Pertanika Proceedings 1 (5): 35 - 39 (2025)

PERTANIKA JOURNALS

PERTANIKA PROCEEDINGS

Journal homepage: http://www.pertanika.upm.edu.my/

Disease Classification in Healthcare Big Data Using Deep Learning

R. Hendra Kumar[1]* and Gurram Sunitha[2]

[1]*School of Computing, Mohan Babu University, Tirupati 517102, India*
[2]*Department of CSE, School of Computing, Mohan Babu University, Tirupati 517102, India*

ABSTRACT

Healthcare big data is driven by advancements in medical imaging, electronic health records, wearable devices etc. It offers opportunities to improve disease diagnosis and classification. This research focuses on addressing challenges by developing a deep learning architecture designed for healthcare data complexities. Suitable optimization algorithm is identified for fine-tuning both the feature extraction and classification models, ensuring they perform at their highest potential. A thorough investigation of different learning pipelines supported identification of the optimal deep model. The simulation result highlighted the greater efficiency of the proposed method over other existing techniques under various measures. The proposed approach achieved an accuracy of 96.93%.

Keywords: Big data, deep learning, disease classification, healthcare, medical imaging

INTRODUCTION

The growth of digital data has significantly expanded big data applications (Ahmed et al., 2023). Hospitals, clinics, research centers etc. produce terabytes of information daily. This data includes structured formats (patient records etc.), unstructured formats (clinical notes, imaging results etc.) (Ara & Mifa, 2024).

Disease classification is an important use of big data in healthcare (Khang et al., 2024). Deep learning (DL) models can process large datasets to identify complex patterns. The effectiveness of these models depends on data quality, data diversity, and architecture design.

Performing disease classification on big data involves several challenges,one of the challenges is heterogeneity (Guo & Chen, 2023). Heterogeneity in big data complicates its integration and analysis. Despite these challenges, it provides valuable opportunities for improving healthcare (Gupta & Kumar, 2023). This research focuses on addressing a

ARTICLE INFO

Article history:
Received: 29 August 2025
Published: 31 October 2025

DOI: https://doi.org/10.47836/pp.1.5.007

E-mail addresses:
hendra.sagar @gmail.com (R. Hendra Kumar)
gurramsunitha@gmail.com (Gurram Sunitha)
* Corresponding author

few challenges by creating a DL architecture. This research specifically intends to design DL architecture to handle healthcare data complexities.

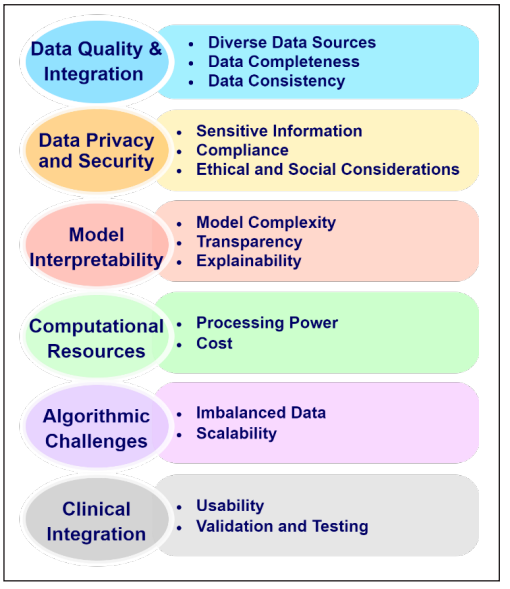

Figure 1. Multi-faceted challenges of big healthcare data classification using deep learning

Literature Review

Big healthcare data consists of diversity of information from sources like patient records, clinical trials, lab tests, pharmacy data, health monitoring applications etc. (Chao et al., 2023). Such data has the potential to enhance patient care, customize treatments etc. However, managing and analyzing data is challenging due to its volume, variety, processing speed etc. (Thayyib et al., 2023).

Figure 1 illustrates key challenges in classifying big healthcare data using DL (Khanna et al., 2023). DL techniques efficiently process and analyze massive healthcare data (Kumari et al., 2023). Thereby, uncovering patterns and making more accurate predictions.

METHODOLOGY

The proposed model for "Disease Classification of Healthcare Big Data using Deep Learning" consists of several key sub-processes designed to address the challenges associated with large-scale healthcare data (Solfa & Simonato, 2023).

Data Preprocessing

It includes data cleaning to remove any noise, managing missing values and normalizing the data to bring it to a common scale. Given the heterogeneity and volume of healthcare data, preprocessing is crucial to improve quality and consistency of big data, making it ready for further processing. Let $X = \{ x1, x2, \ldots, xn \}$ be the set of raw data samples, where each xi is a vector of features $xi = \{ xi1, xi2, \ldots, xim \}$. Let the pre-processed data be $X' = \{ x1', x2', \ldots, xn' \}$.

DL-Based Feature Subset Selection

Most relevant features are selected from the dataset using DL techniques. This step aims to reduce dimensionality of data. Yet retaining the most critical information necessary for disease identification. This step helps in improving the efficiency and accuracy of the

subsequent classification task. The DL model $f\theta$ is trained to select a subset of features. Importance score for each feature is configured as $W = \{w1, w2, \ldots, wm\}$, where wj indicate the importance of feature j. Let λ be threshold parameter that control the number of selected features. The objective is to select a subset of features S such that in Equation 1.

$$S = \{x_j^{'} | w_j > \lambda\} \quad [1]$$

DL-Based Disease Detection

In this step, DL models are used to classify the data and identify the presence of diseases. The models are trained on the pre-processed and feature-reduced dataset, enable them to learn complex patterns that may indicate specific diseases. Let X_S denote the dataset after feature selection. The goal is to classify these samples into disease categories. DL model is defined in Equation 2.

$$\widehat{y_i} = f_\theta\ (x_i^S) \quad [2]$$

where $\widehat{y_i}$ is predicted label for i^{th} sample, and x_i^S is the feature vector of i^{th} sample with selected features.

Optimization Techniques for Hyperparameter Tuning

Optimization techniques are applied for DL model's hyperparameter tuning. By fine-tuning the hyperparameters, the model's accuracy and generalization ability can be significantly improved. Let θ represent the set of hyperparameters in the DL model. The optimization problem can be formulated as shown in Equation 3.

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\, L(\theta) \quad [3]$$

Here, optimization techniques are used to find θ^* that minimizes the loss function $L(\theta)$.

This methodology integrates DL with advanced data processing and optimization techniques to achieve high accuracy and efficiency in disease classification.

RESULTS

Proposed DL architecture for disease classification was evaluated using healthcare dataset. The results provide valuable insights into the model's effectiveness, accuracy and efficiency in classification. The performance evaluation of the proposed approach with relevant metrics is presented in Table 1.

Table 1
Comparative analysis of proposed deep model

Model	Accuracy
VGG16	77.90
VGG19	75.44
Inceptionv3	73.68
Proposed Deep Model	96.93

CONCLUSIONS

This research focused on developing a deep learning architecture for disease classification of healthcare big data. Suitable optimization algorithm was identified for fine-tuning both the feature extraction and classification models, ensuring they perform at their highest potential. A thorough investigation of different learning pipelines supported identification of the optimal deep model. The simulation result highlighted the greater efficiency of the proposed method over other existing techniques under various measures.

ACKNOWLEDGEMENT

The authors express their sincere gratitude to Mohan Babu University, India, for providing facilities and supports throughout the research.

REFERENCES

Ahmed, A., Xi, R., Hou, M., Shah, S. A., & Hameed, S. (2023). Harnessing big data analytics for healthcare: A comprehensive review of frameworks, implications, applications, and impacts. *IEEE Access, 11,* 112891-112928. https://doi.org/10.1109/ACCESS.2023.3323574

Ara, A., & Mifa, A. F. (2024). Integrating artificial intelligence and big data in mobile health: A systematic review of innovations and challenges in healthcare systems. *Global Mainstream Journal of Business, Economics, Development & Project Management, 3*(1), 1-16.

Chao, K., Sarker, M. N. I., Ali, I., Firdaus, R. R., Azman, A., & Shaed, M. M. (2023). Big data-driven public health policy making: Potential for the healthcare industry. *Heliyon*, 9(9), Article e19681. https://doi.org/10.1016/j.heliyon.2023.e19681

Guo, C., & Chen, J. (2023). Big data analytics in healthcare. In Y. Nakamori (Ed.) *Knowledge technology and systems: Toward establishing knowledge systems science* (pp. 27-70). Springer.

Gupta, N. S., & Kumar, P. (2023). Perspective of artificial intelligence in healthcare data management: A journey towards precision medicine. *Computers in Biology and Medicine*, *162*, Article 107051. https://doi.org/10.1016/j.compbiomed.2023.107051

Khang, A., Abdullayev, V., Ali, R. N., Bali, S. Y., Mammadaga, G. M., & Hafiz, M. K. (2024). Using big data to solve problems in the field of medicine. In *Computer vision and AI-integrated IoT technologies in the medical ecosystem* (pp. 407-418). CRC Press.

Khanna, D., Jindal, N., Singh, H., & Rana, P. S. (2023). Applications and challenges in healthcare big data: A strategic review. *Current Medical Imaging*, *19*(1), 27-36. https://doi.org/10.2174/1573405618666220308113707

Kumari, J., Kumar, E., & Kumar, D. (2023). A structured analysis to study the role of machine learning and deep learning in the healthcare sector with big data analytics. *Archives of Computational Methods in Engineering, 30*(6), 3673-3701. https://doi.org/10.1007/s11831-023-09915-y

Solfa, F. D. G., & Simonato, F. R. (2023). Big data analytics in healthcare: Exploring the role of machine learning in predicting patient outcomes and improving healthcare delivery. *International Journal of Computations, Information and Manufacturing*, *3*(1), 1-9. https://doi.org/10.54489/ictim.v3i1.235

Thayyib, P. V., Mamilla, R., Khan, M,. Fatima, H., Asim, M., Anwar, I., Shamsudheen, M. K., & Khan, M. A. (2023). State-of-the-art of artificial intelligence and big data analytics reviews in five different domains: A bibliometric summary. *Sustainability*, *15*(5), Article 4026. https://doi.org/10.3390/su15054026