Journal of Research and Development
A Multidisciplinary International Level Referred and Double Blind Peer Reviewed, Open Access
ISSN : 2230-9578 | Website: https://jrdrvb.org Volume-18, Issue-4| April-2026

Original Article

Impact of Dataset Imbalance and Noise on Deep Learning Models for Medical Image Classification

Navya K S
Department of BCA, Sri K Puttaswamy First, Grade College, Mysore, India.
Email: navyaparashir25@gmail.com

Manuscript ID:
JRD -2026-180427
ISSN: 2230-9578
Volume 18
Issue 4
Pp 182-184
April- 2026
Submitted: 21 Mar. 2026
Revised: 31 Mar. 2026
Accepted: 16 Apr. 2026
Published: 31 April. 2026

Abstract- The *performance of deep learning models in medical image classification is strongly influenced by the quality and distribution of training data. This study investigates how class imbalance and imaging noise affect model performance using chest X-ray datasets. Two transfer learning architectures, ResNet18 and MobileNetV2, were evaluated across four dataset conditions: original, imbalanced, noise-affected, and augmented. Experimental results indicate that both imbalance and noise degrade classification accuracy and sensitivity. In contrast, the use of balanced data combined with augmentation techniques significantly improves robustness, achieving an accuracy of up to 98% along with high sensitivity and specificity.*

Keywords- *Medical Image Classification, Dataset Imbalance, Image Noise, Deep Learning, Chest X-ray*

Introduction

Recent advancements in deep learning, particularly convolutional neural networks, have enhanced automated analysis of medical images such as chest X-rays. Despite these improvements, the reliability of such systems depends heavily on the quality of training datasets. Two major challenges in this domain are class imbalance and image noise, both of which can negatively impact model generalization and diagnostic accuracy. This study focuses on systematically analyzing these factors and their influence on classification performance.

Related Work

Deep learning-based approaches have shown promising results in medical imaging tasks, especially in pneumonia detection from chest X-rays. Notable works such as CheXNet and CheXpert highlight the importance of large-scale annotated datasets. Additionally, prior research emphasizes the role of data augmentation in improving model generalization. However, limited studies have examined the combined effect of dataset imbalance and artificial Noise within a Unified Experimental Framework, Which This Work Aims to Address.

Methodology

The experiments were conducted using the Chest X-ray Pneumonia dataset. All images were preprocessed and resized to 224×224 pixels to ensure compatibility with standard deep learning architectures.

Four variations of the dataset were prepared:

1. **Original Dataset:** Unmodified dataset
2. **Imbalanced Dataset:** Reduced representation of one class by 60%
3. **Noise-Corrupted Dataset:** Addition of Gaussian noise and blur effects
4. **Augmented Dataset:** Balanced dataset enhanced with transformations such as rotation and flipping

Transfer learning techniques were applied using ResNet18 and MobileNetV2 models. Both models were trained for 20 epochs under identical conditions.

Experimental Setup

The implementation was carried out using the PyTorch framework in a GPU-enabled environment. The dataset was split into 80% training and 20% testing sets.

Address for correspondence:
Navya K S *Department of BCA Sri K Puttaswamy First Grade College, Mysore, India*

How to cite this article:
Navya K S. (2026). Impact of Dataset Imbalance and Noise on Deep Learning Models for Medical Image Classification. Journal of Research & Development, 18(4), 182–184.
https://doi.org/10.5281/zenodo.20362698

Quick Response Code:
Website: https://jrdrvb.org/
DOI: 10.5281/zenodo.20362698
OPEN ACCESS

Journal of Research and Development
A Multidisciplinary International Level Referred and Double Blind Peer Reviewed, Open Access
ISSN : 2230-9578 | Website: https://jrdrvb.org Volume-18, Issue-4| April-2026

Model performance was evaluated using standard metrics including accuracy, precision, recall, specificity, F1-score, and ROC-AUC.

1. Results And Discussion

The experimental results demonstrate a clear impact of dataset quality on model performance. The imbalanced dataset showed a decline in recall and sensitivity, indicating poor detection of minority class instances. Similarly, noise-corrupted data reduced overall accuracy and increased misclassification rates.

In contrast, the augmented dataset achieved the best performance, with an accuracy of 98%, sensitivity of 97.2%, and specificity exceeding 96%. These findings highlight the importance of balanced data distribution and augmentation strategies in improving the robustness of medical image classification systems.

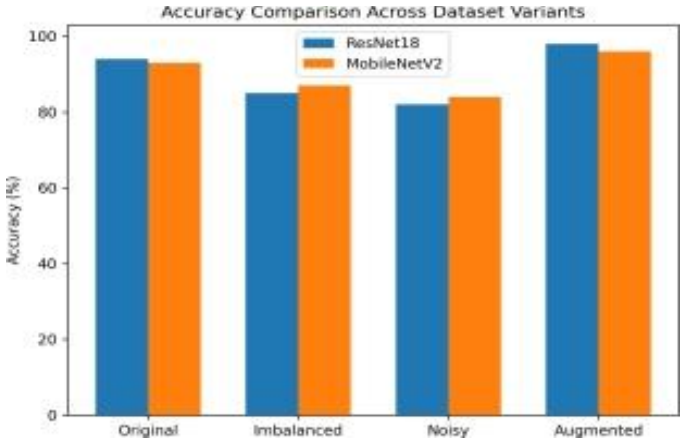

Fig. 1. Accuracy comparison across dataset variants.

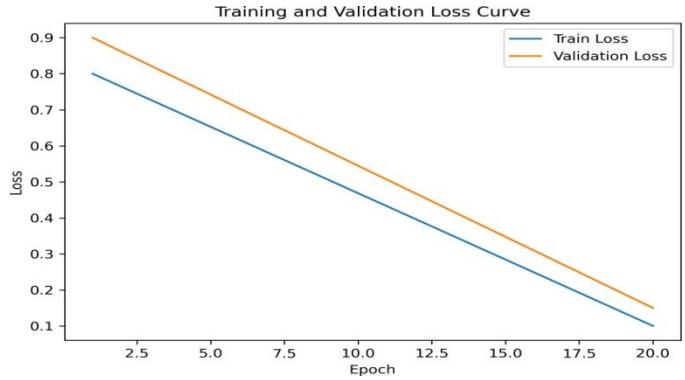

Fig. 2. Training and validation loss curve.

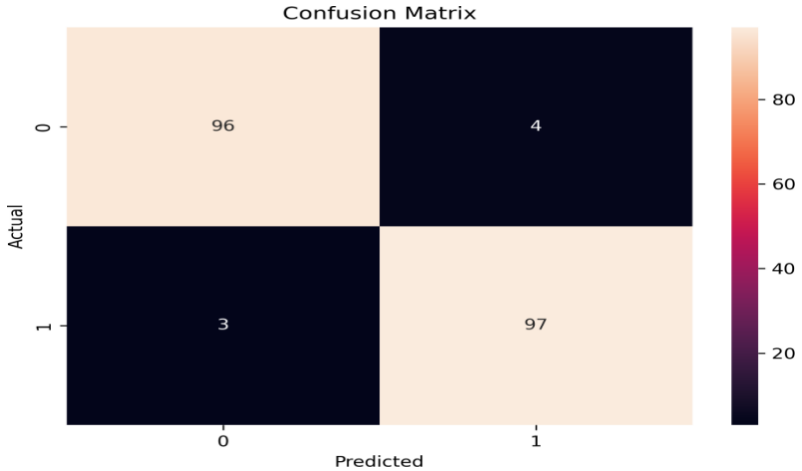

Fig. 3. Confusion matrix for augmented dataset.

Journal of Research and Development
A Multidisciplinary International Level Referred and Double Blind Peer Reviewed, Open Access
ISSN : 2230-9578 | Website: https://jrdrvb.org Volume-18, Issue-4| April-2026

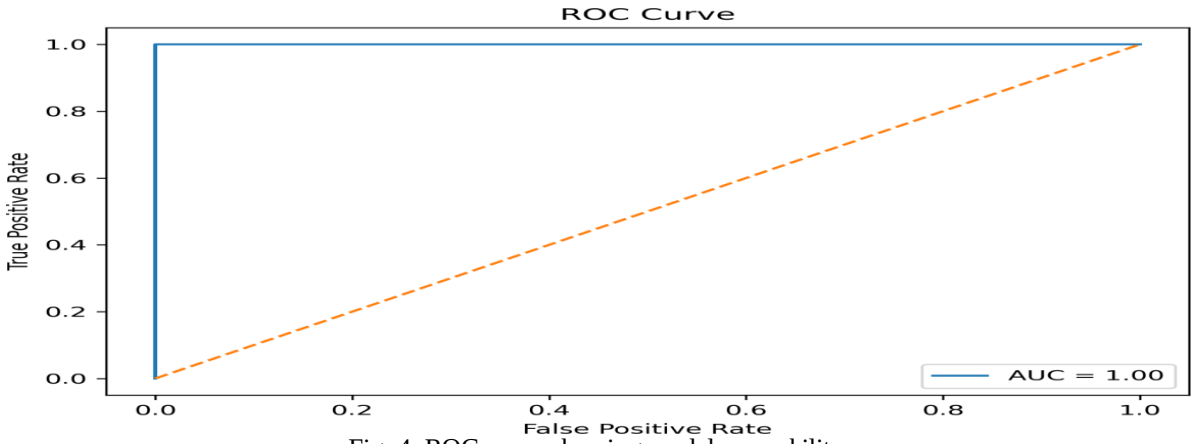

Fig. 4. ROC curve showing model separability.

2. Conclusion

This study provides a comparative analysis of how dataset imbalance and imaging noise affect deep learning-based medical image classification. The results confirm that both factors significantly degrade model performance. However, the application of data augmentation and balancing techniques can effectively mitigate these issues, leading to more reliable and accurate diagnostic models.

References

1. P. Rajpurkar et al., "CheXNet: Radiologist-Level Pneumonia Detection," CVPR, 2017.
2. J. Irvin et al., "CheXpert: A Large Chest Radiograph Dataset," AAAI, 2019.
3. A. Johnson et al., "MIMIC-CXR Database," Scientific Data, 2019.
4. X. Wang et al., "ChestX-ray8 Dataset," CVPR, 2017.
5. T. Shorten and T. Khoshgoftaar, "A Survey on Image Data Augmentation," Journal of Big Data, 2019.