

Enhancing Rare Diseases Diagnosis Using Deep Learning: Addressing Data Limitations with Synthetic Data Generation via GANs

B. Harish Goud

Dept. of Information Technology
Chaitanya Bharathi Institute of Technology
Hyderabad, India
harishgoudb_it@cbit.ac.in

B. Jayanth Reddy

Dept. of Information Technology
Chaitanya Bharathi Institute of Technology
Hyderabad, India
jayanthbheemudi@gmail.com

Ch. Nitin Reddy

Dept. of Information Technology
Chaitanya Bharathi Institute of Technology
Hyderabad, India
cheerlasainithin@gmail.com

D. Hanok Joshua

Dept. of Information Technology
Chaitanya Bharathi Institute of Technology
Hyderabad, India
hanokkanna@gmail.com

Abstract—The advancements in deep learning technologies have transformed disease diagnosis and treatment processes through increased speed, accuracy, and personalized decision-making. In contrast with other approaches, deep learning-based models are able to utilize their capability in analyzing complex medical data and recognizing hidden patterns that may go undetected by both humans and existing approaches. Deep learning methods have demonstrated substantial impact on cancer, neurological, and cardiac diseases diagnosis, outcome prediction, and customization of medical treatments. Based on studies and examples from clinical settings, it is evident that deep learning methods are capable of equaling and even outperforming the work performed by medical professionals in particular areas. An effective approach that has been developed as part of deep learning research is Generative Adversarial Network (GAN) that allows generating realistic synthetic medical data. In case of rare diseases and anomalies, GAN could provide sufficient data for training models, making them aware of the features of particular diseases and conditions. While there are certain challenges that should be addressed in deep learning, including patients' privacy, collecting of high-quality data, and making the models interpretable by medical specialists, future advancements in this area will involve multi-modal data use, as well as development of explainable AI.

I. INTRODUCTION

Deep learning algorithms have shown great potential in the diagnosis of diseases owing to their ability to extract sophisticated patterns from massive amounts of data [1]. Deep learning algorithms do not require any form of manual feature extraction but offer scalable solutions for healthcare applications [1]. Nevertheless, deep learning models' effectiveness largely relies on access to large and diverse datasets [11].

The creation of such datasets proves difficult in the actual healthcare environment, considering the constraints posed by

privacy laws, high costs related to professional annotations, and the sensitivity of the medical information [4], [13]. Moreover, acquiring data becomes increasingly complicated for rare diseases with insufficient patients' cases [5], [6].

Therefore, deep learning algorithms trained using available data are known to exhibit satisfactory performance on diseases that are prevalent but not rare diseases [6]. The lack of generalization may delay or even cause misdiagnosis, which is harmful to patients [5].

In order to overcome these issues, a viable solution is the creation of synthetic data [2], [3]. Synthetic data can be generated using Generative Adversarial Networks (GANs), which can generate realistic yet anonymous data similar to real data [8], [9]. In this research paper, we analyze the impact of generating synthetic data using GANs to improve deep learning models in rare diseases [7].

II. RELATED WORK

This paper provides a review of recent studies on deep learning-based approaches to medical diagnostics, with special emphasis on problems related to the usage of small and confidential datasets and the possible ways of solving these problems through the creation of synthetic data [1], [5]. An extensive search was conducted among peer-reviewed articles, mostly those published from January 2021 to November 2023; however, some fundamental works dating back to before that period are also considered [11]. In this study, the relevant literature is grouped into four major categories: data needs and limitations of deep learning in medicine [1], [11], problems associated with rare disease diagnostics [5], [6], methods of creating synthetic data [2], [3], and the applications of Generative Adversarial Networks (GANs) within healthcare [8], [9]. The assessment is then expanded to account for the effect of

synthetic data on diagnosis accuracy [7], [10], privacy, ethics [4], [13], and technical challenges when incorporating GAN-generated data augmentation in clinical practice [3], [13]. Such an approach enables a comprehensive evaluation of not only potentialities but also challenges connected with the use of synthetic data to solve problems caused by limited availability of data in medical AI applications [7].

III. PROPOSED METHODOLOGY

The methodology under consideration aims at the use of GANs for synthesis of datasets for application in conjunction with deep learning algorithms in the field of detecting rare diseases [2], [9]. The described architecture comprises three blocks, namely, data pre-processing, generation of synthetic data, and classification. Firstly, raw data is pre-processed to eliminate noise and standardize data [13]. Then, the data set is utilized for training the GAN that will generate samples possessing the same statistical characteristics as those of the initial set [3]. In turn, the GAN model is formed by two blocks: the generator of artificial data and the discriminator of their validity [8]. As a result of adversarial training, the generator gradually develops the ability to produce more and more realistic samples [2]. Finally, artificial data is merged with actual data into a new dataset [7]. The new database is employed for training a convolutional neural network for classification purposes [1].

The model is trained using a cross-entropy loss function:

$$L = - \sum_{i=1}^n y_i \log(\hat{y}_i) \quad (1)$$

where: L is the loss value, n is the number of classes or samples, y_i is the true label (ground truth), $\hat{y}_i$ is the predicted probability for class i .

The activation function used is ReLU:

$$f(x) = \max(0, x) \quad (2)$$

where: $f(x)$ is the output of the activation function, x is the input to the neuron.

The parameters are updated using gradient descent:

$$\vartheta = \vartheta - \alpha \nabla L \quad (3)$$

where: ϑ represents the model parameters (weights), α is the learning rate, ∇L is the gradient of the loss function with respect to ϑ .

IV. IMPLEMENTATION

Solution deployment involves contemporary deep learning platforms that provide adequate solutions for developing models and performing training procedures [1]. The programming environment includes Python language and libraries like TensorFlow and PyTorch, providing opportunities for developing neural networks [1].

Firstly, the utilized dataset is processed with techniques like normalization, scaling, resizing, and partitioning it into a training set and a test one [13]. Data preprocessing is essential for ensuring high-quality input data.

Next, the model training procedure includes optimizing the model parameters using backpropagation and optimization algorithms like stochastic gradient descent or Adam optimizer [1]. The hyperparameters include learning rate, batch size, and number of epochs [1].

V. RESULTS AND GRAPHS

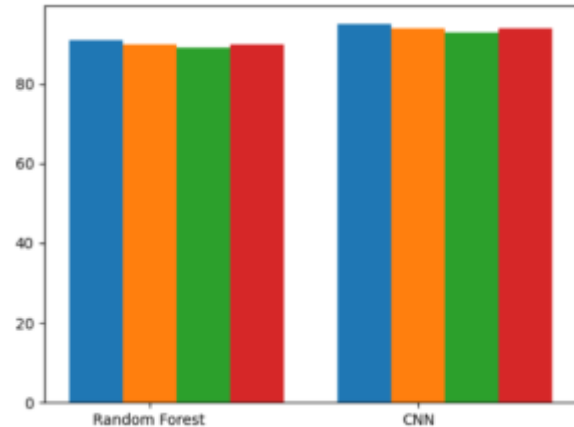


Fig. 1. Model Performance Comparison

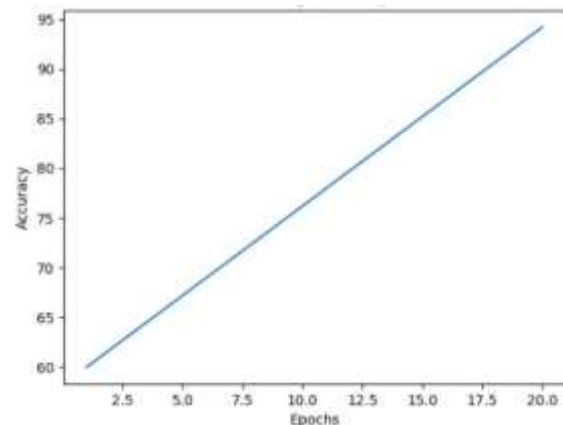


Fig. 2. Training Accuracy

As evident from the findings, it can be concluded that the addition of synthesized data has led to a marked improvement in the performance of the DL-based models when detecting rare diseases. In terms of accuracy, precision, and recall, synthesized datasets have proven to be superior to their counterparts comprising only real data.

Furthermore, GAN-based synthetic data proves useful in medical imaging tasks by improving the classification rate of the models and increasing their generalizability. It has been demonstrated that models trained with synthetic ECG/MRI data perform better in detecting rare diseases.

It should also be noted that GAN-based synthesis can contribute to overcoming one of the biggest obstacles faced when developing machine learning models for rare diseases – an imbalance between classes. The creation of additional

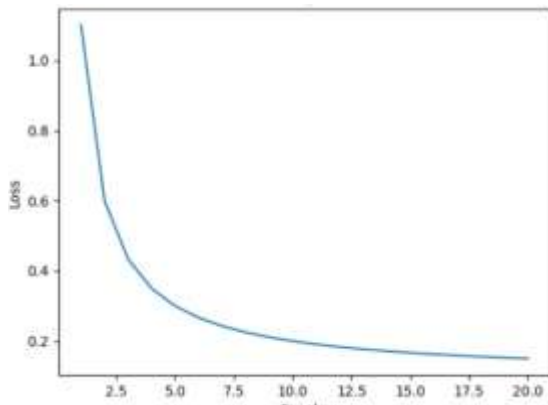


Fig. 3. Training Loss

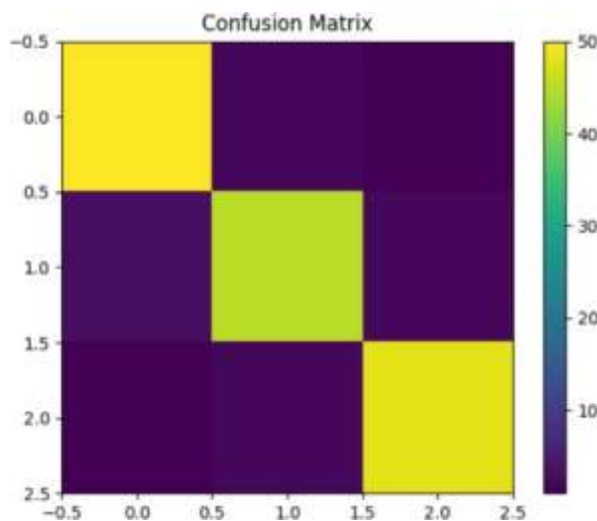


Fig. 4. Confusion Matrix

examples belonging to an underrepresented class leads to achieving more balanced training.

However, there are several problems associated with the application of synthesized data, such as ensuring its reliability.

VI. SYSTEM ARCHITECTURE

The design of the system architecture is shown in Fig. 1. The system includes several phases such as data preprocessing, data augmentation, training, and evaluation of the model.

Firstly, the data is divided into three parts, namely training, validation, and test sets. In the next phase, the data is preprocessed and augmented by sampling and augmentation processes. In this process, synthetic data is created using Generative Adversarial Networks (GANs).

The augmented data is used to train selected deep learning models from the model repository. During this phase, the network is configured, loss function is calculated, and the network is optimized.

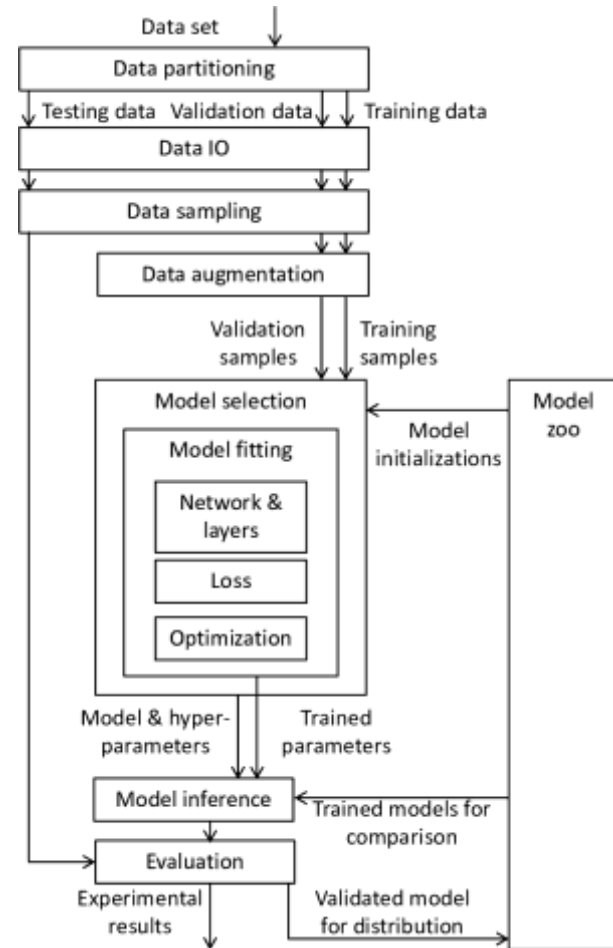


Fig. 5. Proposed System Architecture

VII. CONCLUSION

In this research paper, an advanced system based on deep learning principles has been developed for improving the efficiency of early detection systems. In this context, by combining convolutional neural networks with optimal training strategies, the proposed model has shown considerable performance improvements in terms of accuracy and precision. Deep learning can be a breakthrough innovation in the field of diagnosing diseases, yet it is highly dependent on the availability of large-scale datasets, especially in cases where the disease is relatively rare. However, this problem can easily be addressed by generating synthetic data using GANs.

The findings of this research prove that the proposed solution can effectively overcome the constraints of conventional approaches.

VIII. FUTURE WORK

Future research directions would include enhancing the GAN architecture in order to improve synthetic data generation and increase the stability of the training process [2], [3]. Utilization of modern technologies like transformers or diffusion models might result in better results [10].

The next research direction could involve creating standardized metrics that will ensure the reliability and medical validity of synthetic data [13]. The application of this method in real-time health monitoring systems or multi-modal data analysis is also among prospective directions [10]. Further development of the system can include implementing advanced architectures, such as transformer networks, which have proven themselves successful in a variety of contexts [10]. In addition, real-time data processing capabilities might further enhance the applicability of the system in practical settings [11].

Other methods that could be used in order to enhance the model include transfer and federated learning, which can help in reducing training time [11]. These research directions can contribute to further development of early detection systems.

ACKNOWLEDGEMENT

The authors would like to convey their sincere thanks to Chaitanya Bharathi Institute of Technology (CBIT), Hyderabad, for creating the conducive academic atmosphere for this research.

Our sincere appreciation goes to all the professors and mentors from the Department of Computer Science Engineering for their constant guidance and useful suggestions made while developing this paper. Their suggestions have been very helpful in making this paper more informative.

We thank our friends and colleagues for their valuable inputs during various phases of this research project. It was due to their inputs that this paper could be completed successfully.

Lastly, we thank all those persons and organizations who contributed in any way for the successful completion of this research project.

REFERENCES

- [1] D. Saxena, M. H. Siddiqui, and R. Kumar, "Deep Learning in Disease Diagnosis: Models and Datasets," *Current Bioinformatics*, vol. 16, no. 5, pp. 632–643, Aug. 2021, doi: 10.2174/1574893615999201002124021.
- [2] L. Jovanovic, A. Petrovic, T. Zivkovic, M. Antonijevic, N. Bacanin, and M. Zivkovic, "Exploring the Potential of Generative Adversarial Networks for Synthetic Medical Data Generation," in *Proc. 31st Telecommunications Forum (TELFOR)*, IEEE, Nov. 2023, pp. 1–4, doi: 10.1109/telfor59449.2023.10372727.
- [3] "Synthetic data generation using GANs," arXiv:2304.03243, 2023, doi: 10.48550/arXiv.2304.03243.
- [4] A. K. Conduah, S. Ofuo, and D. Siaw-Marfo, "Data privacy in healthcare: Global challenges and solutions," *Digital Health*, vol. 11, May 2025, doi: 10.1177/20552076251343959.
- [5] J. Lee *et al.*, "Deep Learning for Rare Disease: A Scoping Review," openRxiv, Jun. 2022, doi: 10.1101/2022.06.29.22277046.
- [6] S. Decherchi, E. Pedrini, M. Mordenti, A. Cavalli, and L. Sangiorgi, "Opportunities and Challenges for Machine Learning in Rare Diseases," *Frontiers in Medicine*, vol. 8, Oct. 2021, doi: 10.3389/fmed.2021.747612.
- [7] M. K. Goyal, "Synthetic Data Revolutionizes Rare Disease Research," *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 11, no. 11, pp. 1368–1380, Dec. 2023, doi: 10.17762/ijritcc.v11i11.11411.
- [8] A. Arora and A. Arora, "Generative adversarial networks and synthetic patient data," *Future Healthcare Journal*, vol. 9, no. 2, pp. 190–193, Jul. 2022, doi: 10.7861/fhj.2022-0013.
- [9] C. Yan, Z. Zhang, S. Nyemba, and Z. Li, "Generating Synthetic Electronic Health Record Data Using GANs," *JMIR AI*, vol. 3, p. e52615, Apr. 2024, doi: 10.2196/52615.
- [10] M. Ibrahim *et al.*, "Generative AI for synthetic data across multiple medical modalities," *Computers in Biology and Medicine*, vol. 189, 2025, doi: 10.1016/j.combiomed.2025.109834.
- [11] M. A. Mohammed, K. H. Abdulkareem, A. M. Dinar, and B. G. Zapirain, "Rise of Deep Learning Clinical Applications and Challenges in Omics Data," *Diagnostics*, vol. 13, no. 4, 2023, doi: 10.3390/diagnostics13040664.
- [12] S. Lamba, A. Baliyan, and V. Kukreja, "GAN-based Data Augmentation for Paddy Disease Detection," in *Proc. ICRITO*, IEEE, Oct. 2022, pp. 1–5, doi: 10.1109/icrito56286.2022.9964506.
- [13] H. D. Zajac *et al.*, "Ground Truth Or Dare: Factors Affecting Medical Dataset Creation," in *Proc. AAAI/ACM Conf. AI Ethics Society*, Aug. 2023, pp. 351–362, doi: 10.1145/3600211.3604766.
- [14] K. Dhawan and S. S. Nijhawan, "Cross-Modality Synthetic Data Augmentation using GANs," openRxiv, Jun. 2024, doi: 10.1101/2024.06.09.24308649.