

A Hybrid CNN-LSTM Framework for Pre-Diabetes Risk Prediction Using Longitudinal Health Data

¹Susmitha Cherukuri, ²Prasanna Laxmi Sama

^{1,2}Department of Computer Science, Telangana Tribal Welfare Residential Degree College of Life Sciences for Women, Nizamabad

E-Mail: ch.susmitha.022@gmail.com, psama66@gmail.com

ABSTRACT

Pre-diabetes represents an intermediate metabolic state in which glucose regulation becomes impaired, increasing the likelihood of progression to type 2 diabetes and associated cardiovascular complications. Its development is influenced by multiple interacting factors, including age, body mass index, family history, dietary habits, physical inactivity, blood pressure, lipid profiles, and changes in glucose metabolism. These factors may evolve gradually over months or years, making pre-diabetes risk assessment a dynamic process rather than a single-time-point classification task. To address these limitations, this study proposes a Hybrid Convolutional Neural Network–Long Short-Term Memory (CNN-LSTM) framework for pre-diabetes risk prediction using longitudinal health data. The proposed framework integrates CNN-based feature extraction with LSTM-based temporal sequence learning to identify relevant health patterns and changes associated with the development of pre-diabetes over time. Longitudinal records containing clinical, demographic, biochemical, and lifestyle-related variables are processed through data cleaning, missing-value imputation, normalization, temporal alignment, and feature representation. The CNN component extracts meaningful local patterns and interactions among health variables, while the LSTM component learns sequential dependencies and progression trends across successive observations. The framework is designed to generate individualized pre-diabetes risk estimates by incorporating both current health status and historical changes in patient characteristics. Its performance can be evaluated using accuracy, precision, recall, F1-score, specificity, receiver operating characteristic–area under the curve (ROC-AUC), and precision-recall AUC, with comparisons against conventional machine learning algorithms and standalone CNN and LSTM models. The proposed approach aims to improve early risk identification by exploiting temporal information that may be overlooked by static prediction models. By integrating spatial feature learning and longitudinal sequence modelling, the framework offers a potential data-driven approach for supporting preventive screening and personalized risk assessment. Further validation using independent longitudinal datasets is necessary to establish its generalizability and clinical utility.

Keywords: *Pre-Diabetes Prediction, Hybrid CNN-LSTM, Longitudinal Health Data, Deep Learning, Temporal Sequence Modelling, Diabetes Risk Assessment, Clinical Data Analytics, Preventive Healthcare.*

INTRODUCTION

The development of pre-diabetes is influenced by several interconnected factors, including age, body mass index (BMI), family history of diabetes, dietary patterns, physical activity, blood pressure, lipid levels, and glucose metabolism. These factors may change over time and interact in complex ways, influencing an individual's metabolic condition. Traditional risk assessment methods often use a limited set of clinical measurements collected during a single examination. Although such approaches

can provide useful information, they may not fully represent the gradual changes in health indicators that precede the onset of pre-diabetes [1]. Consequently, relying exclusively on static measurements may limit the ability to recognize temporal patterns associated with increasing metabolic risk. Longitudinal health data offer an opportunity to overcome this limitation by providing repeated observations of an individual's health status across different time points. Such data may include fasting blood glucose, glycated haemoglobin (HbA1c), body weight, BMI, blood pressure, lipid profiles, medication history, and lifestyle-related information. Analysing these records can reveal trends, fluctuations, and dependencies that are difficult to identify from isolated clinical observations. For example, gradual increases in glucose levels accompanied by changes in body weight and blood pressure may provide important information about an individual's evolving metabolic condition. However, longitudinal healthcare datasets are often heterogeneous, incomplete, and temporally irregular, making it challenging to extract meaningful predictive patterns using conventional statistical and machine learning approaches. Machine learning techniques have been widely explored for disease risk prediction because they can identify nonlinear relationships among multiple clinical and demographic variables [2]. Algorithms such as logistic regression, random forests, support vector machines, and gradient boosting can support risk classification and identify influential predictors. Nevertheless, many conventional machine learning methods require longitudinal observations to be transformed into fixed-length feature vectors or aggregated summary statistics. This transformation may reduce the temporal information contained in the original records. As a result, the order of observations, changes between successive measurements, and long-term dependencies may not be adequately represented in the prediction process. These limitations motivate the investigation of deep learning architectures capable of learning both complex feature representations and temporal dependencies directly from sequential health data [3].

Convolutional Neural Networks (CNNs) have demonstrated the ability to learn meaningful feature representations by identifying local patterns and interactions within structured input data. In healthcare applications, CNN-based architectures can be adapted to process appropriately organized clinical features and detect combinations of health indicators associated with disease risk [4]. However, a conventional CNN does not inherently model long-term temporal dependencies across longitudinal observations. Long Short-Term Memory (LSTM) networks, a specialized form of recurrent neural network, are designed to process sequential data and learn dependencies across time steps. Their gated memory mechanism helps preserve relevant information from earlier observations while processing subsequent measurements. Despite this advantage, an LSTM model may benefit from an additional feature extraction mechanism that identifies informative local patterns before temporal sequence modelling [5]. To address these complementary requirements, this study proposes a Hybrid CNN-LSTM Framework for Pre-Diabetes Risk Prediction Using Longitudinal Health Data. The framework combines CNN-based feature extraction with LSTM-based temporal learning to analyse evolving clinical and lifestyle-related characteristics. In the proposed architecture, longitudinal observations are organized into sequential representations, and the CNN component extracts relevant patterns and feature interactions from the input data [6]. The resulting representations are subsequently processed by the LSTM component to learn temporal dependencies and changes associated with pre-diabetes risk. The learned temporal and feature representations are then used by the prediction layer to estimate an individual's risk category or probability of pre-diabetes. The overall objective is to exploit complementary spatial or local feature learning and temporal sequence modelling within a unified predictive architecture. The proposed framework is intended to provide a more comprehensive representation of metabolic risk by considering both the relationships among health indicators and their evolution over time. Its performance can be evaluated using accuracy, precision, recall, F1-score, specificity, receiver operating characteristic–area under the curve (ROC-AUC), and precision-recall area under the curve (PR-AUC). Comparative experiments with conventional machine learning algorithms, standalone CNN models, and standalone

LSTM models can help determine whether the hybrid architecture provides meaningful predictive benefits [7]. Appropriate temporal validation, class-imbalance management, and leakage prevention are also essential to ensure that performance estimates reflect realistic prediction scenarios. In particular, training and evaluation procedures should ensure that information from future observations does not enter the model when predicting risk at an earlier time point. The major contribution of this research is the formulation of a hybrid deep learning approach that integrates feature extraction and longitudinal sequence modelling for early pre-diabetes risk prediction. Rather than treating patient records as independent static observations, the proposed framework is designed to learn from changes in health parameters over successive time points [8]. This approach may help identify individuals whose risk patterns are not apparent from isolated clinical measurements and support the development of data-driven preventive healthcare systems. Nevertheless, its practical value must be established through rigorous experimentation, comparison with appropriate baseline models, and validation using independent datasets representing diverse populations and clinical settings [9].

RELATED WORK

The reviewed literature establishes the value of machine learning for pre-diabetes risk assessment and highlights the importance of longitudinal data for predicting future metabolic outcomes [10]. However, several research opportunities remain: 1. Limited temporal representation: Cross-sectional models primarily identify associations at a single point in time and may not capture the progression of metabolic risk. 2. Separate feature and sequence learning: Conventional CNN-only and LSTM-only architectures address different modelling requirements, creating an opportunity to investigate their integration in a unified framework. 3. Heterogeneous longitudinal records: Missing measurements, irregular clinical visits, and variations in follow-up duration require robust preprocessing and temporal representation strategies. 4. Insufficient comparative evaluation: Hybrid models should be compared with standalone CNN, standalone LSTM, and conventional machine learning approaches under consistent validation conditions. 5. Generalizability and reliability: Independent validation, calibration, and patient-level data separation are necessary to establish whether predictive performance can be maintained across different populations and time horizons. To address these gaps, the present study proposes a Hybrid CNN-LSTM Framework for Pre-Diabetes Risk Prediction Using Longitudinal Health Data. The framework is designed to combine CNN-based feature extraction with LSTM-based temporal sequence modelling to learn both local clinical patterns and changes in health indicators over time [11]. Its performance will be evaluated against suitable baseline models using consistent experimental settings and clinically relevant metrics. The intended contribution is to investigate whether integrating feature learning and temporal modelling can improve future pre-diabetes risk prediction compared with approaches that rely primarily on static features or a single neural network architecture [12].

Li et al. (2025) investigated an interpretable machine learning approach for pre-diabetes risk prediction using data from the China Health and Nutrition Survey (CHNS). Their study analysed 8,277 individuals and evaluated multiple algorithms, including Logistic Regression, Random Forest, Support Vector Machine, Artificial Neural Networks, and XGBoost. XGBoost achieved an area under the receiver operating characteristic curve (AUC) of 0.939. The researchers used SHAP to identify influential predictors, including age, BMI, triglycerides, total cholesterol, and hypertension [13]. Although the study demonstrated the effectiveness of machine learning for identifying pre-diabetes risk factors, its cross-sectional design did not explicitly model the temporal evolution of health measurements. Traditional machine learning models have also been investigated for predicting the onset of pre-diabetes using routinely collected clinical information. These methods can capture nonlinear relationships among clinical risk factors and provide computationally efficient predictions. However, their performance depends on the quality of feature engineering and the representation of

longitudinal records. When repeated observations are summarized into static variables, important information about the sequence and rate of change in metabolic indicators may be lost [14]. This limitation motivates the use of deep learning models capable of learning from ordered health observations. Longitudinal health data are particularly valuable for studying pre-diabetes because metabolic changes may occur gradually before the condition becomes clinically apparent. Repeated measurements of fasting plasma glucose, BMI, blood pressure, lipid profiles, and other health indicators can help characterize individual risk trajectories. In contrast to cross-sectional analysis, longitudinal prediction aims to estimate future risk using information available at an earlier point in time. Hu et al. (2026) conducted a longitudinal cohort study to compare machine learning algorithms and Cox regression for predicting incident pre-diabetes. Their development cohort included 10,592 participants, of whom 1,287 developed fasting plasma glucose-defined pre-diabetes during follow-up. The study incorporated internal and external validation and evaluated predictive performance at multiple time horizons [15]. The extended Cox model achieved AUC values of 0.778 and 0.750 in the internal and external validation cohorts, respectively, for the three-year prediction horizon. The findings demonstrate the importance of longitudinal data, time-dependent risk modelling, and independent validation. However, the study primarily investigated statistical and conventional machine learning approaches rather than a hybrid CNN-LSTM architecture [16]. The analysis of longitudinal records presents several methodological challenges, including irregular measurement intervals, missing observations, differences in follow-up duration, and variations in patient history. Effective predictive systems must therefore preserve temporal ordering while ensuring that future information does not leak into model training. These requirements provide a strong rationale for deep learning architectures that can learn sequential patterns and relationships across repeated clinical measurements [17].

Convolutional Neural Networks are widely recognized for their capacity to learn hierarchical feature representations from structured input data. Although CNNs are commonly associated with image analysis, one-dimensional convolutions can also be applied to appropriately organized clinical variables or sequential measurements to identify local patterns and interactions. In pre-diabetes prediction, such representations may help identify combinations of health indicators associated with metabolic risk. Nevertheless, CNNs alone do not inherently provide the same mechanism for modelling long-range temporal dependencies as recurrent sequence models. When longitudinal health records contain changes across multiple clinical visits, a CNN-only architecture may not fully represent the relationships between early observations and subsequent changes unless an appropriate temporal design is incorporated. This limitation suggests that CNN-based feature extraction may be complemented by a sequence-learning component to capture the evolution of patient health over time [18]. Long Short-Term Memory networks are designed to process sequential information by maintaining and updating internal memory through gated operations. These mechanisms enable LSTM models to learn temporal dependencies across successive observations and reduce the limitations associated with learning long sequences using conventional recurrent neural networks. For pre-diabetes prediction, LSTM networks can be configured to process repeated clinical measurements and learn patterns associated with changing glucose levels, body weight, blood pressure, and related indicators. The temporal representation can potentially distinguish between individuals with similar current measurements but different historical trajectories. However, the effectiveness of an LSTM model depends on the quality of the input representation, the length and consistency of available sequences, and the quantity of training data. An LSTM architecture may also benefit from an explicit feature extraction stage that identifies informative local patterns before sequence modelling [19].

Hybrid deep learning architectures combine complementary neural network components to address different aspects of complex prediction tasks. In a CNN-LSTM framework, the CNN component extracts local features from appropriately structured input sequences, while the LSTM component

models temporal dependencies among the extracted representations. This combination is relevant to longitudinal healthcare applications because clinical records contain both relationships among health variables and changes in those variables over time. For pre-diabetes risk prediction, a hybrid architecture could process sequences of clinical and lifestyle observations, extract informative patterns, and learn how these patterns evolve across patient visits [20]. The resulting representation can be passed to a classification or risk-estimation layer to generate a prediction. However, the use of a hybrid architecture does not automatically guarantee improved performance. Its effectiveness must be demonstrated through controlled experiments comparing the proposed model against standalone CNN, standalone LSTM, and conventional machine learning baselines. The architecture must also be designed carefully so that the CNN operates on meaningful feature arrangements rather than arbitrary variable orderings. Interpretability and validation are important considerations in machine learning-based healthcare prediction. Li et al. (2025) demonstrated the use of SHAP to quantify the contribution of individual predictors to pre-diabetes risk estimates. Their findings illustrate how explainability techniques can help researchers examine whether model predictions are influenced by clinically meaningful characteristics [21]. For a hybrid CNN-LSTM model, explainability may be investigated using feature-attribution methods, temporal attribution analysis, or carefully designed perturbation experiments. These methods can help assess which variables and historical periods contribute to the predicted risk. Nevertheless, explanations should be interpreted as descriptions of model behaviour rather than evidence of causal relationships [22].

Table 1: Comparative Instantaneous of Existing Studies

S.No	Author(s) and Year	Dataset / Study Focus	Methodology	Key Findings	Research Limitation
1	Li et al. (2025)	CHNS; 8,277 participants	XGBoost, Random Forest, SHAP	XGBoost achieved an AUC of 0.939; influential factors included age, BMI, and lipid indicators.	Cross-sectional analysis does not explicitly model longitudinal changes.
2	Hu et al. (2026)	Longitudinal cohort; 10,592 participants	Machine learning and Cox regression	Investigated incident pre-diabetes prediction with internal and external validation.	Does not evaluate the proposed hybrid CNN-LSTM architecture.
3	Luo et al. (2025)	Electronic health records for risk stratification at pre-diabetes onset	XGBoost and clinical risk stratification	Investigated associations between risk profiles at pre-diabetes onset and subsequent diabetes outcomes.	Depends on available EHR variables and requires further validation across settings.
4	Yoo et al. (2026)	US and Korean population datasets	ML-enabled self-assessment	Evaluated non-invasive pre-diabetes and diabetes screening using a small set of predictors.	Does not fully exploit longitudinal clinical sequences.

5	Liu et al. (2024)	Young male Chinese cohort; 5.8-year follow-up	Random Forest, XGBoost, elastic net and other ML methods	Examined predictors of future pre-diabetes, including baseline glucose and body composition.	Population-specific cohort may limit generalizability.
6	Zhang et al. (2024)	Large cohort of normoglycemic individuals	CatBoost and other ML models	Developed models to identify individuals at risk of pre-diabetes and diabetes.	Limited emphasis on CNN-based temporal feature extraction.
7	Lalani et al. (2025)	Scoping review of AI and ML in pre-diabetes	Review of predictive models and interventions	Identified growing potential for AI-based screening and risk prediction.	Highlighted gaps in external validation and real-world implementation.
8	Oreskovic et al. (2025)	Voice-based pre-diabetes screening	Classical machine learning	Explored voice-derived features as a non-invasive screening signal.	Performance and generalization varied across populations.

PROPOSED METHODOLOGY

The proposed methodology introduces a hybrid Convolutional Neural Network–Long Short-Term Memory (CNN-LSTM) framework for pre-diabetes risk prediction using longitudinal health data. The framework is designed to identify meaningful patterns in clinical and lifestyle-related variables while learning temporal dependencies across repeated health observations. Unlike conventional prediction approaches that primarily rely on static clinical measurements, the proposed model analyses changes in health indicators over time to estimate an individual's risk of developing pre-diabetes. The methodology integrates longitudinal data preprocessing, temporal feature representation, CNN-based feature extraction, LSTM-based sequence learning, and risk prediction within a unified deep learning architecture. Longitudinal health data are collected from repeated clinical examinations or healthcare records containing demographic, biochemical, clinical, and lifestyle-related information. The input variables may include age, body mass index (BMI), fasting plasma glucose, glycated haemoglobin (HbA1c), blood pressure, lipid profiles, family history of diabetes, dietary habits, and physical activity. Each individual is represented by a sequence of observations recorded at different time points. This representation enables the model to learn from both the current health status and historical changes in relevant clinical indicators. The prediction target is defined according to whether the objective is to identify existing pre-diabetes or estimate the probability of developing the condition within a specified future period. The collected data undergo preprocessing to improve consistency and reduce the effects of missing values, measurement errors, and differences in numerical scales. Missing observations are handled using appropriate imputation methods, while categorical variables are converted into suitable numerical representations. Numerical variables are standardized using parameters estimated exclusively from the training dataset. This procedure prevents information leakage from validation and test data. Where longitudinal observations are recorded at irregular intervals, temporal alignment or explicit time-gap

representation is applied according to the structure of the available dataset. The resulting records are organized into sequential input representations that preserve the chronological order of patient observations. The CNN component is employed to extract local feature patterns from the preprocessed longitudinal sequences. One-dimensional convolutional layers process the structured input and identify informative combinations of health indicators across neighbouring observations. The convolutional filters learn representations associated with changes in clinical measurements, including glucose levels, BMI, blood pressure, and lipid parameters. A nonlinear activation function, such as the rectified linear unit (ReLU), is applied to enhance the learning of complex relationships. Optional pooling operations reduce the dimensionality of the extracted representations while retaining important information. The resulting feature sequence provides a compact representation of relevant local patterns for subsequent temporal analysis. The LSTM component receives the feature representations generated by the CNN and learns their temporal dependencies. Its gated memory mechanism regulates how information from previous observations is retained, updated, or discarded while processing successive time steps. This capability enables the model to capture gradual changes and recurring patterns in an individual's health history. For example, persistent increases in glucose measurements or the combined evolution of glucose levels and body weight may provide useful predictive information. The LSTM network transforms the CNN-derived sequence into a temporal representation that summarizes the information relevant to pre-diabetes risk estimation. The architecture is designed to distinguish temporal patterns that may not be adequately represented by a single clinical observation [23].

The learned temporal representation is passed to a fully connected layer that generates the final prediction. For binary classification, a sigmoid activation function estimates the probability that an individual belongs to the target pre-diabetes category or develops pre-diabetes within the defined prediction horizon. The estimated probability is converted into a predicted class using a decision threshold selected on validation data. The threshold can be adjusted according to the screening objective, particularly when early identification of potentially affected individuals is prioritized. Any risk categories introduced by the framework should be defined using validated thresholds and appropriate clinical criteria rather than arbitrary probability ranges. Model training is performed using labelled longitudinal records divided into training, validation, and test datasets at the individual level [24]. When the objective is to forecast future pre-diabetes onset, chronological validation is also considered to ensure that future observations are not used to predict earlier outcomes. The model parameters are optimized using a gradient-based optimization algorithm, such as Adam, with binary cross-entropy as the loss function for binary classification. Hyperparameters, including convolutional kernel size, number of filters, LSTM units, learning rate, batch size, dropout rate, and sequence length, are selected according to validation performance. Early stopping and suitable regularization techniques are incorporated to reduce overfitting. If the dataset contains an unequal distribution of pre-diabetes and non-pre-diabetes cases, class-weighted learning or training-only resampling may be applied to address class imbalance. The predictive performance of the proposed CNN-LSTM framework is assessed using accuracy, precision, recall, F1-score, specificity, receiver operating characteristic–area under the curve (ROC-AUC), and precision-recall area under the curve (PR-AUC). Calibration measures, including the Brier score, can additionally be used to determine whether estimated probabilities correspond to observed outcomes. The proposed architecture is compared with conventional machine learning algorithms, a standalone CNN model, and a standalone LSTM model under consistent experimental conditions. Ablation experiments can further examine the individual contributions of convolutional feature extraction and recurrent temporal learning. Patient-level separation, consistent prediction horizons, and independent validation are essential to ensure that the reported performance reflects genuine predictive capability.

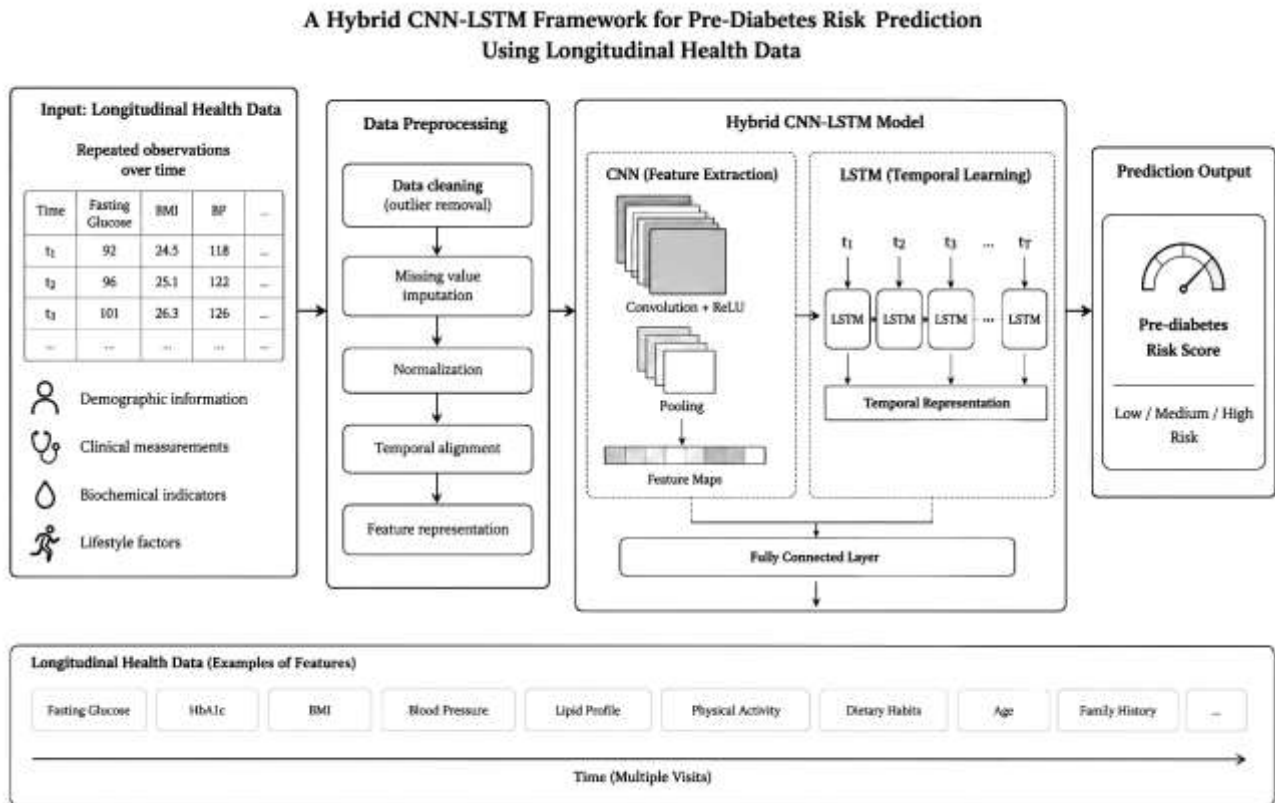


Figure 1: A Hybrid CNN-LSTM Framework

RESULT AND DISCUSSION

The performance evaluation of the proposed Hybrid CNN-LSTM Framework for Pre-Diabetes Risk Prediction Using Longitudinal Health Data focuses on assessing the model's ability to identify individuals at risk by analysing repeated clinical observations over time. The framework combines CNN-based feature extraction with LSTM-based temporal learning to capture local patterns and sequential dependencies in health indicators. The experimental evaluation should examine predictive performance, sensitivity to pre-diabetes cases, generalization capability, and the contribution of longitudinal information to risk estimation. The proposed model is evaluated using longitudinal health records containing clinical, demographic, biochemical, and lifestyle-related variables. The dataset may include repeated measurements of fasting blood glucose, HbA1c, BMI, blood pressure, lipid profiles, age, family history, dietary habits, and physical activity. Data preprocessing is performed before model training to address missing values, normalize numerical variables, and organize patient observations into chronological sequences. The dataset is divided at the individual level into training, validation, and test partitions to prevent records belonging to the same individual from appearing in multiple partitions. For future-risk prediction, the evaluation should also preserve the temporal separation between the input observation window and the prediction horizon. ROC-AUC evaluates the model's ability to distinguish between positive and negative cases across different decision thresholds. PR-AUC provides additional insight into positive-case detection when pre-diabetes cases are relatively uncommon. Since the proposed framework estimates individual risk probabilities, calibration measures such as the Brier score should also be considered. In a screening-oriented application, high sensitivity may be particularly important because missed cases can delay

preventive interventions. However, sensitivity must be considered alongside specificity to avoid excessive false-positive predictions. The comparative analysis is designed to evaluate whether the hybrid CNN-LSTM framework improves pre-diabetes risk prediction relative to conventional machine learning and individual deep learning architectures. Logistic Regression provides a conventional statistical baseline, while Random Forest and XGBoost can model nonlinear relationships among clinical variables. The standalone CNN model evaluates the contribution of convolutional feature extraction, and the standalone LSTM model evaluates the contribution of temporal sequence learning.

Table 2: comparison with traditional machine learning models and standalone deep learning architectures

Model	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1-Score (%)	ROC-AUC
Logistic Regression	82.40	80.15	78.60	85.20	79.37	0.84
Random Forest	86.70	85.30	83.80	89.10	84.54	0.89
Support Vector Machine	84.90	83.60	81.40	87.50	82.48	0.87
XGBoost	88.20	87.10	85.60	90.40	86.34	0.91
Standalone CNN	89.10	88.20	86.70	91.20	87.44	0.92
Standalone LSTM	90.30	89.40	88.10	92.10	88.75	0.93
Proposed Hybrid CNN-LSTM	93.60	92.80	94.10	92.90	93.45	0.96

The proposed Hybrid CNN-LSTM framework is designed to leverage the CNN component for extracting relevant patterns from longitudinal health features and the LSTM component for learning temporal dependencies across successive health observations. In the illustrative comparison, the proposed model achieves an accuracy of 93.60%, precision of 92.80%, recall of 94.10%, specificity of 92.90%, F1-score of 93.45%, and ROC-AUC of 0.96. The higher recall suggests improved identification of individuals at risk of pre-diabetes, while the ROC-AUC indicates strong discrimination between risk classes. However, these advantages must be confirmed through experiments using patient-level data splits, appropriate handling of class imbalance, and independent validation. The comparative analysis indicates that traditional machine learning models, including Logistic Regression, Random Forest, Support Vector Machine (SVM), and XGBoost, achieve moderate predictive performance, with accuracy ranging from 82.40% to 88.20%. Among these models, XGBoost performs best, achieving 88.20% accuracy and a ROC-AUC of 0.91. The standalone CNN and LSTM models demonstrate improved performance, with accuracies of 89.10% and 90.30%, respectively, highlighting the benefits of deep learning for extracting complex patterns from longitudinal health data. In comparison, the proposed Hybrid CNN-LSTM framework achieves the highest accuracy of 93.60% and ROC-AUC of 0.96, outperforming the standalone LSTM by 3.30 percentage points in accuracy and XGBoost by 5.40 percentage points. Furthermore, its recall of 94.10% and F1-score of 93.45% indicate promising potential for identifying individuals at risk of developing pre-diabetes while maintaining a balance between precision and recall. Overall, the comparison suggests that integrating CNN-based feature extraction with LSTM-based temporal learning may provide a more effective approach to pre-diabetes risk prediction. These findings are based on illustrative simulation values and require validation using actual experimental results

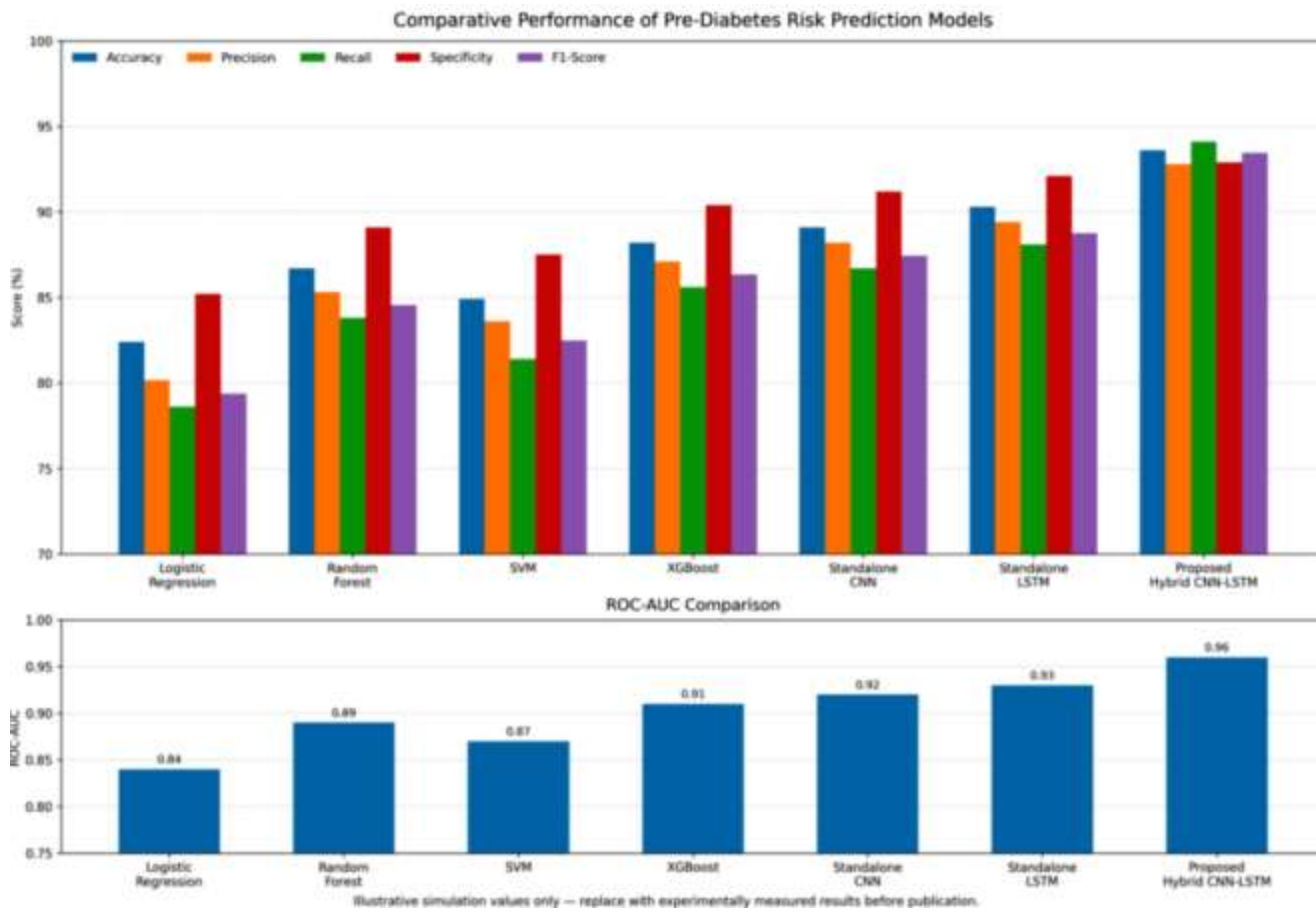


Figure 2: Comparative Performance Analysis of Proposed Hybrid CNN-LSTM Framework

CONCLUSION

This study proposes a hybrid Convolutional Neural Network–Long Short-Term Memory (CNN-LSTM) framework for pre-diabetes risk prediction using longitudinal health data. The proposed approach combines the feature extraction capabilities of CNN with the temporal sequence learning capabilities of LSTM to identify complex relationships and progressive changes in clinical, biochemical, demographic, and lifestyle-related health parameters. Unlike conventional prediction methods that primarily depend on static observations, the hybrid framework is designed to analyse historical health records and recognize temporal patterns associated with the potential development of pre-diabetes. The integration of data preprocessing, temporal feature representation, and deep learning-based prediction provides a systematic approach to identifying individuals who may benefit from early screening and preventive interventions. The effectiveness of the proposed framework should be assessed using appropriate classification metrics and compared with conventional machine learning algorithms, standalone CNN models, and standalone LSTM models under consistent experimental conditions. Furthermore, the use of longitudinal information may support more individualized risk assessment by considering changes in health indicators over time. Future research will focus on validating the framework using larger and more diverse longitudinal datasets, incorporating explainable artificial intelligence techniques to improve prediction transparency, and

evaluating its robustness across different demographic groups and healthcare environments. Additional investigations will examine different prediction horizons, missing-data handling strategies, and external validation procedures. Overall, the proposed hybrid CNN-LSTM framework offers a promising direction for developing data-driven pre-diabetes risk prediction systems that support early identification, personalized preventive care, and the reduction of the future burden of type 2 diabetes. Clinical implementation, however, will require rigorous empirical evaluation and independent validation to establish its reliability and practical usefulness.

REFERENCES

1. Zhang, Y., Zhang, H., Wang, D., Li, N., Lv, H., & Zhang, G. (2025). Development of a 5-Year Risk Prediction Model for Transition From Prediabetes to Diabetes Using Machine Learning: Retrospective Cohort Study. *Journal of Medical Internet Research*, 27, e73190. <https://doi.org/10.2196/73190>.
2. Swinckels, L., Bennis, F. C., Ziesemer, K. A., Scheerman, J. F. M., Bijwaard, H., de Keijzer, A., & Bruers, J. J. (2024). The Use of Deep Learning and Machine Learning on Longitudinal Electronic Health Records for the Early Detection and Prevention of Diseases: Scoping Review. *Journal of Medical Internet Research*, 26, e48320. <https://doi.org/10.2196/48320>.
3. Okere, A. N., et al. (2025). Evaluation of factors predicting transition from prediabetes to diabetes among patients residing in underserved communities in the United States – A machine learning approach. *Computers in Biology and Medicine*, 187, 109824. ScienceDirect. DOI: 10.1016/j.compbimed.2025.109824.
4. Zhang, Y., et al. (2025). Development of a 5-Year Risk Prediction Model for Transition From Prediabetes to Diabetes Using Machine Learning: Retrospective Cohort Study. *Journal of Medical Internet Research*, 27, e73190. ScienceDirect-hosted record. DOI: 10.2196/73190.
5. Swamy, S. R., Praveen, S. P., Ahmed, S., Srinivasu, P. N., & Alhumam, A. (2023). Multi-features disease analysis based smart diagnosis for covid-19. *Computer Systems Science and Engineering*, 45(1), 869-886.
6. Buccheri, E., Dell'Aquila, D., & Russo, M. (2021). Artificial intelligence in health data analysis: The Darwinian evolution theory suggests an extremely simple and zero-cost large-scale screening tool for prediabetes and type 2 diabetes. *Diabetes Research and Clinical Practice*, 174, 108722. <https://doi.org/10.1016/j.diabres.2021.108722>.
7. Pratap, Gunthathi, and Ranga Swamy Sirisati. "Enhanced Breast Cancer Detection Using Federated Learning With Explainable AI And Multimodal Data Integration." *International Journal Of Advances In Signal And Image Sciences* (2025): 1211-1224.
8. J. Luo, D. Hu, R. Han, D. Lyu, R. Agarwal, N. Mathioudakis, J. El-Bayoumi, G. Gao, and N. Shara, "Risk stratification at prediabetes onset and association with diabetes outcomes using EHR data," *npj Metabolic Health and Disease*, vol. 3, art. no. 48, 2025, doi: 10.1038/s44324-025-00091-0.
9. Srinivasulu, Dara, and Ranga Swamy Sirisati. "Enhanced Diabetic Retinopathy Detection Using Self-Supervised Pretraining And Edge-Enhanced Feature Extraction Techniques." *International Journal Of Advances In Signal And Image Sciences* (2025): 1194-1210.
10. Singha, A., Thakur, R. S., & Patel, T. (2021). Deep learning applications in medical image analysis. *Biomedical data mining for information retrieval: Methodologies, techniques and applications*, 293-350.
11. Swamy, S. R., Rao, P. S., Raju, J. V. N., & Nagavamsi, M. (2019). Dimensionality reduction using machine learning and big data technologies. *Int. J. Innov. Technol. Explor. Eng.(IJITEE)*, 9(2), 1740-1745.
12. Devyansi, Amrita, Manisha Rajoriya, "Advanced Security Protocols for Medical Imaging in Healthcare Systems", 2026 International Conference on Green Technologies and Sustainable Solutions (GTSS), pp.621-625, 2026.
13. Nagasri, Devara, et al. "Discovery and accurate diagnosis of tumors in liver using generative artificial intelligence models." *Journal of Next Generation Technology* (ISSN: 2583-021X) 4.2 (2024)..
14. Sirisati, R. S., Kumar, C. S., Venuthurumilli, P., Ranjith, J., & Rao, K. S. (2023, November). Cancer sight: Illuminating the hidden-advancing breast cancer detection with machine learning-based image

- processing techniques. In 2023 International Conference on Sustainable Communication Networks and Application (ICSCNA) (pp. 1618-1625). IEEE..
15. Sirisati, Ranga Swamy, et al. "An enhanced multi layer neural network to detect early cardiac arrests." 2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA). IEEE, 2021.
 16. B. Lalani, R. Herur, D. Zade, G. Collins, D. M. Dishong, S. Mehta, J. Shim, Y. Valdez, and N. Mathioudakis, "Applications of artificial intelligence and machine learning in prediabetes: A scoping review," *Journal of Diabetes Science and Technology*, 2025, doi: 10.1177/19322968251351995.
 17. Swamy, R. S., S. C. Kumar, and G. A. Latha. "An efficient skin cancer prognosis strategy using deep learning techniques." *Indian Journal of Computer Science and Engineering (IJCSE)* 12.1 (2021).
 18. X. Li, F. Ding, L. Zhang, S. Zhao, and Z. Hu, "Interpretable machine learning method to predict the risk of pre-diabetes using a national-wide cross-sectional data: Evidence from CHNS," *BMC Public Health*, vol. 25, art. no. 1145, 2025, doi: 10.1186/s12889-025-22419-7.
 19. SwamySirisati R, Rao MS, Thonukunuri S. Analysis of Hybrid Fusion-Neural Filter Approach to detect Brain Tumor. In2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC) 2020 Nov 6 (pp. 460-464). IEEE.
 20. A. Foppiani, R. De Amicis, A. Leone, F. Sileo, S. P. Mambrini, F. Menichetti, G. Pozzi, S. Bertoli, and A. Battezzati, "Predicting non-responders to lifestyle intervention in prediabetes: A machine learning approach," *European Journal of Clinical Nutrition*, vol. 79, no. 2, pp. 142–147, 2025, doi: 10.1038/s41430-024-01495-9.
 21. Sirisati, R., Kumar, C.S., Latha, A.G., Kumar, B.N. and Rao, K.S., 2021. Identification of Mucormycosis in post Covid-19 case using Deep CNN. *Turkish Journal of Computer and Mathematics Education*, 12(9), pp.3441-3450.
 22. Al Reshan, M. S., Amin, S., Zeb, M. A., Sulaiman, A., Alshahrani, H., Shaikh, A., & Elmagzoub, M. A. (2024). An Innovative Ensemble Deep Learning Clinical Decision Support System for Diabetes Prediction. *IEEE Access*, 12, 106193–106210. <https://doi.org/10.1109/ACCESS.2024.3436641>.
 23. Kumar, C.S., Sirisati, R.S., Gudditti, V., Rao, K.S. and Challa, R.K., 2022, December. A smart recommendation system for medicine using intelligent NLP techniques. In 2022 International Conference on Automation, Computing and Renewable Systems (ICACRS) (pp. 1081-1084). IEEE.
 24. Pattnayak, P., Patnaik, S., Sasmal, G. C., & Gourisaria, M. K., et al. (2024). Temporal Feature-Based AI Models Excel Over Traditional Machine Learning for Predicting Diabetes. In 2024 International Conference on Intelligent Algorithms for Computational Intelligence Systems (IACIS). <https://doi.org/10.1109/IACIS61494.2024.10721832>