Original Paper

Comparison of Deep Learning Approaches Using Chest Radiographs for Predicting Clinical Deterioration: Retrospective Observational Study

Mahmudur Rahman[1], PhD; Jifan Gao[2], MS; Kyle A Carey[3], MPH; Dana P Edelson[3], MD, MS; Askar Afshar[1], MS; John W Garrett[2,4], PhD; Guanhua Chen[2], PhD; Majid Afshar[1,2], MD, MSCR; Matthew M Churpek[1,2], MD, MPH, PhD

[1]Department of Medicine, University of Wisconsin-Madison, Madison, WI, United States

[2]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI, United States

[3]Department of Medicine, University of Chicago, Chicago, IL, United States

[4]Department of Radiology, University of Wisconsin-Madison, Madison, WI, United States

Corresponding Author:

Matthew M Churpek, MD, MPH, PhD
Department of Medicine
University of Wisconsin-Madison
610 Walnut St
Madison, WI, 53792
United States
Phone: 1 608-262-9564
Email: mchurpek@medicine.wisc.edu

Abstract

Background: The early detection of clinical deterioration and timely intervention for hospitalized patients can improve patient outcomes. The currently existing early warning systems rely on variables from structured data, such as vital signs and laboratory values, and do not incorporate other potentially predictive data modalities. Because respiratory failure is a common cause of deterioration, chest radiographs are often acquired in patients with clinical deterioration, which may be informative for predicting their risk of intensive care unit (ICU) transfer.

Objective: This study aimed to compare and validate different computer vision models and data augmentation approaches with chest radiographs for predicting clinical deterioration.

Methods: This retrospective observational study included adult patients hospitalized at the University of Wisconsin Health System between 2009 and 2020 with an elevated electronic cardiac arrest risk triage (eCART) score, a validated clinical deterioration early warning score, on the medical-surgical wards. Patients with a chest radiograph obtained within 48 hours prior to the elevated score were included in this study. Five computer vision model architectures (VGG16, DenseNet121, Vision Transformer, ResNet50, and Inception V3) and four data augmentation methods (histogram normalization, random flip, random Gaussian noise, and random rotate) were compared using the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC) for predicting clinical deterioration (ie, ICU transfer or ward death in the following 24 hours).

Results: The study included 21,817 patient admissions, of which 1655 (7.6%) experienced clinical deterioration. The DenseNet121 model pretrained on chest radiograph datasets with histogram normalization and random Gaussian noise augmentation had the highest discrimination (AUROC 0.734 and AUPRC 0.414), while the vision transformer having 24 transformer blocks with random rotate augmentation had the lowest discrimination (AUROC 0.598).

Conclusions: The study shows the potential of chest radiographs in deep learning models for predicting clinical deterioration. The DenseNet121 architecture pretrained with chest radiographs performed better than other architectures in most experiments, and the addition of histogram normalization with random Gaussian noise data augmentation may enhance the performance of DenseNet121 and pretrained VGG16 architectures.

JMIR AI 2025;4:e67144; doi: 10.2196/67144

Keywords: chest X-ray; critical care; deep learning; chest radiographs; radiographs; clinical deterioration; prediction; predictive; deterioration; retrospective; data; dataset; artificial intelligence; AI; chest; patient; hospitalized

Introduction

Clinical deterioration is common in hospitalized patients and can lead to adverse outcomes, including increased morbidity and mortality if not identified and managed properly [1]. The early detection of patient deterioration and timely intervention can improve patient outcomes [2]. Various early warning scores (EWS) have been developed to identify the deterioration risk by monitoring different clinical variables, and the implementation of machine-learning EWS, such as the electronic cardiac arrest risk triage (eCART) score, has been associated with improved mortality [3-6]. Current EWS rely on structured data, such as vital signs and laboratory values, to predict clinical deterioration and ignore other data modalities that could potentially enhance prediction accuracy [7]. This results in lower detection and higher false-positive rates for these scores that could be mitigated by incorporating additional modalities [8].

Because respiratory failure is a common cause of clinical deterioration, the use of computer vision models with chest radiographs is a promising direction for improving EWS performance [9]. Although traditional computer vision models have historically been used to analyze chest radiographs, prior work on chest radiographs is limited to identifying specific diagnoses [10-12]. In some recent studies, chest radiographs are used to detect lung disease [[13,14]], acute respiratory distress syndrome [15], pneumonia [16,17], tuberculosis [18,19], and COVID-19 [20]. However, to facilitate other tasks with comprehensive machine understanding, chest X-ray interpretation models are being more commonly used with the help of computer vision and transformer-based natural language processing models [21,22]. The advancements in predictive analytics with deep learning methods have led to increased capabilities to extract meaningful information from medical images, including chest radiographs [23]. However, deep learning models have never been trained with chest radiographs to predict clinical deterioration outside the intensive care unit (ICU). There are numerous deep learning architectures for chest radiograph prediction models, such as VGG16, ResNet50, DenseNet121, and Vision Transformer, and the performance of these models is unknown for this specific task. Additionally, there are different data augmentation techniques available to further enhance the performance of a vision model by improving model generalization, but it is unknown whether these data augmentation techniques would improve the performance of the prediction model for this task.

To address these knowledge gaps, the objective of this study was to compare different computer vision architectures and augmentation methods with chest radiographs for predicting clinical deterioration. Our training pipeline incorporates extensive hyperparameter tuning through Bayesian optimization and validates the generalizability of models in a separate hold-out test set. The findings of our experiments have important implications for researchers developing computer vision deep learning models for clinical applications with chest radiographs.

Methods

Ethical Considerations

The study protocol was reviewed and approved by the University of Wisconsin Institutional Review Board (approval #2019-1258). This study was a secondary analysis of limited HIPAA data from hospital electronic health records. The study was approved with a waiver of informed consent.

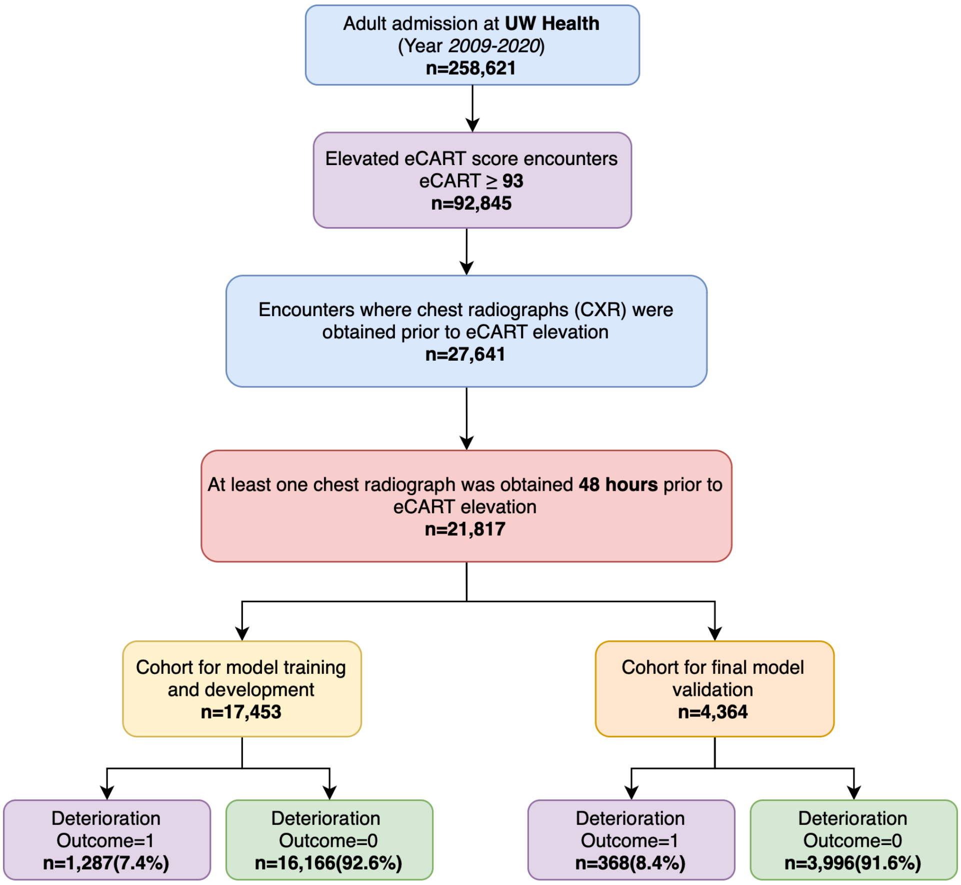
All direct identifiers of patients whose data were used in this study were de-identified prior to analysis to ensure participants privacy and confidentiality. Minimal necessary identifiable information was accessed or stored during the study beyond possible HIPAA data in clinical notes, radiological images, and real dates.

Participants did not receive any compensation for this data analysis, as no new data were collected and no direct contact with participants occurred.

Study Population and Data Collection

All adult patients (age ≥18 years) hospitalized at the University of Wisconsin Health System (UW Health) between 2009 and 2020 with an elevated eCART score ≥93 (which is the threshold used in clinical practice at UW Health) on the medical-surgical wards were eligible for inclusion in this retrospective cohort study. The eCART score [3] is a validated EWS currently in clinical practice and cleared by the Food and Drug Administration that combines demographics, vital signs, and laboratory results in a gradient-boosted machine model to predict future clinical deterioration. The rationale for only including patients with an elevated score is based on creating an enriched cohort where chest radiograph models can enhance the prediction and mitigate the false-positive alerts from these scores. Furthermore, this simplifies the prediction task to a single time point, making it more feasible to compare multiple models and augmentation strategies. Patients with a chest radiograph within 48 hours before the first elevated eCART score were included in the study. Available anterior-posterior or posterior-anterior views were included in the study cohort. In addition to chest radiographs, additional study variables that were collected included patient demographics, admission time, vital signs, laboratory values, patient location, and discharge disposition, which were all collected via the clinical research data warehouse. Figure 1 shows the patient encounter flow chart for inclusion into the analytic cohort.

Figure 1. Study inclusion criteria flow diagram.


Adult admission at **UW Health**
(Year *2009-2020*)
n=258,621
Elevated eCART score encounters
eCART ≥ **93**
n=92,845
Encounters where chest radiographs (CXR) were obtained prior to eCART elevation
n=27,641
At least one chest radiograph was obtained **48 hours** prior to eCART elevation
n=21,817
Cohort for model training and development
n=17,453
Cohort for final model validation
n=4,364
Deterioration Outcome=1
n=1,287(7.4%)
Deterioration Outcome=0
n=16,166(92.6%)
Deterioration Outcome=1
n=368(8.4%)
Deterioration Outcome=0
n=3,996(91.6%)

Outcome

The study outcome of clinical deterioration was defined as a direct ward-to-ICU transfer or ward death within 24 hours of the time of the patient’s first elevated eCART score.

Data Preprocessing

The chest radiograph closest to (but before) the time of the elevated eCART score was used to predict the corresponding deterioration outcome. To address variations in image acquisition and processing protocols, all radiographs were rescaled to a uniform size of 224×224 pixels using nearest neighbor interpolation. Additionally, to address the variabilities in imaging exposure levels, pixel intensity values were normalized to a range of [0, 1] by applying min-max scaling. The clinical deterioration outcome (ie, ICU transfer or mortality within 24 hours from the prediction time point) was encoded as binary labels, with one-hot encoding used for the binary prediction task. These preprocessing steps ensured the creation of a high-quality robust dataset for training deep learning models to predict clinical deterioration from chest radiographs.

Model Development

For the prediction task, computer vision deep learning models were trained and optimized with the dataset created from the cohort. Five publicly available computer vision models were compared for our task: (1) VGG16 [24], (2) DenseNet121 [25], (3) Vision Transformer [26], (4) ResNet50 [27], and (5) Inception V3 [28]. DenseNet121 is a convolutional neural network notable for its dense connections between layers, improving efficiency and reducing risk of overfitting, and VGG16 is known for its simplicity using a series of convolutional layers with small filters followed by max pooling layers. The Vision Transformer model is based on the transformer architecture and uses the self-attention mechanism to process the images. The main rationale of adopting these computer vision models for clinical deterioration tasks is that they are widely used in other chest radiograph detection tasks in clinical setups [29-31]. In addition, these models are easy to implement, and various pretrained weights are readily available. As clinical tasks require fine-grained image understanding for different tasks, these models provide that performance with a manageable model size. However, the main shortcoming of using these models is they do not provide any generalized image understanding for explainability.

We used two different versions of the VGG16 architecture, one using randomly initialized weights (without pretraining) and the other using model weights pretrained on ImageNet [32]. Two different versions of the DenseNet121 architecture were also used: one with model weights pretrained on Imagenet [32] and one pretrained on publicly available radiograph datasets [33]. Specifically, the radiograph datasets used for pretraining consisted of the following datasets: NIH aka Chest X-ray14 [34], PC aka PadChest [35], CheX aka CheXpert [36], MIMIC-CXR [37], OpenI [38], Google [39], and RSNA Pneumonia Detection Challenge [40]. For the Vision Transformer model, we trained two models without any pretrained weights of two different sizes, one with 12 transformer blocks and another with 24 transformer blocks. We employed batch normalization layers after every block to ensure the stability of the optimization process during the model training. Figure 2 presents the overall structure of this study.

For each of the above architectures, we compared them with and without different preprocessing and data augmentation approaches. These included histogram normalization, random rotation (±15 degrees), horizontal flipping, and the addition of random Gaussian noise. Briefly, histogram normalization addresses the regional discrepancy of exposure levels in the case of some images. Additionally, given the presence of noise and artifacts during the acquisition of the radiographs, random Gaussian noise, which was implemented as 0.1 probability with 0 mean and 0.1 standard deviation, may make the models more robust to noise in the input image samples. Figure 3 shows the examples of all the augmentation methods we have used in this work.

Figure 2. Overall structure of this study. *VGG16, DenseNet121, ResNet50, and Inception V3 models were trained from randomly initialized weights and pretrained weights. Other models were trained with randomly initialized weights only.

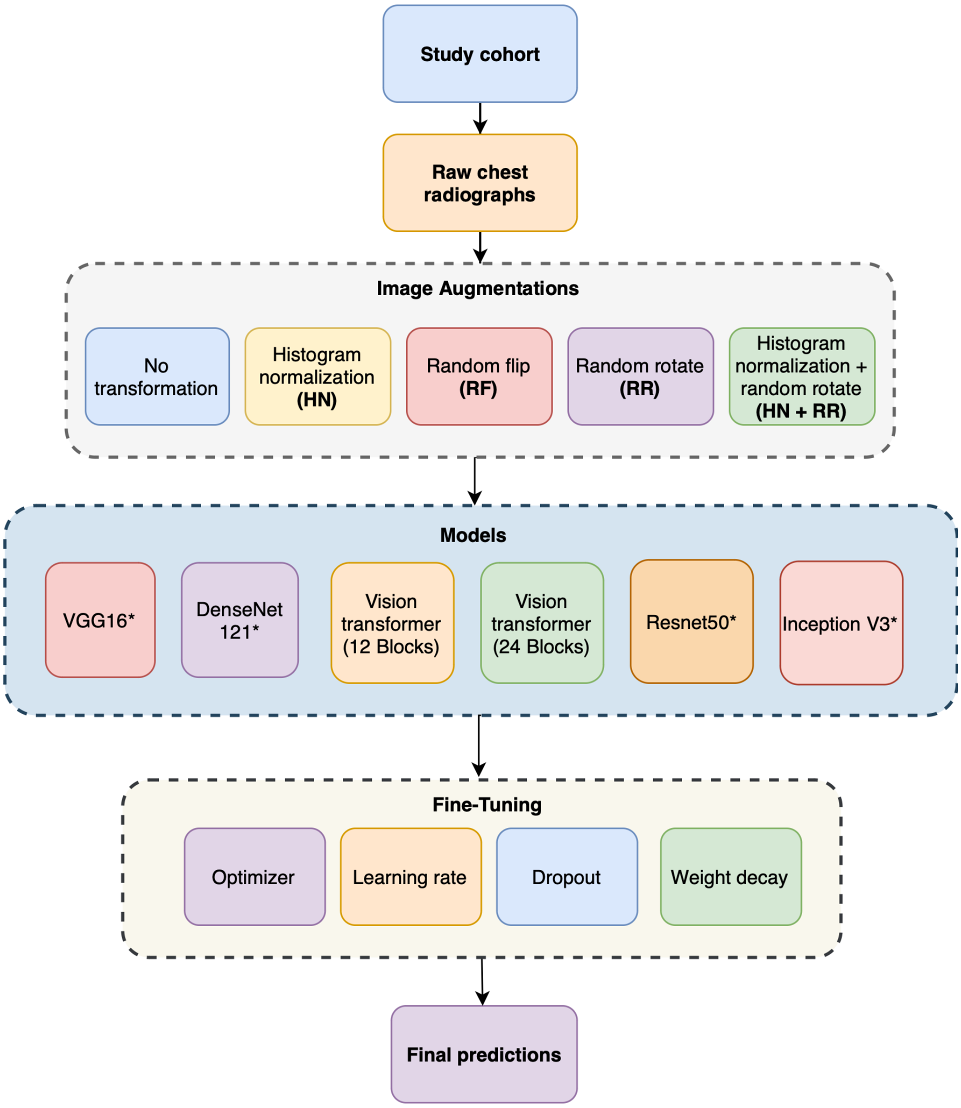
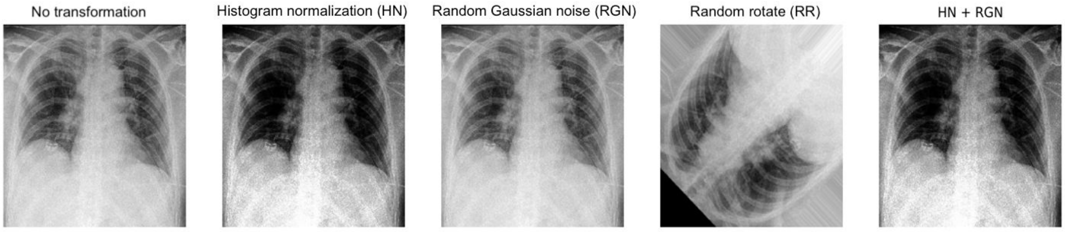


Figure 3. Examples of different image augmentation methods we have utilized. HN: histogram normalization; RGN: random Gaussian noise.



We used the Bayesian optimization algorithm to find the optimal hyperparameters that maximize the area under the receiver operating characteristic curve (AUROC). Details of the hyperparameters are presented in Multimedia Appendix 1. To make the training procedure faster, we used Ray Tune [41] to parallelize the hyperparameter search process in a multi-GPU environment. We trained the model with a randomly selected 60% of the encounters in the dataset and validated it with the development set consisting of 20% of the encounters to optimize hyperparameters and determine the final settings. The remaining 20% of the encounters were completely separated for independent final model evaluation of the optimized models as a test set. We trained the models for 20 epochs and decreased the learning rate by a factor of 0.5 in every epoch. During the training, early stopping was used if the validation AUROC failed to improve in three consecutive epochs. We used Adam mini-batch gradient descent optimization with a batch size from the search space of 32, 64, and 128.

Model Evaluation

All combinations of image augmentations and deep learning computer vision architectures for the clinical deterioration task were evaluated using the test dataset. Predicted probabilities for the deterioration outcome were calculated for every encounter during the evaluation. Model discrimination was assessed using the AUROC and its 95% CI, calculated via the DeLong method [42] as the primary metric and the area under the precision-recall curve (AUPRC) as the secondary metric. The p-values of the AUROC scores are presented in Multimedia Appendix 1. As P<.001 in all cases, our AUROC scores are statistically significant.

Data cleaning and cohort selection with descriptive analysis were conducted using Stata version 16.1 (StataCorp). We used Python version 3.8.10, along with the Monai framework version 1.2.0 (NVIDIA) and Pytorch version 2.0.0 (Facebook) to develop the deep learning models. Additionally, the AUROC score and its 95% CI were calculated using FastDeLong implementation from VMAF (Video Multimethod Assessment Fusion; Netflix) [43].

Results

Cohort Characteristics

A total of 258,621 admissions occurred during the study period, and 92,845 had an elevated eCART score. Of these, for 21,817 admissions, a chest radiograph was obtained within 48 hours of the time of the elevated score and was included in the analysis (Figure 1). The characteristics of the final cohort are presented in Table 1. The patients in the final cohort had a median age of 63 (IQR 52-74) years, with a higher likelihood of being male (56.1%, 12,249/21,817); 5.7% were black (1252/21,187). The median time to eCART score elevation from admission was 21.8 (7.1-47.6) hours and the median time to eCART score elevation from the last radiograph was 9 (7.1-47.6) hours. About 7.5% (1655/21,817) of the encounters had an outcome event, including 4.1% (893/21,817) cases of in-hospital death.

Table 1. Population characteristics of the study cohort (N=21,817).

Variable	Value
Age, years, median (IQR)	63 (52-74)
Female, n (%)	9568 (43.9)
Black race, n (%)	1252 (5.7)
Elevated eCART[a] score, n (%)	1655 (7.59)
Time to the elevated eCART score from admission, hours, median (IQR)	21.8 (7.1-47.6)
Time to the elevated eCART score from the last radiograph, hours, median (IQR)	9.0 (7.1-47.6)
In-hospital mortality, n (%)	893 (4.1)

[a] eCART: electronic cardiac arrest risk triage

Model Discrimination

The model performance AUROC and AUPRC values for all models across all image augmentation methods are presented in Tables 2 and 3, respectively, and the 95% CI of the AUROC and AUPRC are presented in Multimedia Appendix 1. Additionally, receiver operating characteristic (ROC curves and precision-recall curves are shown in Figures 4 and 5, respectively.

Table 2. Model performance area under the receiver operating characteristic curve (AUROC) with the validation dataset across different model architectures, pretrained weights, and image augmentation methods.

Model	Pretrained weights	No transformation	Histogram normalization (HN)	Random flip	Random Gaussian noise (RGN)	Random rotate	HN + RGN	Average AUROC score[a]
VGG16	Random init	0.694	0.723	0.698	0.701	0.674	0.712	0.700
VGG16	ImageNet	0.712	0.717	0.692	0.710	0.689	0.719	0.707
DenseNet121	ImageNet	0.683	0.701	0.672	0.700	0.678	0.716	0.692

Model	Pretrained weights	No transformation	Histogram normalization (HN)	Random flip	Random Gaussian noise (RGN)	Random rotate	HN + RGN	Average AUROC score[a]
DenseNet121	Radiographs	0.723	0.716	0.713	0.696	0.701	0.734	0.714
ResNet50	Random init	0.588	0.684	0.629	0.678	0.638	0.651	0.645
ResNet50	ImageNet	0.715	0.707	0.694	0.694	0.669	0.712	0.700
Inception V3	Random init	0.691	0.672	0.671	0.661	0.703	0.690	0.681
Inception V3	ImageNet	0.714	0.712	0.710	0.706	0.686	0.713	0.707
Vision Transformer (12 Blocks)	Random init	0.661	0.648	0.617	0.652	0.623	0.652	0.642
Vision Transformer (24 Blocks)	Random init	0.654	0.663	0.609	0.651	0.598	0.662	0.640
Average Score over models[b]	—[c]	0.684	0.694	0.671	0.685	0.666	0.696	—
Average Improvement[d]	—	—	0.010	–0.013	0.001	–0.028	0.012	—

[a]The average AUROC score is for a particular model over different augmentation methods

[b]The "Average score over models" row presents the average AUROC score of a particular augmentation method over different models.

[c]"—" indicates not applicable.

[d]The "Average improvement" row shows the average AUROC improvement of an augmentation method over the baseline score without any transformation.

Table 3. Model performance area under the precision-recall curve (AUPRC) scores with the validation dataset across different model architectures, pretrained weights, and image augmentation methods.

Model	Pretrained weights	No transformation	Histogram normalization (HN)	Random flip	Random Gaussian noise (RGN)	Random rotate	HN + RGN	Average AUPRC score[a]
VGG16	Random init	0.346	0.398	0.329	0.349	0.320	0.378	0.353
VGG16	ImageNet	0.371	0.403	0.306	0.343	0.311	0.389	0.354
DenseNet121	ImageNet	0.321	0.373	0.360	0.355	0.365	0.379	0.359
DenseNet121	Radiographs	0.395	0.326	0.338	0.360	0.358	0.414	0.365
ResNet50	Random init	0.135	0.229	0.147	0.243	0.215	0.174	0.191
ResNet50	ImageNet	0.405	0.378	0.357	0.320	0.288	0.344	0.349
Inception V3	Random init	0.319	0.247	0.247	0.304	0.343	0.339	0.300
Inception V3	ImageNet	0.440	0.340	0.421	0.399	0.361	0.369	0.388
Vision Transformer (12 Blocks)	Random init	0.205	0.189	0.143	0.209	0.139	0.204	0.182
Vision Transformer (24 Blocks)	Random init	0.187	0.219	0.121	0.177	0.118	0.196	0.170
Average score over models[b]	—[c]	0.313	0.310	0.277	0.306	0.282	0.319	—
Average improvement[d]	—	—	–0.003	–0.036	–0.007	–0.031	0.006	—

[a]The average AUPRC score is for a particular model over different augmentation methods

[b]The "Average score over models" row presents the average AUROC score of a particular augmentation method over different models.

[c]"—" indicates not applicable.

[d]The "Average improvement" row shows the average AUROC improvement of an augmentation method over the baseline score without any transformation.

Figure 4. Receiver operating characteristic (ROC) curve of the best-performing models in every network architecture. Actual AUROC values are included in the corresponding label. HN: histogram normalization; RGN: random Gaussian noise; AUROC: area under the receiver operating characteristic curve.

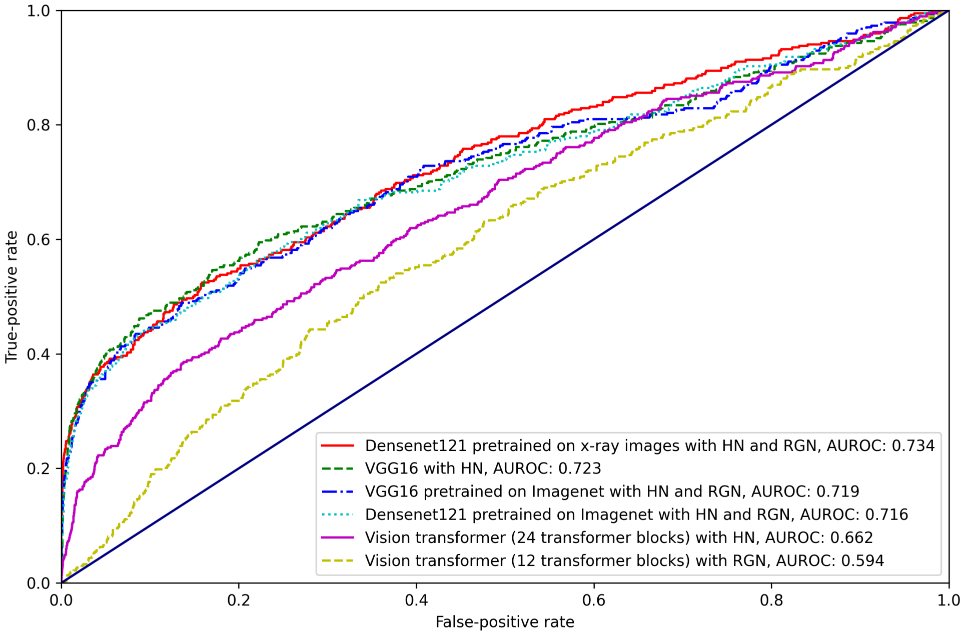
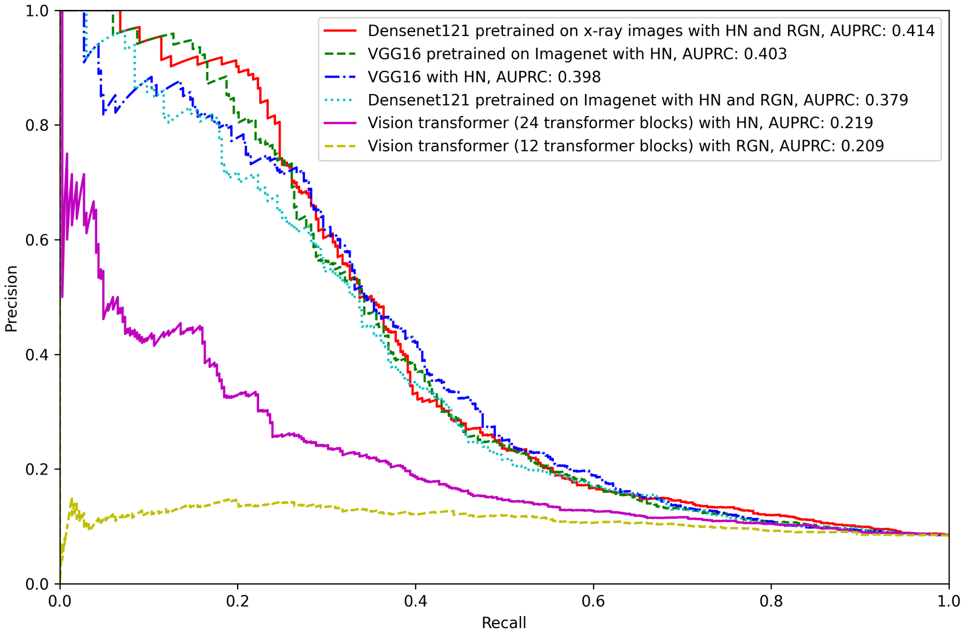


Figure 5. Precision-recall curves of the best-performing models in every network architecture. Actual AUPRC values are included in the corresponding label. Best viewed in color. HN: histogram normalization; RGN: random Gaussian noise; AUPRC: area under the precision-recall curve.



Across all architectures and augmentation combinations, the DenseNet121 model pretrained with chest radiographs and augmented with histogram normalization and Gaussian noise had the highest AUROC (0.734) across all the models. Similarly, when averaged across all augmentation methods, the DenseNet121 models pretrained with chest radiographs had a higher average discrimination than any other architecture in terms of the AUROC (0.714). The vision transformer architectures (12 and 24 transformer blocks) performed similarly to each other on average and had worse average AUROC than other models (0.642 and 0.640 for 12 and 24 transformer blocks, respectively). In terms of the AUPRC, DenseNet121 pretrained with chest radiographs and augmented with histogram normalization and Gaussian noise also had the highest performance (0.414). Accordingly, compared with other models, Inception V3 pretrained with ImageNet had the highest AUPRC (0.388) on average.

In terms of the image augmentation methods, the histogram normalization with random Gaussian noise image augmentations had the best mean AUROC (0.696) when averaged across all architectures, followed by histogram normalization augmentation alone (0.694). The random rotate augmentation had the worst average performance in terms of the AUROC (0.666). In terms of the AUPRC, histogram normalization with random Gaussian noise image augmentations also had the highest average AUPRC (0.319) across the models, and the models with no transformation alone had the next highest average AUPRC of 0.310. Unlike the AUROC results, the random flip augmentation had the worst AUPRC among all the four other augmentation methods.

Discussion

Principal Findings and Comparison With Previous Works

In this retrospective study with over 20,000 hospital admissions, we compared three deep learning computer vision architectures and four image augmentation methods for the early detection of clinical deterioration. We found that the DenseNet121 model pretrained on different publicly available chest radiographs had better discrimination than the VGG16 and Vision Transformer models based on the average AUROC metric. Among different image augmentation methods, a combination of histogram normalization and random Gaussian noise augmentations achieved higher AUROCs and AUPRCs on average than random flip and random rotate transformation. In all of the cases, we found that random flip and random rotate transformation lowered the discrimination compared to the baseline model in terms of both AUROC and AUPRC metrics. To the best of our knowledge, this is the first study to compare different computer vision models and image augmentation methods for predicting clinical deterioration outside the ICU. These findings have important implications in the field of using deep learning models to correctly identify patients showing clinical deterioration and to improve existing EWS applications in health systems.

Although DenseNet121 pretrained on chest radiographs achieved the maximum discrimination with histogram normalization and random Gaussian noise data augmentation, our investigation found multiple models exhibiting competitive performance across different data augmentation methods considering the AUROC. This may be due to our extensive hyperparameter search with Bayesian optimization that enables all models to achieve similar performances. Overall, the pretrained models performed better with respect to the models trained from scratch. This is consistent with the existing literature, as pretrained models already learned the fundamental building blocks of features (eg, lines and shades) from large number of images of the pretrained dataset [44,45]. However, as the VGG16 model was pretrained on the ImageNet [32] dataset, which is a collection of thousands of general-purpose images, and our dataset only contains chest radiographs, there may be a domain gap present in this scenario that prohibits the maximum benefits of the pretraining network. To analyze and mitigate that domain gap, we compared the performance of the DenseNet121 network pretrained on ImageNet and on a collection of the radiograph dataset. In almost all of the cases, DenseNet121 pretrained on radiographs outperformed the DenseNet121 model pretrained on ImageNet in terms of both the AUROC and AUPRC metrics. These experimental results proved our hypothesis and provided important insights into the use of pretrained networks with chest radiograph datasets. A prior study involving the classification of chest radiographs also found DenseNet networks achieving superior performance [10], which aligns with our findings. For example, Alhudhaif et al found that DenseNet201 achieved the highest discrimination in determining COVID-19 pneumonia with chest radiographs [10]. However, another work by Sitaula et al found that the VGG-16 model performed better than the DenseNet121 model for the classification of COVID-19 chest radiographs [11]. This discrepancy may be explained by differences in hyperparameter settings and the use of pretrained weight initialization. They tuned the hyperparameters manually, whereas we tuned the hyperparameters automatically with Bayesian optimization. As the DenseNet121 network is deeper than VGG-16 in terms of the number of layers, better hyperparameter tuning may enable DenseNet121 to learn more complex relationships without overfitting, hence achieving better performance than the VGG-16 network. Although DenseNet121 has more layers than VGG-16, DenseNet121 has fewer parameters than VGG-16 (7.98M vs 138.36M parameters). This parameter efficiency may reduce the risk of overfitting, which is important in medical imaging applications where datasets are often small. We also found that the Vision Transformer model underperformed in almost all the cases compared to other CNN-based models in the clinical deterioration prediction task. This finding contrasts with the recent success of Vision Transformer in general computer vision tasks [46]. However, in the case of classification tasks with radiographs, the lack of pretraining may harm the performance of the Vision Transformer models [47]. For the networks where we compared performance with random initialization and a pretrained model, in most of the cases, the pretrained model performed better than the randomly initialized one. This could be the main cause for the underperformance of the Vision Transformer models in our work, as we trained it from scratch.

In this study, we found that the models trained with histogram normalization combined with random Gaussian noise among different image augmentation methods achieved better performance, exhibiting the highest AUROC four times and the highest AUPRC three times for different architectures with different combinations of pretraining methods. However, the other two augmentation methods, random flip and random rotate, actually worsened the performance. Our findings align with the existing literature presenting performance improvements with histogram normalization and Gaussian random noise. Gielczyk et al showed that the combination of histogram normalization and Gaussian random noise achieved higher performances than the baseline method in detecting COVID-19 and pneumonia with chest radiographs [48]. However, this can be task dependent involving the useful features of that particular task. Lakhani et al presented a deep convolutional neural network for determining the presence and position of endotracheal tubes where random rotation and random flip augmentation achieved higher performances over the baseline values [12]. As that task was geometry-dependent, regularization introduced by random rotate and random flip augmentation might improve the performance. In contrast, our task of predicting clinical deterioration is not geometry dependent and hence did not benefit from geometric transformations like random rotate and random flips. These insights might be helpful in selecting appropriate image augmentation techniques in models involving chest radiographs.

Strengths

Our study has several important strengths. First, our study cohort consisted of elevated-risk patients with an eCART score ≥93. Predicting deterioration in these patients is more challenging due to their rapid and unpredictable progressions compared to lower-risk patients. Second, we compared multiple deep learning architectures to evaluate their efficacy in predicting clinical deterioration. This comparative approach allows for a more robust understanding of a model's performance in this context. Furthermore, by testing different data augmentation methods, the study explores ways to improve model performance. This aspect is crucial for enhancing the generalizability and robustness of the models. Incorporating Bayesian optimization with a large search space provides the models to achieve the most optimal performance.

Limitations

Our study also has some limitations. First, we only considered the latest radiograph for our models to avoid bias and complexity. Although we reasoned that the latest radiograph conveys the most updated features of patients, prior radiographs and trends over time might carry important features for the model to predict clinical deterioration. Second, we focused on a few popular deep learning architectures with four different augmentation methods. Although recent studies have introduced numerous computer vision architectures, a more comprehensive study would be difficult considering our study's dataset size. Third, in the deterioration prediction model, we only considered the features on chest radiographs. Incorporating other modalities, such as structured data and clinical notes, could improve the accuracy and robustness of our models and will be an interesting future work. Finally, even though our study is the largest of its kind, this was a single-center study, and future studies in other centers are needed to evaluate the external validity of our models.

Conclusion

Our study demonstrates that the DenseNet121 model pretrained on chest radiographs often outperforms VGG16 and the Vision Transformer model with chest radiographs for the prediction of clinical deterioration. Furthermore, we found that model performance improves with histogram normalization along with random Gaussian noise augmentation in most models in terms of both the AUROC and AUPRC metrics. These results show that accurate prediction of patient clinical deterioration is feasible by utilizing chest radiographs while offering valuable insights into the use of computer vision-aided risk prediction.

Acknowledgments

This work was supported by the National Institute of Health (NIH) National Heart, Lung, and Blood Institute grant number R01HL157262 (MMC), and NIH National Library of Medicine grant number 1R01LM013151 (JWG).

Data Availability

The data used in this study were acquired from The University of Wisconsin health systems following approval from the Institutional Review Board. The data use agreements prohibit sharing data due to regulatory and legal constraints, and therefore, the data cannot be shared publicly.

Authors' Contributions

MMC, MA, DPE, and GC conceptualized the study. MMC, MA, and JWG managed data collection and Institutional Review Board approvals. KAC, AA, and MR led data preprocessing and cleaning. MR and JG were responsible for data analysis and methodology. MR prepared the original draft, and all authors participated in revising and editing the article.

Conflicts of Interest

MCM and DPE have a patent issued (#11,410,777) for risk stratification algorithms for hospitalized patients. DPE is employed by and has an equity interest in AgileMD, San Francisco, CA. The other authors have declared no potential conflicts of interest.

Multimedia Appendix 1

Supplementary tables and figures.
[DOCX File (Microsoft Word File), 42 KB-Multimedia Appendix 1]

References

1. Padilla RM, Mayo AM. Clinical deterioration: a concept analysis. J Clin Nurs. Apr 2018;27(7-8):1360-1368. [doi: 10.1111/jocn.14238] [Medline: 29266536]
2. Vincent JL, Einav S, Pearse R, et al. Improving detection of patient deterioration in the general hospital ward environment. Eur J Anaesthesiol. May 2018;35(5):325-333. [doi: 10.1097/EJA.0000000000000798] [Medline: 29474347]
3. U.S. food and drug administration. ECARTv5 clinical deterioration suite. URL: https://www.accessdata.fda.gov/cdrh_docs/pdf23/K233253.pdf [Accessed 2025-03-25]

4. Subbe CP, Kruger M, Rutherford P, Gemmel L. Validation of a modified Early Warning Score in medical admissions. QJM. Oct 2001;94(10):521-526. [doi: 10.1093/qjmed/94.10.521] [Medline: 11588210]
5. Prytherch DR, Smith GB, Schmidt PE, Featherstone PI. ViEWS--Towards a national early warning score for detecting adult inpatient deterioration. Resuscitation. Aug 2010;81(8):932-937. [doi: 10.1016/j.resuscitation.2010.04.014] [Medline: 20637974]
6. Smith GB, Prytherch DR, Meredith P, Schmidt PE, Featherstone PI. The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. Resuscitation. Apr 2013;84(4):465-470. [doi: 10.1016/j.resuscitation.2012.12.016] [Medline: 23295778]
7. Fang AHS, Lim WT, Balakrishnan T. Early warning score validation methodologies and performance metrics: a systematic review. BMC Med Inform Decis Mak. Jun 18, 2020;20(1):111. [doi: 10.1186/s12911-020-01144-8] [Medline: 32552702]
8. Xia C, et al. A multi-modality network for cardiomyopathy death risk prediction with CMR images and clinical information. Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. 2019;11765:577-585. [doi: 10.1007/978-3-030-32245-8_64]
9. Rawal K, Sethi G, Walia GK. Impact of machine learning and deep learning in medical image analysis. In: Rabie K, Karthik C, Chowdhury S, Dutta PK, editors. Deep Learning in Medical Image Processing and Analysis. 2023:187-199. [doi: 10.1049/PBHE059E_ch11]
10. Alhudhaif A, Polat K, Karaman O. Determination of COVID-19 pneumonia based on generalized convolutional neural network model from chest X-ray images. Expert Syst Appl. Oct 15, 2021;180:115141. [doi: 10.1016/j.eswa.2021.115141] [Medline: 33967405]
11. Sitaula C, Hossain MB. Attention-based VGG-16 model for COVID-19 chest X-ray image classification. Appl Intell (Dordr). 2021;51(5):2850-2863. [doi: 10.1007/s10489-020-02055-x] [Medline: 34764568]
12. Lakhani P. Deep convolutional neural networks for endotracheal tube position and X-ray image classification: challenges and opportunities. J Digit Imaging. Aug 2017;30(4):460-468. [doi: 10.1007/s10278-017-9980-7] [Medline: 28600640]
13. Al-qaness MAA, Zhu J, AL-Alimi D, et al. Chest X-ray images for lung disease detection using deep learning techniques: a comprehensive survey. Arch Computat Methods Eng. Aug 2024;31(6):3267-3301. [doi: 10.1007/s11831-024-10081-y]
14. Alshmrani GMM, Ni Q, Jiang R, Pervaiz H, Elshennawy NM. A deep learning architecture for multi-class lung diseases classification using chest X-ray (CXR) images. Alexandria Engineering Journal. Feb 2023;64:923-935. [doi: 10.1016/j.aej.2022.10.053]
15. Sjoding MW, Taylor D, Motyka J, et al. Deep learning to detect acute respiratory distress syndrome on chest radiographs: a retrospective study with external validation. Lancet Digit Health. Jun 2021;3(6):e340-e348. [doi: 10.1016/S2589-7500(21)00056-X] [Medline: 33893070]
16. Sharma S, Guleria K. A systematic literature review on deep learning approaches for pneumonia detection using chest X-ray images. Multimed Tools Appl. Aug 2023;83(8):24101-24151. [doi: 10.1007/s11042-023-16419-1]
17. Ibrahim AU, Ozsoz M, Serte S, Al-Turjman F, Yakoi PS. Pneumonia classification using deep learning from chest X-ray images during COVID-19. Cognit Comput. Jan 4, 2021;16(4):1-13. [doi: 10.1007/s12559-020-09787-5] [Medline: 33425044]
18. Vats S, Sharma V, Singh K, et al. Incremental learning-based cascaded model for detection and localization of tuberculosis from chest x-ray images. Expert Syst Appl. Mar 2024;238:122129. [doi: 10.1016/j.eswa.2023.122129]
19. Sharma V, Gupta SK, Shukla KK. Deep learning models for tuberculosis detection and infected region visualization in chest X-ray images. Intelligent Medicine. May 2024;4(2):104-113. [doi: 10.1016/j.imed.2023.06.001]
20. Sunnetci KM, Alkan A. Biphasic majority voting-based comparative COVID-19 diagnosis using chest X-ray images. Expert Syst Appl. Apr 15, 2023;216:119430. [doi: 10.1016/j.eswa.2022.119430] [Medline: 36570382]
21. Chen Z, et al. A vision-language foundation model to enhance efficiency of chest X-ray interpretation. arXiv. Preprint posted online on 2024. URL: http://arxiv.org/abs/2401.12208 [Accessed 2024-09-25] [doi: 10.48550/arXiv.2401.12208]
22. Cid YD, Macpherson M, Gervais-Andre L, et al. Development and validation of open-source deep neural networks for comprehensive chest x-ray reading: a retrospective, multicentre study. Lancet Digit Health. Jan 2024;6(1):e44-e57. [doi: 10.1016/S2589-7500(23)00218-2] [Medline: 38071118]
23. Meedeniya D, Kumarasinghe H, Kolonne S, Fernando C, Díez I la T, Marques G. Chest X-ray analysis empowered with deep learning: a systematic review. Appl Soft Comput. Sep 2022;126:109319. [doi: 10.1016/j.asoc.2022.109319] [Medline: 36034154]
24. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv. Preprint posted online on 2015. URL: http://arxiv.org/abs/1409.1556 [Accessed 2024-09-24] [doi: 10.48550/arXiv.1409.1556]

25. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. Presented at: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Jul 21-26, 2017:2261-2269; Honolulu, HI. [doi: 10.1109/CVPR.2017.243]
26. Dosovitskiy A, et al. An image is worth 16x16 words: transformers for image recognition at scale. arXiv. Preprint posted online on 2021. URL: http://arxiv.org/abs/2010.11929 [Accessed 2025-03-25] [doi: 10.48550/arXiv.2010.11929]
27. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. arXiv. Preprint posted online on 2015. URL: https://arxiv.org/abs/1512.03385 [Accessed 2025-03-25]
28. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. Presented at: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Jun 26 to Jul 1, 2016; Las Vegas, NV, USA. 2016.[doi: 10.1109/CVPR.2016.308]
29. Ahmed F, Abbas S, Athar A, et al. Identification of kidney stones in KUB X-ray images using VGG16 empowered with explainable artificial intelligence. Sci Rep. Mar 14, 2024;14(1):6173. [doi: 10.1038/s41598-024-56478-4] [Medline: 38486010]
30. Islam M, Hannan T, Sarker L, Ahmed Z. COVID-densenet: A deep learning architecture to detect COVID-19 from chest radiology images. In: Saraswat M, Chowdhury C, Mandal CK, Gandomi AH, editors. Presented at: Proceedings of International Conference on Data Science and Applications; Feb 7, 2023:397-415; [doi: 10.1007/978-981-19-6634-7_28]
31. Duong LT, Le NH, Tran TB, Ngo VM, Nguyen PT. Detection of tuberculosis from chest X-ray images: Boosting the performance with vision transformer and transfer learning. Expert Syst Appl. Dec 2021;184:115519. [doi: 10.1016/j.eswa.2021.115519]
32. Deng J, Dong W, Socher R, Li LJ. ImageNet: A large-scale hierarchical image database. Presented at: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops); Jun 20-25, 2005:248-255; Miami, FL. [doi: 10.1109/CVPR.2009.5206848]
33. Cohen JP, et al. TorchXRayVision: A library of chest X-ray datasets and models. arXiv. Preprint posted online on 2021. [doi: 10.48550/ARXIV.2111.00595]
34. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. Presented at: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Jul 21-26, 2017:3462-3471; Honolulu, HI. [doi: 10.1109/CVPR.2017.369]
35. Bustos A, Pertusa A, Salinas JM, de la Iglesia-Vayá M. PadChest: A large chest x-ray image dataset with multi-label annotated reports. Med Image Anal. Dec 2020;66:101797. [doi: 10.1016/j.media.2020.101797] [Medline: 32877839]
36. Irvin J, Rajpurkar P, Ko M, et al. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. 2019;33(1):590-597. [doi: 10.1609/aaai.v33i01.3301590]
37. Johnson AEW, Pollard TJ, Berkowitz SJ, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data. Dec 12, 2019;6(1):317. [doi: 10.1038/s41597-019-0322-0] [Medline: 31831740]
38. Demner-Fushman D, Kohli MD, Rosenman MB, et al. Preparing a collection of radiology examinations for distribution and retrieval. J Am Med Inform Assoc. Mar 2016;23(2):304-310. [doi: 10.1093/jamia/ocv080] [Medline: 26133894]
39. Majkowska A, Mittal S, Steiner DF, et al. Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. Radiology. Feb 2020;294(2):421-431. [doi: 10.1148/radiol.2019191293] [Medline: 31793848]
40. RSNA pneumonia detection challenge. URL: https://www.kaggle.com/competitions/rsna-pneumonia-detection-challenge [Accessed 2025-03-25]
41. Liaw R, Liang E, Nishihara R, Moritz P, Gonzalez JE, Stoica I. Tune: A research platform for distributed model selection and training. arXiv. Preprint posted online on 2018. [doi: 10.48550/arXiv.1807.05118]
42. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. Sep 1988;44(3):837-845. [doi: 10.2307/2531595] [Medline: 3203132]
43. Netflix. GitHub - Netflix/vmaf: Perceptual video quality assessment based on multi-method fusion. URL: https://github.com/Netflix/vmaf/ [Accessed 2025-03-25]
44. He K, Girshick R, Dollar P. Rethinking imagenet pre-training. Presented at: 2019 IEEE/CVF International Conference on Computer Vision (ICCV); Oct 27 to Nov 2, 2019:4917-4926; Seoul, Korea (South). URL: https://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=8972782 [doi: 10.1109/ICCV.2019.00502]
45. Lee J, Lee EJ. Self-supervised pre-training improves fundus image classification for diabetic retinopathy. In: Kehtarnavaz N, Carlsohn MF, editors. Presented at: Real-Time Image Processing and Deep Learning 2022; Apr 3-7, 2022; Orlando, United States. [doi: 10.1117/12.2632901]

46. Khan S, Naseer M, Hayat M, Zamir SW, Khan FS, Shah M. Transformers in vision: a survey. ACM Comput Surv. Jan 31, 2022;54(10s):1-41. [doi: 10.1145/3505244]
47. Jain A, Bhardwaj A, Murali K, Surani I. A comparative study of CNN, ResNet, and Vision Transformers for multi-classification of chest diseases. arXiv. 2024. [doi: 10.48550/arXiv.2406.00237]
48. Giełczyk A, Marciniak A, Tarczewska M, Lutowski Z. Pre-processing methods in chest X-ray image classification. PLoS ONE. 2022;17(4):e0265949. [doi: 10.1371/journal.pone.0265949] [Medline: 35381050]

Abbreviations

AUPRC: area under the precision-recall curve
AUROC: area under the receiver operating characteristic curve
eCART: electronic cardiac arrest risk triage
EWS: early warning scores
ICU: intensive care unit
UW Health: University of Wisconsin Health System
VMAF: Video Multimethod Assessment Fusion

Edited by Yuankai Huo; peer-reviewed by Shiru Wang, Siqi Lu; submitted 03.10.2024; final revised version received 08.03.2025; accepted 10.03.2025; published 10.04.2025

Please cite as:
Rahman M, Gao J, Carey KA, Edelson DP, Afshar A, Garrett JW, Chen G, Afshar M, Churpek MM
Comparison of Deep Learning Approaches Using Chest Radiographs for Predicting Clinical Deterioration: Retrospective Observational Study
JMIR AI 2025;4:e67144
URL: https://ai.jmir.org/2025/1/e67144
doi: 10.2196/67144