applied sciences

MDPI

Article

Prediction of COVID-19 from Chest CT Images Using an Ensemble of Deep Learning Models

Shreya Biswas [1], Somnath Chatterjee [2], Arindam Majee [1], Shibaprasad Sen [3], Friedhelm Schwenker [4,*] and Ram Sarkar [5]

[1] Department of Electronics and Telecommunication Engineering, Jadavpur University, Kolkata 700032, India; mimigg443@gmail.com (S.B.); majeearindam4@gmail.com (A.M.)

[2] Department of Computer Science and Engineering, Future Institute of Engineering and Management, Kolkata 700150, India; somnathchatterjee796@gmail.com

[3] Department of Computer Science and Engineering, University of Engineering and Management, Kolkata 700160, India; shibubiet@gmail.com

[4] Institute of Neural Information Processing, Ulm University, 89069 Ulm, Germany

[5] Department of Computer Science and Engineering, Jadavpur University, Kolkata 700032, India; ram.sarkar@jadavpuruniversity.in

* Correspondence: friedhelm.schwenker@uni-ulm.de; Tel.: +49-731-502-4159

Abstract: The novel SARS-CoV-2 virus, responsible for the dangerous pneumonia-type disease, COVID-19, has undoubtedly changed the world by killing at least 3,900,000 people as of June 2021 and compromising the health of millions across the globe. Though the vaccination process has started, in developing countries such as India, the process has not been fully developed. Thereby, a diagnosis of COVID-19 can restrict its spreading and level the pestilence curve. As the quickest indicative choice, a computerized identification framework ought to be carried out to hinder COVID-19 from spreading more. Meanwhile, Computed Tomography (CT) imaging reveals that the attributes of these images for COVID-19 infected patients vary from healthy patients with or without other respiratory diseases, such as pneumonia. This study aims to establish an effective COVID-19 prediction model through chest CT images using efficient transfer learning (TL) models. Initially, we used three standard deep learning (DL) models, namely, VGG-16, ResNet50, and Xception, for the prediction of COVID-19. After that, we proposed a mechanism to combine the above-mentioned pre-trained models for the overall improvement of the prediction capability of the system. The proposed model provides 98.79% classification accuracy and a high F_1-score of 0.99 on the publicly available SARS-CoV-2 CT dataset. The model proposed in this study is effective for the accurate screening of COVID-19 CT scans and, hence, can be a promising supplementary diagnostic tool for the forefront clinical specialists.

Keywords: COVID-19; transfer learning; chest CT scan image; ensemble learning

check for updates

Citation: Biswas, S.; Chatterjee, S.; Majee, A.; Sen, S.; Schwenker, F.; Sarkar, R. Prediction of COVID-19 from Chest CT Images Using an Ensemble of Deep Learning Models. *Appl. Sci.* **2021**, *11*, 7004. https://doi.org/10.3390/app11157004

Academic Editor: Syoji Kobashi

Received: 16 June 2021
Accepted: 24 July 2021
Published: 29 July 2021

1. Introduction

The pandemic of COVID-19 is causing a genuine emergency at the moment [1] all over the world. With over 182,006,598 total COVID-19 cases around the world and 3,942,777 deaths already, as indicated by the World Health Organization (WHO) statistics [2], this pandemic poses the biggest medical danger towards mankind till date. Though the death rate is less than 2%, the highly contagious nature of COVID-19 is considered the main concern for the world population. Clinical studies reveal that a COVID-19 infected person may experience a dry cough, muscle pain, headache, fever, sore throat, and mild to moderate respiratory illness. The low accessibility of testing kits poses another serious problem in terms of the efficiency of its detection. At present, the tests for detecting the presence of COVID-19 are performed based on real-time reverse transcription-polymerase chain reaction (RT-PCR) [3]. RT-PCR detection of viral RNA from sputum or nasopharyngeal swabs requires specific hardware and also has relatively low sensitivity. It takes a minimum of 4–6 h to generate results.

Another standard diagnosis of COVID-19 is currently the Nucleic Acid Amplification Testing (NAAT). However, NAAT is resource and time consuming and hence is not widely accessible. Ag-RDTs for SARS-CoV-2, usually in a Lateral Flow Immunoassay (LFI) cassette format, have recently been developed and commercialized [4]. These easy-to-use tests offer rapid case detection. However, its performance is highly variable based on the test characteristics and the population tested. Furthermore, the RDTs are sensitive only for the detection of patients with a high viral load and, hence, cannot be trusted as a major diagnostic method. This is why we choose CT scans in this paper, making our method more reliable.

According to [5], chest X-rays and CT scans of COVID-19 suspected patients may help to diagnose COVID efficiently yet in a fast way using artificial intelligence (AI) techniques, wherein the AI model learns by itself to differentiate COVID CT scans from non-COVID CT scans after studying a set of images. However, the problem with chest X-rays is that they cannot differentiate soft tissues accurately [6] and, hence, cannot be fed into the AI models for an all-round evaluation. To overcome this, CT-scans can be used. Several works [7,8] show considerable success in the application of AI and deep learning (DL) approaches for efficient detection of the disease from chest CT scans.

DL-based approaches [9] accelerate classification by reducing the need for hand-designing features. Deep neural networks (DNN) use abstraction to represent data without any application-specific descriptors, thereby eliminating the manual step of feature engineering and representation required in conventional machine learning (ML) [10,11]. Due to this, DL techniques have been used in computer science fields such as computer vision [12], speech recognition [13], and text processing [14].

In recent years, DL-based models have outperformed traditional statistical and ML-based strategies in most of the tasks. Computer vision and machine perception undoubtedly are some of the most influenced fields and have achieved great heights after the development of convolutional neural networks (CNN).

The field of medical imaging has seen far-reaching effects due to recent progress in computer vision, for detection of diseases such as cancer [15], lung diseases [16] pneumonia [17,18], seizures [19], MERS [20], SARS [21], drug discovery [22] and so forth. During the current COVID-19 pandemic, it has become even more important for such DL-based approaches to be used in real-time. DL-based models can potentially be of very high utility, especially when considering cost, speed, and ease. Many works have been done in the detection of COVID using DL [23–27].

In this era of big data and due to the availability of greater computation power, we can train these neural networks with ease. However, there are certain limitations to CNN models. For example, it requires a lot of time to train these networks to achieve appreciable performance. In addition, the dataset that is being used for training should contain all the variations of samples so that the model could easily generalize unseen data, but this is not always possible. If DL-based models are trained on a smaller dataset, over-fitting may occur, which in turn causes generalization errors. In real-world problems, we rarely have good quality labeled data with a significant number of instances for training a DL-based model. For example, in our present work, data augmentation can not be used since the dataset consists of chest CT-scan images that cannot be flipped, rotated, or sliced as those will not be realistic. To deal with all these problems, a TL-based approach is the most effective solution.

In this work, we have utilized the approach of TL [28] instead of building the network architecture from scratch. We have used three pre-existing models, namely VGG-16 [29], ResNet50 [30] and Xception [31] in our current experimentation. These three models have an innately dissimilar architecture that may abstract unrelated information from the images used for the classification purpose. To make the classification process more efficient, we have also used the ensemble-based learning concept. All of the three mentioned models are combined by the strategy of stacking ensemble inspired from stacked generalization [32] to make the final model generalize and also to perform better on the unseen data.

The rest of this paper has been divided into different sections as follows: Section 2 discusses the literature survey mentioning a few important COVID-19 works described by different authors, and Section 3 highlights the dataset used in the current experiment. The proposed methodology for the prediction of COVID-19 has been detailed in Section 4, the observed outcomes with in-depth analysis have been mentioned in Section 5, and lastly, Section 6 concludes the overall work with some future directions.

2. Related Work

This section concerns some recent works related to the detection of COVID-19 from chest CT scans and X-rays. Reference [7] proposed the use of a DCNN-based approach on CT scans for differentiating between COVID-19 and typical viral pneumonia cases, achieving a 73% recognition accuracy. Reference [17] used a local-attention-based mechanism to distinguish between COVID, influenza, and healthy CT scans. Reference [33] proposed a network structure where DenseNet was used for feature extraction and proposed a DL model called DenseCapsNet to detect COVID-19 from a chest X-ray, achieving 98% accuracy. Reference [34] used a support vector machine (SVM) [35] to classify X-rays. Reference [36] used ML techniques for the extraction of graphical features from chest X-rays for COVID-19 detection. Several studies exist where the CT images are firstly segmented to highlight the ROI after which several strategies are employed to detect and classify them. Reference [37] proposed an approach to improve the segmentation of CT images using a modified U-Net architecture, which eliminates several drawbacks of the conventional U-Net architecture. Reference [38] used TL on DenseNet-121 to classify COVID-19. Their website takes radiology images, outputting the infected regions with an accuracy of 87%.

Recently, Yang et al. [39] released a public dataset consisting of CT scans collected from studies in this domain. They also tested the application of DenseNet to distinguish COVID-19 positive cases from the negatives with an accuracy of 84.7%. Reference [40] developed a spiking neural network (SSN) by copying biological models, achieving a high F_1 score. Chattopadhyay et al. [41] proposed a computationally economical method in which they extracted features of the CT scans, optimizing them by a clustering-based golden ratio optimizer (CGRO), and attained state-of-the-art accuracies on publicly available datasets. Authors in [42] proposed a SqueezeNet based model to distinguish COVID-19 CT images from other images and reported 85% sensitivity and a 0.8333 F_1-Score. Apart from using CT images, Reference [43] explored a lung ultrasound (POCUS) dataset to show the importance and significance for COVID-19 detection. Their DL model (POCOVID-Net) was pre-trained on ImageNet to extract features from the images. They have reported a sensitivity of 0.96 and an F_1-Score of 0.92 using a 5-fold cross-validation.

Sen et al. [44] proposed a bi-stage framework for the recognition of COVID patients from chest CT-scan images. They extracted features from the DL-based models from which relevant features are selected using guided features selection methodology. These produced features are optimized using the Dragonfly algorithm producing a good recognition rate. Karbhari et al. [45] introduced a synthetic chest X-ray generating model termed as auxiliary classifier generative adversarial Network (aCGaN). The images obtained are classified using custom-made DL-based models producing a decent accuracy. They also used the harmony search algorithm to optimize the produced features and retained the classification accuracy. Das et al. [46] used chest X-ray images for the efficient detection of COVID-19 using the VGG-19 model. Instead of directly predicting the class from the DL-based model, a feature extraction technique is employed. Extracted features from the VGG-19 model are fed into traditional machine learning models (logistic regression) to appropriately classify whether it is a COVID-positive or healthy patient. It achieved an accuracy of 99.26%. To evaluate the predictive power of DenseNet201 in COVID-19 identification using CT images, Reference [47] used pre-trained DenseNet201 on ImageNet for features extraction and modified dense layers to get the final output. Reference [48] used pre-trained U-Net for segmentation of 3D lung images. This helped them to predict the infected regions with an accuracy of 95.9%. Reference [49] developed CONVNet to extract chest CT features for detection of COVID with 95% classification accuracy.

From the above-mentioned research works, it can be observed that many authors have worked to serve society by diagnosing a dangerous COVID-19 infection and also achieved good recognition accuracy. However, there is enough scope to improve the prediction accuracy. Hence, in this paper, we have proposed a prediction model by using ensemble learning of three pre-trained models that assures a high accuracy and F_1-score in classifying COVID-19 cases.

3. Database Used

The SARS-CoV-2 CT scan dataset [50] was used in the current experiment for the detection of COVID-19 cases. The dataset consists of a total of 2482 CT-scan images, out of which 1252 are positive for SARS-CoV-2 infection i.e., COVID-19 (+ve), and the remaining 1230 are negative for SARS-CoV-2 infection, i.e., COVID-19 (-ve). These data have been collected from real patients in hospitals from Sao Paulo, Brazil. The aim of this dataset is to encourage the research and development of AI methods that are able to identify SARS-CoV-2 infection through the analysis of CT scans.

Figure 1a,b shows sample CT scan images of COVID-19 (+ve) and COVID-19 (-ve) patients from the mentioned dataset.

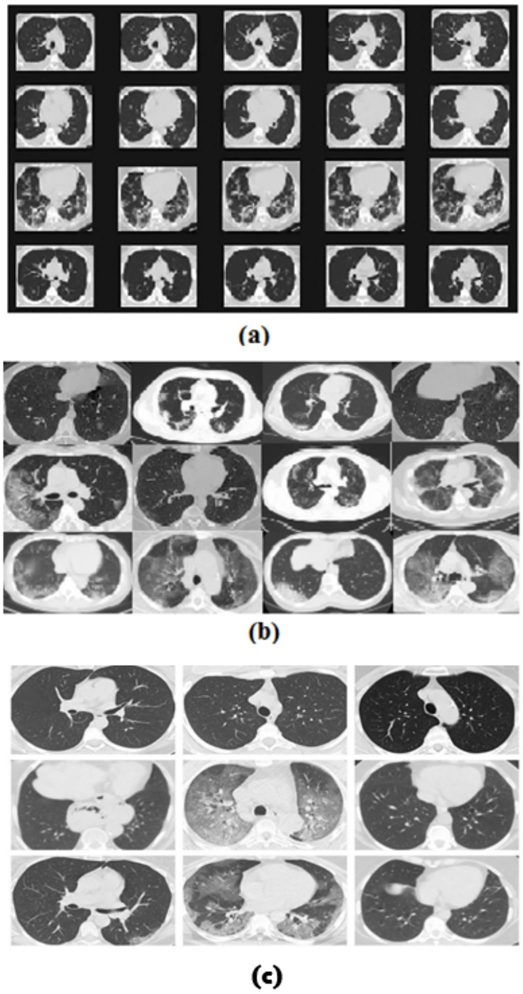

Figure 1. Sample images taken from (**a**) SARS-CoV-2 CT scan dataset, (**b**) COVID-CT database images that are positive for COVID-19, and (**c**) COVID-CT database images that are negative for COVID-19.

4. Proposed Work

A number of research studies have been performed for COVID-19 detection from chest X-ray and CT-scan images to date. CT scans provide low false-positive rates than X-rays [51], which forms the backbone of our proposed work. The main issue was to train a model to give the desired results. To solve this issue, we have used transfer learning. It focuses on storing knowledge gained while solving one problem and applying it to a different but related problem. The new learning task may belong to a different feature space or distribution. In this study, we have a classification task in one domain of interest (chest CT images), while we have considered sufficient training data from a different domain than of the images from ImageNet. In such cases, knowledge transfer, if done successfully, would greatly improve the performance of learning of the new task. It also reduces the training time and helps to avoid overfitting of the data. In the current experiment, we have utilized three pre-trained models, namely VGG-16, ResNet50, and Xception, which are initially trained on the ImageNet dataset consisting of 1.2 million high-resolution images for the ImageNet LSVRC contest to classify into 1000 different classes. When using TL on a new dissimilar dataset, preliminary layers are frozen as they extract generic features, but the later layers in the network extract specific and complex features.

Though many precautions, such as Dropout- [52], Batch Normalization- [53] etc., can be considered to avoid overfitting and to get the desired performance from the DL models, we found that there is room for improvement. Hence, to build an efficient COVID-19 prediction model, we used the concept of ensemble learning. After analyzing different ensemble strategies and their uses, we applied a stacking ensemble most appropriate for this task [32,54]. The flowchart in Figure 2 describes the complete pipeline used in the present work.

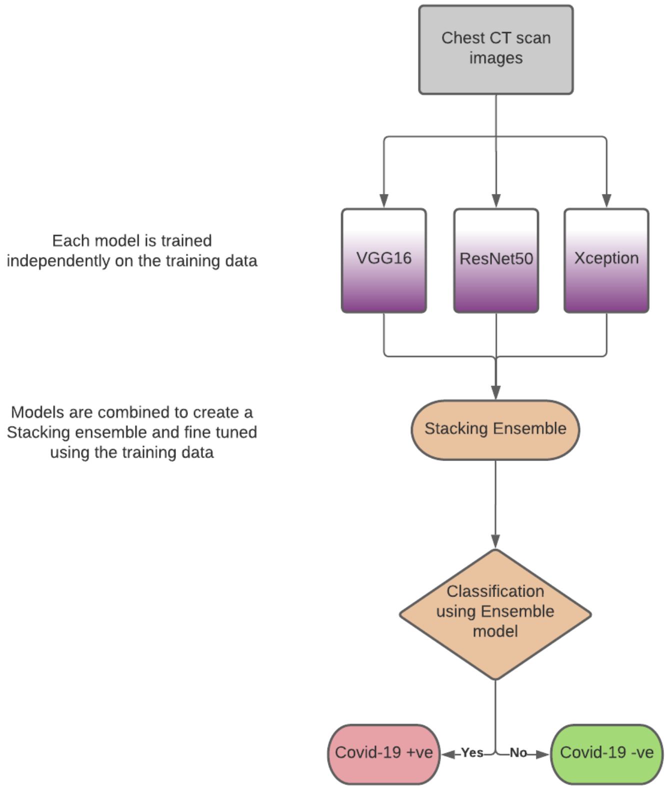
Chest CT scan images
Each model is trained independently on the training data
VGG16
ResNet50
Xception
Models are combined to create a Stacking ensemble and fine tuned using the training data
Stacking Ensemble
Classification using Ensemble model
Covid-19 +ve
Yes
No
Covid-19 -ve

Figure 2. Flowchart of the proposed model.

4.1. Architecture of the Proposed Model

This section briefly describes the different components of the proposed model for the prediction of COVID-19 cases.

VGG-16: This architecture was proposed by Simonyan et al. [29]. VGG-16 was one of the best performing architectures in the ILSVRC challenge 2014. The main specialty of this DCNN is its small kernel size. It uses a kernel of size 3 × 3, which is repeated over 256 and 512 times in the layers. This helps the model to capture localized features peculiar to a particular class and thus improves the classification performance. There exists some drawbacks to using small kernel sizes in VGG architecture. As convolutions used in the VGG model are small, it increases the number of parameters to train. It also uses pooling layers at the appropriate position to eliminate irrelevant features and helps to decrease the complexity of the model.

ResNet50: ResNet is a residual learning framework to ease the training of networks that are substantially deeper than those used previously [30]. The layers are explicitly reformulated as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions. Residual can simply be understood as a subtraction of features learned from the input of that layer. This is done by creating alternate connections, which directly connects the nth layer to the $(n + x)$th layer. This allows the propagation of dominant features much deeper into the network. It also helps to avoid the vanishing gradient problem. These residual networks are easier to optimize and can gain accuracy from considerably increased depth.

Xception: The Xception model architecture [31] can be understood as an extreme customization of inception model architecture. In a traditional DCNN, each layer learns and transforms information obtained from its previous layer, but the usability of this information obtained can be highly influenced by the type of transformation performed. The Xception model takes advantage of this by applying convolutions using different kernel sizes and concatenates them and passes to the next layer. This allows the model to compute the most dominant features and eliminate others. It also maps spatial correlations and cross-channel correlations of the image separately, which is essentially equivalent to an existing operation known as a “depthwise separable convolution”. This makes Xception a highly efficient architecture that learns about several distinctive characteristics and high-level features of the input data, which some simpler models might overlook.

In the present work, all these three models are used with pre-trained weights and biases from the ImageNet dataset for the COVID-19 prediction. Few fully connected layers having rectified linear unit (ReLU) activation and a final classification layer having Sigmoid activation were added so that they work better on the dataset. These models are then used as base-learners in the stacking ensemble architecture.

Here, we have aggregated the submodels, their predictions being blended together to give a better result than the individual predictions. We concatenated the individual models, thereby aggregating them into a single ensemble, and then further added two dense layers, one with ReLu and one with the Softmax activation function. Using CrossEntropy as the loss function, we then trained our stacked model.

4.2. Stacking Ensemble Learning

Ensemble-based systems are formed by combining diverse models together to give robust predictions and to minimize the probability of erroneous predictions [55,56]. This kind of procedure has been extensively used by humans in daily lives, such as asking the opinions of several experts before making a decision, e.g., in the context of medical treatments. Ensemble methods include stacking, as well as boosting and bagging [56], and have been applied successfully in various regression and classification applications [57,58] or in the field of reinforcement learning [59] where an ensemble of agents is voting on the optimal action.

Before stacking ensembles, there were many methods to integrate various models together. For example, in the MAET (model averaging ensemble technique), probabilities from multiple classification models or estimators are combined in equal proportions. Though it is a simple mechanism, it may not be a good proposal regarding the flexibility of the final model. This is because different models will contribute equivalently to the final prediction. An alternative way is to combine the predictions from several models in variable proportion is depending upon a trust score. This trust score or confidence score can be computed depending on the performance of the individual models on unseen training data. This allows better models to contribute more and poor models to contribute less. This technique is called a weighted average ensemble.

Stacked generalization works by deducing the biases of the classifiers with respect to a provided training set. This deduction proceeds by classifying in a second space whose inputs are the predictions of the original classifiers when taught with part of the training set and trying to predict on the holdout dataset for performance assessment. Stacked generalization can be seen as a sophisticated version of cross-validation [32]. In the second feature space, a meta-learner can be trained from scratch to efficiently combine the predictions of the individual models or base models and finally predict the most probable output. This selection of meta-learner greatly impacts the performance of the ensemble model. In the present study, we have used a fully connected neural network as a meta-learner. The aforementioned base models are frozen and only the meta-learner is trained again on the training set to decrease the generalization error and to make better predictions. Figure 3 shows the general architecture of a two-level ensemble model.

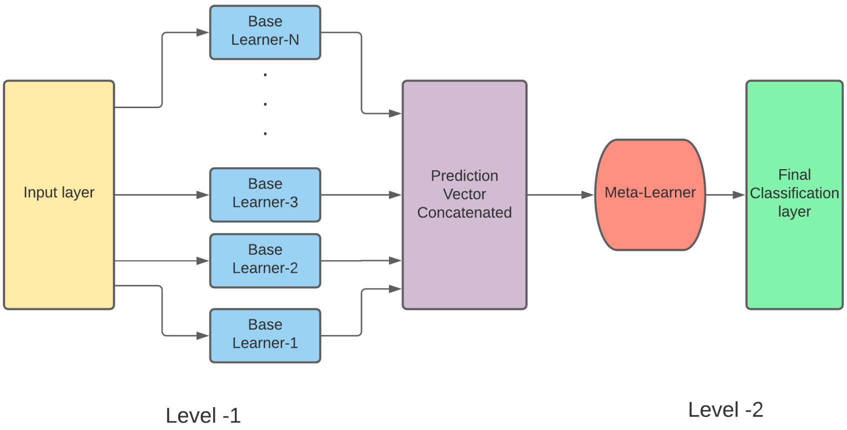

Figure 3. The general architecture of a two-level ensemble model.

5. Experimental Results and Discussion

Training DL models with limited data without overfitting is indeed a challenging task. The proposed TL-based approach resolves the problem with significant improvement in the performance by involving the models VGG-16, ResNet50, and Xception. Afterward, these models are used to create a stacking ensemble classifier for improvement of the overall recognition accuracy. The stacking ensemble classifier used in the present experiment can be thought of as a complex model combining all the mentioned base learners. It has a single input layer that duplicates and distributes the input data into three base learners. These input images are propagated through each base learner separately, and finally, a prediction vector is generated from each base learner predicting the class labels of the input data. These prediction vectors generated from all these three base learners are concatenated and used as a feature for the meta learner. The meta-learner then tries to classify the input data into appropriate classes in an efficient way. In the stacking ensemble model, only the meta-learner is trained on the training data while the base learners are frozen. In the current experiment, the base models are trained for 100 epochs each with batch size 32, and the meta-learner has been trained for 100 epochs on training data. All possible combinations of the base learners have been experimented with to estimate the usage and performance of the ensembled architecture. In the present experiment, the training and test sets are split into an 8:2 ratio.

To evaluate how effective the proposed model is, we have measured its performance in terms of five metrics that include classification accuracy, precision, recall, F_1-score, and area under the curve (AUC). The receiver operating curve (ROC) is a plot of true positive rate (i.e., TPR or sensitivity) versus false-positive rate (FPR).

Equations (1)–(5) have been used to measure the value of FPR, accuracy (a_c), precision (P_r), recall (R_c), and F_1-score (F_s).

$$FPR = \frac{FP}{TN + FP} \tag{1}$$

$$a_c = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$P_r = \frac{TP}{TP + FP} \tag{3}$$

$$R_c = \frac{TP}{TP + FN} \tag{4}$$

$$F_s = 2 \times \frac{P_r \times R_c}{P_r + R_c} \tag{5}$$

where true positive (TP) represents COVID-19 (+ve) patients that are correctly recognized as COVID-19 (+ve). False positive (FP) represents COVID-19 (-ve) patients that are incorrectly recognized as COVID-19 (+ve). True negative (TN) represents COVID-19 (-ve) patients that are correctly recognized as COVID-19 (-ve). Finally, false negative (FN) represents COVID-19 (+ve) patients that are incorrectly recognized as COVID-19 (-ve).

Table 1 reflects a summary of accuracy, precision, recall, F_1-score, and AUC measurements achieved from the mentioned three models for the prediction of COVID-19 on the SARS-CoV-2 CT scan dataset. Figure 4 reflects the ROC curve obtained for the above mentioned three models. From Table 1, we can observe that the VGG-16 model outperforms the other pre-trained models taken into consideration. VGG-16 (with around 138M parameters) has a kernel of size 3 × 3 for all the convolution layers and the size of the kernel in the Maxpool layer is 2 × 2 with a stride of 2. The number of trainable parameters has been reduced by 44.9%. A reduced number of trainable variables means faster learning and is more robust against over-fitting. This might be the reason why it performs better. ResNet50 has around 23 million trainable parameters. It consists of convolution layers with filters of size 3 × 3 (just like VGGNet). Two pooling layers are used throughout the network. Hence, there is a large number of trainable variables, unlike VGG-16. In the Xception model, there is no intermediate rectified linear unit (ReLU) non-linearity. Xception models are also expensive to train due to their modified depth-wise separable convolution as compared to inception models.

Table 1. Details of the performance of the models for the prediction of COVID-19 on the SARS-CoV-2 CT scan dataset.

Model	Accuracy (%)	Precision	Recall	F_1-Score	AUC Score
VGG-16	98.39	0.9839	0.9839	0.9839	0.9839
ResNet50	95.17	0.9517	0.9517	0.9517	0.9515
Xception	94.57	0.9457	0.9457	0.9457	0.9456

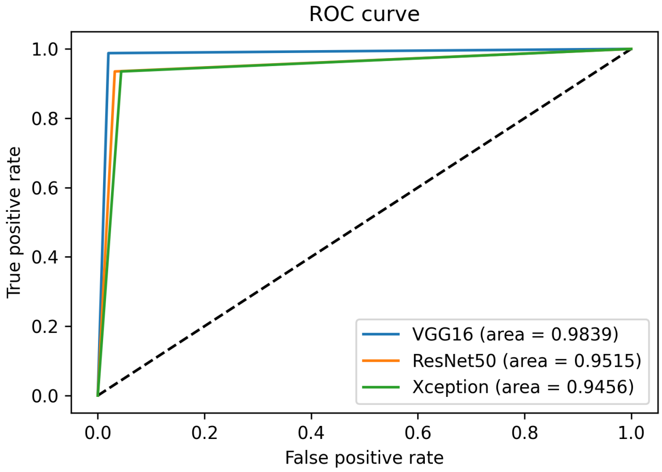

Figure 4. ROC curve obtained for VGG-16, ResNet50, and Xception model when considering the SARS-CoV-2 CT scan dataset.

We have also stacked the different combinations of these three models for the improvement of the model for the prediction of COVID-19. Table 2 reflects the outcomes observed for different combinations of the used models. After analyzing Table 2, it can be said that the ensembling of models (any combination) produces better recognition accuracy than the individual one. The best recognition accuracy of 98.79% was observed when all the three models were ensembled. For this model, the observed classification accuracy, precision, recall, and F_1-score are 98.79%, 0.99, 0.99, and 0.99, respectively. It can be seen from Table 1 that Xception does not perform well considering all the metrics when compared to the other two base models, namely VGG-16 and ResNet50. However, the stacking ensemble consisting of all three base models outperforms all other combinations. We used a fully connected neural network as a meta-learner to provide flexibility to the stacked model and decrease the generalization error. Thus, the neural network combines the predictions obtained from all the base models in such a way that the performance of the overall model improves. The neural network is tuned so that it can ignore the wrong predictions made by the base models and utilizes only those predictions that help improve the classification score. Different CNN models commit errors on different samples; thus, aggregating those together helps to achieve a better accuracy on the test set. Image index 165 belongs to a test set and was originally a COVID image. It is misclassified by VGG-16 as a non-COVID image but properly classified by ResNet50 and Xception. It is also misclassified by both VGG-16+ResNet50 and VGG-16+Xception combinations, but in the stacked model, it is properly classified. Similarly, image index 170 belongs to the test set and is originally a COVID +ve CT-scan, which is misclassified by Xception but accurately recognized by VGG-16 and ResNet50. As all three models are included in the final ensemble architecture, it gets correctly classified by the stacking ensemble model. The aforementioned images can be seen in Figure 5. The complete test set contains 497 samples. The pronounced difference in the performance of the said models on the test set can be visualized from confusion matrices shown in Figure 6. In the final confusion matrix of the stacking ensemble model that can be seen in Figure 7, only four COVID +ve images are incorrectly classified as non-COVID, and two non-COVID samples are recognized as COVID samples. This can be considered as a significant improvement compared to other models.

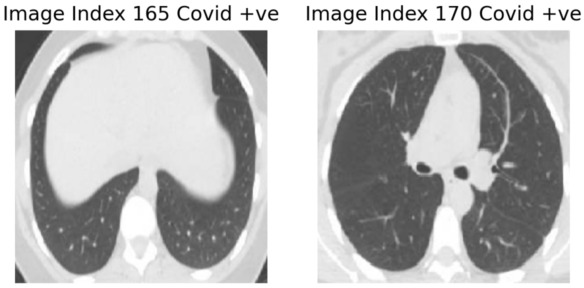

Figure 5. Images misclassified by the base models but accurately recognized by the proposed ensemble framework.

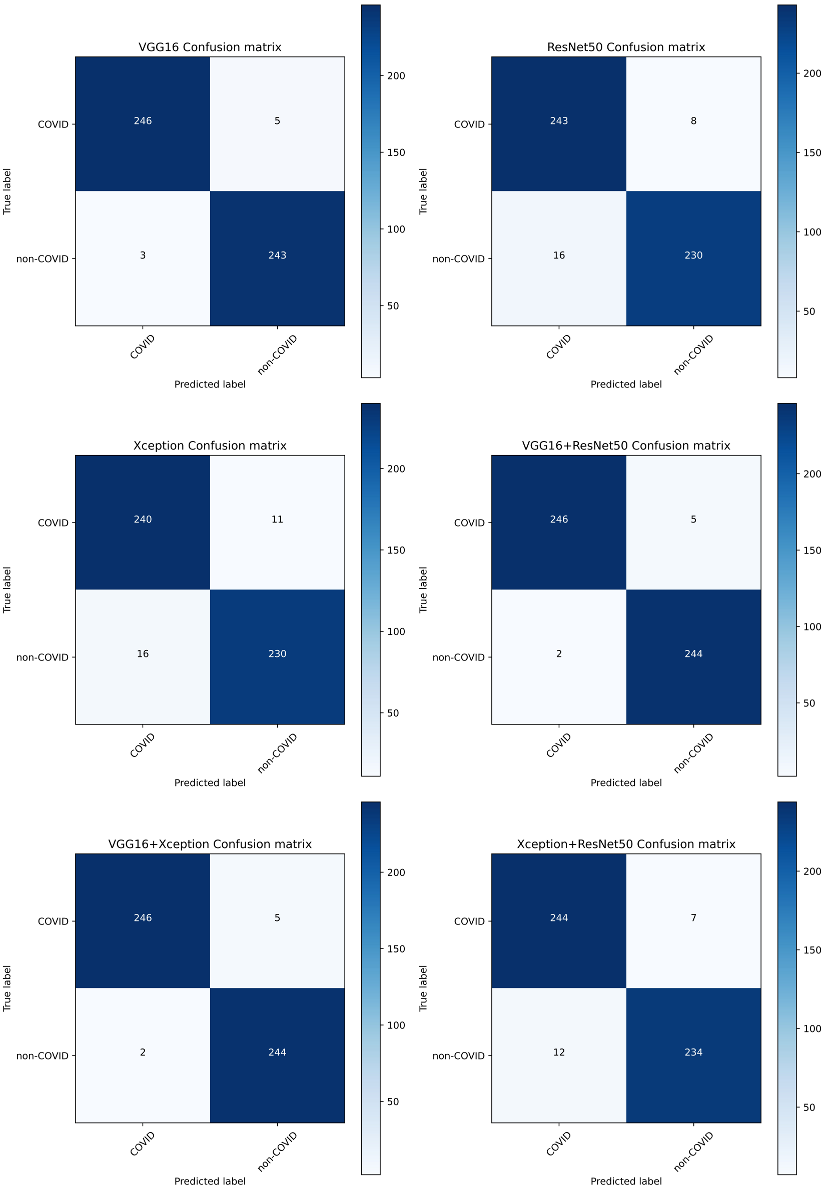

Figure 6. Confusion matrices of base models and different combinations of the ensemble model.

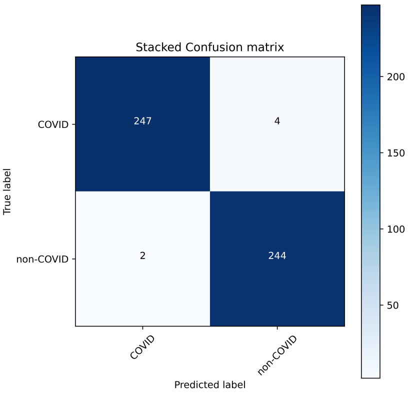

Figure 7. Confusion matrix of the ensemble model consisting of all three base models.

Table 2. Performance of the ensemble models (different combinations) for the prediction of COVID-19 on the SARS-CoV-2 CT scan dataset.

Model	Accuracy (%)	Precision	Recall	F_1-Score	AUC
VGG-16+ResNet50	98.59	0.9859	0.9859	0.9859	0.9860
VGG-16+Xception	98.59	0.9859	0.9859	0.9859	0.9860
ResNet50+Xception	96.18	0.9618	0.9618	0.9618	0.9617
VGG16+ResNet50+Xception	98.79	0.9879	0.9879	0.9879	0.9880

In medical research, especially for critical diseases like COVID-19, it is significantly important to reduce the FP and FN rate as much as possible when designing the prediction model. FN should be as low as possible, because if COVID-19 (+ve) patients get wrongly classified as a COVID-19 (-ve), it may cause otherwise avoidable deaths again, and misclassification of COVID-19 (-ve) cases with COVID-19 (+ve) (FP) may lead to unnecessary emotional disruption for an individual. Hence, it is also essential to lower the number of FP cases. For exhaustive testing, we have performed 5-fold cross-validation for the best performing model (when combined all three models) on the total dataset. In this situation, the overall accuracy gets increased to 98.79%. Figure 8 shows the ROC curve obtained when all the three models are combined.

After analyzing the outcomes, we observed that the measurements of FN (2) and FP (3) in our proposed ensemble model are significantly less. Hence, the proposed model can be an alternative choice to rapid COVID-19 testing kits.

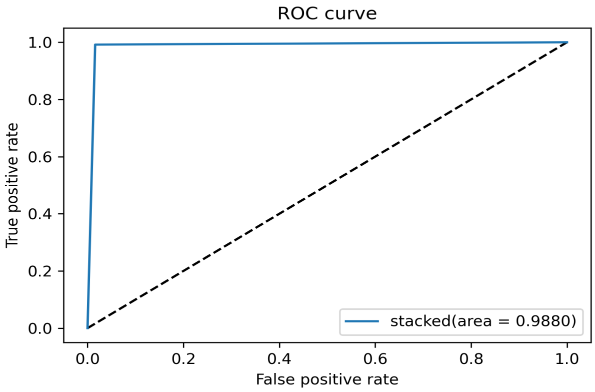

Figure 8. ROC curve obtained for all the combined models when considering SARS-CoV-2 CT scan dataset.

5.1. Grad-CaM Visualization of CT-Scans

Trusting a computer-aided architecture for predicting whether a patient suffers from a disease without proper explanation or reasoning should not be done. DL models have provided several breakthroughs in varied domains of applications. However, when these systems fail to provide an accurate result, it is difficult to say why this has happened. The reason being that DL-based models cannot be segregated into smaller intuitive units, making it difficult to understand. Before adapting these intelligent systems to our daily lives, we need an understanding of the models as to why they predict what they predict. The transparency and the ability to understand the models are useful in many ways. It helps to decode the feature extraction ability, which, in turn, highlights the discriminating features among categories that may not be visible to human eyes. In such cases, Grad-CaM visualization can be used to find the cause of wrong classification.

When the use case of a computer-aided framework is so sensitive, visual inspection is truly required. To create such visualizations, Grad-CaMS [60] or gradient-weighted class activation mapping can be utilized. It uses the gradients of any target concept flowing into the final convolutional layer to produce a coarse localization map, highlighting the important regions in the image for predicting the concept. To obtain the class discriminative localization map of width u and height v for any class c, at first, the gradient of the score for the class c is computed. These gradients flowing back are global average-pooled to obtain the neurons' importance weights for the target class. After calculating the importance for the target class c, a weighted combination of activation maps is performed followed by ReLU activation. This results in a coarse heatmap of the same size as that of the convolutional feature maps. ReLU is applied to see the positive gradient have influence on the class of interest. As mentioned earlier, three DL-based models are used as base learners in this study. We tried to generate the Grad-CaM images to highlight the most important areas of an image, which influence the prediction made by the base learners. Figure 9 shows the produced Grad-cam images along with the original CT-scan images on the left.

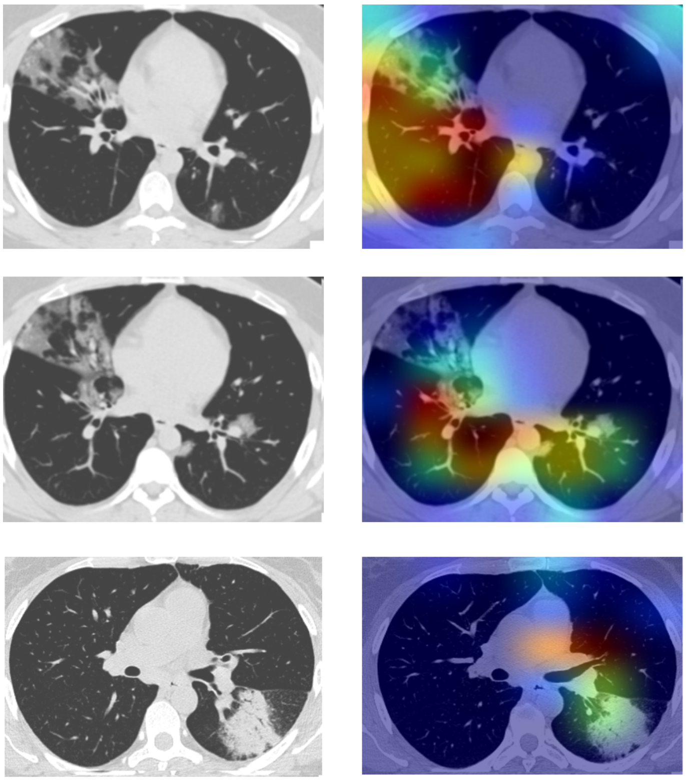

Figure 9. Grad-CaM visualizations generated from the base learners highlighting the region of interest.

5.2. Statistical Analysis

To prove that our ensemble model performs better than the individual base models, we have performed the statistical hypothesis testing using the paired sample *t*-test, also known as Student's *t*-test [61]. It is one of the trusted statistical tests performed to find if there exists a significant difference between two sets of observations. It only allows testing the pairs of observations, and a conclusion can be drawn depending upon the significance level and obtained *p*-values. Therefore, we have compared all the independent DL models and the ensemble model, selecting two at a time with the final stacking ensemble framework. The null hypothesis is assumed as there exists no significant difference between the performance of the individual models and the ensemble framework. On the other hand, an alternative hypothesis states there is a statistically significant difference between the performances of the two compared models. All the statistical trials are performed at a significance level of 5% or the alpha is 0.05. The obtained *p*-values are shown in Table 3. From the resulting *p*-values, which are all less than 0.05, we reject our assumed null hypothesis. This proves that the better performance of the ensemble framework is statistically significant and not a result obtained by chance.

Table 3. The *p*-values of the models obtained after Student's *t*-test.

	VGG16	**ResNet50**	**Xception**	**VGG16+ResNet50**	**VGG16+Xception**	**ResNet50+ Xception**
Stacked model	0.00201	0.01810	0.01207	0.00402	0.0039	0.01207

Table 4 compares the proposed model for the prediction of the COVID-19 disease with a few related past methods on the SARS-CoV-2 CT scan dataset. From this table, it can be observed that for the prediction of COVID-19, Soares et al. [50] used an explainable DL method while Jaiswal et al. [47] used DenseNet201, a pre-trained DCNN model. The authors mentioned in [47,50] achieved 97.38% and 96.25% recognition accuracies, respectively. In contrast, our proposed model successfully detects COVID-19 cases with 98.79% correct classification accuracy, which is higher than in [47,50].

Table 4. Comparison of the proposed model with some past methods for the prediction of COVID-19 on the SARS-CoV-2 CT scan dataset.

Approach	**Accuracy (%)**	**Precision**	**Recall**	**F_1-Score**	**AUC**
eXplainable DL approach	97.38	0.9916	0.9553	0.9731	0.9736
DenseNet201based deep TL	96.25	0.9629	0.9629	0.9629	-
Proposed Method	98.79	0.9879	0.9879	0.9879	0.9880

5.3. Additional Experiments

We have performed some additional experiments using the publicly available pneumonia dataset (https://www.kaggle.com/anaselmasry/covid19normalpneumonia-ct-images, accessed on 20 July 2021). As the CT scans from the pneumonia dataset are incorporated in the database, eventually the domain of the study shifts from a two-class classification problem to a three-class classification problem. This helps to evaluate the robustness and reliability of the proposed stacking ensemble methodology. As mentioned previously, the dataset considered in this study consists of a total of 2492 CT-scan images. Additionally, 1000 samples from pneumonia CT scans are merged with the existing database. This creates three categories of images, namely patients infected by COVID-19, healthy persons and patients infected by bacterial or viral pneumonia. The merged dataset is divided into 80% training data and 20% validation data. The three transfer learning models are trained for 100 epochs each after modifying the final classification layer for the three-class classification problem. These transfer learning models are used as base learners to create the stacking ensemble framework. The obtained results are shown in Table 5. From the results, it can be seen that the ensemble framework outperforms the individual models in all metrics considered and achieves an accuracy of 98.85%. This validates the performance of the ensemble framework. In other words, we can say that if there is an increase in the number of classifying categories, the model performs as expected not only in terms of accuracy but also in the other metrics (precision, recall, F_1 score).

Table 5. Results obtained after conducting experiments on the 3-class CT scan database (COVID-19 affected, normal and pneumonia-affected).

Model	Accuracy	Precision	Recall	F_1-Score
VGG-16	98.71	0.9871	0.9871	0.9871
ResNet-50	97.70	0.9770	0.9770	0.9770
Xception	97.13	0.9716	0.9713	0.9713
Stacked Model (VGG-16+Xception+ResNet-50)	98.85	0.9894	0.9894	0.9892

6. Conclusions

In this paper, we used three models—VGG-16, ResNet50, and Xception—for the prediction of the COVID-19 disease and achieved 98.39%, 94.57%, and 96.17% recognition accuracies. We proposed an ensemble-based learning approach by combining the power of these said three models in search of increasing prediction capability and achieved 98.79% accuracy. We achieved better recognition accuracy for the prediction of COVID-19 cases as compared to the methods mentioned in Table 4. However, we observed a few error cases through false positives and false negatives. Looking into the error cases, we have analyzed that the lack of ample historical COVID data and the poor quality of some images may be the probable causes. As more and more COVID-19 cases are being identified all over the world, larger datasets are being generated. In the future, we will apply our proposed model to these datasets to test the robustness of the model, trying to improve the prediction of the model. We have also planned to enhance the quality of the COVID-19 images by some pre-processing techniques and use some well-established pre-trained CNN models to have better features at the initial stage to improve the overall recognition accuracy. The python implementation of the above framework is available at https://github.com/somnath796/Covid-19_stacking_ensemble (accessed on 20 July 2021).

Author Contributions: Conceptualization, S.B., S.C. and A.M.; software, S.B., S.C. and A.M.; validation, S.S., F.S. and R.S.; writing—original draft preparation, S.B., S.C. and A.M.; writing—review and editing, S.S., F.S. and R.S; supervision, S.S., F.S. and R.S; project administration, R.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Experiment used publicly available datasets.

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

References

1. Available online: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it (accessed on 17 May 2021).
2. Available online: https://www.who.int/emergencies/diseases/novel-coronavirus-2019 (accessed on 17 May 2021).
3. Emery, S.L.; Erdman, D.D.; Bowen, M.D.; Newton, B.R.; Winchell, J.M.; Meyer, R.F.; Tong, S.; Cook, B.T.; Holloway, B.P.; McCaustland, K.A.; et al. Real-time reverse transcription-polymerase chain reaction assay for SARS-associated coronavirus. *Emerg. Infect. Dis.* **2004**, *10*, 311–316. [CrossRef] [PubMed]
4. Available online: https://www.who.int/news-room/articles-detail/sars-cov-2-antigen-detecting-rapid-diagnostic-test-implementation-projects (accessed on 29 June 2021).

5. Singh, D.; Kumar, V.; Vaishali.; Kaur, M. Classification of COVID-19 patients from chest CT images using multi-objective differential evolution-based convolutional neural networks. *Eur. J. Clin. Microbiol. Infect. Dis. Off. Publ. Eur. Soc. Clin. Microbiol.* **2020**, *39*, 1379–1389. [CrossRef]
6. Tingting, Y.; Junqian, W.; Lintai, W.; Yong, X. Three-stage network for age estimation. *CAAI Trans. Intell. Technol.* **2019**, *4*, 122–126. [CrossRef]
7. Wang, S.; Kang, B.; Ma, J.; Zeng, X.; Xiao, M.; Guo, J.; Cai, M.; Yang, J.; Li, Y.; Meng, X.; Xu, B. A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19). *medRxiv* **2020**. [CrossRef]
8. Elaziz, M.A.; Hosny, K.M.; Salah, A.; Darwish, M.M.; Lu, S.; Sahlol, A.T. New machine learning method for image-based diagnosis of COVID-19. *PLoS ONE* **2020**, *15*, e0235187. [CrossRef] [PubMed]
9. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
10. Lowe, D.G. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
11. Winder, S.A.J.; Brown, M. Learning local image descriptors. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8.
12. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
13. Deng, L.; Li, J.; Huang, J.T.; Yao, K.; Yu, D.; Seide, F.; Seltzer, M.; Zweig, G.; He, X.; Williams, J.; Gong, Y.; Acero, A. Recent advances in deep learning for speech research at Microsoft. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 8604–8608. [CrossRef]
14. Torfi, A.; Shirvani, R.A.; Keneshloo, Y.; Tavvaf, N.; Fox, E. Natural Language Processing Advancements by Deep Learning: A Survey. *arXiv* **2020**, arXiv:2003.01200.
15. Aboutalib, S.S.; Mohamed, A.A.; Berg, W.A.; Zuley, M.L.; Sumkin, J.H.; Wu, S. Deep learning to distinguish recalled but benign mammography images in breast cancer screening. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **2018**, *24*, 5902–5909. [CrossRef]
16. Bharati, S.; Podder, P.; Mondal, M.R.H. Hybrid deep learning for detecting lung diseases from X-ray images. *Inform. Med. Unlocked* **2020**, *20*, 100391. [CrossRef]
17. Xu, X.; Jiang, X.; Ma, C.; Du, P.; Li, X.; Lv, S.; Yu, L.; Ni, Q.; Chen, Y.; Su, J.; et al. A Deep Learning System to Screen Novel Coronavirus Disease 2019 Pneumonia. *Engineering* **2020**, *6*, 1122–1129. [CrossRef]
18. Stephen, O.; Sain, M.; Maduh, U.J.; Jeong, D.U. An Efficient Deep Learning Approach to Pneumonia Classification in Healthcare. *J. Healthc. Eng.* **2019**, *2019*, 4180949. [CrossRef]
19. Kuhlmann, L.; Lehnertz, K.; Richardson, M.P.; Schelter, B.; Zaveri, H.P. Seizure prediction—Ready for a new era. *Nat. Rev. Neurol.* **2018**, *14*, 618–630. [CrossRef]
20. Al-Turaiki, I.M.; Alshahrani, M.; Almutairi, T. Building predictive models for MERS-CoV infections using data mining techniques. *J. Infect. Public Health* **2016**, *9*, 744–748. [CrossRef]
21. Mateos, P.A.; Balboa, R.F.; Easteal, S.; Eyras, E.; Patel, H.R. PACIFIC: A lightweight deep-learning classifier of SARS-CoV-2 and co-infecting RNA viruses. *Sci. Rep.* **2020**. [CrossRef]
22. Jiménez, J.; Škalič, M.; Martínez-Rosell, G.; Fabritiis, G.D. KDEEP: Protein-Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. *J. Chem. Inf. Model.* **2018**, *58*, 287–296. [CrossRef] [PubMed]
23. Ghadezadeh, M.; Asadi, F. Deep Learning in the Detection and Diagnosis of COVID-19 Using Radiology Modalities: A Systematic Review. *J. Healthc. Eng.* **2021**, 1–10. [CrossRef] [PubMed]
24. Desai, S.B.; Pareek, A.; Lungren, M.P. Deep learning and its role in COVID-19 medical imaging. *Intell. Based Med.* **2020**, *3–4*, 100013. [CrossRef] [PubMed]
25. Dey, S.; Bhattacharya, R.; Malakar, S.; Mirjalili, S.; Sarkar, R. Choquet fuzzy integral-based classifier ensemble technique for COVID-19 detection. *Comput. Biol. Med.* **2021**, *135*, 104585. [CrossRef] [PubMed]
26. Bandyopadhyay, R.; Basu, A.; Cuevas, E.; Sarkar, R. Harris Hawks optimisation with Simulated Annealing as a deep feature selection method for screening of COVID-19 CT-scans. *Appl. Soft Comput.* **2021**, *111*, 107698. [CrossRef]
27. Kundu, R.; Basak, H.; Singh, P.K.; Ahmadian, A.; Ferrara, M.; Sarkar, R. Fuzzy rank-based fusion of CNN models using Gompertz function for screening COVID-19 CT-scans. *Sci. Rep.* **2021**, *11*. [CrossRef] [PubMed]
28. Pan, S.J.; Yang, Q. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345–1359. [CrossRef]
29. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2015**, arXiv:1409.1556.
30. 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]
31. Chollet, F. Xception: Deep Learning With Depthwise Separable Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1800–1807. [CrossRef]
32. Wolpert, D.H. Stacked generalization. *Neural Netw.* **1992**, *5*, 241–259. [CrossRef]
33. 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**, arXiv:2003.11055.
34. Sethy, P.K.; Behera, S.K. Detection of Coronavirus Disease (COVID-19) Based on Deep Features. *Preprints* **2020**. [CrossRef]
35. Hearst, M.; Dumais, S.; Osuna, E.; Platt, J.; Schölkopf, B. Support vector machines. *IEEE Intell. Syst. Their Appl.* **1998**, *13*, 18–28. [CrossRef]

36. Alqudah, A.M.; Qazan, S.; Alquran, H.H.; Abuqasmieh, I.; Alqudah, A. Covid-2019 Detection Using X-ray Images and Artificial Intelligence Hybrid Systems. *Preprints* **2020**. [CrossRef]
37. Seo, H.; Huang, C.; Bassenne, M.; Xiao, R.; Xing, L. Modified U-Net (mU-Net) With Incorporation of Object-Dependent High Level Features for Improved Liver and Liver-Tumor Segmentation in CT Images. *IEEE Trans. Med. Imaging* **2019**, *39*, 1316–1325. [CrossRef]
38. Sarker, L.; Islam, M.; Hannan, T.; Ahmed, Z. COVID-DenseNet: A Deep Learning Architecture to Detect COVID-19 from Chest Radiology Images. *Preprints* **2020**. [CrossRef]
39. Yang, X.; He, X.; Jinyu Zhao, Y.Z.; Zhang, S.; Xie, P. COVID-CT-dataset: A CT scan Dataset about COVID-19. *arXiv* **2020**, arXiv:2003.13865.
40. Garain, A.; Basu, A.; Giampaolo, F.; Velasquez, J.D.; Sarkar, R. Detection of COVID-19 from CT scan images: A spiking neural network-based approach. *Neural Comput. Appl.* **2021**. [CrossRef] [PubMed]
41. Chattopadhyay, S.; Dey, A.; Singh, P.; Geem, Z.; Sarkar, R. COVID-19 Detection by Optimizing Deep Residual Features with Improved Clustering-Based Golden Ratio Optimizer. *Diagnostics* **2021**, *11*, 315. [CrossRef] [PubMed]
42. Polsinelli, M.; Cinque, L.; Placidi, G. A light CNN for detecting COVID-19 from CT scans of the chest. *Pattern Recognit. Lett.* **2020**, *140*, 95–100. [CrossRef]
43. Born, J.; Brändle, G.; Cossio, M.; Disdier, M.; Goulet, J.; Roulin, J.; Wiedemann, N. POCOVID-Net: Automatic Detection of COVID-19 From a New Lung Ultrasound Imaging Dataset (POCUS). *arXiv* **2021**, arXiv:2004.12084.
44. Sen, S.; Saha, S.; Chatterjee, S.; Mirjalili, S.; Sarkar, R. A bi-stage feature selection approach for COVID-19 prediction using chest CT images. *Appl. Intell.* **2021**, 1–16. [CrossRef]
45. Karbhari, Y.; Basu, A.; Geem, Z.W.; Han, G.T.; Sarkar, R. Generation of Synthetic Chest X-ray Images and Detection of COVID-19: A Deep Learning Based Approach. *Diagnostics* **2021**, *11*, 895. [CrossRef]
46. Das, S.; Roy, S.D.; Malakar, S.; Velásquez, J.D.; Sarkar, R. Bi-Level Prediction Model for Screening COVID-19 Patients Using Chest X-ray Images. *Big Data Res.* **2021**, *25*, 100233. [CrossRef]
47. Jaiswal, A.; Gianchandani, N.; Singh, D.; Kumar, V.; Kaur, M. Classification of the COVID-19 infected patients using DenseNet201 based deep transfer learning. *J. Biomol. Struct. Dyn.* **2020**, 1–8. [CrossRef]
48. Zheng, C.; Deng, X.; Fu, Q.; feng Zhou, Q.; Feng, J.; Ma, H.; Liu, W.; Wang, X. Deep Learning-based Detection for COVID-19 from Chest CT using Weak Label. *medRxiv* **2020**. [CrossRef]
49. Li, L.; Qin, L.; Xu, Z.; Yin, Y.; Wang, X.; Kong, B.; Bai, J.; Lu, Y.; Fang, Z.; Song, Q.; et al. Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT. *Radiology* **2020**, *296*, E65–E71. [CrossRef] [PubMed]
50. Soares, E.; Angelov, P.; Biaso, S.; Froes, M.H.; Abe, D.K. SARS-CoV-2 CT-scan dataset: A large dataset of real patients CT scans for SARS-CoV-2 identification. *medRxiv* **2020**. [CrossRef]
51. Ai, T.; Yang, Z.; Hou, H.; Zhan, C.; Chen, C.; Lv, W.; Tao, Q.; Sun, Z.; Xia, L. Correlation of Chest CT and RT-PCR Testing for Coronavirus Disease 2019 (COVID-19) in China: A Report of 1014 Cases. *Radiology* **2020**, *296*, E32–E40. [CrossRef] [PubMed]
52. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
53. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the 32nd International Conference on International Conference on Machine Learning—Volume 37, ICML'15, Lille, France, 6–11 July 2015; pp. 448–456.
54. Sagi, O.; Rokach, L. Ensemble learning: A survey. *WIREs Data Min. Knowl. Discov.* **2018**, *8*, e1249. [CrossRef]
55. Schwenker, F.; Roli, F.; Kittler, J. (Eds.) *Multiple Classifier Systems*; Springer International Publishing: Cham, Switzerland, 2015. [CrossRef]
56. Bellmann, P.; Thiam, P.; Schwenker, F., Multi-classifier-Systems: Architectures, Algorithms and Applications. In *Computational Intelligence for Pattern Recognition*; Pedrycz, W., Chen, S.M., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 83–113. [CrossRef]
57. Kächele, M.; Thiam, P.; Palm, G.; Schwenker, F.; Schels, M. Ensemble Methods for Continuous Affect Recognition: Multi-Modality, Temporality, and Challenges. In Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, AVEC '15, Brisbane, Australia, 26 October 2015; pp. 9–16. [CrossRef]
58. Schwenker, F.; Dietrich, C.; Thiel, C. Learning of Decision Fusion Mappings for Pattern Recognition. *Int. J. Artif. Intell. Mach. Learn.* **2006**, *6*, 17–21.
59. Faußer, S.; Schwenker, F. Neural Network Ensembles in Reinforcement Learning. *Neural Process. Lett.* **2015**, *41*, 55–69. [CrossRef]
60. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 618–626. [CrossRef]
61. Kalpić, D.; Hlupić, N.; Lovrić, M., Student's t-Tests. In *International Encyclopedia of Statistical Science*; Lovric, M., Ed.; Springer: Berlin/Heidelberg, Germany, 2011; pp. 1559–1563. [CrossRef]