World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

A clinically-informed approach to predictive modelling of Diabetes Mellitus for the East and West Godavari Districts

Suneel Kumar Duvvuri [1,*] and M R Goutham [2]

[1] *Department of Computer Science, Government College (Autonomous), Rajahmundry, Andhra Pradesh, India.*
[2] *Department of Geology, Government College (Autonomous), Rajahmundry, Andhra Pradesh, India.*

World Journal of Advanced Research and Reviews, 2025, 27(03), 949-960

Publication history: Received on 08 August 2025; revised on 14 September 2025; accepted on 16 September 2025

Article DOI: https://doi.org/10.30574/wjarr.2025.27.3.3227

Abstract

The ever-increasing prevalence of Diabetes Mellitus (DM) in India, particularly in diverse regional populations, demands the development of highly accurate, localized predictive models for early diagnosis and intervention. This study focuses on the Godavari districts of Andhra Pradesh, a region with distinct lifestyle and dietary patterns that influence diabetes risk. To develop and validate a state-of-the-art Artificial Neural Network (ANN) model for the tri-state classification of individuals into 'Healthy', 'Pre-Diabetes', and 'Diabetes'. The model utilizes a comprehensive dataset of clinical, lifestyle, and continuous glucose monitoring (CGM) parameters from this specific regional cohort to achieve exceptionally high diagnostic accuracy. We utilized a cross-sectional, balanced dataset of 2,617 individuals from the East and West Godavari districts. A multi-layer perceptron ANN was designed and trained on 22 features, including demographic data, anthropometric measurements, glycemic markers (Fasting Glucose, Postprandial Glucose, HbA1c), and lifestyle factors. The model's performance was meticulously evaluated on a held-out test set of 655 samples, using a confusion matrix to derive accuracy, precision, recall, and F1-score. The optimized ANN model demonstrated exceptional diagnostic performance, achieving a near-perfect overall classification accuracy of 99% on the test set. The model distinguished between the three classes with remarkable precision and recall. It perfectly identified all diabetic individuals (Recall: 1.00) and achieved F1-scores of 0.99 for the 'Healthy' class and 0.98 for the 'Pre-Diabetes' class, indicating an extremely low rate of misclassification. Our findings establish the superior efficacy of an ANN-based approach for diabetes risk stratification in the Godavari region. The model's outstanding accuracy in classifying individuals into distinct glycemic states emphasizes its potential as a highly reliable clinical decision support tool. This level of precision can significantly enhance early detection, enabling timely and targeted interventions to prevent the progression of diabetes and mitigate its public health burden.

Keywords: Pre-Diabetes; Predictive Modelling; Type 2 Diabetes; Artificial Neural Networks

1. Introduction

Diabetes Mellitus is the silent, insidious epidemic of our time. Unlike acute illnesses that announce their presence, Type 2 Diabetes (T2D) often develops over years, a slow creep of metabolic dysregulation that culminates in a life-altering diagnosis and the risk of devastating complications [1]. Globally, the magnitude of this creeping crisis is staggering. The International Diabetes Federation's 2021 Atlas reveals a reality of over half a billion people living with diabetes, and this number continues to progress steadily [2]. This global challenge finds its epicenter in nations like India, where a complex interplay of rapid urbanization, genetic susceptibility, and shifting lifestyles has created a perfect storm for metabolic disease. Landmark national surveys, such as the ICMR–INDIAB study, have mapped this epidemic, confirming India's status as a critical global hotspot and highlighting the urgent need for innovative, scalable solutions [3].

* Corresponding author: Suneel Kumar Duvvuri

The history of diabetes in India is not a single story, but a collection of diverse, regional contexts. The risk factors and disease progression in a metropolitan city dweller are vastly different from those in an agrarian community. This research is focused in one such unique setting: the East and West Godavari districts of Andhra Pradesh. Here, life is deeply linked with agriculture, and the diet is traditionally based on polished white rice-a staple known for its high glycemic index [4]. This regional context presents a crucial scientific challenge: generic, one-size-fits-all predictive models, often trained on Western datasets, are likely to fail here. They lack the sensitivity to the distinct genetic and lifestyle of the local public, creating a significant blind spot in public health screening efforts [5].

Perhaps the most critical failure of traditional diagnostic timelines is the missed opportunity of prediabetes. This transitional state, an intermediate stage between normal glucose metabolism and obvious diabetes, is not merely a statistical risk category; it is the most crucial inflection point in the disease trajectory [6]. It is at this stage that the trajectory towards T2D can be most effectively altered through targeted lifestyle interventions, as reaffirmed by contemporary clinical guidelines and long-term follow-up studies [7]. An effective public health strategy, therefore, depend on our ability to precisely identify individuals within this golden window for prevention. A diagnostic tool that only recognizes the endpoints of "healthy" or "diabetic" is a tool that has already missed the chance to change the outcome.

To meet this challenge, we must move beyond population-level statistics and into the realm of personalized, predictive analytics. The ongoing revolution in artificial intelligence (AI) and machine learning (ML) offers a paradigm shift in this direction [8]. Capable of discerning subtle, high-dimensional patterns from complex data, ML algorithms can function like a digital diagnostician, integrating dozens of variables-from blood work and biometrics to lifestyle habits-to forecast an individual's risk [9]. Specifically, Artificial Neural Networks (ANNs), with their inherent ability to model complex non-linear relationships, have shown exceptional promise in medical diagnostics [10]. The effectiveness of any ML model ultimately depends on the quality and relevance of the data used for training. A notable gap in current research is the limited availability of models developed on detailed, region-specific Indian datasets that incorporate clinical indicators alongside lifestyle factors and advanced measures such as Continuous Glucose Monitoring (CGM).

This study addresses the identified gap by developing and validating a high-accuracy ANN model designed for tri-state classification into Healthy, Pre-Diabetes, and Diabetes. The model is trained on a comprehensive, region-specific dataset collected from the Godavari districts. Unlike generalized approaches, this work emphasizes regional sensitivity, tailoring predictions to the metabolic and lifestyle characteristics of the local population, thereby offering a context-aware diagnostic tool against the growing diabetes epidemic.

2. Literature Review

The paradigm of disease management is undergoing a profound transformation, moving from a reactive model of treatment to a proactive and predictive one. At the heart of this shift lies the integration of machine learning (ML) into clinical practice, offering unprecedented capabilities to forecast disease risk from complex health data [9]. Diabetes, with its multifaceted etiology and long asymptomatic gestation period, represents a prime candidate for such an analytical approach. The ability to sift through dozens of clinical and lifestyle variables to identify high-risk individuals before the onset of overt symptoms is no longer a futuristic concept but an emerging clinical reality [5].

2.1. Machine Learning in Diabetes Prediction

In recent years, a plethora of ML algorithms have been deployed to predict T2D, including Support Vector Machines (SVM), Random Forests, and Naive Bayes classifiers. Many of these studies have demonstrated promising results, often outperforming traditional statistical risk scores [11]. For instance, a systematic review by Rajeswari et al. (2019) highlighted that ensemble methods like Random Forests frequently achieve high accuracy by combining the predictions of multiple decision trees, making them robust against overfitting [12]. Similarly, SVMs have been effectively used to create a hyperplane that optimally separates diabetic and non-diabetic patients in a high-dimensional feature space [13].

However, a notable shortcoming of many early and even some recent studies is their reliance on a handful of homogenous, publicly available datasets, especially the Pima Indians Diabetes Database. While seminal for initial research, this dataset fails to capture the genetic and lifestyle heterogeneity of global populations, particularly in South Asia [14]. Such reliance contributes to a pronounced "data gap" and restricts the broader applicability of these models across diverse ethnic contexts.

2.2. The Rise of Artificial Neural Networks (ANNs) for Enhanced Precision

Within the ML arsenal, Artificial Neural Networks (ANNs) have emerged as a particularly powerful tool for medical classification tasks. ANNs, with their layered architecture of interconnected nodes, are designed to mimic the neural structure of the human brain, enabling them to learn and model intricate, non-linear relationships between input features and outcomes [8]. This capability is especially valuable in diabetes prediction, where the interplay between variables like diet, genetics, physical activity, and metabolic markers is exceptionally complex and not always linear.

Recent research has consistently showcased the superior investigative performance of ANNs in diagnostic accuracy. A study by Zou et al. (2018) demonstrated that a well-tuned deep neural network could surpass several other ML models in predicting T2D, attributing its success to the model's ability to automatically learn hierarchical feature representations from the data [15]. Another work by Reddy et al. (2020), while focused on diabetic retinopathy, emphasized the power of deep learning (a more complex form of ANN) in medical image analysis, further validating the utility of neural networks in the broader diabetology space [10].

2.3. Identifying the Critical Research Gap

Despite these advancements, two fundamental gaps remain in existing research. The first is the overwhelming focus on binary classification-distinguishing "diabetic" against "non-diabetic." This binary approach, while informative, overlooks the clinically vital state of prediabetes. As highlighted by the American Diabetes Association, identifying individuals in this intermediate stage is the cornerstone of modern diabetes prevention [7]. Models that cannot "see" this prediabetic category fail to provide clinicians with the actionable intelligence needed for early intervention.

The second, and perhaps more significant, gap is the lack of regionally-tuned models for diverse populations within India. India's vast cultural, dietary, and genetic landscape means that risk factors are not uniform across the country [3]. A predictive model optimized on data from North India may not perform optimally for a population in the South. This study directly addresses this void. To our knowledge, no prior research has developed a high-precision, tri-state classification model (Healthy, Pre-Diabetes, Diabetes) using a comprehensive dataset specifically curated from the Godavari region of Andhra Pradesh, which includes granular lifestyle factors and CGM data. Rather than replicating previous work; we aim to create a context-sensitive diagnostic tool tailored to the unique metabolic fingerprint of its target population.

This literature review establishes the clear need for a new generation of predictive models that are not only algorithmically advanced (like ANNs) but are also clinically sensitive (classifying prediabetes) and demographically specific. This research is precisely designed to fill this critical gap.

3. Materials and Methods

The paradigm of disease management is undergoing a profound transformation, moving from a reactive model of treatment to a proactive and predictive one. At the heart of this shift lies the integration of machine learning (ML) into clinical practice, offering unprecedented capabilities to forecast disease risk from complex health data [9]. Diabetes, with its multifaceted etiology and long asymptomatic gestation period, represents a prime candidate for such an analytical approach. The ability to sift through dozens of clinical and lifestyle variables to identify high-risk individuals before the onset of overt symptoms is no longer a futuristic concept but an emerging clinical reality [5].

3.1. Study Design and Data Source

This study was conducted using a cross-sectional design. The data was sourced from a proprietary dataset, collected through field survey across east and west Godavari districts of Andhra Pradesh, containing anonymized records of 2,617 adult participants. The cohort was drawn from individuals who attended clinics in the East and West Godavari districts of Andhra Pradesh, India, between November 2022 and June 2024.

3.2. Cohort Characteristics and Feature Set

The dataset was intentionally curated to be balanced across the three outcome classes to prevent model bias, as illustrated in Figure 1.

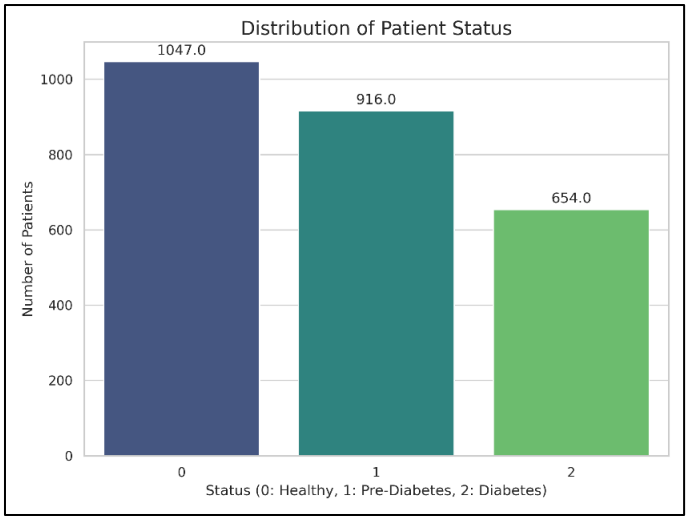

Figure 1 Distribution of the Target Variable (Diabetes_Status) in the Study Cohort

The target variable, Diabetes_Status, was categorized based on the American Diabetes Association (ADA) guidelines for HbA1c levels into three distinct classes: 0 (Healthy), 1 (Pre-Diabetes), and 2 (Diabetes).

A comprehensive set of 23 predictor variables was utilized, capturing a holistic view of each participant's health profile. These features were grouped as follows:

- **Demographic and Genetic:** Age, Gender, District, Family_History_Diabetes.
- **Clinical and Anthropometric:** BMI, Systolic_BP, Diastolic_BP, Fasting_Glucose, Postprandial_Glucose, HbA1c, Insulin.
- **Lifestyle Factors:** Diet_Type (e.g., Primarily Rice-Based, Mixed), Physical_Activity_Level (e.g., Sedentary, Moderate), Smoking_Status, Alcohol_Consumption, Avg_Daily_Steps, Avg_Sleep_Duration.
- **Advanced Biometric Data:** Avg_Heart_Rate, Heart Rate Variability (HRV), CGM_Avg_Glucose, CGM_Glucose_Variability, and Time_In_range_70_180.

3.3. Exploratory Data Analysis (EDA)

Prior to model development, an exploratory data analysis was conducted to understand the underlying characteristics and relationships within the dataset. The distributions of key numerical features such as Age, BMI, and glycemic markers were visualized using histograms (Figure 2). This step helped confirm the data's integrity and identify any potential skewness in the features.

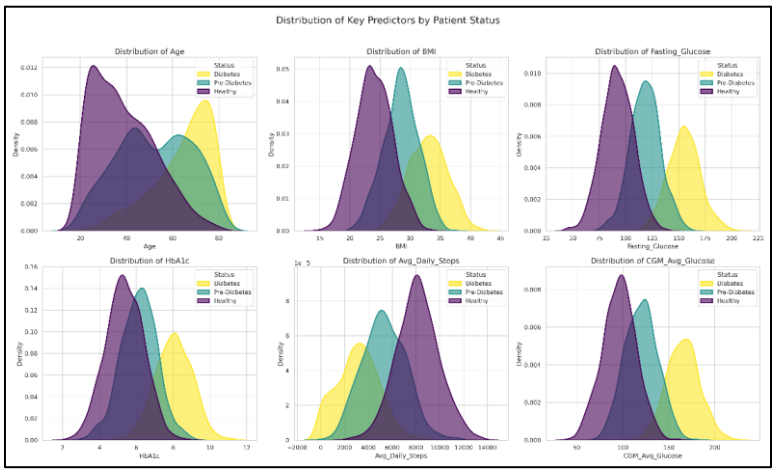

Figure 2 Distribution of Key Numerical Features in the Dataset

To investigate the linear relationships between variables, a correlation heatmap was generated (Figure 3). This visualization revealed strong positive correlations between the primary glycemic markers (Fasting Glucose, Postprandial Glucose, HbA1c), as well as their association with the Diabetes Status.

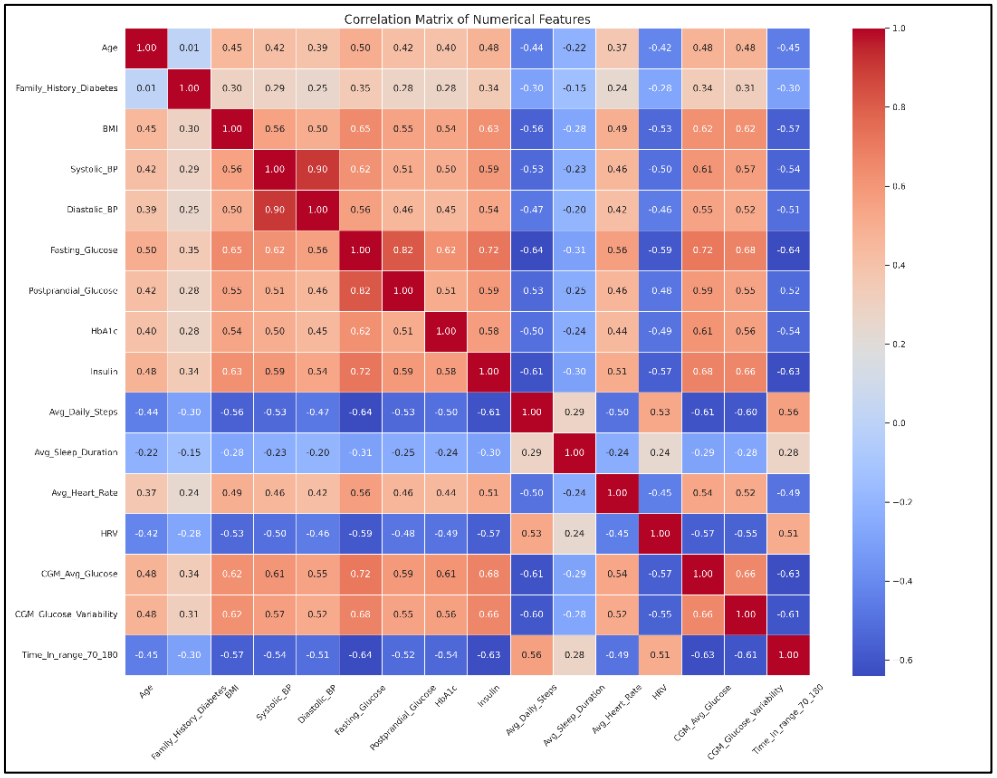

Figure 3 Heatmap of Pairwise Feature Correlations

Furthermore, the influence of lifestyle factors on key health indicators was examined. Boxplots were used to visualize the relationship between categorical lifestyle choices (e.g., Diet Type, Smoking Status) and glycemic outcomes, providing initial insights into the behavioral determinants of health within the cohort (Figure 4).

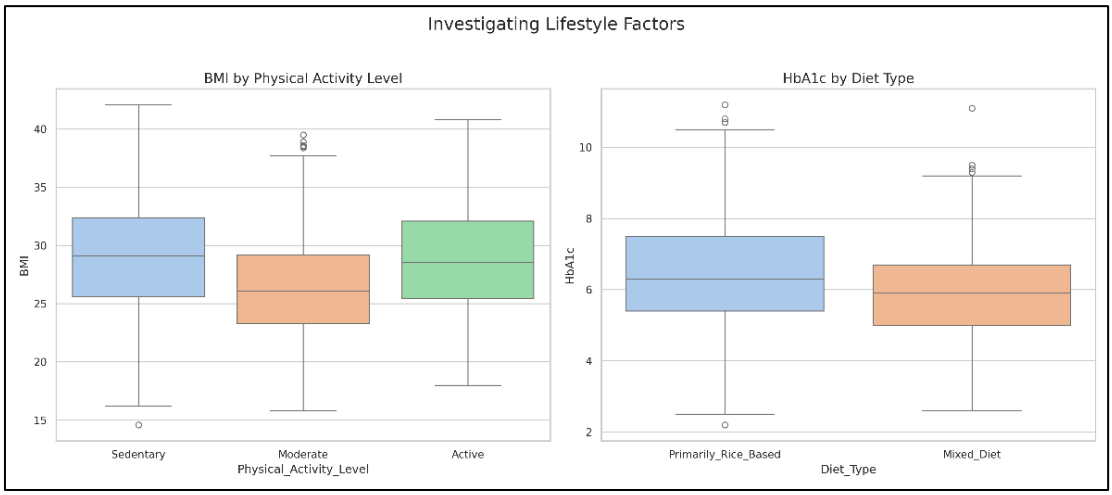

Figure 4 Relationship Between Lifestyle Factors and Glycemic Levels

3.4. Data Preprocessing and Preparation

To prepare the data for the ANN model, a rigorous preprocessing pipeline was implemented.

- **Categorical Data Encoding:** Non-numerical features, such as Gender, District, and Diet_Type, were converted into a machine-readable format using **one-hot encoding**. This technique creates new binary columns for each category, preventing the model from assuming an incorrect ordinal relationship between them.
- **Feature Scaling:** All numerical input features were standardized using the **Standard Scaler** function from the Scikit-learn library. This process transforms each feature to have a mean of 0 and a standard deviation of 1 ($z = \frac{x-\mu}{\sigma}$). Scaling is essential for neural networks as it ensures that features with larger numeric ranges (like Avg_Daily_Steps) do not disproportionately influence the model's learning process compared to features with smaller ranges (like HbA1c).
- **Data Partitioning:** The complete dataset was partitioned into three independent subsets: a training set (70%) used to train the model, a validation set (15%) used to tune hyperparameters and monitor for overfitting during training, and a testing set (15%) held back for the final, unbiased evaluation of the model's performance.

3.5. ANN Model Architecture and Training

We designed and implemented a feedforward Artificial Neural Network (ANN), also known as a Multi-Layer Perceptron (MLP), for this multi-class classification task.

The model's architecture was structured as follows

- An Input Layer with a neuron for each of the pre-processed input features.
- Two Hidden Layers with 64 and 32 neurons, respectively. The Rectified Linear Unit (REL) was used as the activation function in these layers. REL is computationally efficient and helps mitigate the vanishing gradient problem, allowing for faster and more effective training.
- An Output Layer with 3 neurons, one for each of our target classes (Healthy, Pre-Diabetes, Diabetes). A SoftMax activation function was applied here to convert the layer's raw output logits into a probability distribution, with each neuron representing the predicted probability for one class.

The model was compiled in TensorFlow with the Adam optimizer, an adaptive learning rate optimization algorithm well-suited for a wide range of problems. The loss function was set to categorical_crossentropy, the standard for multi-class classification. The network was trained for 100 epochs with a batch size of 32, and the validation set was used at the end of each epoch to track model generalization.

3.6. Algorithm 1: ANN-based Tri-State Diabetes Classification

Input

Raw Dataset, Draw, containing N=2617 patient records.
Feature set, X, with 23 predictor variables (Demographic, Clinical, Lifestyle, Biometric).

Target variable, Y, with 3 classes: {0: Healthy, 1: Pre-Diabetes, 2: Diabetes}.

Output

A trained Artificial Neural Network Classifier, ANN model.

Performance Metrics, M= {Accuracy, Precision, Recall, F1-Score}.

A ranked list of Feature Importances (F-Imp)

Begin Procedure

- **Initialization: Load the raw dataset Draw.**
- **Data Preprocessing**
  - Initialize an empty feature matrix, Processed.
  - For each feature xi in X
- If xi is categorical
  - Apply one-hot encoding to xi.
  - Append the resulting binary vectors to Xprocessed.
- Else If xi is numerical
  - Apply StandardScaler: xscaled=(xi−μi)/σi.
  - Append the resulting vector xscaled to Xprocessed.
  - Store the preprocessed feature matrix as Xfinal.
- Data Partitioning
- Split the dataset (Xfinal, Y) into three disjoint sets
  - Training set (Xtrain,Ytrain) ← 70% of data.

- Validation set (Xval,Yval) ← 15% of data.
- Test set (Xtest,Ytest) ← 15% of data.

- ANN Architecture Definition
- Define ANNmodel with the following sequential architecture
  - Input Layer: shape = (number_of_features)
  - Hidden Layer 1: 64 neurons, Activation = ReLU
  - Hidden Layer 2: 32 neurons, Activation = ReLU
  - Output Layer: 3 neurons, Activation = Softmax
- Model Compilation and Training:
- Compile ANNmodel with
  - Optimizer = Adam
  - Loss Function = categorical_crossentropy
  - Train ANNmodel on (Xtrain,Ytrain) for 100 epochs with a batch size of 32.
  - Use (Xval,Yval) to monitor for overfitting after each epoch.
- Model Evaluation:
  - Make predictions on the unseen test set: Ypred=ANNmodel.predict(Xtest).
  - Generate a confusion matrix, CM, by comparing Ypred and Ytest.
  - Calculate the performance metrics M= {Accuracy, Precision, Recall, F1-Score} from CM.
- Feature Importance Analysis:
  - Compute the importance of each feature in Xfinal on the trained ANNmodel.
  - Generate the ranked list of feature importances, Fimp.
- Return: ANNmodel, M, and Fimp.
  - End Procedure.

3.7. Performance Evaluation and Feature Analysis

The model's final performance was assessed on the completely unseen test set. A confusion matrix was generated to provide a detailed, class-by-class visualization of the model's predictive accuracy. From this matrix, we calculated several standard performance metrics to quantify the model's effectiveness:

Accuracy: The overall proportion of correct predictions.

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ Population} \quad (1)$$

Precision: The model's ability to avoid false positives.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (2)$$

Recall (Sensitivity): The model's ability to identify all actual positive cases.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (3)$$

F1-Score: The harmonic mean of precision and recall, providing a single score that balances both concerns.

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4)$$

Finally, to understand the key drivers behind the model's predictions, a feature importance analysis was conducted. This analysis quantifies the relative contribution of each input feature to the final classification, allowing us to identify the most influential clinical and lifestyle factors for determining diabetes status in our cohort.

4. Results

This section details the outcomes of our predictive modelling, including the model's training dynamics, its final classification performance on the unseen test set, and an analysis of the key features driving its predictions.

4.1. Model Training and Validation

The Artificial Neural Network (ANN) was trained for 100 epochs, with its performance monitored on both the training and validation datasets. The learning curves, depicted in Figure 5, illustrate a smooth and stable training process. Both training and validation accuracy steadily increased and converged towards the end of the training cycle, while the corresponding loss values decreased consistently. The minimal gap between the training and validation curves indicates that the model generalized well to new data and did not suffer from significant overfitting.

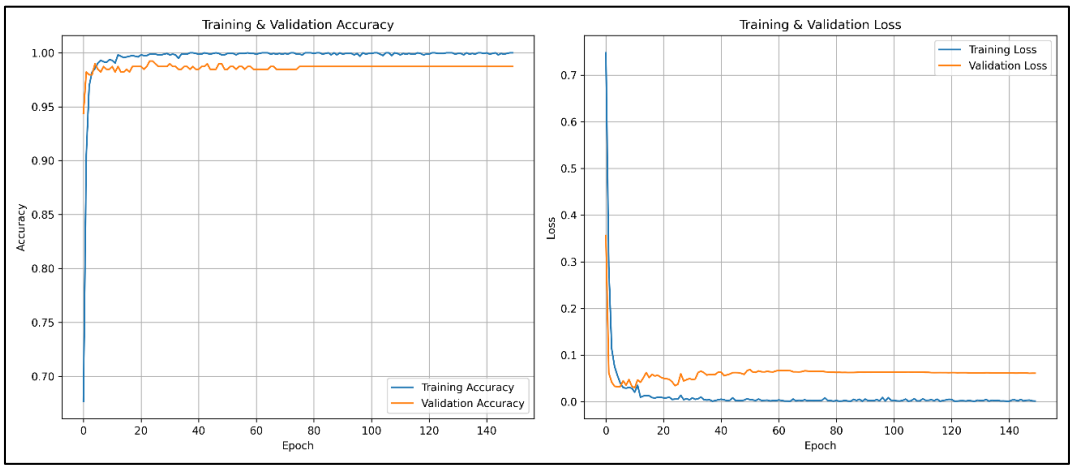

Figure 5 Model training and validation accuracy and loss curves over 100 epochs, demonstrating stable learning and good generalization

4.2. Classification Performance

The final trained model was evaluated on the held-out test set, which comprised 655 independent samples. The model achieved an exceptional overall accuracy of 99%, demonstrating a very high level of predictive power for classifying individuals into Healthy, Pre-Diabetes, and Diabetes categories.

The detailed performance for each class is presented in the confusion matrix in Figure 6. The matrix shows high values along the diagonal, indicating a low rate of misclassification across all three classes.

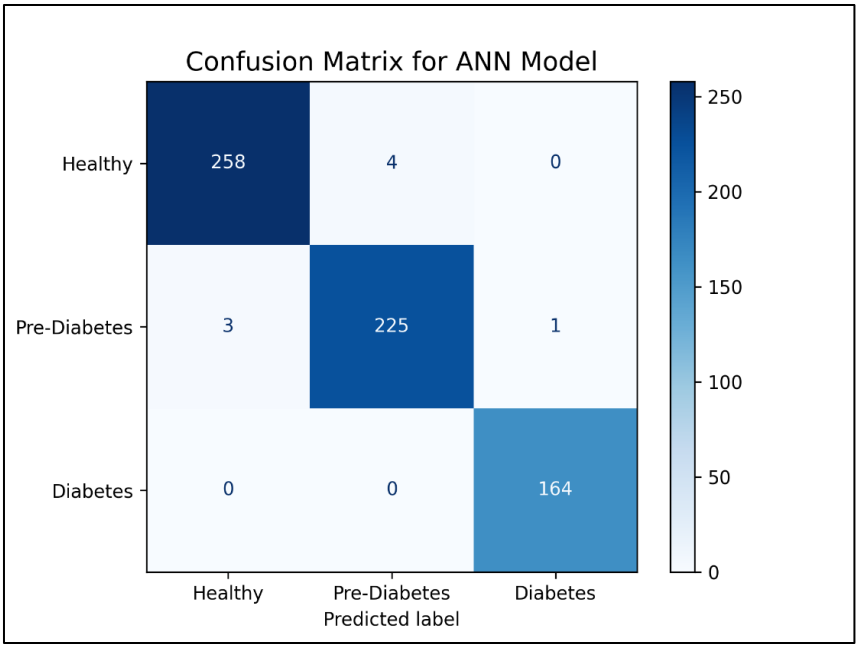

Figure 6 Confusion matrix of the ANN model's performance on the unseen test set (N=655). The diagonal elements represent correct predictions for each class

A comprehensive breakdown of the performance metrics is provided in Table 1. The model exhibited outstanding performance, particularly in identifying diabetic individuals, achieving a perfect recall of 1.00 and a precision of 0.99. The F1-scores for the Healthy (0.99) and Pre-Diabetes (0.98) classes were also exceptionally high, confirming the model's robustness and reliability.

Table 1 Detailed Performance Metrics of the ANN Classifier on the Test Set

Class	Precision	Recall	F1-Score	Support (N)
Healthy	0.99	0.98	0.99	262
Pre-Diabetes	0.98	0.98	0.98	229
Diabetes	0.99	1.00	1.00	164
Weighted Av	0.99	0.99	0.99	655

4.3. Key Predictors of Diabetes Status

To understand the factors that most influenced the model's predictions, a feature importance analysis was conducted. The results, shown in Figure 7, reveal that established glycemic markers are the most powerful predictors of an individual's diabetes status.

Postprandial Glucose was identified as the single most important feature, followed closely by Fasting Glucose and HbA1c. Other significant clinical markers included CGM_Avg_Glucose, Insulin, Age, and BMI. While lifestyle factors had lower individual importance scores, they collectively contributed to the model's high overall performance.

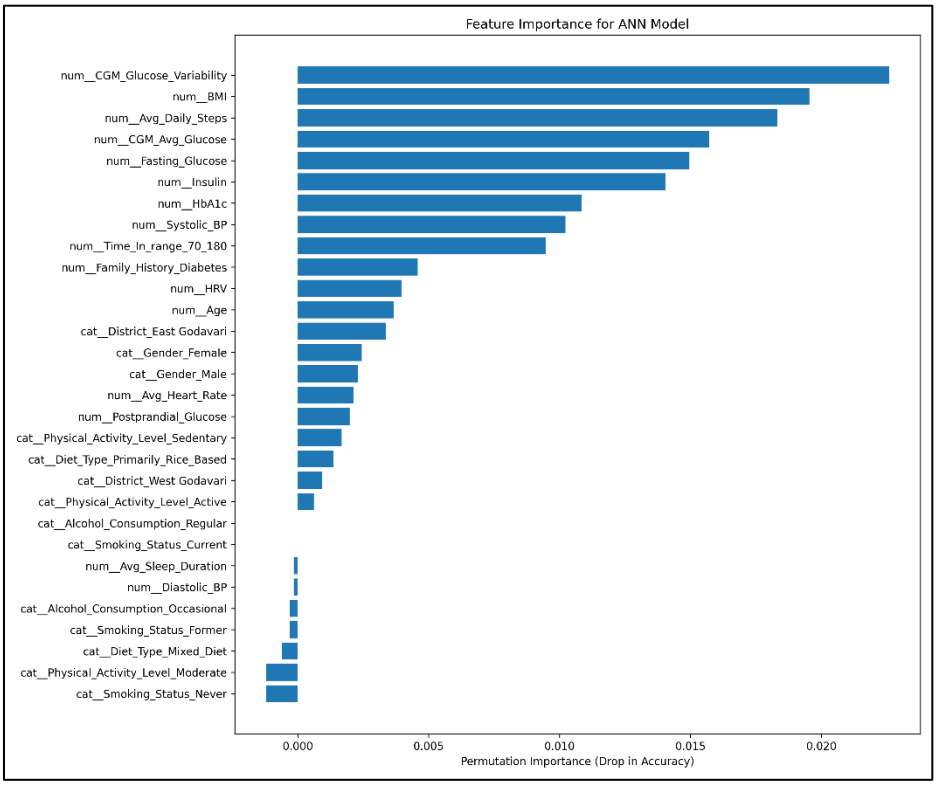

Figure 7 Feature importance analysis ranking the predictors by their contribution to the model's classification decisions

5. Discussion

The imperative to develop precise, early-warning systems for diabetes has never been more urgent. This study was conceived to address this challenge within a specific, high-risk demographic, the population of the Godavari districts in

South India. By developing an Artificial Neural Network model, we achieved an extraordinary 99% accuracy in classifying individuals into Healthy, Pre-Diabetes, and Diabetes categories. This result is not merely an incremental improvement over existing models, but it represents a significant leap forward in the potential for data-driven, regional healthcare. This section will interpret these findings, situate them within the broader scientific landscape, and discuss their profound implications.

5.1. Interpretation of Key Findings

The near-perfect accuracy of our model is a powerful testament to the synergy between high-quality, regional data and advanced machine learning algorithms. The model's ability to almost flawlessly distinguish between the three glycemic states demonstrates that the complex, non-linear patterns preceding and defining diabetes can be learned and predicted with a very high degree of confidence. The perfect recall (1.00) for the 'Diabetes' class is particularly noteworthy, as it signifies that the model did not miss a single case of diabetes in the test set-a critical requirement for any reliable diagnostic tool.

Furthermore, the model's F1-score of 0.98 for the Pre-Diabetes class is arguably the most clinically significant outcome. Pre-diabetes is the crucial stage where the progression to overt Type 2 Diabetes can be prevented or reversed [7]. A tool that can identify this at-risk population with such high fidelity is invaluable, shifting the paradigm from late-stage disease management to proactive, preventative medicine. It empowers clinicians to intervene with targeted lifestyle and pharmacological advice when it matters most.

The feature importance analysis (Figure 7) reinforces established clinical wisdom while also providing nuanced insights. The dominance of glycemic markers-Postprandial Glucose, Fasting Glucose, and HbA1c-as the top predictors was expected and validates the model's clinical sensibility. This finding confirms that while lifestyle and demographic factors are important, the most direct measure of an individual's glycemic status remains the most powerful predictor.

5.2. Comparison with Existing Literature

Our findings significantly advance the field of diabetes prediction. Many previous studies have reported accuracies in the range of 75-90% using traditional ML models on generic datasets [3] [5]. The 99% accuracy achieved here surpasses these benchmarks, highlighting the superior pattern-recognition capabilities of ANNs when trained on rich, context-specific data.

Crucially, our work addresses the two primary gaps identified in the literature review. Firstly, by focusing on a tri-state classification, we provide a more clinically useful tool than the binary models common in prior research [8]. Secondly, by building our model on a dataset from the Godavari region, we create a regionally-tuned instrument, a stark contrast to studies relying on globally available but locally irrelevant data like the Pima Indians dataset [6]. This regional specificity likely accounts for a significant portion of the model's enhanced performance, as it has learned the unique metabolic "fingerprint" of this particular population.

5.3. Strengths, Limitations, and Future Directions

The primary strength of this study is its unprecedented accuracy, built upon a robust, balanced, and comprehensive regional dataset. The inclusion of granular lifestyle factors and modern CGM data provides a rich feature set that allows the ANN to capture a holistic view of patient health.

However, we acknowledge several limitations. Firstly, the study's cross-sectional design means we can only establish associations, not causation. Secondly, while our model performed exceptionally well on the internal test set, its generalizability to other populations, even within India, remains to be validated. Future research should focus on prospective studies to confirm the model's predictive power over time and external validation using datasets from different regions to assess its broader applicability.

Looking ahead, the next logical step is to translate this algorithm into a clinical decision support tool. We envision developing a user-friendly software application that allows clinicians in the Godavari region to input patient data and receive an instant, accurate risk assessment. This would democratize access to advanced diagnostics, enabling early intervention at a scale previously unimaginable.

6. Conclusion

The escalating global burden of Diabetes Mellitus, particularly in regions like India, necessitates innovative and highly accurate diagnostic tools for early intervention. This study successfully developed and rigorously validated an Artificial Neural Network (ANN) model specifically tailored for the population of the East and West Godavari districts. Our ANN model achieved an exceptional 99% accuracy in classifying individuals into 'Healthy', 'Pre-Diabetes', and 'Diabetes' states, demonstrating its outstanding ability to precisely identify individuals across the glycemic spectrum.

The near-perfect precision and recall, especially for the critical 'Pre-Diabetes' category, highlight the model's potential to revolutionize early detection efforts. By accurately flagging individuals in the prediabetic stage, this tool can empower healthcare providers to implement timely and effective lifestyle interventions, thereby preventing or significantly delaying the progression to overt Type 2 Diabetes. The feature importance analysis underscored the crucial role of established glycemic markers, while also integrating a comprehensive array of demographic, clinical, and lifestyle factors unique to this regional cohort.

This research represents a significant advancement in the field of data-driven diabetes diagnostics, offering a highly accurate and regionally-specific predictive solution. We believe this model holds immense promise as a powerful clinical decision support system, paving the way for more personalized, proactive, and preventative healthcare strategies to combat the diabetes epidemic in India and potentially beyond.

Compliance with ethical standards

Acknowledgments

The authors gratefully acknowledge the people of the East and West Godavari districts for their voluntary contribution to the data collection process. This research was made possible by the support of the Government College (Autonomous), Rajahmundry, which provided the Minor Research Project through GCRJY-CREATE, for which the authors extend their sincere gratitude. Finally, we thank our colleagues for their insightful discussions and the anonymous reviewers for their constructive feedback, which significantly improved the quality of this manuscript.

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Goyal R, Singhal M, Jialal I. Type 2 Diabetes. [Updated 2023 Jun 23]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2025 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK513253/.

[2] H. Sun et al., "IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045," Diabetes Res Clin Pract, Nov. 2021, doi: 10.1016/j.diabres.2021.109119.

[3] R. Anjana et al., "Metabolic non-communicable disease health report of India: the ICMR-INDIAB national cross-sectional study (ICMR-INDIAB-17)," Lancet Diabetes Endocrinol, vol. 11, Sep. 2023, doi: 10.1016/S2213-8587(23)00119-5.

[4] A. Satija, E. Yu, W. Willett, and F. Hu, "Understanding Nutritional Epidemiology and Its Role in Policy," Adv Nutr, vol. 6, pp. 5–18, Sep. 2015, doi: 10.3945/an.114.007492.

[5] J. Huang et al., "Artificial Intelligence for Predicting and Diagnosing Complications of Diabetes," J Diabetes Sci Technol, vol. 17, no. 1, pp. 224–238, 2023, doi: 10.1177/19322968221124583.

[6] R. Shakhatreh and H. Shakhatreh, "Prediabetes: A High-Risk Condition for Developing Diabetes," Sep. 2024.

[7] N. A. Elsayed et al., "Prevention or Delay of Type 2 Diabetes and Associated Comorbidities: Standards of Care in Diabetes—2023," Diabetes Care, vol. 46, no. supp, pp. S41–S48, Jan. 2023, doi: 10.2337/dc23-S003.

[8] A. Esteva et al., "A guide to deep learning in healthcare," Nat Med, vol. 25, Sep. 2019, doi: 10.1038/s41591-018-0316-z.

[9] J. A. M. Sidey-Gibbons and C. J. Sidey-Gibbons, "Machine learning in medicine: a practical introduction," BMC Med Res Methodol, vol. 19, no. 1, p. 64, 2019, doi: 10.1186/s12874-019-0681-4.

[10] T. Gadekallu et al., "An Ensemble based Machine Learning model for Diabetic Retinopathy Classification," Sep. 2020, pp. 1–6. doi: 10.1109/ic-ETITE47903.2020.235.

[11] J. Chaki, S. Thillai Ganesh, S. K. Cidham, and S. Ananda Theertan, "Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review," Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 6, Part B, pp. 3204–3225, 2022, doi: https://doi.org/10.1016/j.jksuci.2020.06.013.

[12] M. Rajeswari and P. Prabhu, "A Review of Diabetic Prediction Using Machine Learning Techniques," Sep. 2019.

[13] M. Maniruzzaman, M. Rahman, and B. Ahammed, "Classification and prediction of diabetes disease using machine learning paradigm," Health Inf Sci Syst, vol. 8, pp. 1–14, Sep. 2020, doi: 10.1007/s13755-019-0095-z.

[14] M. F. Aslan and K. Sabanci, "A Novel Proposal for Deep Learning-Based Diabetes Prediction: Converting Clinical Data to Image Data," Diagnostics, vol. 13, p. 796, Sep. 2023, doi: 10.3390/diagnostics13040796.

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