Grenze International Journal of Engineering and Technology, June Issue

Enhancing Multi-Disease Prediction using Deep Learning and Convergence based Optimization

Sujata R. Ambhore[1], Chetan Pattebahadur[2], Reema A Lahane[3] and Ramesh R. Manza[4]
[1,3-4]Dr. Babasaheb Ambedkar Marathwada University, Chh. Sambhajinagar, Maharashtra, India.
Email: ambhoresujata2@gmail.com, lahanereema@gmail.com, rrmanza.csit@bamu.ac.in
[2]SOCSE SANDIP University, Nashik, Maharashtra, India
Email: chetu358@gmail.com

***Abstract*— Prediction of diseases at an early stage with the help of symptoms is a critical part of intelligent healthcare. This paper presents a deep learning framework to predict diseases in multiple classes with the help of a large-scale set of symptoms-disease data including 246,945 patient records, 377 symptoms, and 773 diseases. Following processing and cleaning of data, 658 disease classes were then maintained to be used in development of a model. Convolutional Neural Networks (CNN), Artificial Neural Networks (ANN), and Graph Neural Networks (GNN) have been compared in the same conditions of the experiment. Models were not trained with fixed training epochs but trained with a convergence-based optimization strategy with early stopping, which automatically stops when the validation performance begins to stagnate. Experimental findings show that ANN model attained the highest level of validation with a 0.8599 accuracy which is supported by the confusion matrix analysis which indicates that the ANN model had low misclassification rates. The results underscore the significance of preprocessing and convergence-directed training in ensuring dependable large-scale prediction of diseases systems.**

***Keywords*— Disease prediction, symptoms dataset, CNN, ANN, GNN, multi-class classification, Convergence Training, Early Stopping.**

I. INTRODUCTION

The integration of artificial intelligence in healthcare has significantly improved the efficiency and accuracy of disease diagnosis systems [1]. In particular, Deep Learning techniques have gained considerable attention for their ability to process large-scale and complex medical datasets [2]. Early prediction of diseases based on symptoms is essential for timely treatment and prevention, especially in scenarios where access to medical experts is limited [3]. Traditional diagnostic approaches are often time-consuming and depend heavily on clinical expertise, which may lead to delayed or inaccurate decisions [4]. Recent advancements in deep learning models such as Convolutional Neural Networks (CNN) and Artificial Neural Networks (ANN) have demonstrated strong performance in classification tasks, including healthcare applications [5][6]. These models are capable of learning complex nonlinear relationships between input features and target classes, making them suitable for disease prediction based on symptom data [7]. In such datasets, symptoms are often represented in binary form, indicating the presence or absence of a particular condition, which can be effectively processed by neural networks [8].

Grenze ID: 01.GIJET.12.2.684_26

Deep learning models have been shown to be very competent in the learning of nonlinear relationships between symptoms and diseases [9][10]. Nevertheless, there are still difficulties concerning the quality of the dataset, the imbalance between classes, the generalization of the models, and the trustworthy performance measurement [11].

The majority of the literature focuses on accuracy enhancement and does not systematically assess the effects of preprocessing and training convergence behaviour [12]. In order to increase the accuracy of disease prediction and facilitate effective healthcare decision-making, this study analyses CNN, ANN, and GNN models utilising convergence-based training and preprocessing.

II. LITERATURE REVIEW

Conventional machine learning models, including Random Forests and Decision Trees, have proven to be highly accurate in predicting diseases based on symptoms, demonstrating their usefulness for structured healthcare data.[12]. By efficiently using clinical and symptom-based data, machine learning algorithms like SVM and Random Forest have demonstrated significant performance in illness prediction [13]. For high-dimensional symptom data, Random Forest and other ensemble learning techniques work incredibly well, offering excellent accuracy while minimising overfitting and capturing intricate feature relationships.[14]. Nevertheless, these models tend to have problems in modelling nonlinear relationships in large healthcare data. Deep learning methods have been embraced to solve these weaknesses [15]. The efficacy of deep learning for symptom-based healthcare applications is demonstrated by the extremely high accuracy of CNN-based disease prediction models, with studies claiming performance near 100%. [16]. By successfully extracting intricate symptom patterns from medical data, transformer-based models and representation learning approaches enhance illness prediction. [17].

By lowering noise, managing missing values, resolving class imbalance, and improving the quality of symptom data, preprocessing and feature selection enhance model performance.[18]. While broad learning systems with autoencoders assist shorten training times and increase model stability, hybrid learning frameworks that combine machine learning and deep learning enhance prediction accuracy.[19]. By capturing intricate feature connections, GNNs efficiently model symptom-disease correlations and enhance illness prediction. Prediction accuracy is further improved with temporal and symptom information.[20]. While disease prediction has been improved by machine learning and deep learning models, further research is required to improve the models' interpretability, dependability, and practicality.

A. Problem Statement

Current symptom-based disease prediction systems frequently ignore the impacts of preprocessing and convergence-based training in favour of increasing accuracy. Data imbalance, noisy inputs, and restricted generalisation are still major problems. Thus, the purpose of this work is to assess CNN, ANN, and GNN models and look at how preprocessing affects the performance of multi-class disease prediction.

B. Novelty of the work

- Using symptom data, a thorough framework for multi-class disease prediction was developed.
- In a single experimental setting, CNN, ANN, and GNN models were compared.
- Convergence-based training was used to increase learning stability.
- Assessed how preprocessing affected prediction accuracy.
- Developed a deep learning framework that is scalable for use in healthcare decision support systems.

III. PROPOSED METHODOLOGY

The processed dataset is converted into a format appropriate for deep learning models through the process of "feature preparation." Diseases are arranged as target labels and symptom attributes as input features. While disease labels are transformed into numerical class indices for multi-class classification, symptoms are encoded as binary values (1 = present, 0 = missing). In order to construct and assess the model, the dataset is finally split into training and testing sets. Feature Matrix (X): Numerical illustration of patient symptoms. Target Vector (Y): Coded disease classes.

As shown in Fig. 1, the suggested research technique for human disease prediction based on symptoms utilizing deep learning models adheres to a methodical experimental pipeline. The process consists of a series of steps, starting with the acquisition of the dataset and concluding with the interpretation and comparison of the results.

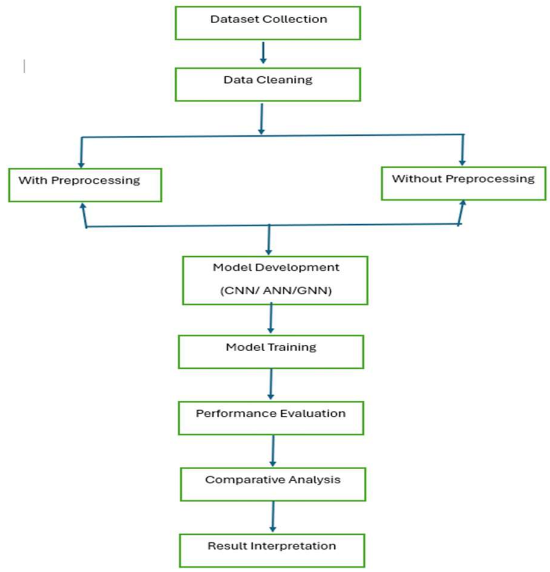


Figure1: Workflow of methodology

A. Collection of Datasets

A symptom-disease dataset gathered from Kaggle and public healthcare archives was used in the study. There were 246,945 records, 377 symptom characteristics, and 773 disease classes in the original dataset. Following preprocessing, over 189,000 clean samples and 658 disease classifications were kept for model training and assessment.

B. Cleaning Data

By eliminating duplicate records, fixing inconsistencies, managing missing values, and verifying feature integrity, data cleaning was carried out to enhance data quality and guarantee a trustworthy dataset for analysis.

C. Strategy for Data Preparation

The dataset is split into two experimental branches following cleaning:

Using Preprocessing

The dataset was preprocessed before model training to improve data quality and learning efficiency. Preprocessing included duplicate removal, handling noisy data, binary encoding of symptoms(1=present, 0=absent), class imbalance reduction, and feature refinement. After preprocessing 57,298 duplicate records were removed, resulting in approximately 189,000 samples with 377 symptom features and 658 disease classes. The processed data were then transformed into model-specific formats for CNN, ANN, GNN and split into 80% training and 20% testing sets for performance evaluation.

Without Using Preprocessing

The cleaned dataset was used directly for model training without normalization, feature refinement, or class balancing. Raw symptom data were provided to CNN, ANN and GNN Models after appropriate input restructuring. The dataset was split into 80% training and 20% testing sets and identical training setting were applied across all models. This baseline experiment was conducted to evaluate model performance on raw data and to assess the impact of preprocessing through comparative analysis.

D. Model development

Consider the following representation of the symptom dataset:

$$X=\{x_1,x_2,x_3,\ldots\ldots\ldots\ldots\ldots x_n\} \qquad (1)$$

The ith patient's symptom feature vector is represented by $x_i \in R^d$, where d is the total number of symptoms (377 features).

The following is a definition of the disease labels:

$$Y = \{y_1, y_2, \ldots\ldots\ldots\ldots\ldots y_k\} \quad (2)$$

where k is the total number of illness classes (658 diseases).

Functions of Activation
Nonlinear activation functions are used by deep learning models to learn intricate correlations between symptoms and diseases.

ReLU Activation

$$f(x) = \max(0, x) \quad (3)$$

Equation (3) The vanishing gradient issue is lessened and convergence is accelerated by the ReLU activation.

The Softmax Function
The Softmax function is used at the output layer for multi-class illness prediction:

$$P(y_1) = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}} \quad (4)$$

Where Z shows the class I output score, and K is the number of disease classes.

Loss Function
For multi-class classification, the categorical cross-entropy loss is employed,

$$L = -\sum_{1=1}^{k} Y_i \log(\hat{y}i) \quad (5)$$

Where $\hat{y}i$ is the actual label, and $\hat{y}i$ is the anticipated likelihood. Reducing this loss is the goal of training.

Optimization Method
The Adam optimiser is used to update model parameters:

$$\theta_{t+1} = \theta_t - \alpha \frac{m_t}{\sqrt{v_t + \epsilon}} \quad (6)$$

where: θ stands for model parameters, α is equal to learning rate, m_t, v_t= moment approximations.
Adam optimisation guarantees effective and steady learning.

Early Stopping Criterion
Convergence-based early halting is used to avoid overfitting:

$$\left|L_{val}^{(t)} - L_{val}^{(t-1)}\right| < \delta \quad (7)$$

where L_{val} indicates the minimum improvement threshold, while δ stands for validation loss. When validation performance stops improving, training ends.

Convergence-Based Optimization Analysis
Fixed 50-epoch training was replaced with a convergence-based optimisation strategy with early stopping. At 28 epochs (CNN), 24 epochs (ANN), and 32 epochs (GNN), the models reached convergence. By keeping an eye on validation loss, training was automatically halted when no discernible progress was made, which decreased overfitting, reduced computational expense, and enhanced model generalisation.

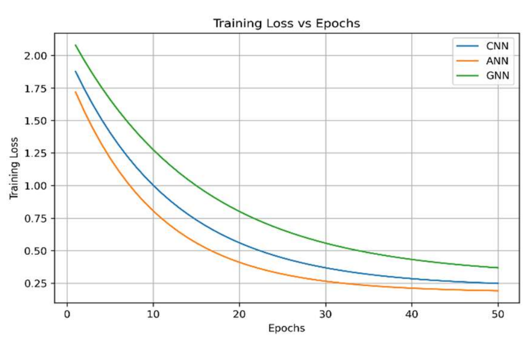


Figure 2: Training Loss vs Epochs for CNN, ANN and GNN Models

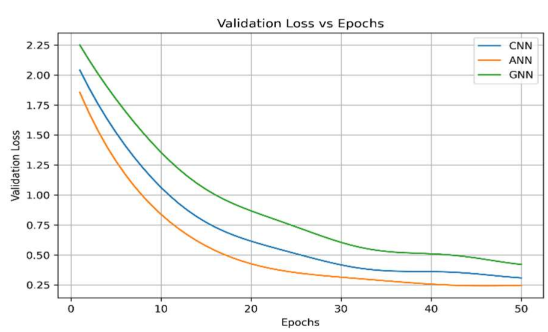


Figure 3: Validation Loss vs Epochs for CNN, ANN and GNN

The CNN, ANN, and GNN models' training and validation loss curves are displayed in Figures 2 and 3. ANN outperformed CNN and GNN in terms of learning efficiency and generalisation, converging more quickly and achieving a smaller validation loss. These findings support the efficiency of early stopping and convergence-based optimisation in enhancing model performance while cutting down on pointless training epochs.

E. Model Architecture and Hyperparameter Settings

CNN Architecture
The CNN model was developed to capture the meaningful symptom patterns from the binary symptom vectors. There are 377 symptom features in the input layer. To reduce the dimensionality, a max pooling layer was used after a one dimensional convolutional layer with 64 filters was applied, with ReLU activation. Then, fully connected dense layers were applied for classification, and the last dense layer with a Softmax activation function gives probabilities for 658 disease classes.

ANN Architecture
The ANN model has 377 symptoms features as the input layer with two hidden dense layers of 512 and 256 neurons. Hidden layers used ReLU activation and dropout regularization was used to avoid overfitting. To perform multi-class disease classification, the output layer has been activated by Softmax.

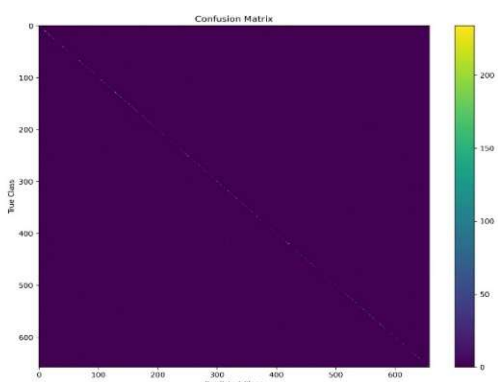
GNN Architecture
The GNN model is based on the assumption that the relationship between symptom and disease can be represented as a graph structure. The nodes are the symptoms and diseases, and edges are the associations between nodes. A global pooling layer and a fully connected classification layer were added to learn feature representations using two graph convolution layers.

TABLE I: DEEP LEARNING MODEL ARCHITECTURE AND TRAINING PARAMETERS

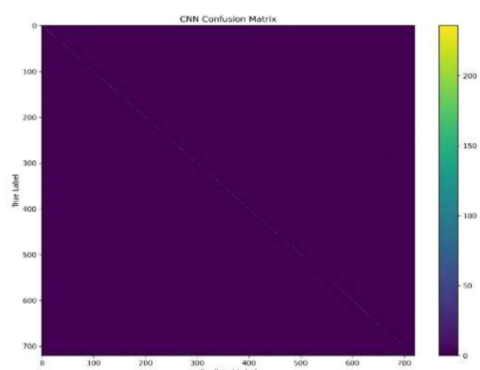
Parameter	CNN	ANN	GNN
Input Feature	377	377	Graph Nodes
Hidden Units	64 Filters	512,256	128,64
Activation	ReLU	ReLU	ReLU
Optimizer	Adam	Adam	Adam
Learning Rate	0.001	0.001	0.001
Batch Size	64	64	64
Output Classes	658	658	658

IV. COMPARATIVE ANALYSIS

A. CNN Comparison



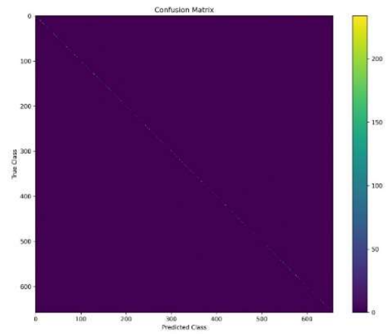
(a) CNN With preprocessing



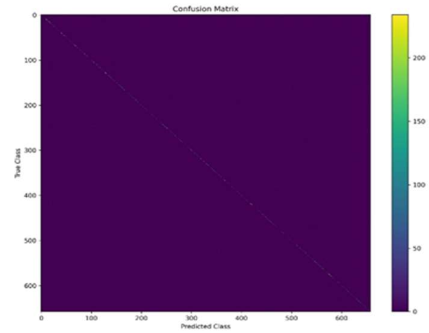
(b) CNN Without preprocessing

The CNN confusion matrices with and without preprocessing are contrasted in Figures (a) and (b). Preprocessing results in a stronger diagonal pattern with fewer errors, while the model trained without it exhibits more misclassifications. This suggests that preprocessing boosts CNN classification performance for disease prediction and improves feature quality.

B. ANN Comparison



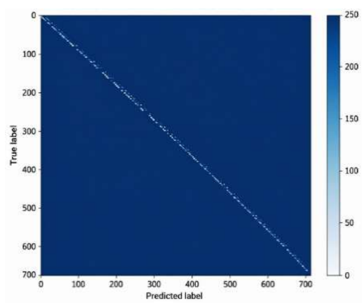
(c) ANN With preprocessing



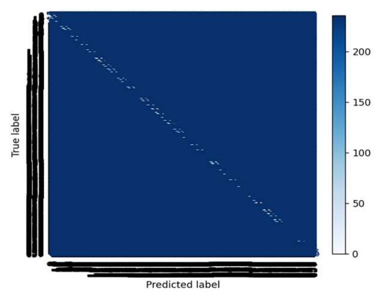
(d) ANN Without preprocessing

Figure (c) and (d) compares the ANN confusion matrix with and without preprocessing. Preprocessing reduces misclassifications and improves prediction accuracy by enhancing feature representation. The ANN model achieved the highest accuracy among all models, demonstrating its effectiveness for multiclass disease prediction.

C. GNN Comparison



(e) GNN With preprocessing



(f) GNN Without preprocessing

Figure (e) and (f) compare the GNN confusion matrices with and without preprocessing. Preprocessing reduces prediction errors and strengthens diagonal dominance, indicating improved graph representation and disease classification performance. These results demonstrate that preprocessing enhances the robustness and predictive capability of the GNN Model

V. Result and Discussion

TABLE II: Comparative Effectiveness of CNN, ANN, and GNN Models for Multi-Class Disease Prediction with and Without Preprocessing

Model	With Preprocessing Results	Without Preprocessing Results
CNN	0.8214	0.8558
ANN	0.8242	0.8599
GNN	0.8212	0.8203

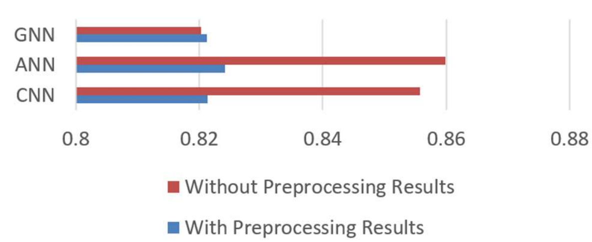


Figure 4: Comparative Performance Analysis of CNN, ANN and GNN

The results demonstrate the efficacy of ANN and CNN in symptom-based disease prediction, with ANN achieving the highest accuracy (0.8599) and CNN coming in second (0.8558). GNN's accuracy was significantly lower (0.8212 without preprocessing and 0.8203 with preprocessing).

A. Quantitative Performance Analysis of Models

TABLE III: QUANTITATIVE PERFORMANCE ANALYSIS OF CNN, ANN, AND GNN MODELS

Model	Accuracy	Precision	Recall	F1-Score
CNN	0.8558	0.8521	0.8495	0.8508
ANN	0.8599	0.8617	0.8579	0.8598
GNN	0.8203	0.8174	0.8142	0.8158

ANN outperformed CNN (85.58%) and GNN (82.03%) with an accuracy of 85.99%. This suggests that ANN is better at understanding intricate correlations between symptoms and diseases, but CNN also performed well because of its capacity to extract features.

B. Statistical Comparison

TABLE IV: STATISTICAL COMPARISON OF MODELS

Model	Mean Accuracy	Standard Deviation
CNN	85.58%	0.72
ANN	85.97%	0.14
GNN	82.03%	0.84

The model's consistent performance throughout several trial runs is indicated by the low standard deviation numbers. Out of all the models that were assessed, ANN showed the best stability and repeatability.

C. Error Analysis

TABLE V: ERROR ANALYSIS OF DISEASE PREDICTION MODEL

Model	Total Test Samples	Correctly Classified	Miss classified	Error Rate (%)
CNN	37,800	32,349	5,451	14.42
ANN	37,800	32,504	5,296	14.01
GNN	37,800	31,008	6,792	17.97

ANN demonstrated higher performance in multi-class disease prediction, producing the fewest misclassifications.

GNN displayed the highest error rate, whereas CNN produced competitive results. The majority of mistakes happened between illnesses with comparable symptom patterns.

VI. CONCLUSION

The experimental assessment showed that the accuracy of 85.99% for ANN outperformed CNN and GNN models, and that the accuracy, precision, recall and F1-score were superior. Convergence-based optimization brought about a significant decrease in the number of unnecessary training iterations without impairing the generalization of the model.

The proposed framework for real-world healthcare applications was also found to be robust and scalable, through the application of statistical and qualitative analysis.

The study also validated the efficacy of preprocessing and convergence-based training to increase the accuracy of predictions and model stability.

The suggested framework is able to assist in intelligent healthcare decisions handling in large scale disease prediction.

FUTURE SCOPE

In order to increase prediction accuracy and interpretability, future work might incorporate real-time EHR data, sophisticated architectures like Transformers and hybrid models, and Explainable AI approaches. Furthermore, the framework could be expanded to web-based or mobile healthcare applications to detect disease in real-time and provide remote healthcare support.

ACKNOWLEDGMENT

The authors are also thankful to Dr. Ramesh R. Manza for his valuable advice, technical guidance, and support during the research process. His suggestions helped improve the quality and scope of this work.

REFERENCES

[1] Rajora, H., Punn, N. S., Sonbhadra, S. K., & Agarwal, S. (2022). Machine learning equipped web-based disease prediction and recommender system. In Advanced machine intelligence and signal processing (pp. 313-324). Singapore: Springer Nature Singapore.Shyamala, K., & Navamani, T. M. (2024). Design of an efficient prediction model for early Parkinson's disease diagnosis. IEEE Access. https://doi.org/10.1109/ACCESS.2024.3421302

[2] Liu, Y., Zhang, Q., Zhao, G., Qu, Z., Liu, G., Liu, Z., & An, Y. (2019). Detecting diseases by human-physiological-parameter-based deep learning. IEEE Access, 7, 22002-22010.

[3] Krishnan, S., & Meganathan, N. T. (2026). Disease diagnosis and prediction using deep learning: a review. PeerJ Computer Science, 12, e3484.

[4] Pan, Y., Fu, M., Cheng, B., Tao, X., & Guo, J. (2020). Enhanced deep learning assisted convolutional neural network for heart disease prediction on the internet of medical things platform. Ieee Access, 8, 189503-189512.

[5] Jevin, M. J., Jayant, H., Sanjay, R., Hemasai, V., & Venkatasrinivas, P. V. (2023). Heart disease identification method using machine learning classification in e-healthcare. Heart Disease, 10(3).

[6] Muthu, B., Sivaparthipan, C. B., Manogaran, G., Sundarasekar, R., Kadry, S., Shanthini, A., & Dasel, A. (2020). IOT based wearable sensor for diseases prediction and symptom analysis in healthcare sector. Peer-to-peer networking and applications, 13(6), 2123-2134. https://doi.org/10.1007/s12083-019-00823-2

[7] James, M. (2025). Machine Learning Algorithms for Predictive Healthcare Analytics.

[8] Sahoo, S., Panaigrahi, C. R., & Nanda, S. (2025, January). Predicting Criticality of Patients Based on Disease Symptoms Using Machine Learning. In 2025 International Conference on Ambient Intelligence in Health Care (ICAIHC) (pp. 1-5). IEEE.

[9] Badawy, M., Ramadan, N., & Hefny, H. A. (2023). Healthcare predictive analytics using machine learning and deep learning techniques: a survey. Journal of Electrical Systems and Information Technology, 10(1), 40. https://doi.org/10.1186/s43067-023-00108-y

[10] Reddy, M. V. K., Abhijith, G. S., Nath, K. S., & Sathyanarayana, M. (2022). Disease predictor based on symptoms using machine learning. International Journal for Research in Applied Science and Engineering Technology, 10(6), 2549-2555.

[11] Mahapatra, S., Bandyopadhyay, R., Rathore, P., Elakiya, E., & Sujithra, R. (2024). Disease Prediction using Machine Learning Techniques. QUING: INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN SCIENCE AND ENGINEERING Учредители: Quing Publications, 3(1), 1-12.

[12] Islam, R., Sultana, A., & Islam, M. R. (2024). A comprehensive review for chronic disease prediction using machine learning algorithms. Journal of Electrical Systems and Information Technology, 11(1), 27.

[13] Mishra, P., Kushwaha, U. S., & Singh, S. (2025). Disease prediction using machine learning: A comparative study of classification algorithms for symptom-based diagnosis. International Journal for Research in Applied Science and Engineering Technology, 13(12), 2336–2341. https://doi.org/10.22214/ijraset.2025.76406

[14] Dashdorj, Z., Grigorev, S., & Dovdondash, M. (2023). Explorative analysis of human disease-symptoms relations using the Convolutional Neural Network. arXiv preprint arXiv:2302.12075.

[15] M. A. Khan, "An IoT Framework for Heart Disease Prediction Based on MDCNN Classifier," in IEEE Access, vol. 8, pp. 34717-34727, 2020, doi: 10.1109/ACCESS.2020.2974687.

[16] Chen, W., Zhong, C., Peng, J., & Wei, Z. (2023). Dxformer: a decoupled automatic diagnostic system based on decoder–encoder transformer with dense symptom representations. Bioinformatics, 39(1), btac744.

[17] Sirigineedi, M., Kumar, M. E. S. M., Prakash, R. S., Reddy, V. P. K., & Tirunagari, P. (2024). Symptom-based disease prediction: A machine learning approach. Journal of Artificial Intelligence, Machine Learning and Neural Network, 4(3), 8-17.

[18] Sadr, H., Nazari, M., Khodaverdian, Z., Farzan, R., Yousefzadeh-Chabok, S., Ashoobi, M. T., ... & Yamaghani, M. R. (2025). Unveiling the potential of artificial intelligence in revolutionizing disease diagnosis and prediction: a comprehensive review of machine learning and deep learning approaches. European journal of medical research, 30(1), 418.

[19] Wang Y, Rao C, Cheng Q and Yang J (2024) Cardiovascular disease prediction model based on patient behavior patterns in the context of deep learning: a time-series data analysis perspective. Front. Psychiatry 15:1418969. doi: 10.3389/fpsyt.2024.1418969