Institut Riset dan Publikasi Indonesia (IRPI)
MALCOM: Indonesian Journal of Machine Learning and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/malcom
Vol. 4 Iss. 3 July 2024, pp: 1072-1076
ISSN(P): 2797-2313 | ISSN(E): 2775-8575

Comparative Analysis of Neural Network Architectures for Predicting Chronic Disease Indicators Using CDC's Chronic Disease Indicators Dataset

Gregorius Airlangga

Information System Study Program, Atma Jaya Catholic University of Indonesia, Indonesia

E-Mail: gregorius.airlangga@atmajaya.ac.id

Received May 23th 2024; Revised Jun 24th 2024; Accepted Jul 5th 2024
Corresponding Author: Gregorius Airlangga

Abstract

This research evaluates the performance of three machine learning models—Neural Network (NN), Convolutional Neural Network (CNN), and Recurrent Neural Network (RNN) using Long Short-Term Memory (LSTM) units—in predicting chronic disease indicators using the CDC's Chronic Disease Indicators (CDI) dataset. The study employs a comprehensive preprocessing pipeline and 5-fold cross-validation to ensure robustness and generalizability of the results. The CNN model outperformed both the NN and RNN models across all key performance metrics, achieving an accuracy of 0.6303, precision of 0.6445, recall of 0.6303, and F1 score of 0.5950. The superior performance of the CNN is attributed to its ability to capture spatial hierarchies and interactions within the structured dataset. The findings underscore the importance of selecting appropriate machine learning architectures based on the data characteristics. This research provides valuable insights for public health officials and policymakers to enhance chronic disease monitoring, early detection, and intervention strategies. Future work will explore hybrid models and advanced techniques to further improve predictive performance. This study highlights the potential of CNNs in public health informatics and sets a foundation for further research in this domain.

Keyword: Chronic Disease Indicators, Convolutional Neural Network, Neural Network, Public Health Informatics, Recurrent Neural Network

1. INTRODUCTION

Chronic diseases, including heart disease, cancer, and diabetes, are the leading causes of death and disability globally, posing significant public health challenges. Addressing these challenges requires robust data to inform public health practices and policies. The Centers for Disease Control and Prevention (CDC) has developed a comprehensive dataset, the Chronic Disease Indicators (CDI), to provide a cross-cutting set of 124 indicators for states, territories, and large metropolitan areas [1], [2]. This dataset allows uniform collection and reporting of chronic disease data, offering a valuable resource for public health research and intervention planning. Over the past 15 years, various health-related questions have been assessed across the United States, providing a rich dataset with confidence intervals and demographic stratifications [3]–[5]. The availability of such detailed data presents an opportunity to leverage advanced machine learning techniques to analyze and predict chronic disease trends and outcomes [6]–[8].

The application of machine learning in healthcare has gained considerable attention in recent years. Studies have demonstrated the potential of machine learning models to improve disease diagnosis, predict patient outcomes, and enhance healthcare delivery. Traditional machine learning algorithms, such as logistic regression, decision trees, and support vector machines, have been widely used for predicting chronic diseases [9]–[11]. However, these models often struggle with high-dimensional and complex datasets [12]–[14]. Deep learning, a subset of machine learning, has shown promising results in various healthcare applications. CNNs, originally designed for image recognition, have been adapted for analyzing time-series and structured data, providing significant improvements in performance [15]–[17]. RNNs, particularly Long Short-Term Memory (LSTM) networks, are well-suited for sequential data analysis, making them ideal for capturing temporal patterns in health data [18]. Recent research has explored the integration of CNN and RNN architectures to leverage the strengths of both models, demonstrating improved predictive accuracy and robustness [19]. Chronic diseases continue to burden healthcare systems worldwide, necessitating effective strategies for early detection, prevention, and management. The COVID-19 pandemic has further exacerbated the impact of

chronic diseases, highlighting the need for timely and accurate public health data [20]. Leveraging advanced machine learning techniques to analyze chronic disease indicators can provide valuable insights for policymakers, healthcare providers, and researchers, ultimately contributing to better health outcomes and resource allocation [21].

Current state-of-the-art approaches in chronic disease prediction predominantly utilize deep learning models. CNNs have been successfully applied to classify medical images and detect patterns in physiological signals. RNNs, especially LSTMs, have been effective in modeling temporal dependencies in health data, such as patient monitoring and disease progression [22]. The primary goal of this research is to develop and evaluate machine learning models for predicting chronic disease indicators using the CDI dataset. Specifically, we aim to preprocess and standardize the CDI dataset to ensure high-quality input data for model training, omplement and compare the performance of Deep NN, CNNs, and LSTM models, furthermore we identify the most effective model in terms of predictive accuracy, precision, recall, and F1 score. Lastly, we provide practical recommendations for applying these models in public health practice. The remaining sections of this research article are structured as follows. The Methods section details the data preprocessing steps, model architectures, and evaluation metrics used in this study. The Results section presents the performance comparison of the different models, highlighting the advantages of the each deep learning variant approach. The Discussion section interprets the findings, discusses the implications for public health practice, and identifies potential limitations of the study. Finally, the Conclusion section summarizes the key contributions, outlines future research directions, and emphasizes the significance of integrating advanced machine learning techniques in public health informatics.

2. MATERIALS AND METHOD

This section outlines the systematic procedures employed in this research to develop and evaluate machine learning models for predicting chronic disease indicators using the Chronic Disease Indicators (CDI) dataset. The methodology includes data collection, preprocessing, feature selection, model selection, model training and evaluation, hyperparameter tuning, and performance metrics. The dataset used in this research is the CDC's Chronic Disease Indicators (CDI) dataset, which provides a comprehensive set of 124 indicators for states, territories, and large metropolitan areas in the United States. The dataset spans over 15 years and includes various health-related questions assessed at different times and locations, providing a valuable resource for public health research, dataset can be downloaded from [23]. To preprocess the data, the dataset was loaded into a Pandas DataFrame for manipulation and analysis. Missing values in the 'DataValue' column were filled with the median value to maintain data consistency. Columns with all missing values were dropped from the dataset. Categorical variables were encoded using Label Encoding to convert them into numerical values suitable for machine learning models. Features and target variables were defined, with the target variable 'DataValueTypeID' selected, and other irrelevant columns dropped. The feature set was standardized to ensure that each feature contributes equally to the model's performance. The target variable was encoded using one-hot encoding to prepare it for model training.

Three machine learning models were selected for this research: Neural Network (NN), Convolutional Neural Network (CNN), and Recurrent Neural Network (RNN). Each model was chosen for its unique strengths in handling complex datasets. A traditional neural network with fully connected layers was implemented as presented in equation 1. The architecture consisted of an input layer, hidden layers with ReLU activation, dropout layers for regularization, and an output layer with softmax activation.

$$\text{NN model architecture:} \quad h_i = \sigma(W_i h_{i-1} + b_i) \tag{1}$$

CNN was designed to capture spatial patterns in the data. The architecture included convolutional layers, max-pooling layers, and fully connected layers as presented in equation 2.

$$\text{CNN model architecture:} \quad h_{i,j} = \sigma\left(\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} W_{i,j,m,n} h_{i-m,j-n} + b_{i,j}\right) \tag{2}$$

An RNN, specifically an LSTM network, was implemented to capture temporal dependencies in the data. The architecture included LSTM layers, dropout layers for regularization, and fully connected layers as presented in equation 3.

$$\text{LSTM model architecture:} \quad h_t = \sigma\left(W_f h_{t-1} + U_f x_t + b_f\right) \tag{3}$$

The models were trained and evaluated using 5-fold cross-validation with the StratifiedKFold method to ensure that each fold had a representative distribution of classes. Cross-validation was used to assess the

generalizability of the models. Each model was trained and evaluated using cross-validation, with performance metrics such as accuracy, precision, recall, and F1 score computed for each fold. Performance metrics were calculated as presented in the equation 4-7.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{4}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

Where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively. Hyperparameter tuning was performed to optimize the performance of the models. Grid search and random search techniques were used to identify the best combination of hyperparameters, such as the number of layers, number of neurons per layer, learning rate, batch size, and dropout rate. The optimal hyperparameters were selected based on the model's performance on the validation set.

Each model was evaluated using the cross-validation method, and their performance metrics were compared. Input data for CNN and RNN models was reshaped to match the expected input shapes. The performance of the models was evaluated by computing the mean and standard deviation of each metric to assess the models' generalizability and robustness. This research methodology outlines a comprehensive approach to developing and evaluating machine learning models for predicting chronic disease indicators using the CDI dataset. By systematically preprocessing the data, implementing, and comparing different model architectures, and rigorously evaluating their performance, this study aims to identify the most effective model for analyzing chronic disease data and providing actionable insights for public health practice.

3. RESULTS AND DISCUSSION

The performance of three different machine learning models—Neural Network (NN), Convolutional Neural Network (CNN), and Recurrent Neural Network (RNN) using Long Short-Term Memory (LSTM) units—was evaluated for predicting chronic disease indicators using the Chronic Disease Indicators (CDI) dataset as presented in the table 1. The models were assessed based on accuracy, precision, recall, and F1 score metrics, and the results were obtained using 5-fold cross-validation to ensure robustness and generalizability. The Neural Network (NN) achieved an accuracy of 0.6180 with a standard deviation of 0.0035. The precision was 0.6288 with a standard deviation of 0.0370, recall matched the accuracy at 0.6180 with a standard deviation of 0.0035, and the F1 score was 0.5607 with a standard deviation of 0.0315. These results indicate that the NN model had a moderate ability to correctly classify the chronic disease indicators but struggled to balance precision and recall, as reflected in the lower F1 score.

The NN's performance demonstrates its capability to handle complex datasets, yet it falls short in optimizing both precision and recall simultaneously. The F1 score, which balances precision and recall, being lower than both, suggests that while the model was fairly good at making correct predictions (precision) and identifying true cases (recall), it was less successful when these metrics were considered together. Furthermore, the CNN outperformed the NN in several metrics. The CNN achieved an accuracy of 0.6303 with a standard deviation of 0.0027. The precision was 0.6445 with a standard deviation of 0.0410, recall matched the accuracy at 0.6303 with a standard deviation of 0.0027, and the F1 score was 0.5950 with a standard deviation of 0.0239. These results suggest that the CNN model was better at identifying patterns and features in the CDI dataset, resulting in improved performance across all metrics compared to the NN. Then, the CNN's ability to perform better is likely due to its architecture, which is adept at capturing spatial hierarchies in data through convolutional layers. This capability is particularly useful in structured data, where interactions between features can be spatially oriented. The higher precision and recall indicate that the CNN made more correct predictions and was better at identifying true cases, leading to a higher F1 score as well.

The RNN using LSTM units showed similar performance to the NN but did not surpass the CNN. The RNN achieved an accuracy of 0.6195 with a standard deviation of 0.0029. The precision was 0.6197 with a standard deviation of 0.0165, recall matched the accuracy at 0.6195 with a standard deviation of 0.0029, and the F1 score was 0.5707 with a standard deviation of 0.0376. These metrics indicate that while the RNN was capable of capturing temporal dependencies in the data, it did not significantly outperform the NN and was less effective than the CNN. The LSTM's performance highlights its strength in handling sequential data, as it can retain information over time through its memory units. However, the CDI dataset, while complex, may not

have contained the type of temporal dependencies that LSTMs excel at capturing, which could explain why the RNN did not outperform the CNN. The lower F1 score relative to the CNN suggests that while the RNN was decent at making predictions and identifying true cases, it faced challenges in balancing these aspects effectively.

Comparing the three models, the CNN emerged as the most effective model for predicting chronic disease indicators using the CDI dataset. It achieved the highest accuracy, precision, and F1 score, indicating a superior ability to identify and classify chronic disease indicators accurately. The CNN's architecture, which excels at detecting patterns and interactions between features, likely contributed to its better performance. The NN and RNN models, while effective, did not match the CNN's performance. The NN, with its simpler architecture, struggled to optimize precision and recall simultaneously. The RNN, despite its advanced architecture for handling sequential data, did not find significant temporal patterns in the CDI dataset that would leverage its full potential.

The results of this study highlight the importance of selecting appropriate machine learning architectures based on the nature of the dataset. The CDI dataset, comprising structured public health data, benefitted most from the CNN's capability to capture spatial hierarchies and interactions between features. This finding is consistent with previous research that demonstrates CNNs' effectiveness in handling structured and image-like data. The moderate performance of the NN and RNN models underscores the challenges these architectures face with the CDI dataset. The NN, while flexible and powerful, may require more tuning and optimization to achieve better performance. The RNN's design for sequential data suggests that it might perform better with datasets containing clearer temporal sequences or time-series data. Hyperparameter tuning played a crucial role in optimizing each model's performance. Techniques such as grid search and random search were employed to identify the best combination of hyperparameters, including the number of layers, neurons per layer, learning rate, batch size, and dropout rate. This process helped improve the models' performance, but the inherent strengths and weaknesses of each architecture ultimately determined the outcomes.

The study's findings have significant implications for public health practice. By identifying the most effective machine learning model for predicting chronic disease indicators, public health officials and policymakers can leverage these insights to enhance disease monitoring, early detection, and intervention strategies. The superior performance of the CNN model suggests that it can be effectively applied to analyze complex health datasets, providing valuable predictions and insights.

Future research could explore hybrid models that combine the strengths of CNNs and RNNs to further improve performance. Additionally, incorporating more advanced techniques such as attention mechanisms could help capture more nuanced patterns in the data. Expanding the dataset to include more temporal elements or additional features could also provide a more comprehensive evaluation of these models' capabilities.

Table 1. Classification Report – Deep Learning Architectures

Model	Accuracy	Precision	Recall	F1 Score
Neural Network	0.6180 ± 0.0035	0.6288 ± 0.0370	0.6180 ± 0.0035	0.5607 ± 0.0315
Convolutional Neural Network	0.6303 ± 0.0027	0.6445 ± 0.0410	0.6303 ± 0.0027	0.5950 ± 0.0239
Recurrent Neural Network	0.6195 ± 0.0029	0.6197 ± 0.0165	0.6195 ± 0.0029	0.5707 ± 0.0376

4. CONCLUSION

This research aimed to evaluate and compare the performance of three machine learning models—Neural Network (NN), Convolutional Neural Network (CNN), and Recurrent Neural Network (RNN) using Long Short-Term Memory (LSTM) units—in predicting chronic disease indicators using the Chronic Disease Indicators (CDI) dataset. The study found that the CNN model outperformed both the NN and RNN models across all key performance metrics, including accuracy, precision, recall, and F1 score. The CNN's superior performance can be attributed to its ability to capture spatial hierarchies and interactions between features, making it particularly effective for structured datasets like the CDI. This finding underscores the importance of selecting the appropriate machine learning architecture based on the nature of the data. While the NN and RNN models showed moderate performance, they did not match the efficacy of the CNN, highlighting the challenges these models face with the CDI dataset. The results have significant implications for public health practice. By identifying the most effective machine learning model for predicting chronic disease indicators, public health officials and policymakers can enhance disease monitoring, early detection, and intervention strategies. The insights gained from this study can inform the development of more accurate and reliable predictive models, ultimately contributing to better health outcomes and resource allocation. Future research should explore hybrid models that combine the strengths of CNNs and RNNs to further improve performance. Additionally, incorporating advanced techniques such as attention mechanisms could capture more nuanced patterns in the data. Expanding the dataset to include more temporal elements or additional features could provide a more comprehensive evaluation of these models' capabilities.

ACKNOWLEDGMENTS

We express our deepest gratitude to Atma Jaya Catholic University of Indonesia for the unwavering support and resources provided throughout the duration of this research project. Our sincere thanks also extend to the Information System Study Program, whose rigorous academic environment and collaborative ethos have been instrumental in facilitating our work.

REFERENCES

[1] C. Wilson, "The Possible Inadequacies of Business Practices at Healthcare Clinics that Provide Services to Minority Patients: A Quantitative Study," Northcentral University, 2022.

[2] L. Slawomirski and N. Klazinga, "The economics of patient safety: from analysis to action," 2022.

[3] K. Divaris *et al.*, "Cohort profile: ZOE 2.0—a community-based genetic epidemiologic study of early childhood oral health," *Int. J. Environ. Res. Public Health*, vol. 17, no. 21, p. 8056, 2020.

[4] M. E. Akokuwebe and E. S. Idemudia, "A comparative cross-sectional study of the prevalence and determinants of health insurance coverage in Nigeria and South Africa: a multi-country analysis of demographic health surveys," *Int. J. Environ. Res. Public Health*, vol. 19, no. 3, p. 1766, 2022.

[5] L. J. Beesley *et al.*, "The emerging landscape of health research based on biobanks linked to electronic health records: Existing resources, statistical challenges, and potential opportunities," *Stat. Med.*, vol. 39, no. 6, pp. 773–800, 2020.

[6] Z. Nenova and J. Shang, "Chronic disease progression prediction: Leveraging case-based reasoning and big data analytics," *Prod. Oper. Manag.*, vol. 31, no. 1, pp. 259–280, 2022.

[7] A. Rehman, S. Naz, and I. Razzak, "Leveraging big data analytics in healthcare enhancement: trends, challenges and opportunities," *Multimed. Syst.*, vol. 28, no. 4, pp. 1339–1371, 2022.

[8] M. M. Ahsan, S. A. Luna, and Z. Siddique, "Machine-learning-based disease diagnosis: A comprehensive review," in *Healthcare*, 2022, vol. 10, no. 3, p. 541.

[9] S. Nusinovici *et al.*, "Logistic regression was as good as machine learning for predicting major chronic diseases," *J. Clin. Epidemiol.*, vol. 122, pp. 56–69, 2020.

[10] I. Preethi and K. Dharmarajan, "Diagnosis of chronic disease in a predictive model using machine learning algorithm," in *2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE)*, 2020, pp. 191–196.

[11] K. Arumugam, M. Naved, P. P. Shinde, O. Leiva-Chauca, A. Huaman-Osorio, and T. Gonzales-Yanac, "Multiple disease prediction using Machine learning algorithms," *Mater. Today Proc.*, vol. 80, pp. 3682–3685, 2023.

[12] N. Rajender and M. V. Gopalachari, "An efficient dimensionality reduction based on adaptive-GSM and transformer assisted classification for high dimensional data," *Int. J. Inf. Technol.*, vol. 16, no. 1, pp. 403–416, 2024.

[13] B. Pes, "Learning from high-dimensional and class-imbalanced datasets using random forests," *Information*, vol. 12, no. 8, p. 286, 2021.

[14] E. Capobianco, "High-dimensional role of AI and machine learning in cancer research," *Br. J. Cancer*, vol. 126, no. 4, pp. 523–532, 2022.

[15] B. Yan, M. Aasma, and others, "A novel deep learning framework: Prediction and analysis of financial time series using CEEMD and LSTM," *Expert Syst. Appl.*, vol. 159, p. 113609, 2020.

[16] J. Li, B. Yang, H. Li, Y. Wang, C. Qi, and Y. Liu, "DTDR--ALSTM: Extracting dynamic time-delays to reconstruct multivariate data for improving attention-based LSTM industrial time series prediction models," *Knowledge-Based Syst.*, vol. 211, p. 106508, 2021.

[17] M. A. Morid, O. R. L. Sheng, and J. Dunbar, "Time series prediction using deep learning methods in healthcare," *ACM Trans. Manag. Inf. Syst.*, vol. 14, no. 1, pp. 1–29, 2023.

[18] A. De Santo, A. Galli, M. Gravina, V. Moscato, and G. Sperl\\i, "Deep Learning for HDD health assessment: An application based on LSTM," *IEEE Trans. Comput.*, vol. 71, no. 1, pp. 69–80, 2020.

[19] T. Tian, R. Cooper, J. Deng, and Q. Zhang, "Transforming Investment Strategies and Strategic Decision-Making: Unveiling a Novel Methodology for Enhanced Performance and Risk Management in Financial Markets," *arXiv Prepr. arXiv2405.01892*, 2024.

[20] S. H. Mboweni and P. R. Risenga, "The Impact of The COVID-19 Pandemic on the Management of Chronic Disease in South Africa: A Systematic Review," *Open Public Health J.*, vol. 15, no. 1, 2022.

[21] S. Erismann *et al.*, "How to bring research evidence into policy? Synthesizing strategies of five research projects in low-and middle-income countries," *Heal. Res. Policy Syst.*, vol. 19, pp. 1–13, 2021.

[22] Y. Liu, Z. Zhang, A. J. Yepes, and F. D. Salim, "Modeling long-term dependencies and short-term correlations in patient journey data with temporal attention networks for health prediction," in *Proceedings of the 13th ACM international conference on bioinformatics, computational biology and health informatics*, 2022, pp. 1–10.

[23] Jainaru, "Chronic Disease Indicators." 2023.