bioengineering

MDPI

Article

DeepCOVID-Fuse: A Multi-Modality Deep Learning Model Fusing Chest X-rays and Clinical Variables to Predict COVID-19 Risk Levels

Yunan Wu [1,*], Amil Dravid [2], Ramsey Michael Wehbe [3] and Aggelos K. Katsaggelos [1,2]

[1] Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL 60201, USA; aggk@eecs.northwestern.edu
[2] Department of Computer Science, Northwestern University, Evanston, IL 60201, USA; amildravid2023@u.northwestern.edu
[3] The Division of Cardiology, Department of Medicine and Bluhm Cardiovascular Institute, Northwestern Memorial Hospital, Chicago, IL 60611, USA; ramsey.wehbe@northwestern.edu
* Correspondence: yunanwu2020@u.northwestern.edu

Abstract: The COVID-19 pandemic has posed unprecedented challenges to global healthcare systems, highlighting the need for accurate and timely risk prediction models that can prioritize patient care and allocate resources effectively. This study presents DeepCOVID-Fuse, a deep learning fusion model that predicts risk levels in patients with confirmed COVID-19 by combining chest radiographs (CXRs) and clinical variables. The study collected initial CXRs, clinical variables, and outcomes (i.e., mortality, intubation, hospital length of stay, Intensive care units (ICU) admission) from February to April 2020, with risk levels determined by the outcomes. The fusion model was trained on 1657 patients (Age: 58.30 $\pm$ 17.74; Female: 807) and validated on 428 patients (56.41 $\pm$ 17.03; 190) from the local healthcare system and tested on 439 patients (56.51 $\pm$ 17.78; 205) from a different holdout hospital. The performance of well-trained fusion models on full or partial modalities was compared using DeLong and McNemar tests. Results show that DeepCOVID-Fuse significantly ($p < 0.05$) outperformed models trained only on CXRs or clinical variables, with an accuracy of 0.658 and an area under the receiver operating characteristic curve (AUC) of 0.842. The fusion model achieves good outcome predictions even when only one of the modalities is used in testing, demonstrating its ability to learn better feature representations across different modalities during training.

Keywords: COVID-19; risk level prediction; multi-modality; fusion CNNs; CXRs; clinical variables

check for updates

Citation: Wu, Y.; Dravid, A.; Wehbe, R.M.; Katsaggelos, A.K. DeepCOVID-Fuse: A Multi-Modality Deep Learning Model Fusing Chest X-rays and Clinical Variables to Predict COVID-19 Risk Levels. *Bioengineering* **2023**, *10*, 556. https://doi.org/10.3390/bioengineering10050556

Academic Editor: Mario Petretta

Received: 11 April 2023
Revised: 28 April 2023
Accepted: 2 May 2023
Published: 5 May 2023

1. Introduction

Coronavirus disease 2019 (COVID-19) has been heavily straining the healthcare systems of countries across the world, with over 500 million cases and 6 million deaths as of July 2022 [1]. Reverse-transcription polymerase chain reaction (RT-PCR) is the current gold standard for the diagnosis of COVID-19. However, the use of RT-PCR for COVID-19 diagnosis is limited to authorized, trained clinical laboratory personnel and patients with suspected COVID-19, which can create bottlenecks in the testing process. Results can take more than 24 h to produce, leading to delays in patient care and allocation of resources [2]. Previous studies have shown that chest radiographs (X-rays) and computed tomography (CT) images can reveal COVID-19 features [3], which can be combined with clinical judgment to make a COVID-19 diagnosis. Artificial Intelligence (AI) algorithms have shown great promise in detecting these signatures from chest X-rays and CT scans [4–7], enabling faster and more accurate treatment of suspected patients. The use of CT in this domain is restrictive, particularly due to their high cost and longer processing time, whereas X-ray scanning is more commonly conducted and accessible. Nevertheless, relying solely on chest X-rays may not be sufficient to serve as a diagnostic tool for COVID-19, but they can

instead be used to inform the diagnosis and flag at-risk patients who may need further testing [8]. This is particularly important in resource-limited settings, where the use of chest X-rays can aid in resource allocation, triage, and infection control.

Many AI models have been proposed for the purpose of COVID-19 detection [9–12]. For example, Shaheed et al. developed a computer-aided diagnostic scheme that utilizes extracted features from transformers and a random forest classifier on CXRs for automatic recognition of COVID-19 and pneumonia [13]. However, their use of small and biased public datasets raises skepticism about their deployment [14]. In contrast, DeepCOVID-XR, which proposes an ensemble of convolutional neural networks (CNNs), serves as one of the first works trained and tested on a large clinical dataset for COVID-19 detection [4]. It has been found to be more robust to biases present in models trained on public data. With rapid testing becoming more readily available than at the beginning of the pandemic, more efforts can be made to serve those who are infected. Furthermore, identifying the severity and prognosis of infected individuals can aid in triaging and allocating resources appropriately. However, while many published AI algorithms have focused on better detection of COVID-19, risk stratification of confirmed COVID-19 subjects remains relatively unexplored. To address this gap, we propose DeepCOVID-Fuse, a model that fuses clinical variables from electronic health record (EHR) data and image features from CXRs to categorize infected COVID-19 patients into low-, intermediate-, and high-risk classes. Previous work on fusion and risk prediction has included utilizing the fusion of different image feature representations for COVID detection [15], tackling binary risk prediction with patient characteristic data [16] and clinical features [17], predicting the chance of survival and kidney injury with tabular clinical and biochemistry data [18], and utilizing gene enrichment profiles from blood transcriptome data to stratify COVID-19 patients [19]. Our DeepCOVID-Fuse model ensembles three different architectures with image and tabular clinical data, providing accurate fine-grained risk predictions for COVID-19 patients. Notably, we thoroughly compare the performance of fusion models trained on multiple modalities but tested on one or a subset of modalities, and find that testing the fusion model even with a missing modality still provides more informative predictions than networks trained on a single modality.

2. Materials and Methods

2.1. Patients

This study is based on a cohort of 2085 COVID-19 patients from over 20 different sites across Northwestern Memorial Health Care System. All patients were tested positive from February 2020 to April 2020 and their corresponding electronic health records (EHR) were collected with a positive reverse transcription polymerase chain reaction (RT-PCR). Unlike a previous study [4], which used CXRs of both COVID-19-negative and positive subjects to build models for COVID-19-positive prediction, our study focuses exclusively on the initial CXR taken after the first inpatient admission of each COVID-19-positive subject. Furthermore, we broaden the scope of our investigation by incorporating clinical variables from subjects' EHRs to make risk predictions for COVID-19 subjects. Specifically, each of the COVID-positive patients was categorized into low-, intermediate-, or high-risk classes. These three classes correspond, respectively, to (1) hospital length of stay (LoS) of less than one day, (2) hospital LoS greater than one day but no death or admittance to the ICU, and (3) death or admittance to the ICU, as documented in the patients' EHRs.

2.2. CXRs Acquisition and Preprocessing

CXRs images were preprocessed in accordance with metadata using appropriate windowing operations. The grayscale images were first converted to 3-channel RGB images (with identical R, G, and B planes) as this is the typical input of deep learning models. To remove unnecessary background and focus more on lung features, images were then center-cropped using a UNet-based algorithm [20], which was pre-trained on the public CXR dataset [21,22] to segment lung fields. Finally, all cropped images were resized to a

resolution of 224 × 224 pixels, scaled to a range of 0 to 1 by dividing each pixel value by 255 (8-bit images), and normalized using ImageNet's mean and standard deviation before being fed into the model. This preprocessing was applied to all training, validation, and test sets.

2.3. Clinical Data Processing

Clinical variables were obtained from each subject's EHRs across different categories: basic demographic information, laboratory results, comorbidities, electrocardiogram (ECG), and modified early warning score (MEWS). To preprocess the data, we first matched each subject's first CXR with its temporally closest EHR within 24 h. We then discarded features that were missing more than 40% of their entries. The remaining features were classified into three types for preprocessing, namely, binary, categorical, and continuous, as shown in Supplementary Figure S1. Specifically, for binary features, such as comorbidities, missing values in the training set were set to non-existent. For multi-class features, such as race and smoking status, missing values were set to an additional unknown class, and all classes were converted to one-hot vectors. For continuous features, missing values were imputed using the mean computed from the training set and all features were scaled to the range of 0 to 1 using min–max normalization. The mean value of each clinical feature on the training set was applied to the validation and test sets for normalization. The details of all selected clinical features are provided in Supplementary Table S1.

2.4. Model Details

The DeepCOVID-Fuse is a combination of three fusion neural network architectures that were trained using a weighted ensemble approach to accurately classify COVID-19 patients into three risk categories (i.e., low risk, intermediate risk, and high risk), as shown in Figure 1. Each individual network consists of two branches: the CXR image branch and the clinical variable branch. In comparison to our previous model DeepCOVID-XR [4], where six different networks were ensembled—DenseNet-121 [23], ResNet-50 [24], InceptionV3 [25], Inception-ResNetV2 [26], Xception [27], and EfficientNet-B2 [28]—DeepCOVID-Fuse is designed to balance efficiency with accuracy by utilizing only three CNNs for CXR image processing. These three CNNs are chosen as the CXR image branch to process 224 × 224 chest X-rays, namely EfficientNet-B2, ResNet50, and DenseNet-121. To each network, a fully connected layer was added to adjust the feature dimension of the image branch, followed by a dropout layer to prevent overfitting.

Specifically, the clinical variable branch includes a fully connected layer designed to process 99 clinical features from tabular EHR data, followed by a dropout layer. Overall, the two branches were fused together using a concatenation layer, followed by two fully connected layers and a three-class output node with a softmax activation function for final classification. Further information on the hyperparameters can be found in Supplementary materials. The training process of the entire framework for the two network branches consists of two steps. First, the weights of the image branch were initialized with the corresponding weights from DeepCOVID-XR, while the clinical variable branch was randomly initialized in accordance with TensorFlow standards. During this stage, only the clinical variable branch and the fusion layers were trained, while the convolutional layers of the image branch were frozen. In the second stage, after early stopping concluded the first stage of training, all layers were unfrozen and fine-tuned. Finally, the outputs of each of the three models (i.e., probabilities after softmax) were averaged for the final prediction.

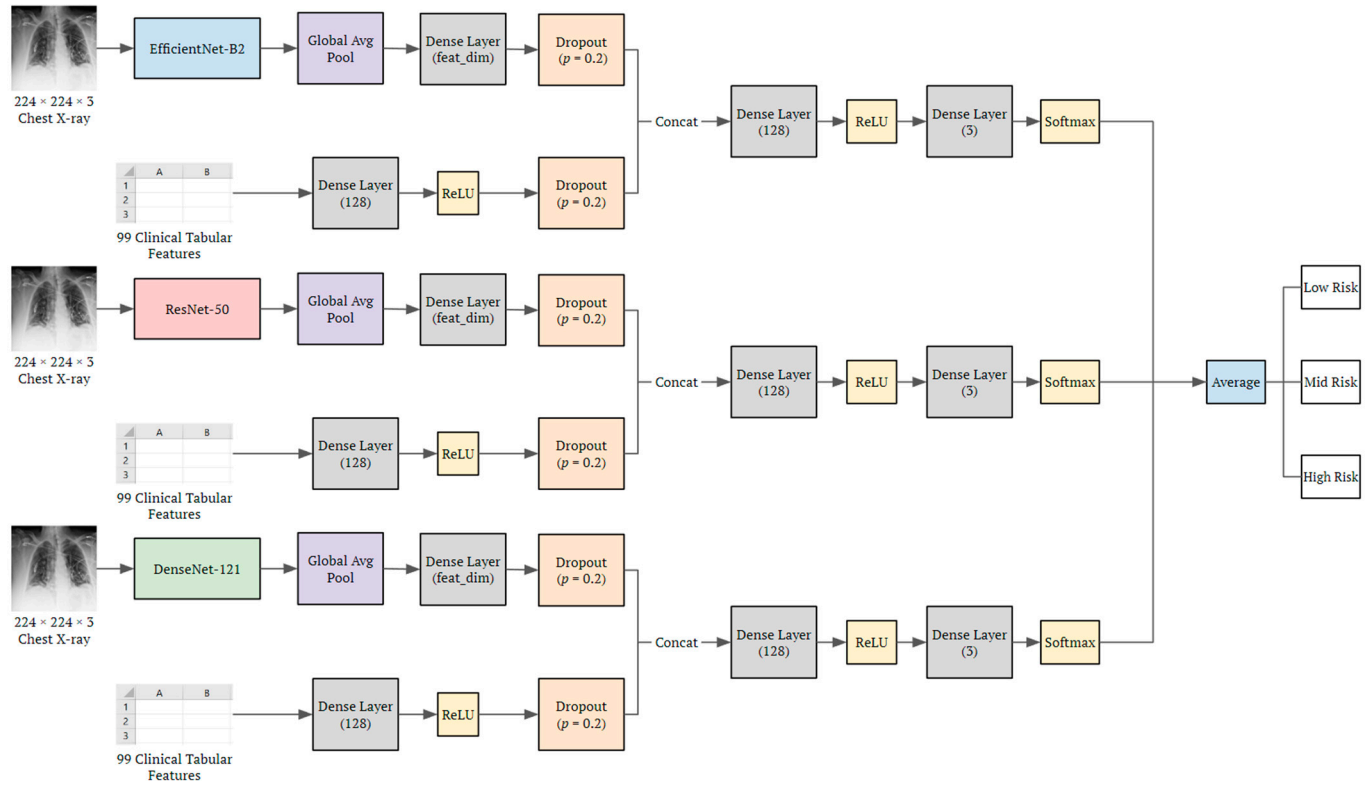

Figure 1. The architecture of the DeepCOVID-Fuse ensemble model. The preprocessed image was fed into three different CNN architectures, followed by a fully connected layer to transform the image feature dimension. The clinical tabular features were fed into a fully connected layer. The features were fused and fed into another fully connected layer, followed by the last classification layer with softmax as the activation function. Feat_dim changes to compare different combinations of features from image and clinical feature branches.

2.5. Statistical Analysis

The performance of different models was evaluated by calculating various metrics such as overall accuracy, precision, recall, F1 score, MCC (The Matthews correlation coefficient), and AUC (area under the receiver operating characteristic (ROC) curve). To ensure reliability, each experiment was run independently five times, and 95% confidence intervals were obtained. MCC is a useful metric for multi-class classification as it considers true positives, true negatives, false positives, and false negatives, making it more suitable for imbalanced classes and providing a comprehensive understanding of the model's performance. McNemar's test [29] was performed for pairs of models to compare the accuracy, precision, recall, and F1 score, and the DeLong test [30] was performed to compare the AUCs of different models. A p-value < 0.05 was considered statistically significant.

3. Results

3.1. Experimental Design

A total of 2085 subjects were included in this study. The demographic distribution details are shown in Supplementary Table S1. A three-class outcome was predicted for each COVID-19 subject, i.e., low risk (L), intermediate risk (I) and high risk (H). The data split follows the approach used in our previous work [4], where the training and validation sets are sourced from multiple institutes, while the test set is obtained from a separate, different institute. Since only the initial CXR after each COVID-19 subject's first inpatient admission was considered, all experiments had a total of 2085 images, of which 1657 (L: 476, I: 663, H: 518; Mean age: 58.30 years ± 17.74 (standard deviation); Female: 807) were used for training and 428 (L: 119, I: 176, H: 133; 56.41 ± 17.03; 190) for validation; the same cohort

applies to clinical features. A separate hold-out test set of 439 subjects (L: 101, I: 193, H: 145; 56.51 ± 17.78; 205) from a different hospital were used to evaluate model performance.

Overall, three types of models were evaluated and compared in this study, including (1) the fusion models trained on CXRs and clinical features with different feature size combinations (feat_dim in Figure 1) from the image and feature branches, (2) the same models trained on CXRs only with the image branch and (3) the models trained on clinical features only with the feature branch. Additionally, for (1), we compared model performances of three fusion models individually, and the ensemble of all three. For (2), to show the effect of the fusion model, we evaluated the fusion model on CXRs as input only. Likewise, for (3), we evaluated the fusion model on clinical features as input only and compared the result with those trained directly on some machine learning algorithms. Further experimental details are included in the Supplementary Materials. All experiments were run independently five times to account for model variability. The models were trained and evaluated using Tensorflow 2.0 in Python 3.6 on a single GPU (NVIDIA TITAN V).

3.2. Performance of DeepCOVID-Fuse

The performance of each individual fusion model and an ensemble of all models on the testing set are compared in Table 1. Overall, the ensemble model significantly outperformed all individual models on this COVID-19 risk prediction task, achieving an accuracy of 0.658, a recall of 0.660, a precision of 0.689, an F1 of 0.660, an MCC of 0.640, and an AUC score of 0.842. Notably, for all individual models, ablation studies of different feature dimensions from CXRs and clinical variables showed that models with higher proportional features from the clinical branch than the CXR branch (i.e., CXRs: clinical = 64:128) achieved better model predictions than equal (i.e., 128:128) or lower fractions (i.e., 1408:128). Furthermore, the fusion model with a DenseNet architecture had the best performance of AUC from 0.814 to 0.824, followed by a ResNet architecture from 0.794 to 0.815 and an EfficientNet architecture from 0.794 to 0.805. In addition, we analyzed the performance of the models across different age groups by categorizing the subjects into four groups: ages 20–40, 40–60, 60–80, and 80–100, as presented in Supplementary Table S2. The findings suggest that our proposed model performs well across different age groups. However, the results indicate that the model performs optimally in the middle age group, followed by the younger and older age groups.

Table 1. Performance of fusion models for risk predictions in confirmed COVID-19 subjects on external test sets with different combinations of latent feature sizes from X-rays and clinical variables.

	EfficientNet			**ResNet**			**DenseNet**			**Ensemble**
Latent feature (X_ray × clinical data)	64 × 128	128 × 128	1408 × 128	64 × 128	128 × 128	2048 × 128	64 × 128	128 × 128	1408 × 128	64 × 128
Accuracy	0.618 [0.600, 0.637]	0.622 [0.606, 0.638]	0.626 [0.590, 0.662]	0.628 [0.610, 0.645]	0.630 [0.620, 0.642]	0.611 [0.589, 0.632]	0.658 [0.650, 0.667]	0.638 [0.622, 0.654]	0.640 [0.632, 0.647]	0.658 *
Recall	0.619 [0.600, 0.639]	0.622 [0.606, 0.638]	0.626 [0.590, 0.662]	0.626 [0.595, 0.656]	0.633 [0.623, 0.642]	0.611 [0.589, 0.632]	0.657 [0.649, 0.666]	0.638 [0.621, 0.655]	0.640 [0.632, 0.647]	0.660 *
Precision	0.649 [0.631, 0.666]	0.648 [0.620, 0.676]	0.675 [0.648, 0.702]	0.665 [0.652, 0.678]	0.675 [0.664, 0.685]	0.652 [0.619, 0.685]	0.671 [0.658, 0.684]	0.641 [0.623, 0.659]	0.647 [0.635, 0.659]	0.689 *
F1	0.616 [0.599, 0.633]	0.619 [0.603, 0.637]	0.623 [0.583, 0.663]	0.626 [0.608, 0.645]	0.627 [0.612, 0.642]	0.607 [0.586, 0.627],	0.658 [0.650, 0.666]	0.638 [0.621, 0.655]	0.639 [0.632, 0.647]	0.660 *
MCC	0.607 [0.603, 0.611]	0.614 [0.606, 0.622]	0.617 [0.609, 0.625]	0.618 [0.612, 0.624]	0.620 [0.615, 0.625]	0.601 [0.594, 0.608]	0.635 [0.629, 0.641]	0.624 [0.619, 0.629]	0.626 [0.620, 0.632]	0.640 *
AUC	0.805 [0.798, 0.812]	0.794 [0.778, 0.811]	0.804 [0.780, 0.827]	0.815 [0.804, 0.826]	0.815 [0.809, 0.820]	0.794 [0.782, 0.807]	0.824 [0.822, 0.826]	0.814 [0.797, 0.831]	0.820 [0.805, 0.836]	0.842 *

Notes: Data in parentheses are 95% CIs from five repeated experimental runs. AUC = area under the receiver operating characteristic curve. Latent feature = Image and clinical feature dimensions when concatenated in a fusion model. * p value < 0.05 denotes the comparisons are statistically significant.

3.3. Comparison of Image-Only with Fusion-Image-Only

To show the importance of fusion, the performance comparison between the model trained and tested on CXR images only (Image-only) and the model trained on the fusion model (i.e., both CXRs and clinical variables) but tested on CXR images only (Fusion-image-

only) are provided in Table 2. Combined with the results in Table 1, the fusion model with additional clinical variables significantly improved COVID-19 risk prediction compared to the Image-only model. Notably, even without the clinical variables, the well-trained fusion model outperformed the Image-only model on the same CXR-only test set. Specifically, for three individual models, the well-trained fusion model improved the accuracy by 0.008~0.011, the recall by 0.004~0.012, the precision by 0.008~0.043, the F1 by 0.007~0.013, the MCC by 0.009~0.020, and the AUC by 0.009~0.016. Additionally, heatmaps generated from the Image-only and Fusion-image-only models using gradient class activation maps (Grad-CAM) are provided in Figure 2 to visualize the salient features of each CXR used by the model for COVID-19 risk level classification. For correct risk-level predictions, these heatmaps highlight abnormalities in the lungs and demonstrate that the fusion model captures more relevant features for classification than the image-only model. In some cases, where the fusion model made the correct classification and the image-only model misclassified, the heatmaps showed different feature patterns, with the former highlighting lung abnormalities and the latter not.

Typically, the clinical features are partially present. Supplementary Table S3 illustrates the results of the well-trained fusion model on CXR images with the proportionally increasing clinical variables. The results showed that, as more clinical features are integrated into the model, its performance of the well-trained fusion model on the test set improves. For example, when 80% of the clinical variables are present, and only 20% are missing at random, the fusion model (e.g., with the DenseNet architecture) achieved an accuracy of 0.645, a recall of 0.647, a precision of 0.656, an F1 of 0.644, an MCC of 0.625, and an AUC of 0.816.

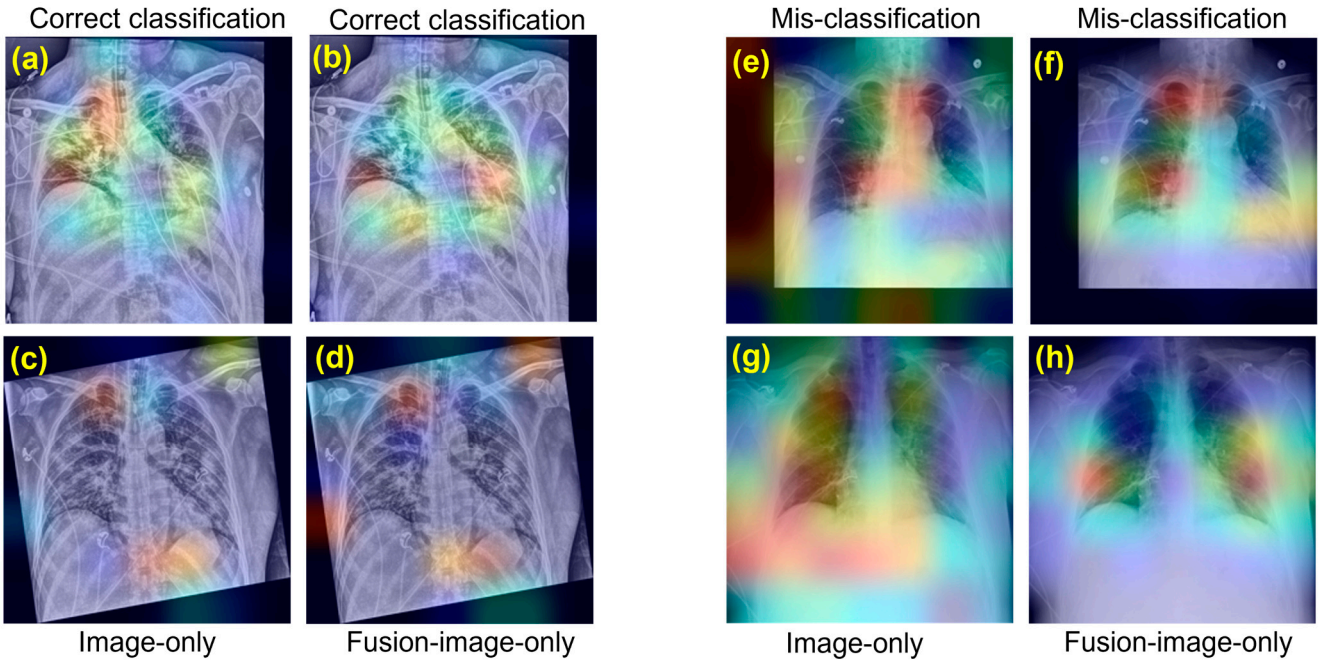

Figure 2. Heatmaps of Gradient class activation maps (Grad-CAM) generated from the model showing the location of important features for high-risk predictions in confirmed COVID-19 patients. The redder the intensity of the heatmap, the more important the feature areas. Heatmaps generated from Image-only models (**a**,**c**,**e**,**g**) and Fusion-image-only models (**b**,**d**,**f**,**h**) are compared for cases of correct (**a**–**d**,**f**,**h**) and incorrect predictions (**e**,**g**). For the same subject using only CXR as model input, the Fusion model made correct predictions, with heatmaps (**f**,**h**) highlighting abnormalities in the lungs, while the Image-only model misclassified the predictions, (**e**,**g**) highlighting unnecessary background.

Table 2. Performance of models for risk predictions in confirmed COVID-19 subjects on external test sets using only CXRs as model input.

COVID-Level	EfficientNet Image-Only	ResNet Image-Only	Densenet Image-Only	Image-Only Ensemble	EfficientNet Fusion-Image-Only	ResNet Fusion-Image-Only	DenseNet Fusion-Image-Only	Fusion-Image-Only Ensemble
Accuracy	0.582 [0.572, 0.591]	0.614 [0.604, 0.624]	0.615 [0.608, 0.622]	0.621 *	0.593 [0.581, 0.606]	0.625 [0.615, 0.634]	0.623 [0.604, 0.641]	0.632 *
Recall	0.581 [0.572, 0.591]	0.616 [0.604, 0.624]	0.616 [0.607, 0.624]	0.619 *	0.593 [0.582, 0.606]	0.625 [0.615, 0.634]	0.620 [0.608, 0.632]	0.629 *
Precision	0.604 [0.594, 0.614]	0.664 [0.645, 0.683]	0.631 [0.627, 0.634]	0.665 *	0.657 [0.646, 0.667]	0.662 [0.643, 0.681]	0.639 [0.623, 0.647]	0.664 *
F1	0.576 [0.567, 0.586]	0.609 [0.595, 0.624]	0.614 [0.606, 0.621]	0.620 *	0.583 [0.567, 0.600]	0.619 [0.607, 0.631]	0.627 [0.611, 0.639]	0.634 *
MCC	0.553 [0.540, 0.566]	0.587 [0.580, 0.594]	0.602 [0.590, 0.614]	0.608	0.562 [0.553, 0.571]	0.607 [0.594, 0.620]	0.613 [0.605, 0.621]	0.618
AUC	0.769 [0.764, 0.774]	0.798 [0.788, 0.808]	0.781 [0.72, 0.792]	0.807 *	0.781 [0.768, 0.796]	0.807 [0.803, 0.811]	0.797 [0.784, 0.806]	0.813 *

Notes: Data in parentheses are 95% CIs from five repeated experimental runs. AUC = area under the receiver operating characteristic curve; Image-only = models trained and tested with CXRs only; Fusion-image-only = well-trained fusion models but tested with CXRs only. * p value < 0.05 denotes the comparisons are statistically significant.

3.4. Comparison of Feature-Only with Fusion-Feature-Only

We performed the same analysis as described above, by comparing the performance of models trained and tested on clinical features only (Feature-only) with models first well-trained on the fusion model but tested on the clinical features (Fusion-Feature-only). As shown in Table 3, even without CXRs as input, the well-trained fusion model significantly outperformed the Feature-only model with an AUC of 0.733. In addition, we compared the neural-network-based models with several machine learning algorithms trained on the same clinical features, including random forests (RM), quadratic discriminant analysis (QDA), and Linear Ridge (LR) classification. Interestingly, RF achieved the best performance, followed by Fusion-Feature-only and LR, while Feature-only had the lowest performance among all metrics. Furthermore, it is still worth noting that the model in Table 1 combining CXRs and clinical variables still outperformed all results in Table 3.

Table 3. Performance of models for risk predictions in confirmed COVID-19 subjects on external test sets using only clinical variables as model input.

COVID-Level	DNN Feature-Only	Fusion Feature-Only	Random Forests	QDA	Linear Ridge
Accuracy	0.440 [0.432, 0.448]	0.539 [0.525, 0.553]	0.560 [0.553, 0.567]	0.526 [0.519, 0.533]	0.536 [0.527, 0.546]
Recall	0.441 [0.430, 0.449]	0.540 [0.526, 0.555]	0.563 [0.554, 0.569]	0.528 [0.517, 0.539]	0.533 [0.525, 0.541]
Precision	0.193 [0.183, 0.214]	0.567 [0.553, 0.582]	0.588 [0.517, 0.671]	0.532 [0.526, 0.538]	0.544 [0.532, 0.556]
F1	0.269 [0.253, 0.280]	0.560 [0.542, 0.577]	0.573 [0.568, 0.581]	0.479 [0.461, 0.496]	0.536 [0.527, 0.545]
MCC	0.243 [0.230, 0.256]	0.541 [0.529, 0.553]	0.562 [0.550, 0.574]	0.435 [0.421, 0.449]	0.507 [0.497, 0.517]
AUC	0.502 [0.481, 0.522]	0.733 [0.730, 0.737]	0.768 [0.759, 0.777]	0.600 [0.587, 0.613]	0.625 [0.613, 0.636]

Notes: Data in parentheses are 95% CIs from five repeated experimental runs. AUC = area under the receiver operating characteristic curve; QDA = Quadratic Discriminant Analysis; Feature-only = models trained and tested with clinical variables only; Fusion-Feature-only = well-trained fusion models but tested with clinical variables only.

4. Discussion

In this study, we proposed DeepCOVID-Fuse, a fusion model that incorporates clinical variables with CXRs to predict future risks of clinically meaningful outcomes in patients diagnosed with confirmed COVID-19. The fusion model was trained and tested using only the first inpatient admission data of each subject, which has great clinical implications for improving our healthcare management system, particularly in intensive care units. DeepCOVID-Fuse achieved an overall accuracy of 0.658 and an AUC of 0.842 on a hold-out testing set from a separate hospital. We further compared this model with models trained on CXR images only or clinical variables only, and evaluated the performance of DeepCOVID-Fuse when only CXR images or clinical variables were available. To the best of our knowledge, our study is the first to demonstrate the effectiveness of a fusion model, which is well-trained on multiple modalities but is capable of achieving a better prediction performance and generating meaningful visual heatmaps when only one or parts of the modalities were available on CXRs and clinical features.

The aim of our work is to assist with resource allocation by addressing a three-class prediction problem, where the level of risk for COVID-19 patients is determined based on their mortality status, need for mechanical ventilation, ICU admission, and hospital length of stay (LoS). The three classes are categorized as low, intermediate, and high risk. As the demand for hospital capacity is reported to be dramatically increasing during the COVID-19 pandemic [31], predicting ventilator usage or ICU admissions in advance will reduce pressure on hospitalization management. In addition, LoS is critical to the allocation of bed capacity, so we chose a 1-day LoS as the separation to differentiate low and intermediate risk, as only patients with a LoS of more than 1 day needed to be allocated a bed. Furthermore, the results in Tables 1 and 2 show that the fusion model with the addition of clinical variables significantly improved risk performances over the model trained only on CXRs, indicating that clinical variables are strongly associated with COVID-19 severity. Meanwhile, the performance of the ensemble fusion models being higher than that of each model individually is consistent with the previous study that showed the ensemble model reduces the generalization error of predictions [4].

In most real-world scenarios, it is common for a modality to be missing or incomplete. As such, the fusion model is not guaranteed to utilize inputs from all modalities, i.e., some COVID-19 patients have either CXRs or a subset of clinical data. One study showed that this can be a limitation of fusion models, as predictions can be overly influenced by the most feature-rich modalities leading to poor generalization [32]. However, our study shows that even if only one or partial modalities are available, well-trained fusion models can still achieve better performances than models trained on that single or partial modality alone, as shown in Tables 2 and 3. Learning correlations across different modalities is the possible explanation for this improved performance. Specifically, since different modalities of a fusion model are simultaneously back-propagated through the loss, they complement each other, so the fusion model is able to learn better latent space representations for each model branch. Therefore, even if only a subset of CXRs or clinical variables are available, fusion models can still play an important role in learning more discriminative features. The experiments in Supplementary Table S3 further show that once the fusion model was well trained, the model performance continued to improve as long as more clinical variables were available in the test set. This can have significant implications for future medical research, as it gives a strong support to the scenario that if images are provided with more usable information during training, i.e., simple features such as age and gender, even if only images are available at testing stage, a better classification prediction can be achieved compared with that using only images to train and test models.

Heatmaps generated by Grad-CAM provide another perspective on the superiority of fusion models in learning feature representations of CXR images compared to image-only models. As shown in Figure 2e–h, when only CXRs were available, the image-only model misclassified a high-risk subject as intermediate, while the well-trained fusion model made the correct prediction. This can be observed from their respective heatmaps, where the

fusion model highlighted discriminative features of the lung, while the other located the wrong area. When making the correct predictions, all heatmaps looked at areas close to the lung, as shown in Figure 2a–d,f–h.

Although previous studies have existed to predict the severity of confirmed COVID-19 patients, our work has a different focus and is unique in many ways. For example, Liang et al. developed a DL-based survival model on a 1D clinical dataset collected at admission to predict the risk of COVID-19 patients being critically ill within 30 days, achieving an overall AUC of above 0.85 [33]. However, they were limited by a lack of clinical datasets, and no imaging data were available. Shamout et al. later proposed a deep learning model using CXRs and routine clinical variables to predict the deterioration risk (i.e., intubation, ICU admission, or mortality) in COVID-19 subjects within 96 h with an AUC of 0.786 [34]. Similarly, Jiao et al. used a DL network combining CXR and clinical data to predict binary outcomes of COVID-19 patient severity (i.e., severe or not), and obtained AUCs ranging from 0.731 to 0.792 [8]. Although two modalities were provided, both studies adopted a late fusion strategy with two independently trained models. In contrast, we trained an end-to-end fusion model that could learn and transfer information between two modalities. A study similar to our work that combined initial CXRs and clinical variables into an end-to-end fusion model to predict mortality in COVID-19 subjects achieved an AUC of 0.82 [35]. However, their model was only trained on 499 subjects with an age range of 21 to 50 years, which may lead to poor model generalization, whereas our model included 2085 subjects of all ages. Another study from Soda et al. developed a multi-branch deep learning framework that combined CXRs and clinical information to predict the clinical outcome (binary: mild or severe) of COVID-19 patients [36]. The model achieved an accuracy of 0.748; however, the study focused solely on the binary outcome with a relatively small sample size of 820 subjects. In addition, Deb et al. proposed a CovSeverity-Net, which uses CXRs to estimate the severity (mild, moderate, severe) of COVID-19 patients [37]. However, unlike our approach, which aims to better assist with resource allocation, Deb et al. included images from all time points, rather than solely using the first inpatient admission of each subject. Most importantly, the focus of this paper is to comprehensively evaluate the statistical and visual performance of fusion models trained on multiple modalities but tested on one or a subset of modalities.

There are some limitations to this study that need to be acknowledged. First, several clinical data in the training dataset are still missing or incomplete. Although we have shown that that not all clinical data is needed in the test set, having a more complete training dataset guarantees a better and more robust model. Second, we did not compare the performance of our fusion model with radiologists, because risk prediction by experts on both CXR and clinical data is challenging and subjective. There is no true, universal ground truth. Next, as shown in Table 3, we found that basic machine learning algorithms, such as random forests, outperformed deep learning-based models, indicating that our fusion model has not yet perfectly extracted features from 1D clinical data. Therefore, future work will explore integrating random forests with deep neural networks to further improve model performance. Lastly, it is worth noting that the current model cannot identify the most relevant features that contribute to patient outcomes, which is important information for clinicians. To address this limitation, we plan to explore new models, such as the merging of random forests with CNNs or incorporating attention mechanisms [38,39], as they may help to predict the importance score of each clinical variable. This information can be valuable for clinicians in understanding the relative importance of different variables in determining patient outcomes.

5. Conclusions

In conclusion, we proposed DeepCOVID-Fuse, a fusion model to predict risk levels in COVID-19 subjects using CXRs and clinical variables obtained at their initial inpatient admission. We showed that models combining both CXRs and clinical features outperformed models with only CXRs or clinical variables. Furthermore, we demonstrated that

the well-trained fusion model was able to achieve good model performance when only single or partial modality was available. We believe that this work demonstrates that it is possible to predict high-risk patients at admission to further benefit hospital triage systems, and also has the potential to promote the use of fusion models in other fields of medical research. Finally, we have made our codes and model weights publicly available to facilitate future research and enable easy comparison of our model's performance with others.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/bioengineering10050556/s1. Supplementary Table S1: Patient characteristics from the training, validation and test set; Supplementary Table S2: Performance of DeepCOVID-Fuse (Ensemble) for risk predictions in confirmed COVID-19 subjects on external test sets in different age groups; Supplementary Table S3: Performance of fusion-image-only models for risk predictions in confirmed COVID-19 subjects on external test sets using CXRs as model input with a random subset (%) of clinical variables (0: no clinical features, 100: full clinical features); Figure S1. The preprocessing of clinical features. Features are classified into three types, binary, multi-class, and continuous with different missing imputation and scaling operations.

Author Contributions: Conceptualization, Y.W., R.M.W. and A.K.K.; Methodology, Y.W. and A.D.; Software, Y.W. and A.D.; Formal Analysis, Y.W. and A.D.; Data Curation, Y.W. and R.M.W.; Writing—Original Draft Preparation, Y.W. and A.D.; Writing—Review and Editing, R.M.W. and A.K.K.; Supervision, A.K.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This retrospective study was approved by the Northwestern institutional review board (STU00212323).

Informed Consent Statement: This study was granted a waiver of Health Insurance Portability and Accountability Act authorization and a waiver of written informed consent.

Data Availability Statement: The code and model weights used in the study are publicly available from the GitHub repository (https://github.com/YunanWu2168/DeepCOVID-Fuse), accessed on 1 May 2023.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. World Health Organization. Coronavirus Disease (COVID-19). 12 October 2020. Available online: https://covid19.who.int/ (accessed on 1 May 2023).
2. Zuckerman, D.M. Emergency Use Authorizations (EUAs) Versus FDA Approval: Implications for COVID-19 and Public Health. *Am. J. Public Health* **2021**, *111*, 1065–1069. [CrossRef] [PubMed]
3. Sverzellati, N.; Ryerson, C.J.; Milanese, G.; Renzoni, E.A.; Volpi, A.; Spagnolo, P.; Bonella, F.; Comelli, I.; Affanni, P.; Veronesi, L.; et al. Chest Radiography or Computed Tomography for COVID-19 Pneumonia? Comparative Study in a Simulated Triage Setting. *Eur. Respir. J.* **2021**, *58*, 2004188. [CrossRef] [PubMed]
4. Wehbe, R.M.; Sheng, J.; Dutta, S.; Chai, S.; Dravid, A.; Barutcu, S.; Wu, Y.; Cantrell, D.R.; Xiao, N.; Allen, B.D.; et al. DeepCOVID-XR: An Artificial Intelligence Algorithm to Detect COVID-19 on Chest Radiographs Trained and Tested on a Large U.S. Clinical Data Set. *Radiology* **2021**, *299*, E167–E176. [CrossRef]
5. Oh, Y.; Park, S.; Ye, J.C. Deep Learning COVID-19 Features on CXR Using Limited Training Data Sets. *IEEE Trans. Med. Imaging* **2020**, *39*, 2688–2700. [CrossRef]
6. Harmon, S.A.; Sanford, T.H.; Xu, S.; Turkbey, E.B.; Roth, H.; Xu, Z.; Yang, D.; Myronenko, A.; Anderson, V.; Amalou, A.; et al. Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets. *Nat. Commun.* **2020**, *11*, 4080. [CrossRef] [PubMed]
7. Wynants, L.; Van Calster, B.; Collins, G.S.; Riley, R.D.; Heinze, G.; Schuit, E.; Bonten, M.M.J.; Dahly, D.L.; Damen, J.A.; Debray, T.P.A.; et al. Prediction models for diagnosis and prognosis of COVID-19: Systematic review and critical appraisal. *BMJ* **2020**, *369*, m1328. [CrossRef] [PubMed]
8. Jiao, Z.; Choi, J.W.; Halsey, K.; Tran, T.M.L.; Hsieh, B.; Wang, D.; Eweje, F.; Wang, R.; Chang, K.; Wu, J.; et al. Prognostication of patients with COVID-19 using artificial intelligence based on chest x-rays and clinical data: A retrospective study. *Lancet Digit. Health* **2021**, *3*, e286–e294. [CrossRef]
9. Castiglioni, I.; Ippolito, D.; Interlenghi, M.; Monti, C.B.; Salvatore, C.; Schiaffino, S.; Polidori, A.; Gandola, D.; Messa, C.; Sardanelli, F. Machine learning applied on chest x-ray can aid in the diagnosis of COVID-19: A first experience from Lombardy, Italy. *Eur. Radiol. Exp.* **2021**, *5*, 7. [CrossRef]

10. Ozturk, T.; Talo, M.; Yildirim, E.A.; Baloglu, U.B.; Yildirim, O.; Acharya, U.R. Automated detection of COVID-19 cases using deep neural networks with X-ray images. *Comput. Biol. Med.* **2020**, *121*, 103792. [CrossRef] [PubMed]
11. Wang, L.; Lin, Z.Q.; Wong, A. COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. *Sci. Rep.* **2020**, *10*, 19549. [CrossRef]
12. Hemdan, E.E.D.; Shouman, M.A.; Karar, M.E. COVIDX-Net: A Framework of Deep Learning Classifiers to Diagnose COVID-19 in X-Ray Images. *arXiv* **2020**. [CrossRef]
13. Shaheed, K.; Szczuko, P.; Abbas, Q.; Hussain, A.; Albathan, M. Computer-Aided Diagnosis of COVID-19 from Chest X-ray Images Using Hybrid-Features and Random Forest Classifier. *Healthcare* **2023**, *11*, 837. [CrossRef]
14. DeGrave, A.J.; Janizek, J.D.; Lee, S.-I. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nat. Mach. Intell.* **2021**, *3*, 610–619. [CrossRef]
15. Bayram, F.; Eleyan, A. COVID-19 detection on chest radiographs using feature fusion based deep learning. *Signal Image Video Process.* **2022**, *16*, 1455–1462. [CrossRef] [PubMed]
16. Quiroz-Juárez, M.A.; Torres-Gómez, A.; Hoyo-Ulloa, I.; León-Montiel, R.D.J.; U'ren, A.B. Identification of high-risk COVID-19 patients using machine learning. *PLoS ONE* **2021**, *16*, e0257234. [CrossRef]
17. Barough, S.S.; Safavi-Naini, S.A.A.; Siavoshi, F.; Tamimi, A.; Ilkhani, S.; Akbari, S.; Ezzati, S.; Hatamabadi, H.; Pourhoseingholi, M.A. Generalizable machine learning approach for COVID-19 mortality risk prediction using on-admission clinical and laboratory features. *Sci. Rep.* **2023**, *13*, 2399. [CrossRef]
18. Aboutalebi, H.; Pavlova, M.; Shafiee, M.J.; Florea, A.; Hryniowski, A.; Wong, A. COVID-Net Biochem: An Explainability-driven Framework to Building Machine Learning Models for Predicting Survival and Kidney Injury of COVID-19 Patients from Clinical and Biochemistry Data. *arXiv* **2022**, arXiv:2204.11210.
19. Daamen, A.R.; Bachali, P.; Grammer, A.C.; Lipsky, P.E. Classification of COVID-19 Patients into Clinically Relevant Subsets by a Novel Machine Learning Pipeline Using Transcriptomic Features. *Int. J. Mol. Sci.* **2023**, *24*, 4905. [CrossRef] [PubMed]
20. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015, Proceedings of the 18th International Conference, Munich, Germany, 5–9 October 2015*; Springer: Cham, Switzerland; pp. 234–241. [CrossRef]
21. Jaeger, S.; Candemir, S.; Antani, S.; Wáng, Y.X.J.; Lu, P.X.; Thoma, G. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. *Quant. Imaging Med. Surg.* **2014**, *4*, 475–477.
22. Shiraishi, J.; Katsuragawa, S.; Ikezoe, J.; Matsumoto, T.; Kobayashi, T.; Komatsu, K.-I.; Matsui, M.; Fujita, H.; Kodera, Y.; Doi, K. Development of a Digital Image Database for Chest Radiographs With and Without a Lung Nodule. *Am. J. Roentgenol.* **2000**, *174*, 71–74. [CrossRef] [PubMed]
23. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269. [CrossRef]
24. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
25. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 Jun. 2016; pp. 2818–2826. [CrossRef]
26. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; Volume 31. [CrossRef]
27. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1800–1807. [CrossRef]
28. Tan, M.; Le, Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the 36th International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
29. McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **1947**, *12*, 153–157. [CrossRef]
30. Delong, E.R.; Delong, D.M.; Clarke-Pearson, D.L. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics* **1988**, *44*, 837–845. [CrossRef]
31. Sacchetto, D.; Raviolo, M.; Beltrando, C.; Tommasoni, N. COVID-19 Surge Capacity Solutions: Our Experience of Converting a Concert Hall into a Temporary Hospital for Mild and Moderate COVID-19 Patients. *Disaster Med. Public Health Prep.* **2022**, *16*, 1273–1276. [CrossRef] [PubMed]
32. Huang, S.-C.; Pareek, A.; Seyyedi, S.; Banerjee, I.; Lungren, M.P. Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines. *NPJ Digit. Med.* **2020**, *3*, 136. [CrossRef] [PubMed]
33. Liang, W.; Yao, J.; Chen, A.; Lv, Q.; Zanin, M.; Liu, J.; Wong, S.; Li, Y.; Lu, J.; Liang, H.; et al. Early triage of critically ill COVID-19 patients using deep learning. *Nat. Commun.* **2020**, *11*, 3543. [CrossRef] [PubMed]
34. Shamout, F.E.; Shen, Y.; Wu, N.; Kaku, A.; Park, J.; Makino, T.; Jastrzębski, S.; Witowski, J.; Wang, D.; Zhang, B.; et al. An artificial intelligence system for predicting the deterioration of COVID-19 patients in the emergency department. *NPJ Digit. Med.* **2021**, *4*, 80. [CrossRef]

35. Kwon, Y.J.; Toussie, D.; Finkelstein, M.; Cedillo, M.A.; Maron, S.Z.; Manna, S.; Voutsinas, N.; Eber, C.; Jacobi, A.; Bernheim, A.; et al. Combining Initial Radiographs and Clinical Variables Improves Deep Learning Prognostication in Patients with COVID-19 from the Emergency Department. *Radiol. Artif. Intell.* **2021**, *3*, e200098. [CrossRef] [PubMed]
36. Soda, P.; D'amico, N.C.; Tessadori, J.; Valbusa, G.; Guarrasi, V.; Bortolotto, C.; Akbar, M.U.; Sicilia, R.; Cordelli, E.; Fazzini, D.; et al. AIforCOVID: Predicting the clinical outcomes in patients with COVID-19 applying AI to chest-X-rays. An Italian multicentre study. *Med. Image Anal.* **2021**, *74*, 102216. [CrossRef] [PubMed]
37. Deb, S.D.; Jha, R.K.; Kumar, R.; Tripathi, P.S.; Talera, Y.; Kumar, M. CoVSeverity-Net: An efficient deep learning model for COVID-19 severity estimation from Chest X-Ray images. *Res. Biomed. Eng.* **2023**, *39*, 85–98. [CrossRef]
38. Mahmud, T.; Alam, J.; Chowdhury, S.; Ali, S.N.; Rahman, M.; Fattah, S.A.; Saquib, M. CovTANet: A Hybrid Tri-Level Attention-Based Network for Lesion Segmentation, Diagnosis, and Severity Prediction of COVID-19 Chest CT Scans. *IEEE Trans. Ind. Inform.* **2021**, *17*, 6489–6498. [CrossRef]
39. Ullah, Z.; Usman, M.; Latif, S.; Gwak, J. Densely attention mechanism based network for COVID-19 detection in chest X-rays. *Sci. Rep.* **2023**, *13*, 261. [CrossRef] [PubMed]