International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 - 6435 (Online) Volume 14, Issue 6, June 2025)

CNN-Based Approach for Classification of Diabetes Using Clinical Data

Gaurav Nayak[1], Divyarth Rai[2]

[1] *Research Scholar SOCST LNCT University, Bhopal,MP*
[2] *Professor Department of Computer Science and Engineering LNCT University Bhopal, India*

Abstract:Diabetes mellitus is a chronic metabolic disorder that poses a significant burden on healthcare systems worldwide. The early diagnosis of diabetes is critical for managing the disease and preventing its complications. With the growing volume of electronic health records (EHRs) and structured clinical datasets, machine learning (ML) and deep learning (DL) methods have emerged as powerful tools for automating disease detection and classification. Convolutional Neural Networks (CNNs), a subclass of deep learning models initially developed for image recognition, have shown great promise in handling structured clinical data for medical diagnosis tasks, including diabetes classification. This survey explores the adaptation of CNN models for classifying diabetes using clinical data. It covers the fundamentals of CNN architecture, reviews recent research contributions, identifies the current challenges, and suggests potential future research directions. The review also examines commonly used datasets, data preprocessing techniques, model evaluation metrics, and performance comparisons with traditional ML methods.

1. Introduction

Diabetes affects over 400 million people globally and remains one of the leading causes of death and disability. The chronic nature of the disease, combined with its silent progression in early stages, highlights the need for accurate and timely diagnostic methods. Traditionally, diabetes diagnosis relies on laboratory tests such as fasting glucose levels, oral glucose tolerance tests, and glycated hemoglobin (HbA1c) values. However, these methods often require multiple visits, may be prone to human error, and are not always available in low-resource settings.

Artificial Intelligence (AI), particularly deep learning, has emerged as a promising approach to support medical diagnosis by automatically learning complex patterns from large datasets. Convolutional Neural Networks (CNNs) have demonstrated significant success in image processing tasks, but recent developments have shown their adaptability to structured, one-dimensional clinical data. This survey presents an in-depth review of CNN-based approaches for diabetes classification using clinical data. The paper aims to provide a detailed understanding of how CNNs can be applied beyond image data and explores their effectiveness compared to other ML techniques.

2. Literature Survey

The application of deep learning in healthcare has rapidly expanded, with CNNs being used not only for medical image analysis but also for tabular and time-series data classification. Several studies have explored the potential of CNNs in diagnosing diabetes by utilizing clinical datasets such as the PIMA Indian Diabetes dataset and real-world electronic health records.

In early work by Alghamdi et al. (2021), a 1D CNN was applied to the PIMA dataset, achieving an accuracy of over 85%, outperforming traditional machine learning models like Logistic Regression and Support Vector Machines. Their model demonstrated that CNNs could effectively capture non-linear relationships among clinical features without manual feature engineering.

Nishat et al. (2022) proposed a hybrid CNN model that integrated data augmentation techniques to address the issue of class imbalance in diabetes datasets. Their model achieved over 90% accuracy and emphasized the

importance of data preprocessing and architecture tuning in improving classification performance.

Liu et al. (2023) introduced a CNN-LSTM hybrid model that combined the spatial learning capability of CNNs with the temporal modeling power of LSTMs to analyze longitudinal patient data. The hybrid approach showed superior performance in predicting the onset of diabetes by capturing both static and dynamic patient information.

Another notable work is by Kaur and Singh (2020), who utilized a deep CNN model on electronic health records and emphasized the importance of using dropout layers and batch normalization to prevent overfitting. Their model achieved robust results across multiple datasets and demonstrated generalizability.

Comparative studies consistently report that CNN-based models outperform traditional classifiers when sufficient data and proper preprocessing techniques are employed. These findings suggest that CNNs are not only applicable to image data but also highly effective for structured tabular data used in clinical diagnosis.

3. CNN Architecture for Clinical Data

CNNs are designed to extract local and hierarchical features through convolutional operations, pooling, and fully connected layers. For clinical data, which is typically one-dimensional and structured in tabular form, CNNs are adapted into 1D architectures.

The basic components of a 1D CNN model for diabetes classification include:

- **Input Layer**: Receives the feature vector, which consists of patient attributes such as age, BMI, glucose level, insulin, blood pressure, and others.
- **Convolutional Layers**: Apply 1D filters that slide over the feature vector to extract local patterns.
- **Activation Functions**: Typically ReLU (Rectified Linear Unit) is used to introduce non-linearity.
- **Pooling Layers**: Perform down-sampling to reduce dimensionality and computation while retaining important features.
- **Fully Connected Layers**: Combine extracted features and perform the final classification using a sigmoid or softmax activation function.

Regularization techniques such as dropout and batch normalization are commonly employed to prevent overfitting. The architecture can be deepened by adding more convolutional and pooling layers, but careful tuning is required to balance performance and computational efficiency.

4. Datasets and Preprocessing

Effective training of CNN models requires high-quality and well-preprocessed datasets. The most commonly used dataset for diabetes classification is the PIMA Indian Diabetes dataset, which includes 768 records with 8 clinical features. Other sources include the UCI Diabetes dataset, MIMIC-III database, and hospital-specific electronic health records.

Preprocessing steps are critical for model performance and include:

- **Normalization**: Scaling features to a uniform range (e.g., 0 to 1) to stabilize learning.
- **Handling Missing Values**: Using imputation methods such as mean substitution or K-nearest neighbors.
- **Data Balancing**: Addressing class imbalance through oversampling techniques like SMOTE or undersampling majority classes.
- **Feature Selection**: Removing irrelevant or redundant features based on correlation analysis or statistical tests.

Preprocessing not only improves the quality of data but also ensures that the CNN can efficiently learn useful representations without being biased or misled by noisy inputs.

5. Evaluation Metrics

Performance of CNN-based classification models is assessed using standard evaluation metrics, including:

- **Accuracy**: Overall correctness of predictions.
- **Precision**: Proportion of true positive predictions among all positive predictions.
- **Recall (Sensitivity)**: Proportion of actual positives correctly identified.
- **F1-Score**: Harmonic mean of precision and recall, useful in imbalanced datasets.
- **AUC-ROC**: Measures the trade-off between sensitivity and specificity.

These metrics provide a comprehensive understanding of the model's diagnostic performance and guide the optimization process.

6. Problem Definition

The core problem addressed in this survey is the accurate classification of diabetes using structured clinical data through CNN-based deep learning models. The traditional ML models often require extensive feature engineering and may fail to capture complex, non-linear patterns in the data. CNNs offer an alternative approach by automatically learning representations and identifying intricate patterns that correlate with diabetes onset or presence.

The challenges associated with this problem include limited availability of large-scale labeled datasets, data quality issues such as missing or inconsistent values, class imbalance, and the need for model interpretability. The goal is to design a CNN-based framework that can efficiently process clinical data and provide reliable, explainable predictions to support medical professionals in diagnosing diabetes.

7. Challenges and Limitations

Despite their advantages, CNN-based models for clinical data face several challenges:

- **Data Scarcity**: Most publicly available datasets are small, which can lead to overfitting.
- **Class Imbalance**: Diabetes datasets often have more non-diabetic than diabetic samples.
- **Interpretability**: CNNs are often considered black boxes, making it difficult to explain predictions.
- **Generalizability**: Models trained on one dataset may not perform well on data from different populations or healthcare systems.
- **Computational Resources**: Training deep CNNs can be computationally intensive and may require specialized hardware.

Addressing these challenges requires innovative approaches, such as transfer learning, explainable AI, and federated learning to build robust and scalable diagnostic systems.

8. Future Directions

Future research should focus on improving model interpretability through tools like SHAP and Grad-CAM, which help visualize the contribution of each input feature. Federated learning can enable collaborative model training across institutions without sharing sensitive patient data, enhancing model generalization and privacy.

Integration of real-time sensor data from wearable devices with clinical records can enable continuous monitoring and early warning systems. Lightweight CNN models that can run on mobile or embedded devices will facilitate deployment in resource-limited settings. Lastly, the development of multimodal deep learning models that combine structured data, images, and textual information from EHRs could further improve prediction accuracy.

9. Conclusion

CNNs have demonstrated substantial promise in the classification of diabetes using structured clinical data.

Their ability to automatically learn complex feature representations makes them superior to traditional machine learning models in many scenarios. However, realizing their full potential requires addressing challenges related to data quality, model interpretability, and generalizability. Through continued research and technological innovation, CNN-based diagnostic tools can become integral components of intelligent healthcare systems, offering early detection and improved disease management for diabetes patients worldwide.

This survey consolidates existing knowledge and sets the stage for future exploration into CNN-based medical diagnosis systems, particularly in the context of chronic diseases such as diabetes. As healthcare moves toward data-driven and personalized medicine, CNNs are poised to play a critical role in shaping the future of diagnostic decision support systems.

References

[1] Warke M, Kumar V, Tarale S, Galgat P, Chaudhari D. Diabetes diagnosis using machine learning algorithms. International Research Journal of Engineering and Technology. 2019; 6(3): 1470-6.

[2] Kavakiotisab I, Tsave O, Salifoglou A, Maglaveras N, Vlahavasa I, Chouvarda I. Machine learning and data mining methods in diabetes research. Computational and Structural Biotechnology Journal. 2017; 15: 104- 16.

[3] Benbelkacem S, Atmani B. Random forests for diabetes diagnosis. International Conference on Computer and Information Sciences. IEEE; 2019.

[4] Sun YL, Zhang DL. Machine learning techniques for screening and diagnosis of diabetes: A survey. Tehnic˘ki Vjesnik. 2019; 26(3): 872- 80.

[5] Maniruzzaman M, Rahman MJ, Ahammed B, Abedin MM. Classification and prediction of diabetes disease using machine learning paradigm. Health Inf Sci Syst. 2020; 8(1): 7. PMID: 31949894 DOI: 10.1007/s13755-019-0095-z [PubMed]

[6] Chaki, J., Ganesh, S. T., Cidham, S. K., & Theertan, S. A. (2020). Machine learning and artificial intelligence-based diabetes mellitus detection and self-management: a systematic review. Journal of King Saud University-Computer and Information Sciences.

[7] Anuar, N. N., Hafifah, H., Zubir, S. M., Noraidatulakma, A., Rosmina, J., Ain, M. N.,& Rahman, A. (2020). Cardiovascular disease prediction from electrocardiogram by using machine learning.

[8] Swapna, G., Vinayakumar, R., & Soman, K. P. (2018). Diabetes detection using deep learning algorithms. ICT express, 4(4), 243-246.

[9] Saeedi, P., Petersohn, I., Salpea, P., Malanda, B., Karuranga, S., Unwin, N., & IDF Diabetes Atlas Committee. (2019). Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas. Diabetes research and clinical practice, 157, 107843.

[10] Shinde, G. R., Kalamkar, A. B., Mahalle, P. N., Dey, N., Chaki, J., & Hassanien, A. E. (2020). Forecasting models for coronavirus disease (COVID-19): a survey of the state-of-the-art. SN Computer Science, 1(4), 1-15.