Multi-Class Disease Prediction with Deep Neural Networks

1st MD. ZOBAYER
BRAC University
Computer Science and Engineering
Dhaka, Bangladesh
md.zobayer@g.bracu.ac.bd
22101518

2nd MD. SHOYEB AKHTER
BRAC University
Computer Science and Engineering
Dhaka, Bangladesh
md.shoyeb.akhter@g.bracu.ac.bd
22101645

3rd MD. IBRAHIM AKANDA
BRAC University
Computer Science and Engineering
Dhaka, Bangladesh
md.ibrahim.akanda@g.bracu.ac.bd
21201233

4th MAHMUDUL HASAN TAMAL
BRAC University
Computer Science
Dhaka, Bangladesh
mahmudul.hasan.tamal@g.bracu.ac.bd
24341131

***Abstract*—Accurate prediction of disease risk is essential for effective healthcare planning and early intervention. Intelligent health risk prediction models built with deep learning architectures offer a powerful tool for physicians to identify patterns in patient data that indicate risks associated with certain types of chronic disease [1]. This project proposes a deep learning–based approach for multi-class disease prediction using a real-world healthcare risk factors dataset. The data set includes demographic and health-related attributes that represent multiple categories of diseases. After data preprocessing was performed, including handling missing values, feature scaling, and label encoding, a Deep Neural Network (DNN) model was developed using TensorFlow and Keras. The model was trained and evaluated using stratified training, validation, and testing splits to ensure reliable performance. Experimental results demonstrate that the proposed DNN achieves strong classification performance across multiple disease classes, indicating its ability to capture complex, non-linear relationships within healthcare data. The results suggest that deep neural networks can serve as effective tools for multi-class disease risk prediction and have substantial potential to support data-driven decision-making in healthcare systems.**

I. Introduction

Noncommunicable diseases (NCDs) or chronic diseases constitute a group of conditions that occur not due to infection but a combination of genetic, physiological, environmental, and behavioral factors, and these conditions result in lasting health consequences and often require long-term treatment and care [2]. Current diagnosis method measures blood pressure using pulse transit time (PTT) derived from two physiological signals, typically ECG and PPG, which enhances stability and has been validated by past studies, but also increases operational complexity, particularly due to ECG acquisition [3]. Traditional diagnostic methodologies often rely on isolated clinical thresholds—such as a single blood pressure reading, BMI calculation or other invasive techniques to diagnose different diseases to assess risk. While effective in specific contexts, these linear methods frequently fail to extract the complex, non-linear interactions between a patient's diverse lifestyle choices, demographic background, and physiological markers.

Delayed or missed diagnoses cause a significant risk to patient outcomes, highlighting an urgent need for advanced, data-driven tools capable of analyzing complex health profiles. This project addresses the challenge by developing a Multi-Class Disease Prediction system using Deep Neural Networks (DNNs), which are uniquely suited to modeling the intricate, high-dimensional dependencies found in medical data. The key objective is to build a robust model capable of classifying seven conditions: Hypertension, Diabetes, Obesity, Asthma, Arthritis, Cancer, and Healthy states. Utilizing the Healthcare Risk Factors Dataset with good accuracy, we engineered 17 clinical and demographic features—such as glucose levels and diet scores—to provide a reliable diagnostic aid. By successfully navigating core healthcare data challenges, including noise reduction and missing value imputation, this study presents a reliable framework for automated disease risk assessment.

II. Methodology

The methodology follows a structured pipeline designed to transform raw healthcare data into a reliable predictive system. Starting with the Dataset and preparation, the study utilizes the Healthcare Risk Factors Dataset, containing 30,000 records. Following a rigorous cleaning process, irrelevant and noisy columns were removed, and records with missing target labels were discarded. Missing numerical values were handled via Median Imputation, while categorical/binary features used Mode Imputation. Data was validated to ensure features fell within realistic medical ranges, resulting in a finalized dataset of 25,500 records.

Identify applicable funding agency here. If none, delete this.

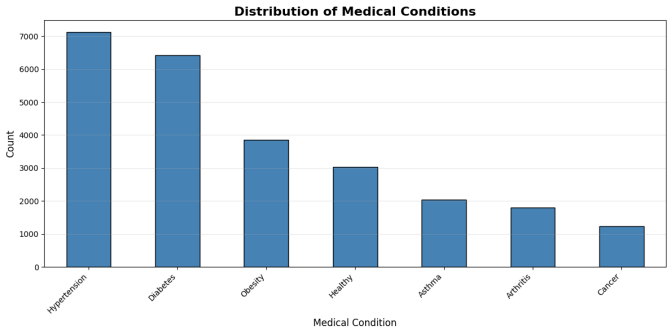

Fig. 1. Distribution of Medical Conditions.

To prepare the data for the neural network, we performed data transformation and feature engineering. Gender was binary encoded, and the target variable was label-encoded into seven discrete classes. Standardization: All 17 features underwent Z-score normalization to ensure that features with different scales (e.g., Age vs. Glucose) contribute equally to the gradient descent process. The model utilizes 17 inputs, which firstly include clinical information such as Age, Glucose, Blood Pressure, BMI, Oxygen Saturation, Cholesterol, Triglycerides, and HbA1c. Secondly, lifestyle information as Physical Activity, Diet Score, Stress Level, Sleep Hours, Smoking, and Alcohol. Finally, Demographic: Gender, Family History, and Length of Stay. A 70% training, 15% validation, and 15% test stratified split was used to maintain class balance. A Deep Neural Network was designed with a 17-neuron input layer and four dense hidden layers with varying neuron counts to balance capacity and complexity. ReLU activation was used in hidden layers. Batch Normalization and Dropout were integrated to stabilize training and prevent overfitting. while 7 neurons with a Softmax activation function for multi-class probability estimation. For optimization, we used the Adam optimizer, and Sparse Categorical Cross-Entropy was our loss function. A maximum of 100 epochs with Early Stopping, Learning Rate Reduction on plateau, and Model-Checkpointing to save the best-performing weights was implemented.

III. Results

The DNN model achieved high diagnostic precision on the unseen test set. For evaluation, we have used multiple metrics as follows:

- Overall Accuracy: 91.87%.
- Log Loss: Low cross-entropy loss indicates high confidence in the predicted classes.
- ROC-AUC: A one-vs-rest approach yielded a high AUC score (averaging $\sim$0.99), indicating excellent class separability.
- Confusion Matrix: Revealed that the model successfully distinguished between most categories, with particularly high F1-scores for "Healthy," "Diabetes," and "Cancer."

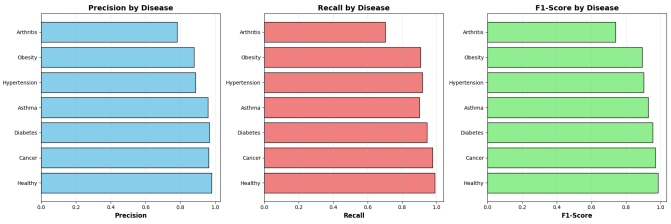

Fig. 2. Per-Class Performance Metrics.

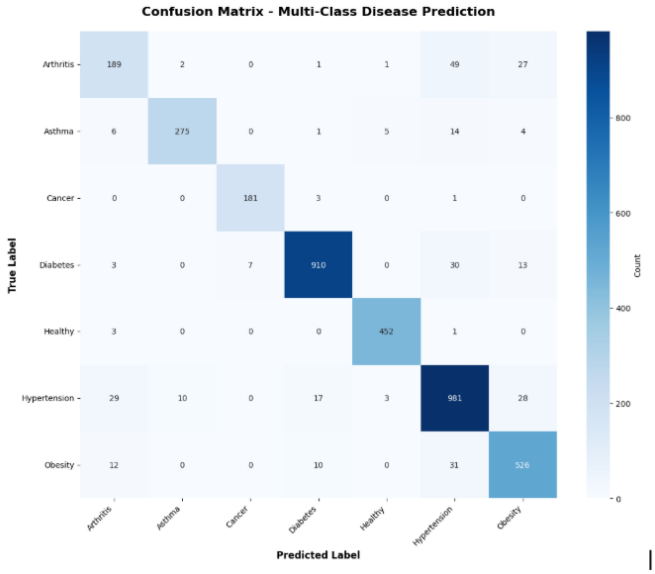

Fig. 3. Confusion Matrix - Multi-Class Disease Prediction.

IV. Discussion

The high accuracy of 91.87% demonstrates that the selected 17 features, combining clinical markers like HbA1c with lifestyle indicators including stress and diet, provide a comprehensive signature for these diseases. The use of a four-layer DNN was justified, as it successfully extracted non-linear relationships that simpler models (like Logistic Regression) might be unable to detect.

The integration of Batch Normalization and Dropout was key in managing the "noise" inherent in healthcare data. Although the model is highly effective, the slight confusion observed between metabolic-related conditions (Obesity and Hypertension) suggests that these conditions share overlapping risk profiles.

Conclusion

This project successfully executed a scalable and reproducible pipeline for multi-class disease prediction. By leveraging Deep Neural Networks and a structured data preparation framework, the system achieved a robust 91.87% accuracy in identifying seven different health states. This framework serves as a powerful proof-of-concept for automated diagnostic aids, potentially allowing healthcare providers to prioritize high-risk patients and implement early measures. Future research could

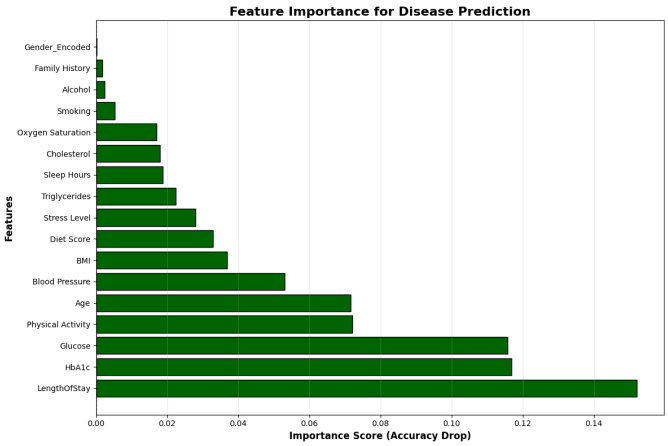

Fig. 4. Feature Importance for Disease Prediction.

involve testing the model on real-world clinical electronic health records (EHR) to validate its external generalizability further.

REFERENCES

[1] A. Maxwell et al., "Deep learning architectures for multi-label classification of intelligent health risk prediction," BMC Bioinformatics, vol. 18, no. S14, p. 523, Dec. 2017, doi: 10.1186/s12859-017-1898-z.

[2] "Multi-Class Classification Method with Feature Engineering for Predicting Hypertension with Diabetes," River Publishers Journals & Magazine — IEEE Xplore, May 01, 2023. https://ieeexplore.ieee.org/abstract/document/10976611

[3] Y. Liang, Z. Chen, R. Ward, and M. Elgendi, "Photoplethysmography and deep learning: Enhancing hypertension risk stratification," Biosensors, vol. 8, no. 4, p. 101, Oct. 2018, doi: 10.3390/bios8040101.