94 - 1.00]**	**0.93**	**0.96**	**0.93**

Footnote: Performance metrics were evaluated on a dataset of 537 instances (17.5% positive for complications), balanced to 443 positive and 443 negative instances using SMOTE. Models were trained on 80% of the data and tested on 20%, with preprocessing including scaling and missing value treatment. Optimal parameters: Random Forest (n_estimators = 100, max_depth = 90), Decision Tree (max_depth = 80), KNN (n_neighbors = 3), ANN (two Dense layers with ReLU, sigmoid output, BCE loss). KNN: K-Nearest Neighbors; ANN: Artificial Neural Network. Best values are bolded.

Citation of this Article:

Debdeep Saha, James Devasia, Jayaprakash Sahoo, & Subitha Lakshminarayanan. (2025). DiREcT AI: Development and Validation of a Machine Learning Tool for Diabetes Complications Risk Education in South Indian Patients. *International Research Journal of Innovations in Engineering and Technology - IRJIET*, 9(7), 71-80. Article DOI https://doi.org/10.47001/IRJIET/2025.907008

****

