(ANPA Conference 2025)
ISSN: 2392-473X(Print), 2738-9537 (Online)

Research Article ©Nepal Physical Society
DOI: https://doi.org/10.3126/jnphyssoc.v11i1.87420

Machine learning and machine learned prediction in chest X-ray images

Shereiff M. Garrett,[1] Abhinav Adhikari,[2] Sarina Gautam,[3] Da'Shawn M. Morris,[1] Laxmi Shah,[4] and Chandra M. Adhikari[1, a)]

[1)] *Department of Chemistry, Physics and Materials Science, Fayetteville State University, Fayetteville, NC 28301, USA*
[2)] *Department of Mathematical and Statistical Sciences, University of Nebraska Omaha, Omaha, NE 68182, USA*
[3)] *Jack Britt High School, Fayetteville, NC 28306, USA*
[4)] *Pediatric Intensive Care Unit, Kanti Children's Hospital, Kathmandu 44600, Nepal*

[a)] *Corresponding author:cadhikari@uncfsu.edu*

Abstract. Machine learning and artificial intelligence are fast-growing fields of research in which data is used to train algorithms, learn patterns, and make predictions. This approach helps to solve seemingly intricate problems with significant accuracy without explicit programming by recognizing complex relationships in data. Taking an example of 5824 chest X-ray images, we implement two machine learning algorithms, namely, a baseline convolutional neural network (CNN) and a DenseNet-121, and present our analysis in making machine-learned predictions in predicting patients with ailments. Both baseline CNN and DenseNet-121 perform very well in the binary classification problem presented in this work. Gradient-weighted class activation mapping shows that DenseNet-121 correctly focuses on essential parts of the input chest X-ray images in its decision-making more than the baseline CNN.

Received: August 10, 2025; **Revised**: November 11, 2025; **Accepted**: November 17, 2025

Keywords: Machine learning, Artificial intelligence, Convolutional neural network, DenseNet121, Pneumonia detection, Grad-CAM

1. INTRODUCTION

Systems that can be expressed and explained by known physical, mathematical, or logical laws are better understood through theoretical models, as they offer precise, interpretable, and generalizable predictions, providing direct insight into cause-and-effect relationships. Some systems are so intricate, with unclear dynamics of their parameters, that presenting input-output relationships in a closed form of a mathematical model is difficult or not feasible at all, or theories are impractical and/or incomplete. As machine learning (ML) can adapt to real-world messiness and make predictions using Artificial intelligence (AI), a data-driven approach can be a wise choice to tackle such a system, provided a sufficiently large dataset is available. Disease diagnosis is one of many areas where ML and AI have great potential to revolutionize the diagnosis process by reviewing immense amounts of images and performing image classification.

Neural networks (NNs), inspired by the human brain, have demonstrated human-level performance across multiple task domains, although the NN is based on statistical measures and relies on human-simulated intelligence programmed and controlled through algorithms. One node of an NN mimics the brain's smallest measured unit, such as a voxel. The input layer of the NN mimics the raw data received by sensory organs and/or the primary cortex, such as the eyes, skin, or ears. NN's hidden layers imitate the intermediate processing of the brain, as done by the neocortex, to extract hierarchical features analogous to the deeper processing steps that occur in the brain. Output layer of NN copies the functionality of motor cortex, aka brain's decision regions, to produce final action or classification akin to behavior/output in brain [1]

The article aims at two objectives. First, we briefly review two different approaches to NNs: the baseline Convolutional Neural Network (CNN) [2] and a densely connected DenseNet-121 CNN [3], used in this study, which takes advantage of an open-source deep-learning framework called PyTorch [4], a Python library. PyTorch utilizes tensors as a fundamental data structure, enabling researchers to modify the network's behavior in real-time.

Second, we implement these neural network techniques to analyze chest X-ray images with the aim of accurately predicting the presence of an ailment in the chest and demonstrating the capabilities of ML and AI.

Pneumonia is one of the leading causes of illness and death worldwide, especially in more vulnerable populations [5, 6, 7]. According to the "National Center for Health Statistics" report of the Centers for Disease Control and Prevention (CDC) for 2023, pneumonia is the eleventh leading cause of death in the USA. For the 0-19 age group, it was listed as the 9th leading cause of death and the second among the disease-related deaths behind COVID-19 in the year 2019 [8]. In 2023, approximately 1.5 million patients visited emergency departments with pneumonia as the primary diagnosis, caused by infectious organisms, and pneumonia resulted in the deaths of 41,210 patients, which is 1.23 per 10,000 population [9].

A Rapid and accurate diagnosis from chest X-rays is critical and can be challenging. It is worth noting that deep Learning has shown considerable promise in automating and supporting the clinical interpretation of X-ray images. Reliable and interpretable AI tools can assist radiologists in accurately identifying any ailment present in images, aiding in diagnosis.

The baseline CNN is a custom-designed neural network that utilizes convolutional layers for feature extraction and fully connected layers for classification. It usually uses a gray image of dimensions $1 \times 32 \times 32$ (or $1 \times 128 \times 128$), or a color image of dimensions $3 \times 32 \times 32$ (or $3 \times 128 \times 128$), where the first numbers 1 and 3 denote the number of channels (1 for gray and 3 for RGB), while 32s and 128s are the height and width of the images in pixels. A densely connected CNN, such as DenseNet121, however, has an input size of $3 \times 224 \times 224$. It features a deep and pre-trained architecture with multiple layers. In this project, we compare the effectiveness of these two deep learning models for pneumonia detection on chest X-ray images, assessing both classification performance and model interpretability.

2. METHODS

Taking the chest X-ray dataset from Kaggle, which contains 5824 JPEG images with two labels, namely, NORMAL and PNEUMONIA [10], we investigate the effectiveness of ML models to predict pneumonia images correctly. This work will support researchers focusing on automating ML/AI-based methods to detect and classify human diseases from medical images.

The dataset contains images that were taken during pediatric patients' routine clinical care. The dataset is moderately imbalanced, having slightly more than 3 times as many images of patients with PNEUMONIA as NORMAL. To correctly address this issue, data augmentation

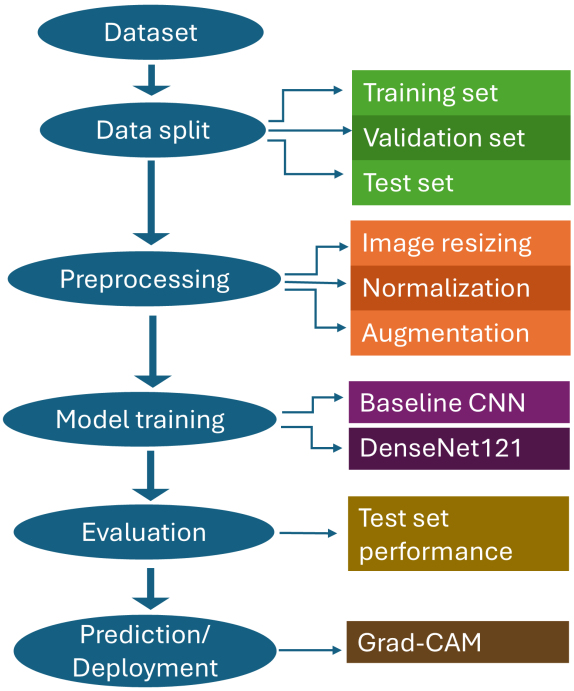

FIGURE 1: Work flow.

was performed. Data augmentation included small random brightness and contrast adjustments (±5%) and rotations (±3°), preserving the medical realism of images. Each image was auto-oriented, resized, and normalized for consistency. All low-quality or unreadable scans of chest X-rays were removed to have a better data-driven analysis. The dataset was split into training, validation, and test sets for fair evaluation, with distributions of 88%, 8%, and 4%, respectively. Preprocessing involves the application of auto-orientation, resizing, normalizing each image to the same size, and images augmentation. These image data were used to train separate baseline CNN and DenseNet 121 deep learning architectures and were evaluated using various metrics, including confusion matrices [11, 12] , receiver operating characteristic (ROC) curves [13], and the area under the curve (AUC) in ROC curves [14]. DenseNet-121 was pretrained on ImageNet, requiring 224×224 RGB input images, while the baseline CNN was trained from scratch at 128×128 resolution as a lightweight benchmark. This distinction highlights architectural learning capacity rather than direct resolution advantage.

The test performance was tested for both models. To interpret the observations and predictions done by these models, the Gradient-weighted Class Activation Mapping (Grad-CAM) [15] model interpretability technique was used. Grad-CAM exploits the gradients of the target class percolating through the final convolutional layer to cre-

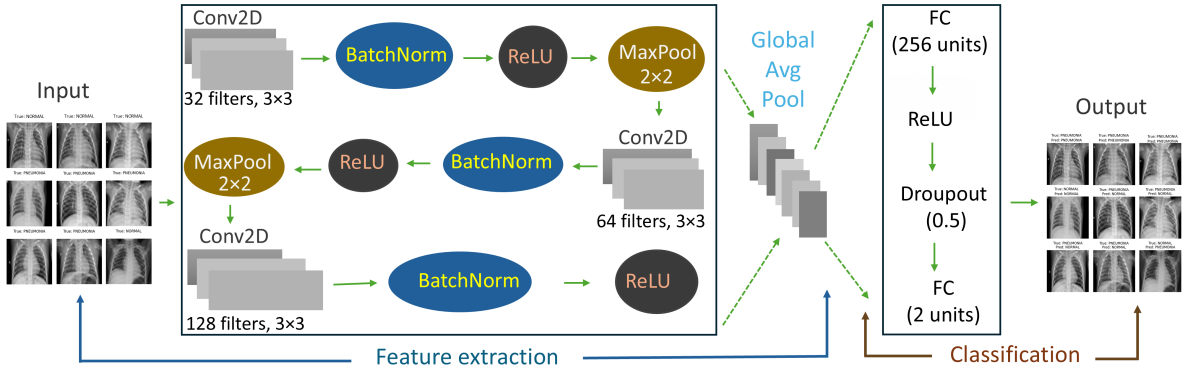

FIGURE 2: A simple architecture of a baseline CNN.

ate a heat-map highlighting important regions in the input image. Figure 1 shows the workflow followed in this binary classification with ML and AI. Before presenting the results, a brief discussion on each of the NNs employed to learn chest X-rays in this article is in order.

Baseline CNN

A baseline CNN is a simple yet effective neural network model used to establish a minimum standard for performance in a specific deep learning task, typically in image classification, object detection, segmentation, and/or segmentation. It usually acts as a benchmark for evaluating more complex models, not only by providing insight into whether deeper models are really helping, but also helps to validate data pipeline, loss functions, and training setup, and in turn provides a starting point for improvement of model via architecture tuning, augmentation, regularization, and hyperparamenter adjustment if needed.

The baseline CNN follows the sequential connectivity in such a way that each layer builds upon the immediately preceding feature map. Consequently, the input in the nth layer can be written as

$$x_n = H_n\left(x_{n-1}\right), \tag{1}$$

where H_n is a composite function applied at nth layer. The basic building block of baseline CNN's H_n function is Conv $\rightarrow$ ReLU $\rightarrow$ Pooling, where Conv and ReLU denote convolution and rectified linear unit, respectively. The Conv extracts spatial patterns such as edges, textures, etc, ReLU enables the model to learn non-linear representations, and Pooling reduces spatial dimensions and overfitting. Figure 2 shows a schematic of the architecture used for baseline CNN in this study. We have used adaptive moment (Adam) optimizer with learning rate of 0.001, batch size of 32, and 30 epoches. Cross-entropy loss function is employed tomeasure the difference between true label and model predicted probability distributions.

DenseNet-121

Unlike baseline CNNs, DenseNet connects each layer to every other layer in a feed-forward fashion such that input x_n to each layer n receives feature maps from all preceding layers [3]. The architecture of each layer, receiving input from all previous layers, promotes feature reuse and efficient gradient flow. DenseNet connectivity formula can be written as

$$x_n = H_n\left(\left[x_0, x_1, x_2, \cdots, x_{n-1}\right]\right), \tag{2}$$

where H_n is a composite function consisting of batch normalization (BN), ReLU, and convolution. DenseNet features a bottleneck design (BN $\rightarrow$ ReLU $\rightarrow$ Conv 1×1) preceding the main 3×3 convolution, aiming to reduce the number of input channels and thereby decrease computational cost while increasing the depth of the feature map without a significant increase in parameters.

The DenseNet-121 architecture consists of 121 layers, comprising an initial convolution, four dense blocks, three transition layers, and a fully connected (FC) output layer. The four dense blocks have 6, 12, 24, and 16 pairs of bottleneck layers with 256, 512, 1024, and 1024 output channels, respectively. The layers count as follows: 1 initial convolution + 2$\times$ (6 dense block-1 + 12 dense block-2 + 24 dense block-3 + 16 dense block-4) + 3 transitions + 1 FC to make 121.

One may wonder whether the baseline CNN can outperform DenseNet-121. A large number of layers makes networks of baseline CNN deeper, which can model complex functions, increasing the receptive field and abstraction level. Thus, theoretically, matching or even exceeding the DenseNet-121 in accuracy from a baseline CNN

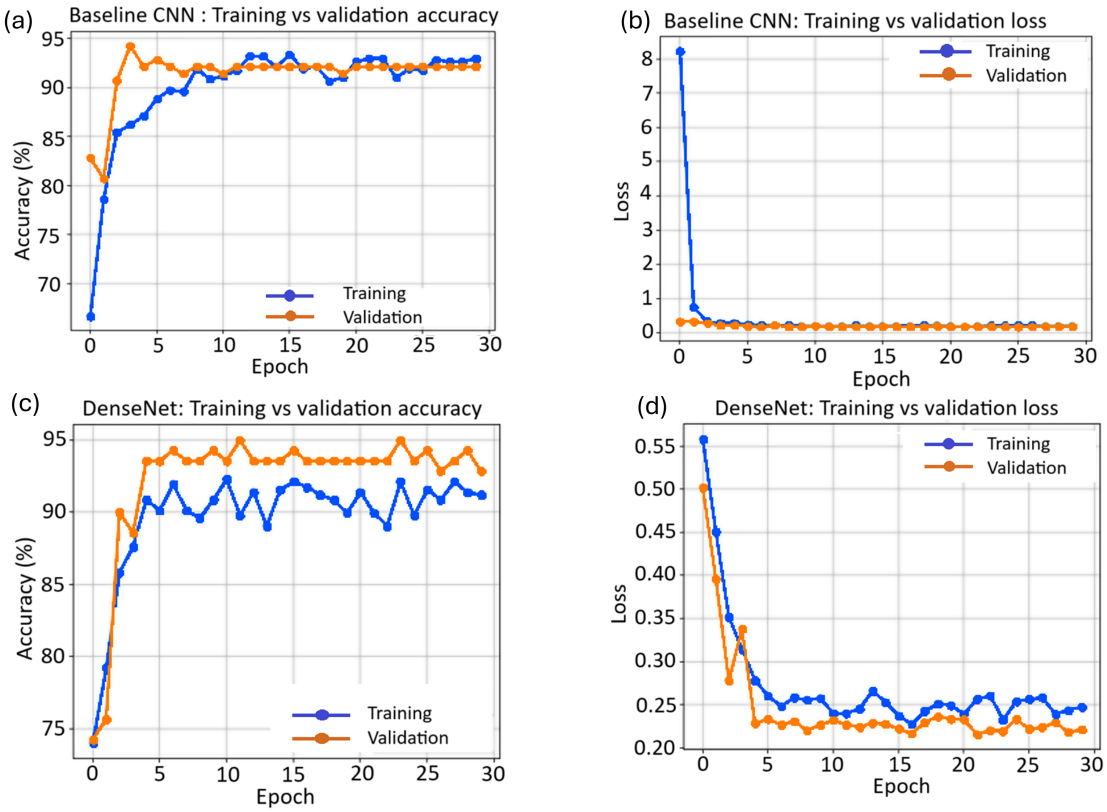

FIGURE 3: Training and validation accuracies and losses as a function of epochs for baseline CNN and Densenet-121.

with a large number of layers is possible. However, a simple deep CNN usually fails due to various reasons. First, gradients are back-propagated layer by layer in deep NN, as a result, the gradients can vanish in early layers, resulting in ineffective learning, or the gradient blows up, making unstable updates. Increasing the depth makes the network weaker at propagating the learning signals back to early layers. Second, each layer of the baseline CNN discards the features of the previous layers, causing it to learn again from scratch. Thus, making the baseline CNN deeper is essentially relearning a similar pattern repeatedly, which wastes learning capacity and makes training even more challenging. Third, increasing the depth in baseline CNN increases the number of convolutional filters, each with its own weights, causing a rapid growth of the memory usage, leading to a slow training process. In summary, simply increasing the layers in the baseline CNN does not make it smarter than DenseNet-121, as increasing the layers introduces many architectural issues that deteriorate its performance. Instead, using dense connections, the DenseNet-121 design performs smartly, maximizing gradient flow, feature reuse, and parameter efficiency.

The model's performance depends on many factors, such as sample size, image labeling complexity, ease of learning features, number of parameters, binary vs. multi-class task, etc. If images are of moderate complexity and the dataset is medium-sized (a few thousand images), a practical diagnostic rule is to evaluate the performance of both models. For a small dataset of low-complexity images with low intra-class variation and low variability, the baseline CNN performs better than Densenet-121. One cannot expect the same across different scenarios when the dataset is large and contains diverse, noisy images.

3. NON-LINEARITY: A KEY IN NEURAL NETWORK

Although linear models are easy to train and interpret due to the proportional or hyperplane relationships between input and output, the patterns that field-collected data in real-world systems often exhibit are complex in nature and non-separable by linear boundaries, making non-linearity a key factor to consider in their study. Without any activation and incorporation of non-linearity, an NN is essentially a single-layer linear model, regardless

of the NN's depth [16]. Consequently, a non-linear model that is absent or suppressed faces severe limitations on its ability to solve real-world challenges, such as image classification, medical diagnosis, and facial detection. Non-linearity enables NNs to evolve into universal function approximators, empowering them to model complex patterns and dependencies that exist among many features, the non-linear decision boundaries the system has, and high-dimensional data relationships that arise from the interactions among many input features.

In NNs, non-linearity is introduced through activation functions such as sigmoid, Tanh, ReLU, Swish, Mish, softplus, exponential linear unit (ELU), etc, which provoke bending and curvature in the NN's computational process. For example, the sigmoid function $f(x) = (1 + e^{-x})^{-1}$ provides S-shaped saturation varying value in the range of 0 to 1. Tanh function $f(x) = \tanh(x) = (e^{x} - e^{-x})/(e^{x} + e^{-x})$ gives a similar-shaped activation to that of sigmoid but ranges from -1 to 1, centering at zero. ReLU function $f(x) = \max(0,x)$ introduces sparse and non-saturating non-negative activation in the range of 0 to ∞. Swish $f(x) = x/(1 + e^{-x})$ and Mish $f(x) = x \cdot \tanh[\ln(1 + e^{x})]$ activation functions give smooth activation in the range of $-\infty$ to ∞. Softplus is a smooth approximation of ReLU. In need of a zero-centered smooth non-saturating function, ELU defined by $f(x) = x$, if $x \geq 0$ and $f(x) = \alpha(e^{x} - 1)$ otherwise, is used. For reference, see [17] for details on activation functions.

Non-linear activation functions embedded in the composite function H_n transform the weighted sum of inputs in a neuron into an output that is no longer linear in relationship. Non-linearity empowers the NN to solve real-world, linearly non-separable, complex patterns, thereby expanding its representational capability. Furthermore, non-linearity enables NN hierarchical feature learning, extracting abstract features from all previous layers in DenseNet-121 or just the previous one in baseline CNN.

4. RESULTS AND DISCUSSION

Data were analyzed using an open-source ML library called PyTorch, developed by Meta AI, used for building and training deep learning models designed from first principles to support an imperative and Python programming style [4]. To diagnose how the model's performance evolves over training epochs, we visualized the comparison of training and validation accuracy and loss versus epochs, as presented in Fig. 3. Accuracy measures the ratio of total correct predictions to the total predictions made by the model. It can be expressed as a percentage by multiplying it by 100. The validation accuracy is slightly improved, resulting in a reduced loss with the DenseNet-121 model compared to the baseline CNN. The loss is a measure of error. It assesses how bad the models' predictions are. Note that accuracy and loss need not sum to unity or 100%. The comparative analysis of Figs. 3(b) and 3(d) shows that DenseNet-121 is marginally better than the baseline CNN.

We have used confusion matrices, also known as error matrices, as a performance evaluation tool for our binary classification problems. The diagonal entries of the matrices (true positive and false negative) represent all instances in which the model correctly predicted the classes, namely, "NORMAL" and "PNEUMONIA". The off-diagonal entries (false positives and false negatives) indicate the confusion level or error in prediction, where one class is mislabeled as another. The baseline CNN correctly predicted 81% of the normal chest X-rays as normal and 93% of the pneumonia cases as pneumonia. 19% of normal cases were falsely predicted as pneumonia, and 7% of the pneumonia cases were also predicted as normal (see Fig. 4). Predicting normal chest X-rays as normal is 12% less for DenseNet-121 than baseline CNN, while the DenseNet-121 predicted pneumonia cases as pneumonia slightly better than baseline CNN.

The ROC curves render a visual and quantitative measure to examine the ability of a model to differentiate between classes across all possible classification thresholds. The ROC curves visualize the variation of true positive rate (TPR) with false positive rate (FPR), where TPR = TP/ (TP+FN) and FPR = FP/ (FP+TN) with TP, FP, TN, and FN being true positive, false positive, true negative, and false negative counts, respectively [18, 19]. TP count measures how many of the predicted positives are truly positive, while FP count measures how many of the predicted positives are actually negative. The area under the curve (AUC) abridges the model's overall ability to distinguish its constituent classes into a single number. Any model with an AUC of less than 0.5 is considered a bad model, while a model with an AUC of 1.0 is a perfect model that flawlessly discriminates its classes. In real-world practice, an AUC of 1.0 is only theoretically feasible, and models with an AUC of 1.0 may indicate issues such as data leakage or overfitting, or an algorithmic error. It is also important to note that ML model struggles or overfit when data is scarce, and overfitting is always a red flag in ML. Any models with an AUC greater than 0.9 are considered excellent-performing models, with higher values indicating better performance. Figure 5 shows the ROC curves for baseline CNN and DenseNet-121 models. The AUC values for both cases fall within the excellent performing range, indicating that both models are working excellently in distinguishing between chest X-rays of normal patients and those with pneumonia.

A NN learns features automatically from raw data, adjusting its internal parameters to map inputs to correct outputs. Grad-CAM is a powerful visualization tool that generates a heatmap highlighting the critical regions in an image, aiding in class prediction. The Grad-CAM is

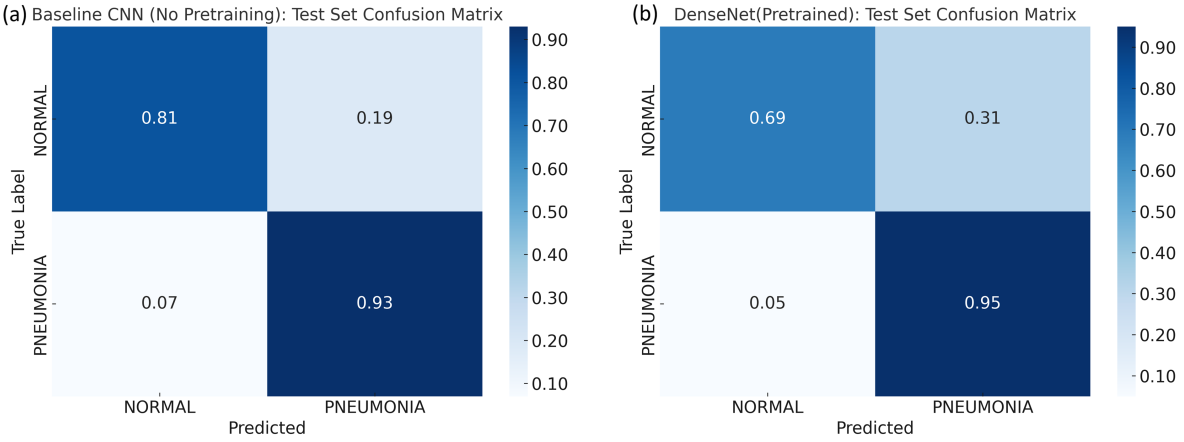

FIGURE 4: Confusion matrices, for (a) baseline CNN and (b) DenseNet-121, showing the values of predicted label for each true label normalizing the values for each class to unity.

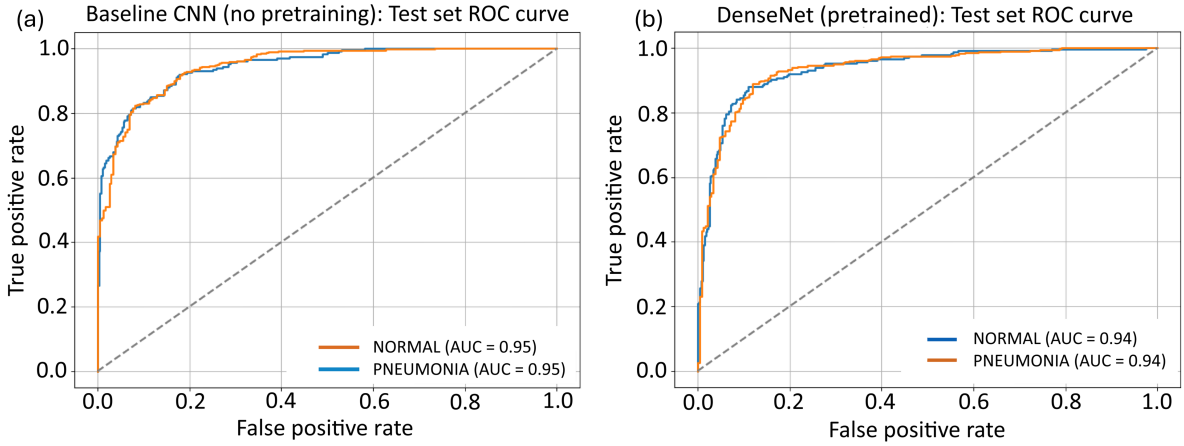

FIGURE 5: Receiver operating characteristic (ROC) curves showing true positive rate (TPR) on the vertical axis and false positive rate (FPR) on the horizontal axis for baseline CNN (a) and DenseNet-121 (b), respectively. Numbers in the legend are area under the curve (AUC). The dashed line is for random guessing with AUC = 0.5.

based on gradients of the class score with respect to feature maps. As the input image goes through the NN, the model collects feature maps from a convolutional layer and evaluates the gradient of the output score of the target with respect to the selected feature maps. The NN computes weights and averages out gradients spatially to get important weights for each feature map. Then, weighted activation maps are created, and an overlay heatmap is applied to the original image. Strongly influential areas are represented by bright red, while less influential or non-influential regions are represented by dark blue. Grad-CAM helps to understand why the model predicted a particular class, which not only reveals which parts of the input image were most important for a class decision, but also builds trust in AI decisions by providing the reason behind the prediction. Both baseline CNN and DenseNet-121 can localize regions of interest in chest

TABLE I: Model performance on the held-out test set (4% test split ≈ 230 images).

models	accuracy (%)	AUC	precision	recall	F1-score
baseline CNN	86.2	0.931	0.911	0.864	0.887
DenseNet-121	84.5	0.935	0.829	0.946	0.884

X-rays, while DenseNet-121 produces more focused and clinically plausible attention maps (see Figs. 6 and 7). In baseline CNN, being trained from scratch, Grad-CAMs are more diffuse and sometimes highlight irrelevant areas, indicating less precise feature learning and, in some instances, misfire even on true positives. In contrast, the DenseNet-121 demonstrates higher consistency in correct predictions and focuses on meaningful lung zones, high-

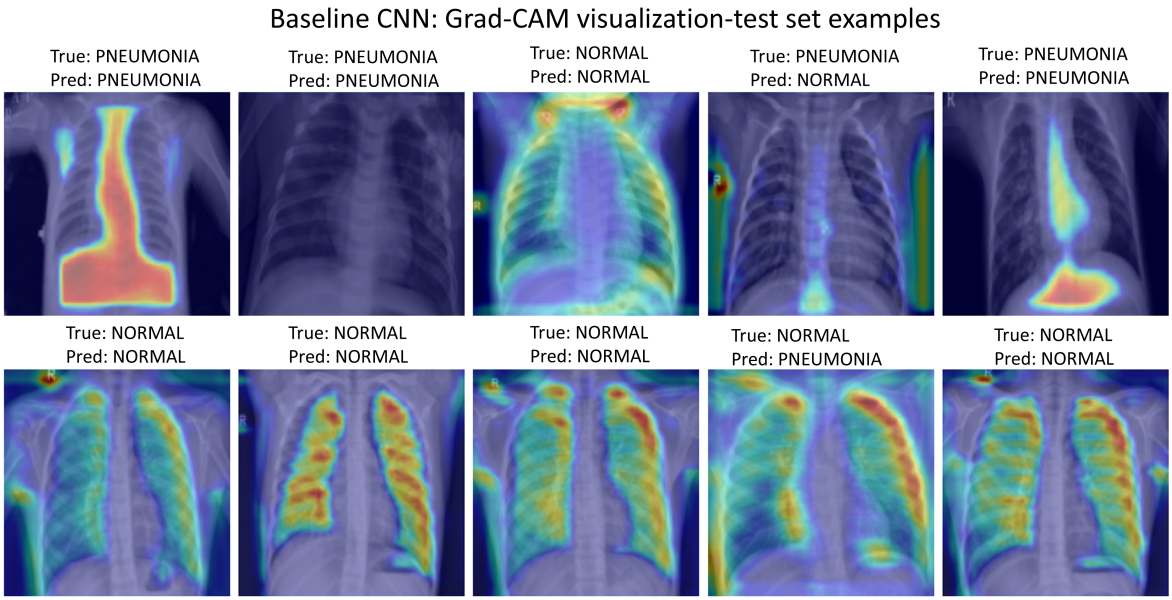

FIGURE 6: Some example images of Grad-CAM visualization in baseline CNN. Bright red spots are strongly influential and dark-blue are less or no influential areas.

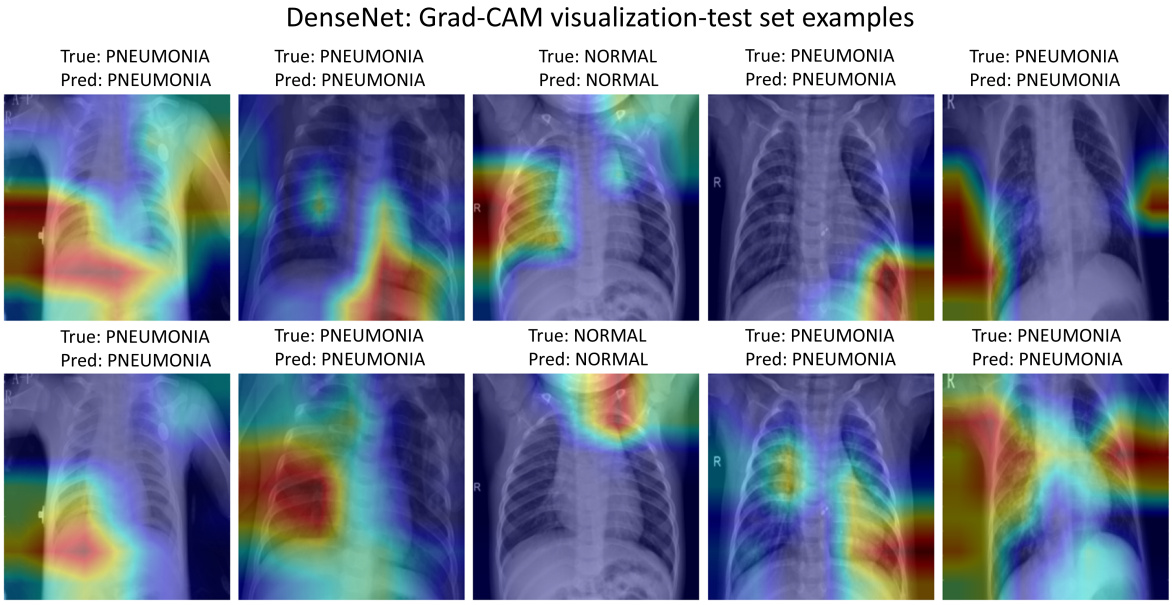

FIGURE 7: Some example images of Grad-CAM visualization in DenseNet-121. Bright red spots are strongly influential and dark-blue are less or no influential areas.

lighting disease-relevant regions more clearly, supporting better interpretability for pneumonia detection.

Table I summarizes the models' performance on the Held-Out Test Set (4% test split $\approx$ 230 images). The baseline CNN performed comparably to DenseNet-121 on our dataset. Being a medical dataset consisting of moderate-complexity images with low intra-class variation and low variability, and binary classification, made the baseline CNN's performance comparable to that of DenseNet-121. Our results are consistent with previous pneumonia detection studies such as see Refs. [20, 21]. This comparison confirms that both our baseline CNN and DenseNet-121 perform competitively with established benchmarks.

NN is a powerful means of learning from data and mak-

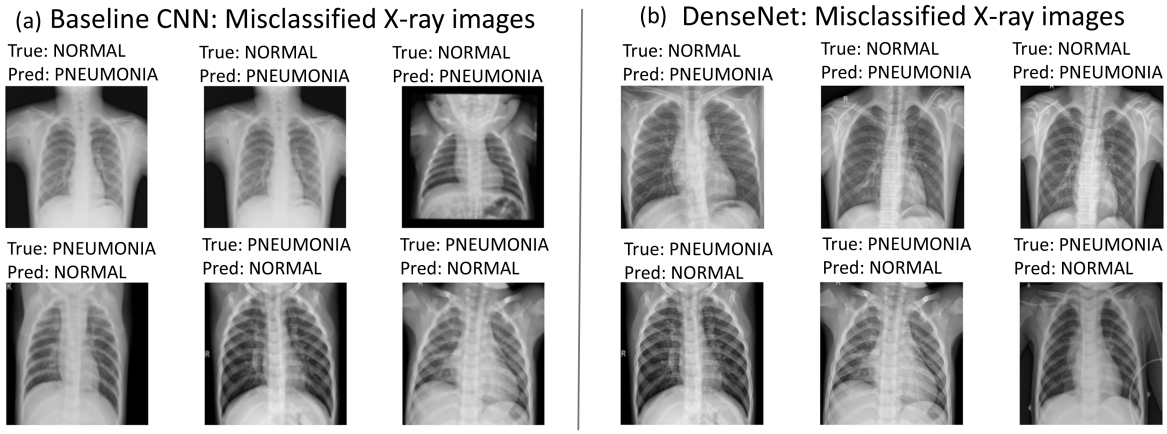

FIGURE 8: Some randomly selected misclassified chest X-ray images from (a) baseline CNN and (b) DenseNet-121 models.

ing decisions to classify classes present in the dataset correctly. However, NN models can still misclassify due to various reasons such as visual and statistical similarity of data classes, imbalanced data, underfitting or overfitting, wrong hyperparameters, sensitivity of the data to noise, class overlap, insufficient data, noisy training data, and so on [22]. Some randomly selected Baseline CNN and DenseNet-121 misclassified chest X-rays are shown in Fig. 8. The chest X-ray dataset is large enough, with more than 5000 images for only two classes. Data has been split into the well-accepted data set distribution in NN. There is no class overlap, and we did not perform hyperparameter tuning in the model. Early stopping and mixed-precision training were used to improve performance and prevent overfitting. The misclassification in our model is mainly due to the inherent visual similarity in the images, despite their actual dissimilarity, as they depict different people taken at different times by different professionals.

It is essential to note that, although deep learning NN models, as presented in this work, such as the baseline CNN and DenseNet-121, can be astonishingly accurate, human experts are still crucial for reviewing, validating, and contextualizing AI results before accepting the predictions made by them, as AI lacks real-world understanding and clinical judgment and cannot make judgments about whether inputs are out-of-distribution. For a given input, models will still provide a prediction. Still, the conditions of the data may have changed, such as the equipment being different or the patient suffering from a different but related disease. Furthermore, the model may make a correct prediction with high confidence, but if it looks at the text or a wrong spot, rather than the organ to focus on, in this case, the lungs. A field expert can recognize exceptions, identify new patterns, and detect critical errors but algorithms may fail on rare or unexpected inputs.

We acknowledge that the small test set and possible patient overlap could influence statistical stability. Future work will apply $k-$fold cross-validation, patient-level deduplication, and bootstrap confidence intervals for more robust testing.

5. CONCLUSION

The architecture of two NN models, namely, baseline CNN consisting of only a few layers and DenseNet-121 comprising 121 layers, is briefly reviewed. A discussion on the power of nonlinearity in NN is presented. Using a publicly available dataset of chest X-rays containing over 5,000 images, the capabilities of baseline CNN and DenseNet-121 NNs were analyzed in detecting images of patients with ailments.

Both the baseline CNN and DenseNet-121 models accurately detected patients with pneumonia. To reach this conclusion, we observed variations in training and validation accuracies and losses as a function of epoch, confusion matrices, and ROC curves. We used Grad-CAMs as a tool to visually explain the predictions of NNs and better understand why the model predicted a certain class for a chosen image. In baseline-CNN, Grad-CAMs are more dispersed and sometimes highlight irrelevant areas, indicating less precise feature learning and, in some instances, misfire even on true positives. The DenseNet-121 demonstrates higher consistency in correct predictions and focuses on meaningful lung zones, highlighting disease-relevant regions more clearly, which supports better interpretability for pneumonia detection, making it more logical from a human perspective.

AI models make predictions even if some inputs may fall into out-of-distribution. Being a data-driven model, ML and ML-based decisions are based on data that trains

the model, rather than the domain knowledge that a human expert adds. NN models serve as powerful tools, assisting in prediction and informed judgment, and human experts make certain that their use is safe, responsible, and intelligent.

EDITORS' NOTE

This manuscript was submitted to the Association of Nepali Physicists in America (ANPA) Conference 2025 for publication in the special issue of the Journal of Nepal Physical Society.

REFERENCES

1. S. Nonaka, K. Majima, S. C. Aoki, and Y. Kamitani, "Brain hierarchy score: Which deep neural networks are hierarchically brain-like?" Science, **24**(9), 103013 (2021). https://doi.org/10.1016/j.isci.2021.103013
2. L. D. Marchi and L. Mitchell, "Hands-on Neural Networks: Learn How to Build and Train Your First Neural Network Model Using Python," 7th ed. (Packt Publishing, Birmingham, UK, 2019).
3. G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, **2261** (2017). https://doi.org/10.1109/CVPR.2017.243.
4. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An imperative style, high-performance deep learning library." In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Advances in Neural Information Processing Systems, **32**, 8024 (2019). https://dl.acm.org/doi/10.5555/3454287.3455008
5. C. P. Bondarchuk, B. Grobman, A. Mansur, and C. Y. Lu, "National trends in pneumonia-related mortality in the United States, 1999-2019. Infectious diseases (London, England), **57**(1), 56 (2025). https://doi.org/10.1080/23744235.2024.2390180
6. C. M. Obianyo, N. Muoghalu, E. M. Adjei, A. N. Njoku, N. L. Okoro, O. B. Sulaiman, I. M. Nwabuokei, and V. U. Barrah, "Trends and Disparities in Pneumonia-Related Mortality in the U.S. Population: A Nationwide Analysis Using the CDC WONDER Data" Cureus, **17**(5), e83371 (2025). https://doi.org/10.7759/cureus.83371
7. L. Santo, S. M. Schappert, J. J. Ashman, "Emergency department visits for influenza and pneumonia: United States, 2016–2018" NCHS Data Brief, no 402. Hyattsville, MD: National Center for Health Statistics. 2021. https://doi.org/10.15620/cdc:102795.
8. S. Flaxman, C. Whittaker, E. Semenova, T. Rashid, R. M. Parks, A. Blenkinsop, H. J. T. Unwin, S. Mishra, S. Bhatt, D. Gurdasani, and O. Ratmann, "Assessment of COVID-19 as the Underlying Cause of Death Among Children and Young People Aged 0 to 19 Years in the US." JAMA network open, **6**(1), e2253590 (2023). https://doi.org/10.1001/jamanetworkopen.2022.53590
9. Centers for Disease Control and Prevention, National Center for Health Statistics. National Vital Statistics System, Mortality 2018-2023 on CDC WONDER Online Database, released in 2024. Accessed at http://wonder.cdc.gov/ucd-icd10-expanded.html on Jul 15, 2025 2:58:34 AM.
10. T. Mohamed, *Chest X-Rays Dataset*. Roboflow Universe, Roboflow, November 2022. Available at: https://universe.roboflow.com/mohamed-traore-2ekkp/chest-x-rays-qjmia. Accessed: July 19, 2025.
11. S. Swaminathan and B. R. Tantri, "Confusion Matrix-Based Performance Evaluation Metrics," African Journal of Biomedical Research, **27** 4023 (2024). https://doi.org/10.53555/AJBR.v27i4S.4345
12. S. V. Stehman, "Selecting and interpreting measures of thematic classification accuracy," Remote Sensing of Environment, **62**, 77 (1997). https://doi.org/10.1016/S0034-4257(97)00083-7
13. M. R. J. Junge and J. R. Dettori, "ROC Solid: Receiver Operator Characteristic (ROC) Curves as a Foundation for Better Diagnostic Tests," Global spine journal, **8**(4), 424 (2018). https://doi.org/10.1177/2192568218778294
14. F. S. Nahm, "Receiver operating characteristic curve: overview and practical use for clinicians," Korean journal of anesthesiology, **75**(1), 25 (2022). https://doi.org/10.4097/kja.21209
15. R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," Int J Comput Vis **128**, 336 (2020). https://doi.org/10.1007/s11263-019-01228-7
16. E. Filippi-Mazzola and E. C. Wit, "Modeling non-linear effects with neural networks in Relational Event Models," Social Networks, **79**, 25 (2024). https://doi.org/10.1016/j.socnet.2024.05.004.
17. S. R. Dubey, S. K. Singh, and B. B. Chaudhuri, "Activation functions in deep learning: A comprehensive survey and benchmark," Neurocomputing, **503**, 92 (2022). https://doi.org/10.1016/j.neucom.2022.06.111
18. A. A. R. A. Omar, B. Soudan, and A. Altaweel, "A comprehensive survey on detection of sinkhole attack in routing over low power and Lossy network for internet of things," Internet of Things, **22**, 100750 (2023). https://doi.org/10.1016/j.iot.2023.100750
19. D.S. Burke, J. F. Brundage, R. R. Redfield, J. J. Damato, C. A. Schable, P. Putman, R. Visintine, and H. I. Kim, "Measurement of the False Positive Rate in a Screening Program for Human Immunodeficiency Virus Infections," N. Engl. J. Med. **319**, 961 (1988). https://www.nejm.org/doi/abs/10.1056/NEJM198810133191501
20. P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M.P. Lungren, A. Y. Ng, "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning" arXiv:1711.05225 (2017). ¯https://doi.org/10.48550/arXiv.1711.05225
21. D.S. Kermany, M. Goldbaum, W. Cai, C.C.S. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, J. Dong, M. K. Prasadha, J. Pei, M. Y.L. Ting, Jie Zhu, C. Li, S. Hewett, J. Dong, I. Ziyar, A. Shi, R. Zhang, L. Zheng, R. Hou, W. Shi, X. Fu, Y. Duan, V. A.N. Huu, C. Wen, E. D. Zhang, C. L. Zhang, O. Li, X. Wang, M. A. Singer, X. Sun, J. Xu, A. Tafreshi, M. A. Lewis, H. Xia, K. Zhang, "Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning," Cell, **172** (5), 1122 (2018). https://doi.org/10.1016/j.cell.2018.02.010
22. M. Nartker, Z. Zhou, and C. Firestone, "When will AI misclassify? Intuiting failures on natural images," Journal of vision, **23**(4), 4 (2023). https://doi.org/10.1167/jov.23.4.4