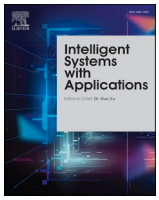
Intelligent Systems with Applications 17 (2023) 200160

Contents lists available at ScienceDirect

Intelligent Systems with Applications

journal homepage: www.journals.elsevier.com/intelligent-systems-with-applications



Multi-modal image classification of COVID-19 cases using computed tomography and X-rays scans

Nida Nasir [a,*], Afreen Kansal [b], Feras Barneih [a], Omar Al-Shaltone [a], Talal Bonny [a,c], Mohammad Al-Shabi [a,d], Ahmed Al Shammaa [e]

[a] Research Institute of Science and Engineering, University of Sharjah, Sharjah, UAE
[b] Department of Statistics, London School of Economics and Political Science, London, UK
[c] College of Computing and Informatics, University of Sharjah, Sharjah, UAE
[d] College of Engineering, University of Sharjah, Sharjah, UAE
[e] Khorfakkan University, Khorfakkan, UAE

ARTICLE INFO

Keywords:
Machine learning
Transfer learning
Adam optimiser
Binary cross entropy loss
Data augmentation

ABSTRACT

COVID pandemic across the world and the emergence of new variants have intensified the need to identify COVID-19 cases quickly and efficiently. In this paper, a novel dual-mode multi-modal approach is presented to detect a covid patient. This has been done using the combination of image of the chest X-ray/CT scan and the clinical notes provided with the scan. Data augmentation techniques are used to extrapolate the dataset. Five different types of image and text models have been employed, including transfer learning. The binary cross entropy loss function and the adam optimizer are used to compile all of these models. The multi-modal is also tried out with existing pre-trained models such as: VGG16, ResNet50, InceptionResNetV2 and MobileNetV2. The final multi-modal gives an accuracy of 97.8% on the testing data. The study provides a different approach to identifying COVID-19 cases using just the scan images and the corresponding notes.

1. Introduction

Late in December 2019, in Wuhan, China, the COVID-19 illness caused by the SARS-CoV-2 coronavirus made its initial appearance (Phan, 2020). All ages, including kids and teenagers, are susceptible to COVID-19 infection, which can lead to life-threatening consequences. As of April 11, 2022, the World Health Organization reported that there had been over 500 million confirmed cases of COVID-19, resulting in 6, 250,000 fatalities. The SARS-CoV-2 virus can transmit through direct touch or through droplets coughed or sneezed out. When COVID-19 affects the respiratory system, it can result in severe pneumonia, which can lead to death (De Miranda & Teixeira, 2020). To detect the SARS-Cov-2, the Reverse transcription polymerase chain reaction (RT-PCR) test is used. This test is relatively complicated and produces less consistent results (Kucirka, Lauer, Laeyendecker, Boon, & Lessler, 2020). Radiography examination by radiologists is an alternative method to visually detect COVID-19 viral infection. However, detecting the infection from X-ray images is challenging and requires a high level of expertise. Clinical diagnosis of X-ray and CT images by radiologists yields an accuracy of 75% (Satia et al., 2013; Wong et al., 2020). Therefore, a quick and more precise method is needed to aid physicians in identifying COVID-19 symptoms.

In the past few years, deep learning (DL) have been widely used in the medical field in detecting area such as hypertension detection (Nasir et al., 2021), diabetic retinopathy detection (Nasir et al., 2022b), epileptic seizure detection (Barneih et al., 2022), sleep apnea detection (Qatmh et al., 2022) and image object detection and image classification (Woźniak, Siłka, & Wieczorek, 2021). In the COVID-19 pandemic, artificial intelligence has been extensively used in areas such as diagnosis, social control, surveillance public health and controlling the COVID-19 patients. To alleviate the significant strain on limited medical resources caused by the COVID-19 pandemic, the most important measures to control the pandemic's spread are rapid diagnosis, accurate prediction, enhanced monitoring, and effective treatments. Many review articles on the subject have been published. However, the findings of these studies are inconclusive, and there is little research systematically assessing the application of AI for COVID-19 in accordance with PRISMA, with the majority of them focusing on aspects such as diagnosis or treatment. Researchers have made significant contributions to the anti-COVID-19 campaign, and the number of COVID-19-related AI models in the

* Corresponding author.
E-mail address: nnasir@sharjah.ac.ae (N. Nasir).

https://doi.org/10.1016/j.iswa.2022.200160
Received 11 August 2022; Received in revised form 21 November 2022; Accepted 27 November 2022
Available online 30 November 2022

literature is rapidly increasing. Artificial intelligence models that have been properly trained can ensure accurate and rapid diagnosis or assist doctors in streamlining the diagnosis and reducing manual labor. By using training data, AI models could detect patients at higher risk, characterize the epidemiology of COVID-19, and model disease transmission. Artificial intelligence-based methods, such as repurposing existing drugs, screening targets as vaccines based on the potential mutation model to SARS-CoV-2, and screening compounds as potential vaccine adjuvants, could aid in the discovery of novel drugs and vaccines. A unique retinal blood vessel categorization approachsis suggested in Dash et al. (2022), this article recommends a combination model of a directed filter and a matched filter for improving atypical retinal images with weak vascular contrasts.

This paper proposes a multi-modal approach to detect whether a patient is COVID-19 positive or not. Along with using the images of the CT scan/X-ray of the patient, the notes that have been jotted down by the doctor/nurse are also considered for the final prediction, which has resulted in better performance and efficient detection of COVID-19 cases. Specific keywords that can only be associated with COVID-19 are very helpful in detection. Along with this, even the problem of small size datasets is resolved by using various data augmentation techniques to increase the number of observations in the data to get results that reflect the real world scenario in terms of covid - non covid cases imbalance.

In this paper, we explore the solution to the problem in coherence to the following contributions:

1. Concatenation of a text and Image model to predict COVID.
2. Comparison of Augmentation results: a) with No-Augmentation, b) with Augmentation on whole data, and c) with Augmentation on training data only.
3. Comparing the performance of benchmark CNNs and Proposed Multi-Modal Approach in classifying the X-ray scans (along with three Image Models).

The novelty of the study of Dual-mode (Text and Image) multi-Model for covid detection. This study will help as a precautionary step towards various ailment detection. The binary cross entropy loss function and the Adam optimizer are used to compile all of these models. The model is trained using the default batch size of 32 and the early stopping criterion and model checkpoint callbacks.

2. Literature review

Using artificial intelligence (AI) and machine learning (ML) techniques, many researchers developed models to diagnose COVID-19 cases from chest X-ray and CT imaging. El Asnaoui & Chawki (2021) detected and classified COVID-19 cases using seven different deep learning models i.e. ResNet50, DenseNet201, MobileNetV2, InceptionResNetV2, InceptionV3, VGG16 and VGG19. The overall accuracy was 82.80%, with InceptionResNetV2 achieving the highest accuracy of 92.18%. Wang, Lin, & Wong (2020a), proposed COVID-Net, which is a CNN used for detecting COVID-19 from X-ray images. The network was trained using the COVIDx dataset, which consists of 13,975 chest X-ray images. The model achieved a testing accuracy of 91%. Authors in Horry et al. (2020), compared different CNN models and then chose and optimized a VGG19 model. Using OpenCV library, they pre-processed the images by applying histogram equalization followed by texture enhancement. Their model was able to detect COVID-19 using chest X-ray images, CT scans and ultrasound with an accuracy of 86%, 84% and 100% respectively.

Zhang et al. (2020) proposed COVID19XrayNet which is a deep learning based model that detects COVID-19 from X-ray images. The model is based on ResNet32 with two layers i.e., smoothing layer and feature extraction layer. The model achieved better results than the original ResNet32 with an accuracy of 91.92%. Authors in Ismael & Şengür (2021) proposed Support Vector Machine (SVM) for COVID19 classification and several pretrained CNN models i.e. VGG16, VGG19, ResNet18, ResNet50 and ResNet101 for feature extraction to achieve an accuracy of 94.7%. The dataset used consists of 380 normal and COVID-19 chest X-ray images. Hemdan et al. Hemdan, Shouman, & Karar (2020), proposed COVIDX-Net which a deep learning framework dedicated to detect COVID-19 using X-ray images. Authors composed a comparative study of other deep learning models including Inception-ResNetV2, InceptionV3, VGG19, ResNetV2, Xception and MobileNetV2. Their study showed that VGG19 and DenseNet19 achieved the highest accuracy of 90%. Authors in Maghdid et al. (2021) combined simple CNNs (single convolution layer followed by batch normalization, rectified linear unit (ReLU) with two fully-connected layer and AlexNet model. The proposed model achieved an accuracy of 94%. Authors in Hall, Paul, Goldgof, & Goldgof (2020) used transfer learning strategy with VGG16. Moreover they used data augmentation to increase the size of the dataset achieving an accuracy of 96.1%. A comparison of state-of-the-art studies has been done in discussion section which compares results of proposed study with existing studies.

3. Methodology

This section discusses the dataset, models and their architectures, and proposed methodology. The dataset has been gone through data augmentation and text analysis. Moreover, the model architectures discussed along with a basic CNN architecture are VGG16, Resnet50, MobileNetV2 and InceptionResnetV2.

3.1. Dataset description

The data used in this article is a public dataset made available by Cohen, Morrison, & Dao (2020a); Cohen et al. (2020c). The data contains images of the chest X-rays and CT scans done on patients who either tested positive for COVID-19 or were suspected of having COVID-19 or other viral/bacterial pneumonia. Along with the images, metadata is made available which contains information about the patient like their sex, age, clinical notes and other additional notes associated with the scan. The 2 columns of the data - clinical notes and other notes are combined into a single column by string concatenation. The data consists of 535 images. Out of these 535 images, clinical notes associated are only available for 485 of them and information about sex and age for only 483 and 440, respectively. The missing values of age are imputed with the mean while the missing values for the sex are filled with the mode. The missing clinical notes are filled with an empty string. The problem of classification in this paper is converted to a binary classification problem, with all other labels except for covid categorized as “Non-Covid”. This results in an imbalanced data consisting of 342 covid cases and remaining 193 as non-covid cases. The data is split into the training, validation and test datasets where the training data is 85%, validation data 10% and testing data 5% of the original data. This is done via random sampling.The noise and blurring of image has been fixed during the pre-processing stage, using denoising function.

3.1.1. Data augmentation

Given the imbalanced nature of the dataset, data augmentation techniques are implemented to make the data balanced and the results compared with the imbalanced data. These data augmentation methods are applied to both the text and the images. The augmentation is done for the texts and images associated to the non-covid patients and is done in two ways, once on the whole data and once just on the training data. The augmentation is done such that the data still remains unbalanced, but with higher number of non-covid cases. So for each non-covid case in the data, two more augmented observations are added.

For the text data, the augmentation is done in 2 ways - by replacing certain number of random words with their synonyms and randomly swapping words within the text. The number of words to be replaced and

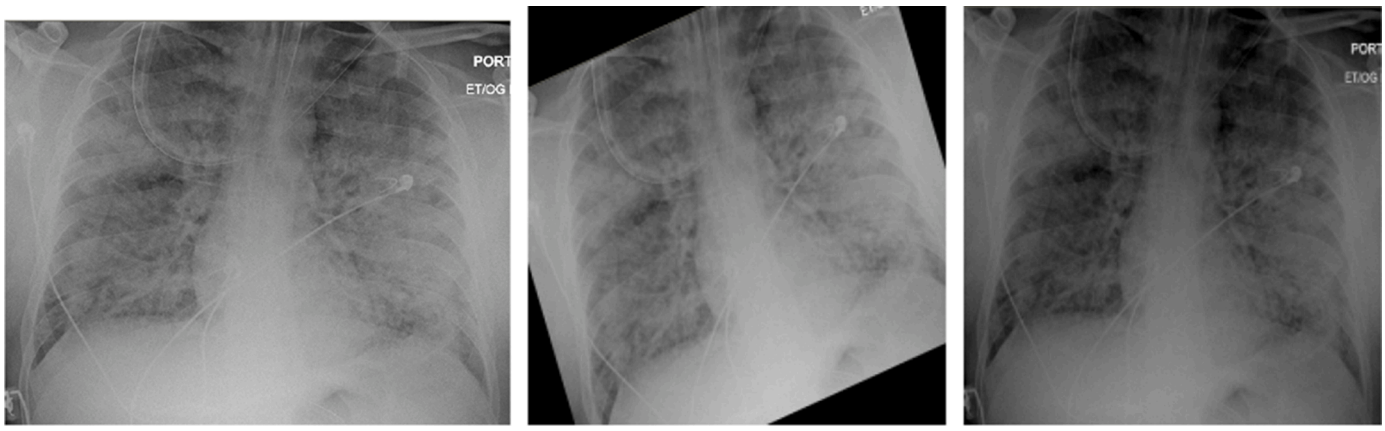


Fig. 1. Image augmentation : *left: Original, middle: Rotated, right: Decreased Brightness.*

swapped is chosen as 15 and the resulting text is the augmented text. For images as well, two types of augmentation is done - rotation and decreasing brightness. Since dealing with medical images, inversion of the images is not possible. The augmented images are rotated by an angle of 20 degrees. Data augmentation is done twice, once on the whole dataset and once only on the training dataset. An example of the augmented images is shown in Fig. 1.

3.1.2. Text analysis

The clinical notes are analyzed by plotting word clouds and top 20 uni-grams, bi-grams and tri-grams. Word clouds are a visual representation that is often used to visualize text data. It breaks down the texts into words and plots the words with varying sizes and colors that represent it's frequency in the data. A word which is much bigger in size in the word cloud is said to be most frequently occurring word in the data while smaller sized words are less frequent. Uni-grams refer to single words alone. Bi-grams refer to pairs of words together while tri-grams refer to groups of 3 words together. All possible such combinations are taken and the most frequently occurring groups of words are then plotted.

3.2. K-fold cross validation

Three different situations are tested to get the best results - when there is no data augmentation, data augmentation only on the training data and data augmentation on the whole data. Out of these three, the best model is chosen and then *K*-fold cross validation is performed to test for the validity of results since the data is small and data splitting is done randomly. For this, *K* is chosen as 10. For each of the 10 iterations, the model was run keeping one fold as testing and the remaining as training data. While running the model, validation data size is chosen as 30% of the training data.

3.3. Model architectures

The state-of-the-art pre-trained networks included in the Keras core library have consistently outperformed Convolutional Neural Networks on the ImageNet challenge. These networks also show a strong ability to generalize to images outside of the ImageNet dataset using transfer learning techniques such as feature extraction and fine-tuning. Four used CNN architectures are discussed below:

3.3.1. VGG16 model

The most distinctive aspect of VGG16 is that it focused on having convolution layers of 3×3 filter with stride one instead of a bunch of hyper-parameters and always utilized the same padding and maxpool layer of 2×2 filter with stride two. Convolution and max pool layers are arranged in this manner throughout the entire architecture. two fully connected layers and a softmax are included as its final features. The 16 in VGG16 stands for the number of weighted layers, which are 16. This network has around 138 million parameters, making it fairly huge (Simonyan & Zisserman, 2014).

3.3.2. ResNet50 model

The introduction of ResNet or residual networks, which are made up of Residual Blocks, has alleviated the problem of training very deep networks. The difference is that there is a direct connection that skips some layers in between (this may vary depending on the model). This connection is known as the 'skip connection,' and it is at the heart of residual blocks. Because of this skip connection, the layer's output is no longer the same. Without this skip connection, the input '$\times$' is multiplied by the layer weights before being multiplied by a bias term. This term is then passed through the activation function, $f()$, and the result is $H(x) = f(x)$. With the addition of the skip connection, the output is now $H(x) = f(x) + x$. This method appears to have a minor flaw when the dimensions of the input differ from those of the output, which can occur with convolutional and pooling layers. When the dimensions of $f(x)$ differ from those of x, one of two approaches can be taken: the skip connection is padded with extra zero entries to increase its dimensions. To match the dimension, the projection method is used, which is accomplished by adding 11 convolutional layers to the input. In this case, the result is $H(x) = f(x) + w1.x$. In this case, we add an extra parameter w1, whereas in the first approach, no extra parameter is added. The skip connections in ResNet solve the problem of vanishing gradient in deep neural networks by allowing the gradient to flow through an alternate shortcut path. Another way that these connections help is by allowing the model to learn the identity functions, which ensures that the higher layer performs at least as well as, if not better than, the lower layer (He, Zhang, Ren, & Sun, 2016).

3.3.3. MobileNetV2 model

In MobileNetV2, there are two different kinds of blocks. A residual block with a stride of one and another one with a stride of two for downsizing. Both sorts of blocks have an 11 convolution with ReLU6 layer as their first layer. A depth wise convolution makes up the second layer, and a further 11 convolutions with no non-linearity make up the third layer. Deep networks are said to only have the power of a linear classifier on the non-zero volume portion of the output domain if ReLU is applied once more (Sandler, Howard, Zhu, Zhmoginov, & Chen, 2018).

3.3.4. InceptionResNetV2 model

The Inception-ResNet-v2 convolutional neural network was trained on over a million images from the ImageNet database. Images may be categorized into 1000 different object categories using the 164-layer

Table 1
Document feature vectors.

Documents	I	like	an	apple	and	bananas	he	ate
Sentence 1	1	1	1	1	1	1	0	0
Sentence 2	0	0	1	1	0	0	1	1

network, including the keyboard, mouse, pencil, and numerous animals. The network has therefore learned in-depth feature representations for a wide range of images. The network outputs a list of estimated class probabilities after receiving a 299 by 299 picture as input. It is made by merging the Residual connection and the Inception structure. In the Inception-Resnet block, multiple convolutional filters of various sizes are merged with residual connections. In addition to avoiding the deterioration problem brought on by deep structures, using residual connections speeds up training. Fig. 5 depicts the fundamental network architecture of Inception-Resnet-v2 (Mahdianpari, Salehi, Rezaee, Mohammadimanesh, & Zhang, 2018).

3.4. Proposed methodology

Two different kinds of model architectures are tried out for the classification problem. In the first, only the images are considered for the basis of classification. Three kinds of basic CNN architectures are implemented for this purpose. In the first image model (Model 1), only one 2D convolutional neural network layer with 16 filters, kernel size (3,3), stride of length 1 and padding of type *same* is used due to the small size of the data and to avoid over-fitting. This layer is then followed by a max pooling layer with pool size (2,2) and stride length two. The output is then flattened and passed onto a dense layer with 64 units and activation function *ReLU* and kernel initializer *he uniform*. This is completed by the final output Dense layer with one unit and activation function *sigmoid*. The second image model (Model 2) follows the same architecture but the difference being that the convolutional layer has 32 filters and the Dense layer has 128 units. The third image model (Model 3)is made more deep by including 3 groups of a convolutional 2D layer, batch normalization, max pooling and a dropout layer, each with increasing number of units, 16, 32 and 64 and dropout percentage 0.2,0.25 and 0.3. This is then passed to a Dense layer with 100 units and finally the output layer. The structure of all the three models are shown in Figs. 6 and 7.

All these models are compiled using the binary cross entropy loss function and *Adam* optimizer. The early stopping criterion and model checkpoint callbacks are used and the model is trained using the default batch size of 32. The best model is saved, wherein best is defined as the model having the highest validation accuracy, and that is used to make predictions for the test data.

The image models alone don't perform very well, and to improve this, a second model architecture is considered, where even the clinical notes associated with the patients' scan is passed as input. A multi-modal approach is utilized to incorporate both the images and text in the input. The images are passed to a separate image model and the text is passed into a text model. The outputs of both these models are concatenated and then passed into a final model which gives the resulting prediction. The model architecture is shown in Fig. 8.

The text is converted to a numeric vector before being passed to the model. Since we are dealing with medical data and few keywords in the notes make a lot of difference in diagnosis, each document is converted to a vector by counting the frequency of each word in the text. For example, if we consider 2 sentences - **I like an apple and bananas** and **He ate an apple**, then both the sentences can be converted to a numeric vector in the following way - All the words from all documents are considered and the vector is formed such that each value in the vector represents the frequency of that corresponding word in the sentence. So the vector form for these 2 sentences are shown in Table 1. This is done for all the texts in the data. As a pre-processing step, the stopwords from the texts are removed. Stopwords refer to all those words that are very frequently used in a sentence but offer no contextual information. These include words like *and, the, I, am, etc*. Along with these, terns frequently used in medical text like *patient, doctor, dr, etc.* are also removed. The words are converted to lower case as well for easier implementation. Once each text is converted to the document feature vectors, these are passed as input to the text model. Although there are more sophisticated text models build like recurrent neural networks and BERT models, for the proposed study, the Bag Of Words approach is used. This is done due to the importance given to specific keywords in the clinical notes written by the nurse/doctor that can help identify symptoms of COVID-19. Since the main focus is on those keywords and it's frequency in the text, BoW model has been used instead of RNNs and other NLP models. To test this hypothesis, an LSTM model is also used to see whether it performs better than the BoW approach or not.

The text model is created as a simple 2 layer deep neural network. Both of the dense layers have 64 units. The output of the final layer is passed on as input to the concatenation layer and the final model. The image model used is the third image model used described formerly. The final concatenated model is just one dense layer with 16 units and activation function *ReLU*. This is followed by a dropout layer and then the final output layer. The model again is compiled using the binary cross entropy loss function and *Adam* optimizer. The same set of callbacks are used to get the best model.

For the LSTM model, the texts are tokenized and padded to create all vectors of the same length. The maximum length of a vector is 451. The LSTM text model is created using an embeddings later and an LSTM layer. The number of output units in the embeddings layer is 10 while the number of units in the LSTM layer is 16.

Four pre-trained models are also tried out as a replacement for the custom build image model. The models tried out were - ResNet 50, InceptionResNetV2, MobileNetV2 and VGG16. The text model and final model, along with the compiling and training conditions, are kept the same. This is done only on the augmented training data.

For all the models trained, the best model is obtained and tested on the testing data. The learning diagnostic curves are plotted for all the models' history - training and validation loss plotted with number of epochs and the training and validation accuracy with number of epochs. Using the best model, predictions on the test data are obtained and the confusion matrix and ROC curve is plotted. The full methodology for the multi-modal is described in Fig. 9.

4. Results

In the notes associated with covid cases, as shown in the word clouds in Fig. 10, the most frequently used words are **chest, bilateral, fever, cough, day, history**. The very frequently used pair of words are **chest radiography, dry cough, shortness breath, oxygen saturation, pleural effusion**. The frequently used groups of 3 words are **normal range elevated, polymerase chain reaction, fever dry cough**. The same for non-covid cases are **night, chest, left, lung, pneumonia, normal, upper lobe, lower lobe, left lung, middle lobe, weight loss** and **right upper lobe, left lower lobe, anteroposterior radiograph obtained, human immunodeficiency virus**. The covid cases presented were noted for the commonly occurring symptoms - fever, cough among others. The top uni-grams, bi-grams and tri-grams are shown in Figs. 11–13.

The image models on their own don't perform very well when only the images are considered for classification. Considering the first image model, it performs poorly when there is no data augmentation or when data augmentation is done only on the training data. It fails to identify many covid cases and misclassifies them as non-covid, resulting in high number of false negatives. The learning curves and results are shown in Fig. 14. When the loss vs. epochs curve is inspected, it can be seen that the validation loss is little higher than the training loss with there is no data augmentation and when the data augmentation is done only on the

Table 2
Performance metrics.

Models		Accuracy	Sensitivity	Specificity	Precision	F1 Score
Model 1	**No Data Aug**	70.37%	72.22%	66.67%	81.25%	76.47%
	Data Aug (Whole)	91.30%	90.91%	91.67%	90.91%	90.91
	Data Aug (Training)	70.37%	66.67%	77.78%	85.71%	75.00%
Model 2	**No Data Aug**	76.92%	83.33%	62.50%	83.33%	83.33%
	Data Aug (Whole)	86.96%	90.91%	83.33%	83.33%	86.96%
	Data Aug (Training)	66.67%	72.22%	55.56%	76.47%	74.29%
Model 3	**No Data Aug**	62.96%	61.11%	66.67%	78.57%	67.75%
	Data Aug (Whole)	91.30%	81.82%	100%	100%	90.00%
	Data Aug (Training)	55.56%	44.44%	77.78%	80.00%	57.14%
Multi-Model	**No Data Aug**	96.30%	100%	88.89%	94.74%	97.30%
	Data Aug (Whole)	97.83%	95.45%	100%	100%	97.67%
	Data Aug (Training)	96.30%	100%	91.74%	88.89%	94.12%
Transfer Learning	**MobileNetV2**	70.37%	66.67%	77.78%	85.71%	75.00%
	ResNet50	96.30%	100%	88.89%	94.74%	97.30%
	InceptionResNetV2	96.30%	100%	88.89%	94.74%	97.30%
	VGG16	92.59%	94.44%	88.89%	94.44%	94.44%

Table 3
Comparison with other studies.

Ref.	Dataset used	Methods/Models	Results	Description
Sahinbas and Catak (2021)	COVID-19 X-ray images + collected 50 positive and 50 negative	CNN	Accuracy in VGG16 got the highest percentage which equals to 80%	Images scaled to 256*22 and later augmented by flipping and different angles. Study presented five pretrained deep CNN models, including VGG16, VGG19, ResNet, DenseNet, and InceptionV3, for transfer learning implementing X-ray images.
Ohata et al. (2020)	"1394 Chest X-ray Images (Pneumonia) with data augmentation (Kermany et al., 2018)	CNN, MLP, and SVM	SVM got the highest accuracy of 98.5%	Used CNNs to extract features, then using the transfer learning approach and categorizing these features with consolidated machine learning methods.
Apostolopoulos and Mpesiana (2020)	1427 X-ray images from Cohen et al. (2020d) without data augmentation	CNN	Highest accuracy of 96.78%	Assessed the effectiveness of CNN designs created in recent years for medical image classification.
Shaik and Cherukuri (2022)	2483 images for SARS-CoV-2 where 1252 of them is diagnosed with the virus (Soares et al., 2020) Dataset for COVID-CT (Zhao et al., 2020) which contains 349 COVID-19 CT and 463 non-COVID-19 images from 216 patients	CNN	Highest accuracy for SARS-CoV-2 = 98.99, Highest accuracy for COVID-CT = 93.33	The study is to present an effective ensemble strategy for identifying SARS-CoV-2 infection in chest CT scan images.
Wang et al. (2020b)	ImageNet dataset, the number of the dataset is 18,567 with using data augmentation	ResNet101 and ResNet152	Accuracy = 96.1%	Their approach attempts to transfer learning, integrate models, and categorize chest X-ray pictures into three categories: normal, COVID-19, and viral pneumonia.
Phankokkruad (2020)	COVID-19 research challenge dataset that contains 323 images without data augmentation (Cohen et al., 2020b)	CNN	Highest accuracy of 97.19%	CNN model with Xception outperforms the VGG16 and Inception-ResNet-V2 models in terms of accuracy.
[This Work]	Image + text Dataset	VGG 16, Resnet 50, MobileNet V2 and Inception-Resnet V2	multi-modal results in a 97.8% accuracy	Multi-Modal approach with data augmentation methods are applied to both the text and the images.

training data. Since data augmentation is done only on the training data, the balance between the classes is different in the validation/testing data, hence resulting in a higher loss and lower accuracy. But addition of more data has resulted in the validation loss and accuracy to be more stable across epochs as compared to when there was no data augmentation done. Even the area under the ROC curve is highest for the case when the class imbalance is consistent across the training, validation and testing datasets.

The second image model doesn't perform very well either giving low accuracy, especially when data augmentation is done only on the training data. This model suffers from the problem of high false positives. Many patients are termed as covid positive despite being negative. The results for this model is shown in Fig. 15. The performance of the second model is much more unstable compared to the first model. The validation loss and accuracy is highly erratic as the model is trained for more epochs. The area under the ROC curve is much lesser too than that of the first model. Consistent with the first model, performance is more unstable in the case of no augmented data compared to when there was additional augmented data added.

The third image model has the lowest performance among all of them, which can be attributed to over-fitting due to a small size dataset. There is very high misclassification of covid patient as not having COVID-19, resulting in a high number of false negatives. The learning curves and results are shown in Fig. 16. The model shows similar unstable behavior as the previous model owing to over-fitting. The model also makes more errors in classification as compared to the previous 2 models. The same can be said for the behavior of the ROC curve.

The third image model is used as the image model for the multi-modal classification. Since it's a complex architecture, the text model is made to be very simple. The multi-modal approach performs very well in classifying the patient as covid positive or not. In all 3 cases, just one case is misclassified. The graphs associated with this model are shown in Fig. 17. The validation and training accuracy reach almost 100% with much more consistent values of loss across epochs, although, there seems to be slight overfitting when the model is trained for more number of epochs. The ROC curve when data augmentation is done on the whole

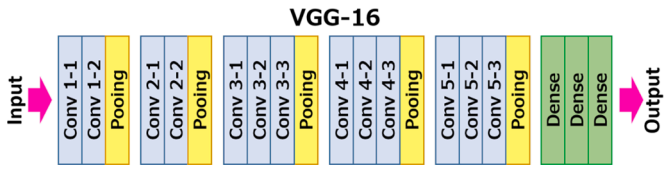


Fig. 2. VGG16 model architecture (Simonyan and Zisserman, 2014).

data is almost perfect, giving an area under the curve value of 0.98. The addition of the text has shot up the performance of the models, than what was obtained by just using the images.

Finally, the results of using the pre-trained models are summarized in Table 2 and the learning curves are shown in Fig. 18. The ResNet50 and InceptionResNetV2 model perform better than the other 2. When MobileNetV2 is used, the performance on the validation data is extremely poor. This can be attributed to the complex structure of the model. When using ResNet50, the validation and training loss are perfectly stable, with few ups and downs in the validation accuracy. When using InceptionResNetV2, the model starts to overfit which can be seen from the sharp upward rising spike in the validation loss and slightly falling validation loss. But this and the ResNet50 model result in only 1 false positive, with all other cases classified correctly. The VGG16 model shows signs of overfitting from the beginning, with an upward rising validation loss curve and a downward validation accuracy curve.

The results of the *K*-Fold cross-validation on the data is summarised in Fig. 19. The average of all the accuracies for each fold is 85.1% and the standard deviation is 14.17%, which accounts to nearly 10 observations being misclassified out of 92, which is the size of the testing data. The results of the LSTM text model is shown in Fig. 20. The testing accuracy obtained is 88.89% with 3 data points misclassified.

5. Discussion

The models are trained on both kinds of data, completely augmented data and augmented training data. This helps us to give us models in both scenarios, when there are a lot of covid cases and during the time when there are less cases. When the augmentation is done on the whole data, the validation and testing data are imbalanced but the majority of cases are non-covid. When the augmentation is done only on the training data, the validation and testing data are also imbalanced but now, the majority are covid cases. In both these cases, the models perform well and only one case is misclassified.

5.1. Comparison with other studies

This section offers an important evaluation of deep learning algorithm for detecting COVID-19 positive cases for some related papers as shown in Table 3, moreover, a compared study from other similar deep learning approaches with our proposed model that was done. And a

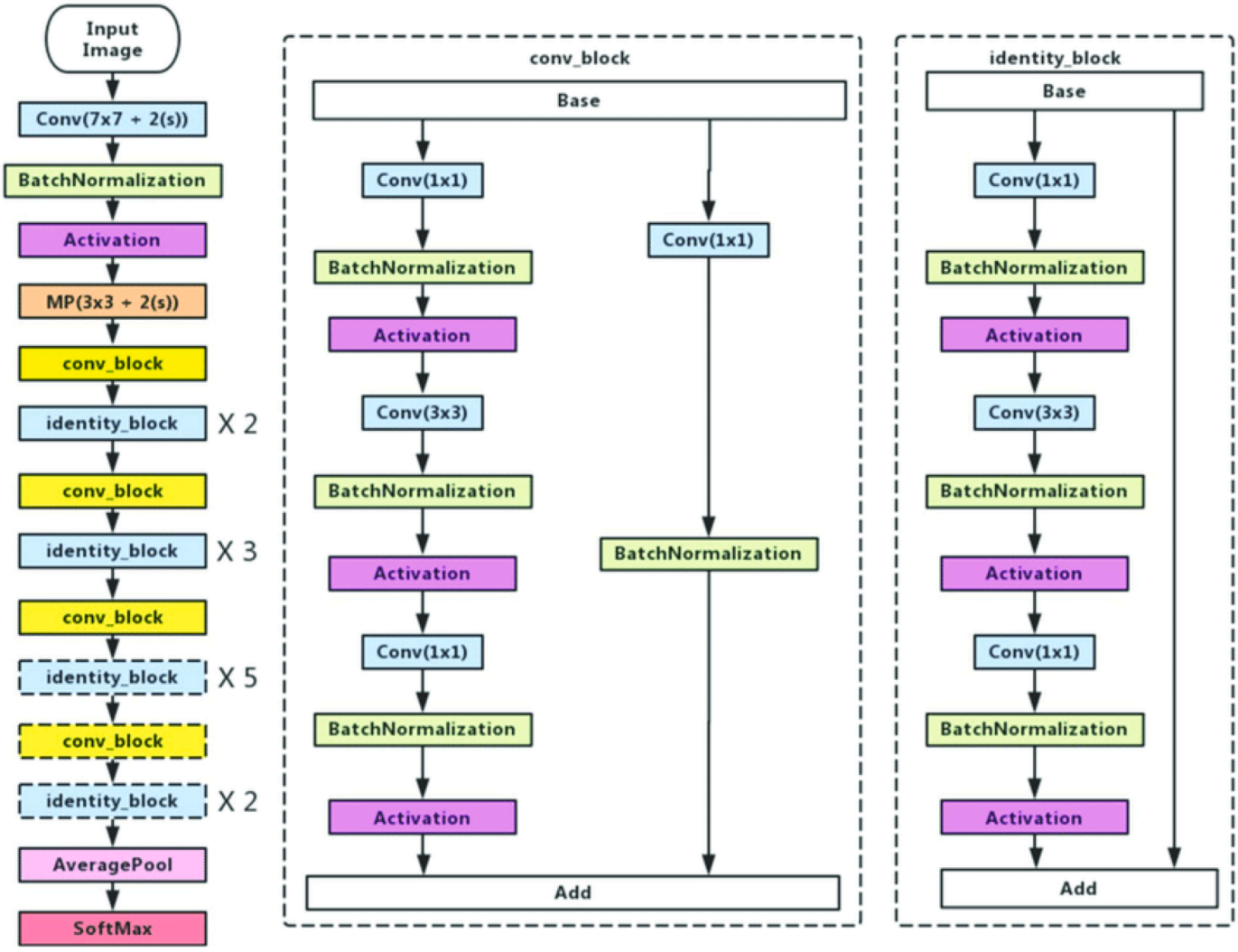


Fig. 3. ResNet50 model architecture (Ji et al., 2019).

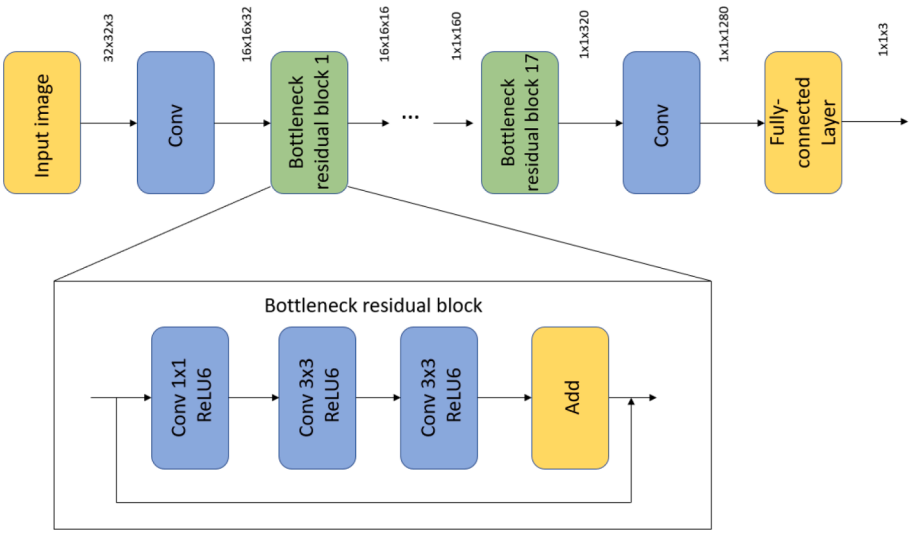


Fig. 4. MobileNetV2 model architecture (Seidaliyeva et al., 2020).

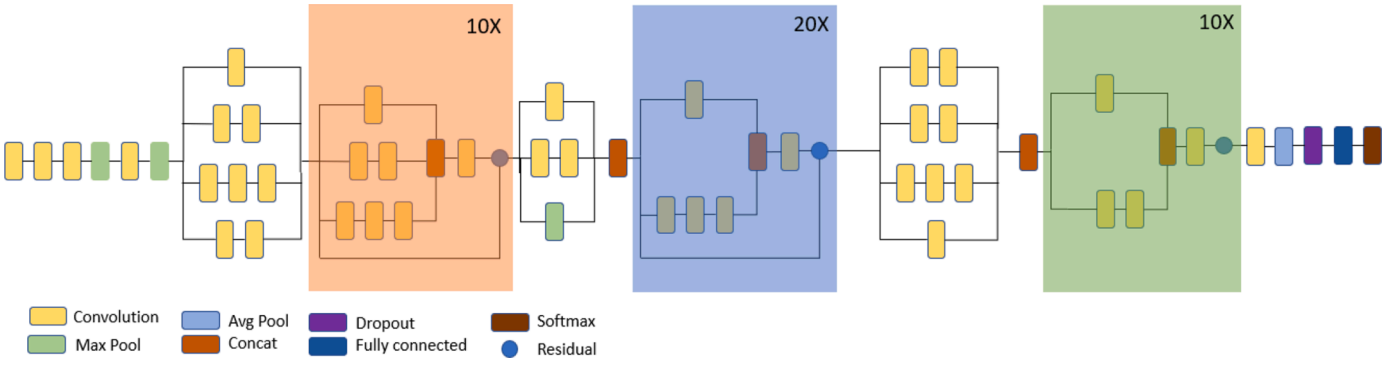


Fig. 5. Compressed InceptionResNetV2 model architecture (Mahdianpari et al., 2018).

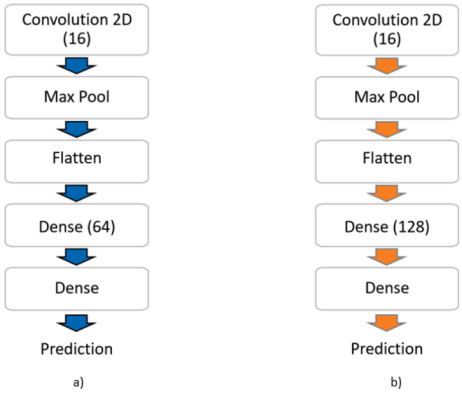


Fig. 6. Different layers of: a) image model 1, and b) image model 2.

discussion table (Table 3) has been created to evaluate our model in terms of others. According to Table 3, the majority of the datasets contained a small quantity of data (limited pictures for training and testing) to create and improve their model. Another notable fact is that the authors' most prevalent techniques for model creation were based on VGG and ResNet. In this paper, authors employed VGG16, Resnet50, MobileNetV2, and InceptionResnetV2 to create the model faster and more reliably so that it may be used as a real-time evaluation tool. All the models of transfer learning are standard and therefore when compared to other studies, their conditions and parameters are same.

5.2. *Tradeoffs of performance metrics*

To improve precision, the model's parameters and hyperparameters can be changed. While adjusting, you may notice that higher precision generally leads to lower recall, and higher recall leads to lower accuracy. Similarly, the recall value of any machine learning model can be altered by adjusting multiple parameters or hyperparameters. A higher or lower

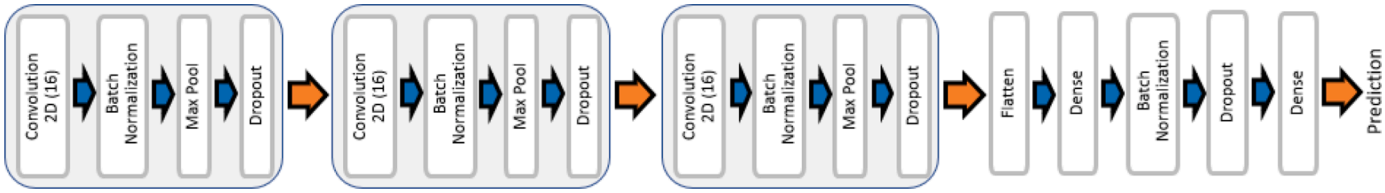


Fig. 7. Different layers of image model 3.

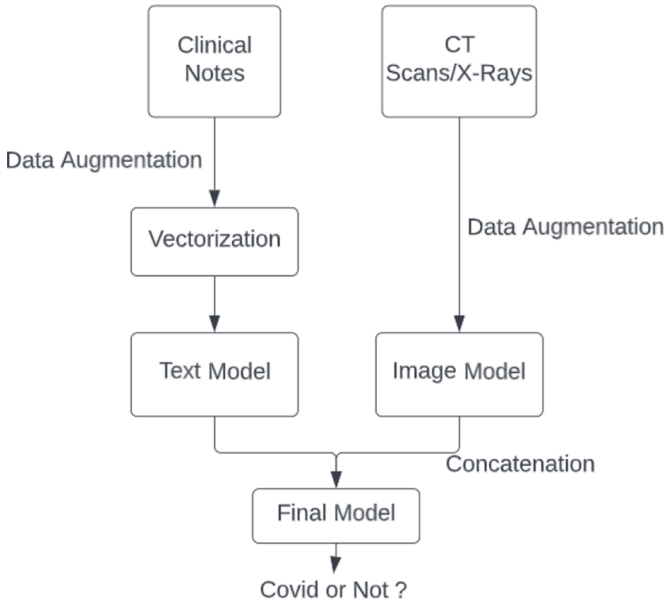


Fig. 8. Multi-modal architecture.

recall for any model has a specific meaning: With a high recall, the majority of positive instances (TP + FN) will be identified (TP). As a result, the number of FP measurements increases while overall accuracy decreases. Assume, however, that the outcome is low recall.In that case, it indicates that there were many FNs (should have been positive but labeled negative), which means that if the results find a positive example, there is a better chance that it is a true positive. Furthermore, while F1 is less intuitive than accuracy, it is usually more advantageous, particularly when the class distribution is unequal. Accuracy improves when the cost of false positives and false negatives is the same. If the cost of false positives and false negatives is significantly different, both Precision and Recall should be considered.

Furthermore, recall and sensitivity are inversely proportional. Susceptible tests yield more positive results in patients who are sick, whereas precise tests reveal no illness in patients who do not have a finding. Sensitivity and specificity should always be considered concurrently to provide a complete diagnosis. Furthermore, accuracy is a good quality measure when datasets are symmetric and the values of false-positive and false-negatives are nearly similar. As a result, other parameters play an important role in determining the performance of a model.

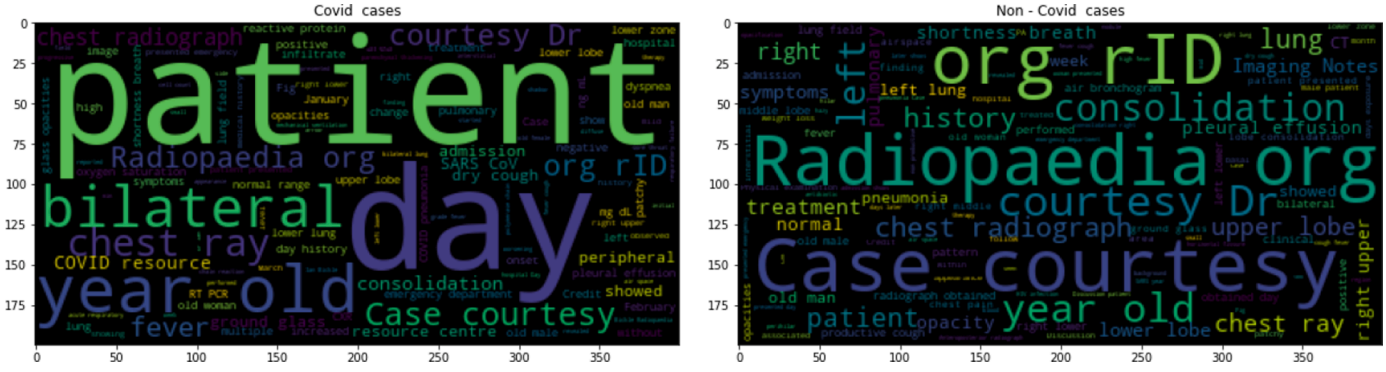


Fig. 9. Metholodogy.

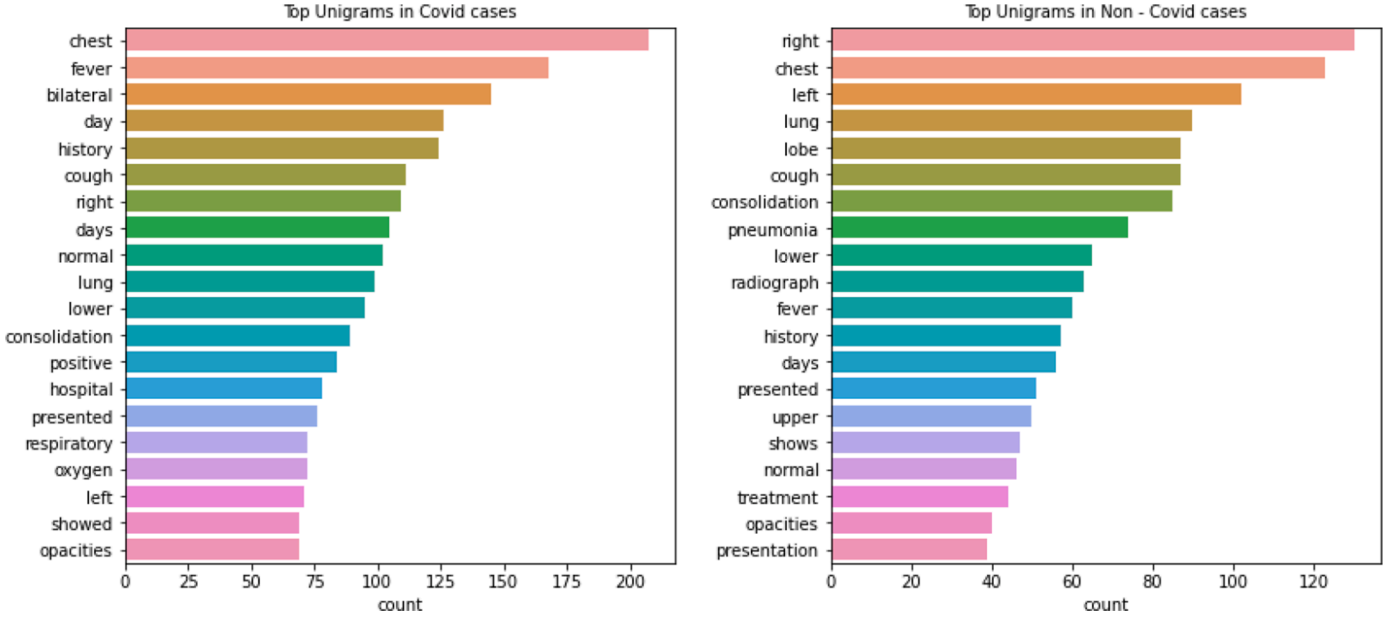


Fig. 10. Word clouds of clinical notes.

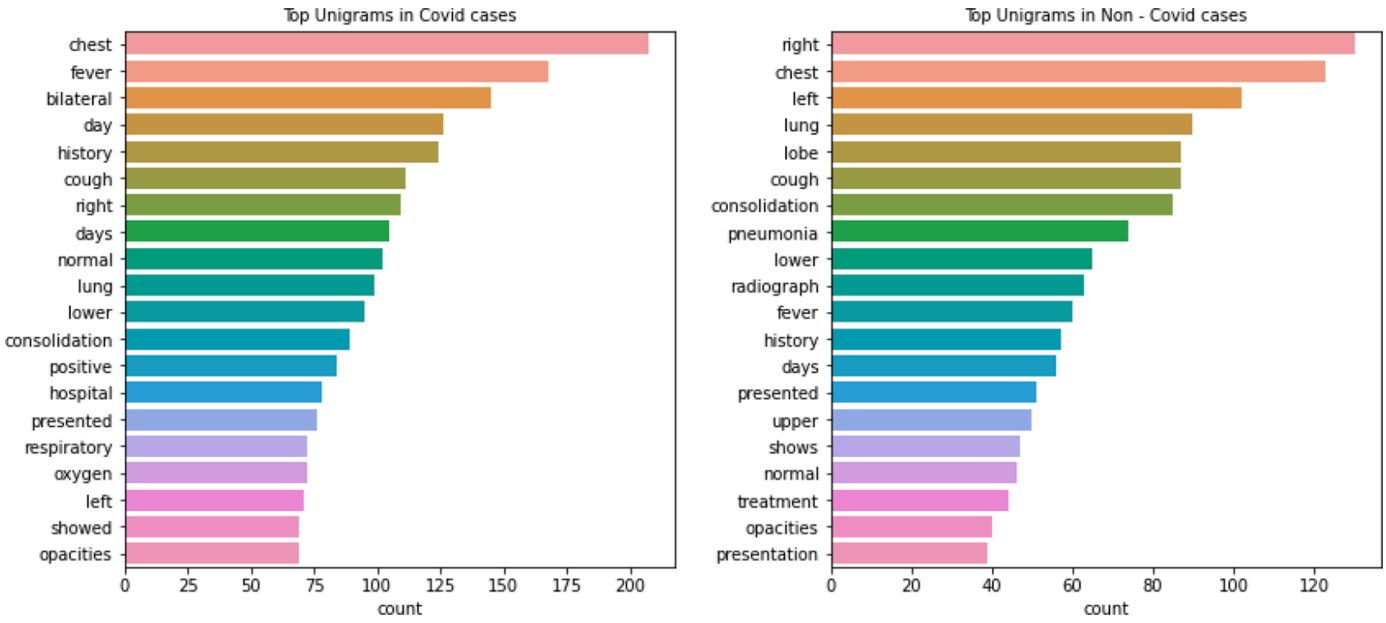


Fig. 11. Top uni-grams of clinical notes.

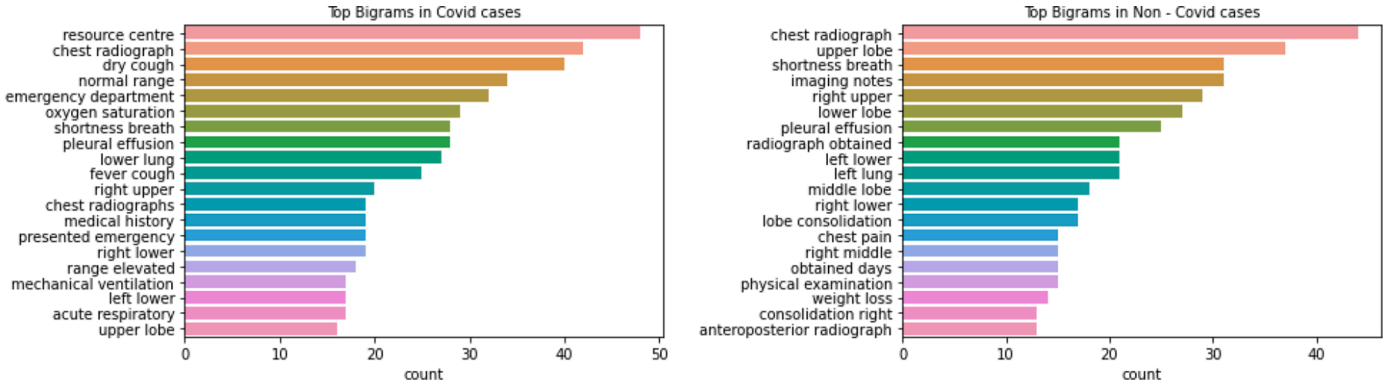


Fig. 12. Top bi-grams of clinical notes.

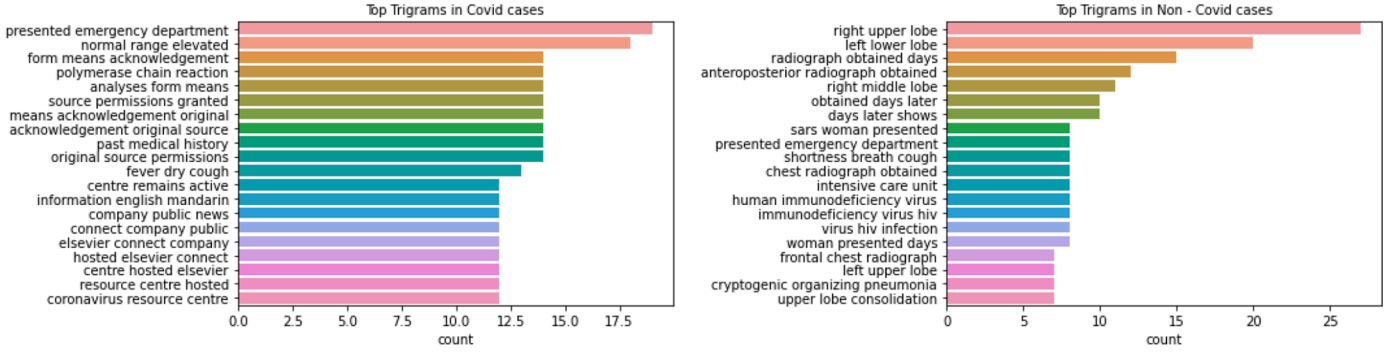


Fig. 13. Top tri-grams of clinical notes.

5.3. Behaviour of models used

Each VGG block is made up of 2D Convolution and Max Pooling layers, as shown in Fig. 2. As the number of layers in CNN increases, so does the model's ability to fit more complex functions. As a result, more layers promise improved performance. This is not to be confused with an Artificial Neural Network (ANN), where increasing the number of layers does not always result in improved performance. The backpropagation algorithm is used to update the weights of a neural network, which makes minor changes to each weight in order to reduce the model's loss. It updates each weight so that it moves in the direction of the decreasing loss.This is simply the gradient of this weight as determined by the chain rule. However, as the gradient flows backward to the initial layers, the value grows with each local gradient. As a result, the gradient becomes smaller and smaller, resulting in very small changes to the initial layers. As a result, the training time is significantly increased. If the local

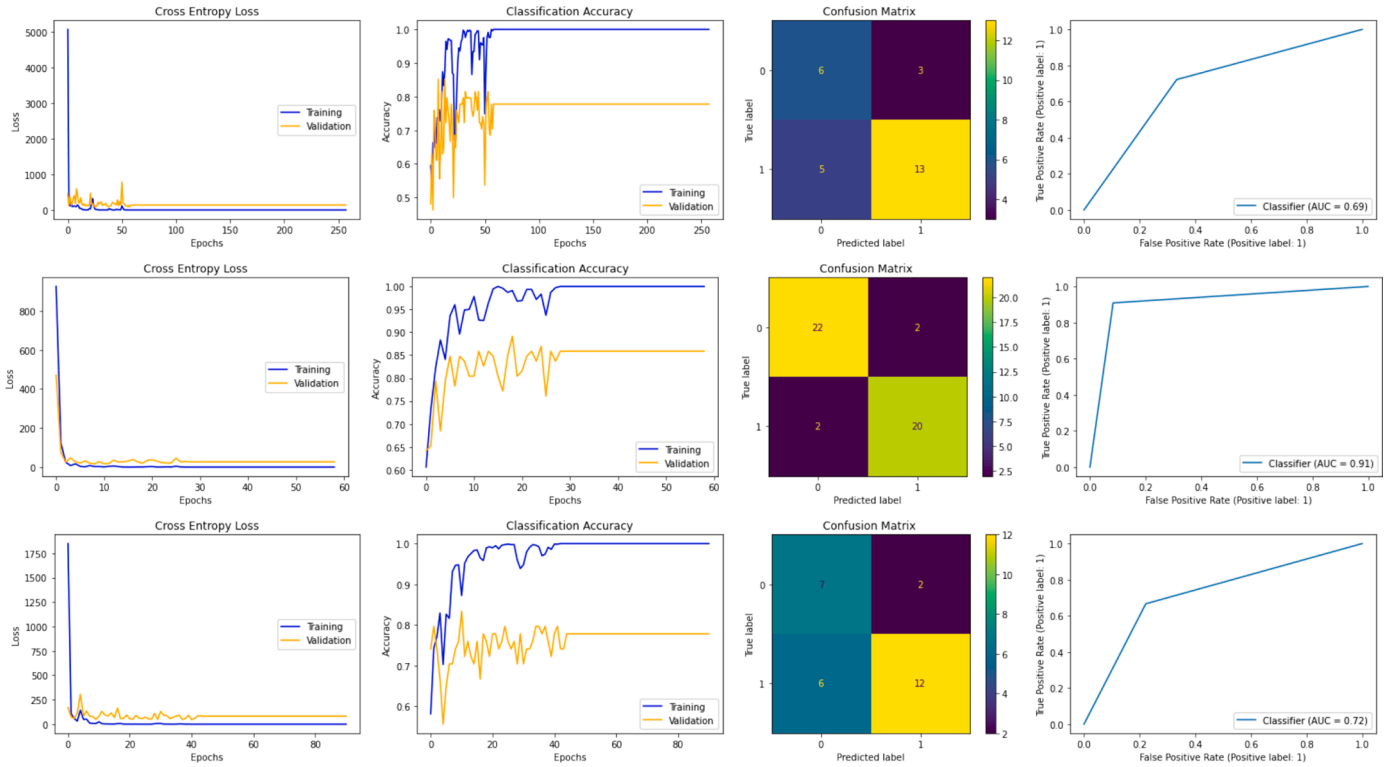


Fig. 14. Image model 1 results [Top-Bottom: No Data Aug, Data Aug (Whole), Data Aug (Training)].

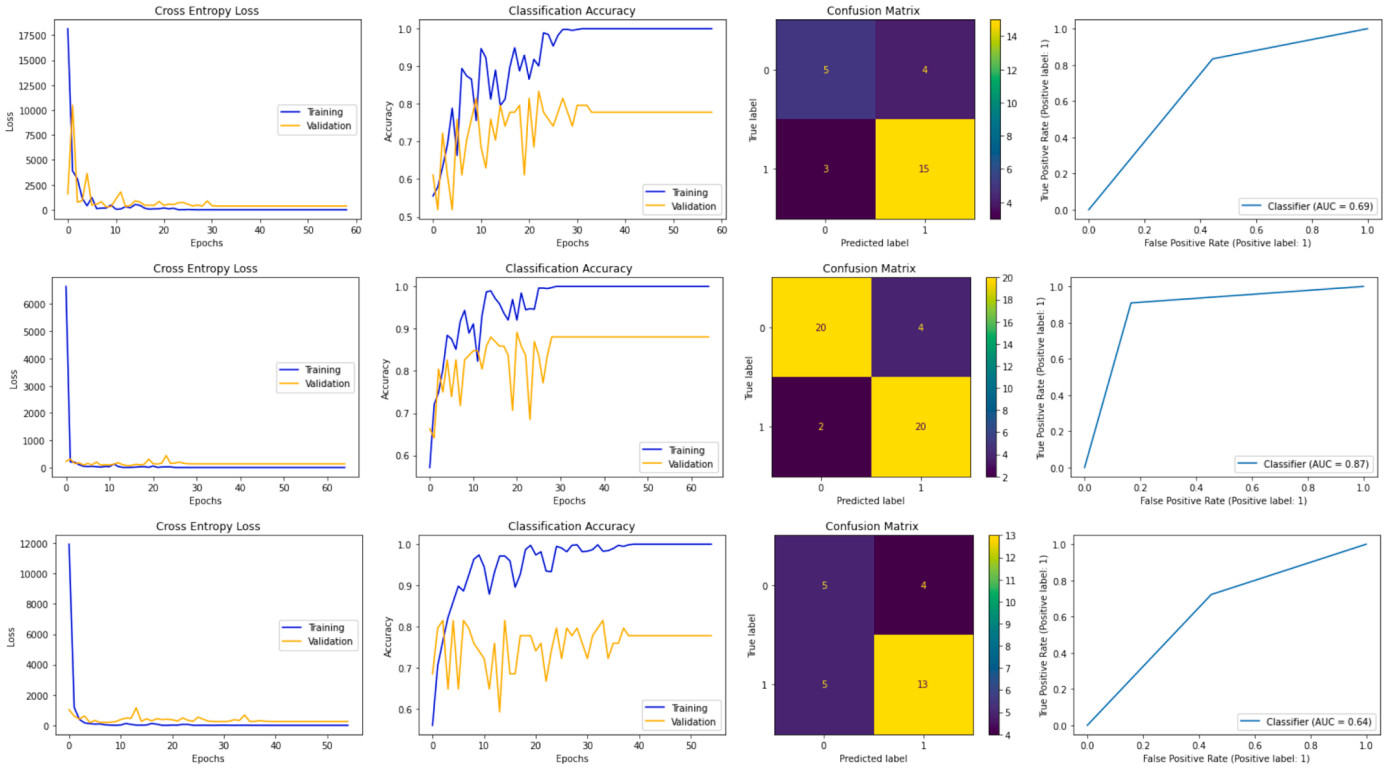


Fig. 15. Image model 2 results [Top-Bottom: No Data Aug, Data Aug (Whole), Data Aug (Training)].

gradient equals one, the problem is solved.

This is where ResNet comes in, as it accomplishes this via the identity function. As a result, as the gradient is back-propagated, its value does not decrease because the local gradient is 1.Deep residual networks (ResNets), such as the popular ResNet-50 model, are another type of 50-layer deep convolutional neural network architecture (CNN), as seen in Fig. 3. A residual neural network converts a plain network into its residual network counterpart by inserting shortcut connections. ResNets are less complex than VGGNets because they have fewer filters. The vanishing gradient problem is not permitted in ResNet. The skip

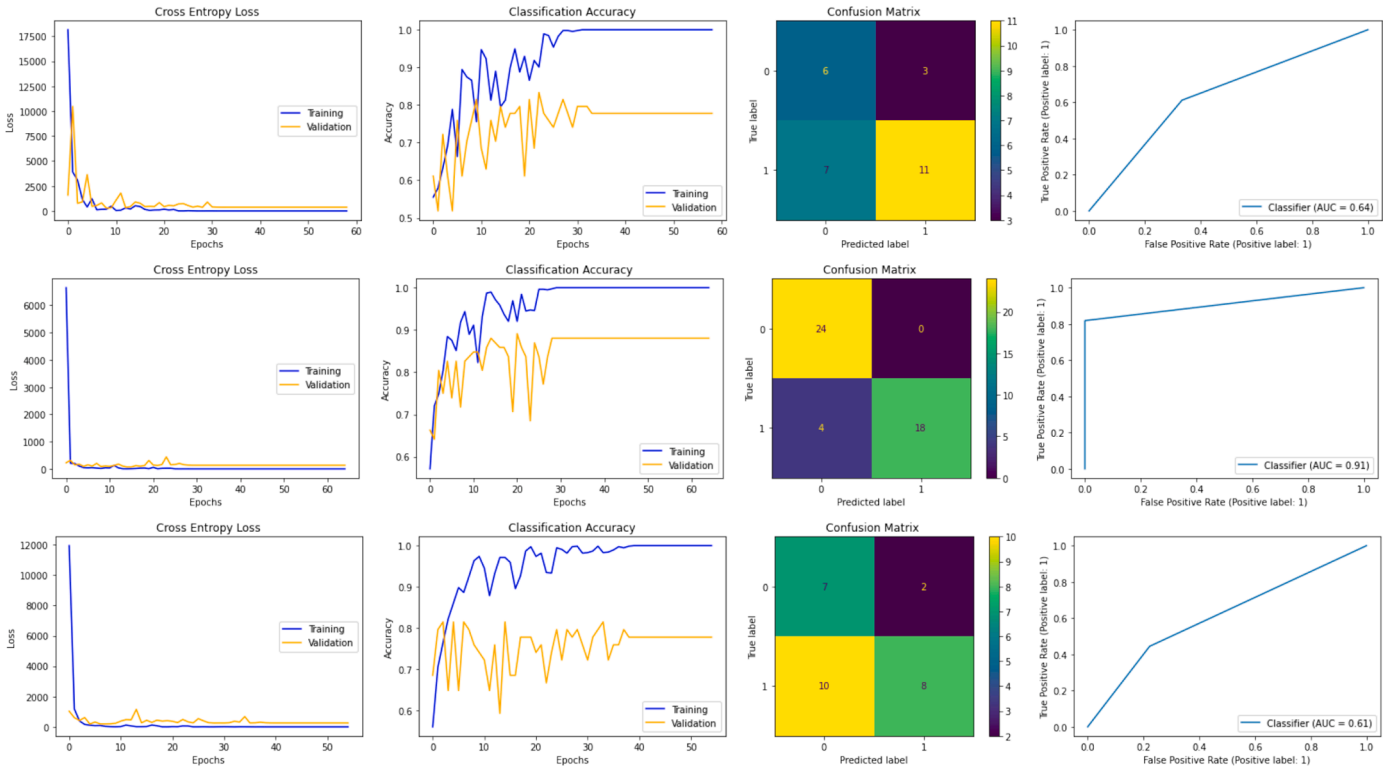


Fig. 16. Image model 3 results [Top-Bottom: No Data Aug, Data Aug (Whole), Data Aug (Training)].

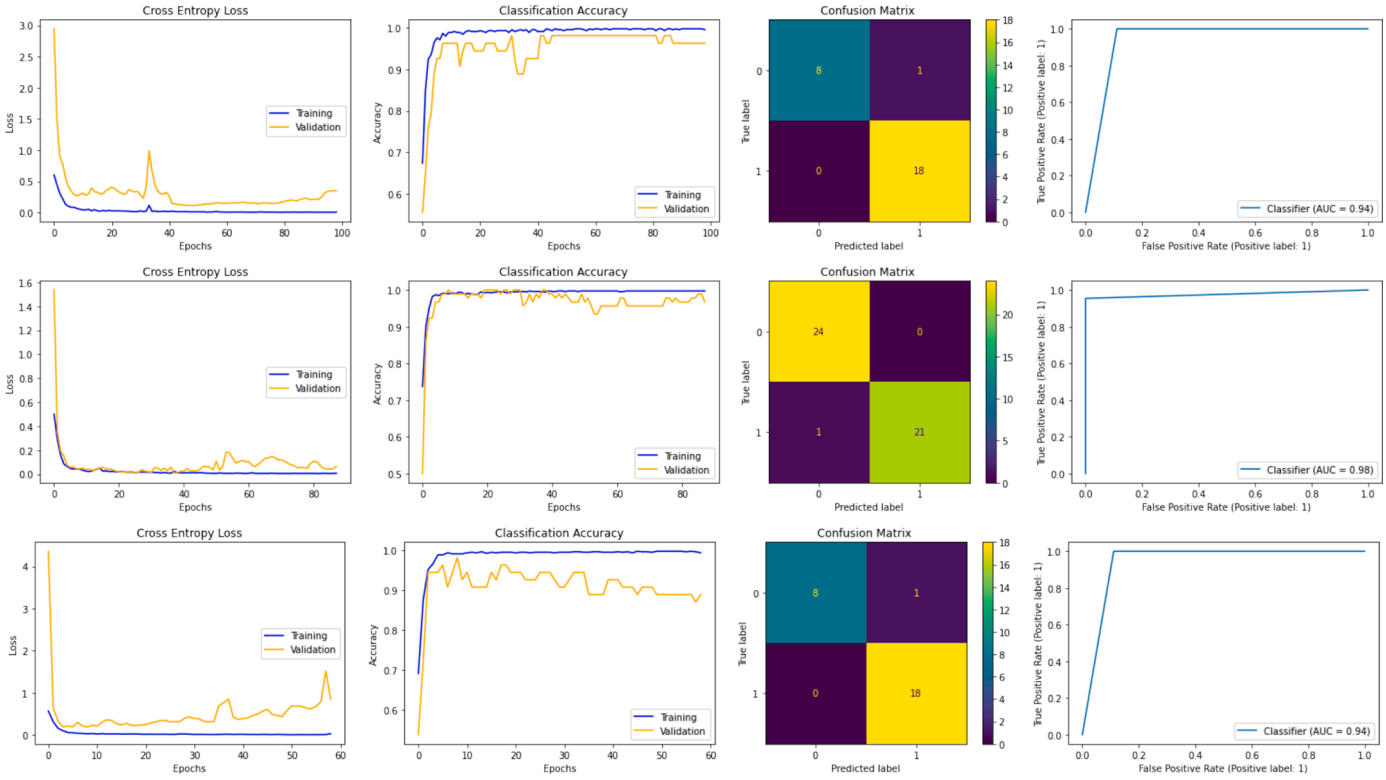


Fig. 17. Multi-modal results [Top-Bottom: No Data Aug, Data Aug (Whole), Data Aug (Training)].

connections function as gradient superhighways, allowing the gradient to flow freely. This is also one of the main reasons why ResNet comes in different versions such as ResNet50, ResNet101, and ResNet152.

Inception was designed to reduce the computational burden of deep neural nets while achieving cutting-edge performance. Because the computational efficiency decreases as the network grows deeper, the authors of Inception were interested in finding a way to scale up neural nets without increasing computational cost. Fig. 5 shows the

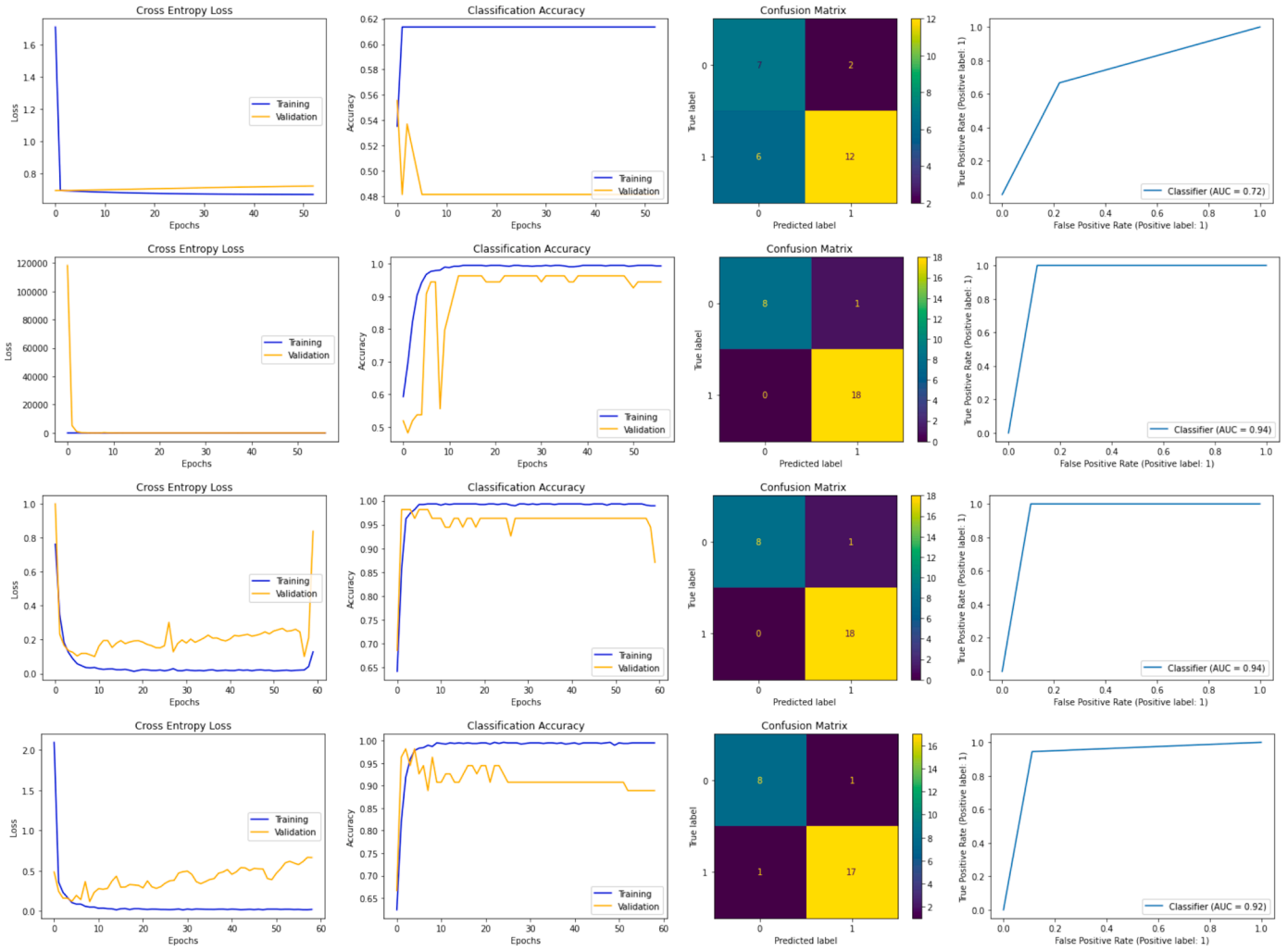


Fig. 18. Transfer learning [Top-Bottom: MobileNetV2, ResNet50, InceptionResNetV2, VGG16].

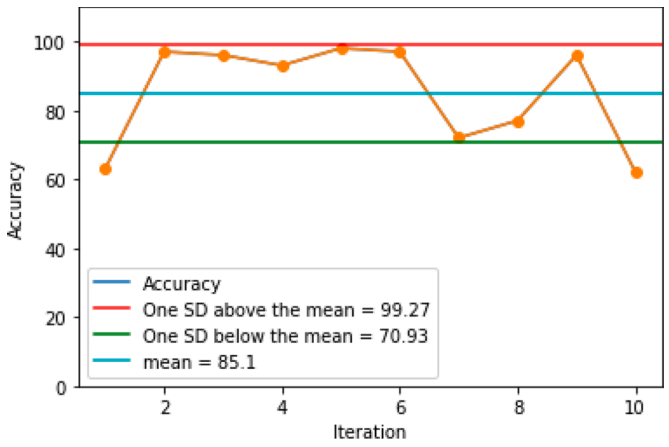


Fig. 19. K-fold cross validation results.

InceptionResNetV2 model architecture. While Inception is concerned with computational cost, ResNet is concerned with computational accuracy. In theory, deeper networks should outperform shallower networks, but in practice, deeper networks outperformed shallower networks due to an optimization problem rather than overfitting. In short, the deeper the network, the more difficult it is to optimize. To achieve higher accuracy, computer vision networks are becoming deeper and more complicated.Deeper networks, on the other hand, come at the expense of size and speed. The object detection task must be able to be performed on a computationally limited platform in real-world applications such as an autonomous vehicle or robotic visions.

MobileNet, a network for embedded vision applications and mobile devices, was created to address this issue. The idea behind MobileNet is to build lighter deep neural networks by using depthwise separable convolutions. The convolution kernel or filter is applied to all of the channels of the input image in a regular convolutional layer by doing a weighted sum of the input pixels with the filter and then sliding to the next input pixels across the images. Only the first layer of MobileNet employs this regular convolution. The depthwise separable convolutions are the next layers, which are a combination of the depthwise and pointwise convolutions. The depthwise convolution convolutions each channel independently. If the image has three channels, the output image will also have three channels. The input channels are filtered using this depthwise convolution.The pointwise convolution follows, which is similar to regular convolution but with a 1x1 filter. The goal of pointwise convolution is to combine the depthwise convolution output channels to create new features. As a result, the computational work required is less than that of regular convolutional networks. The model architecture is shown in Fig. 4. MobileNet outperforms other cutting-edge convolutional neural networks such as VGG16, VGG19, ResNet50, InceptionV3, and Xception. MobileNets are thin deep neural networks that are ideal for mobile and embedded vision applications. It uses depthwise separable convolutions in a streamlined architecture and

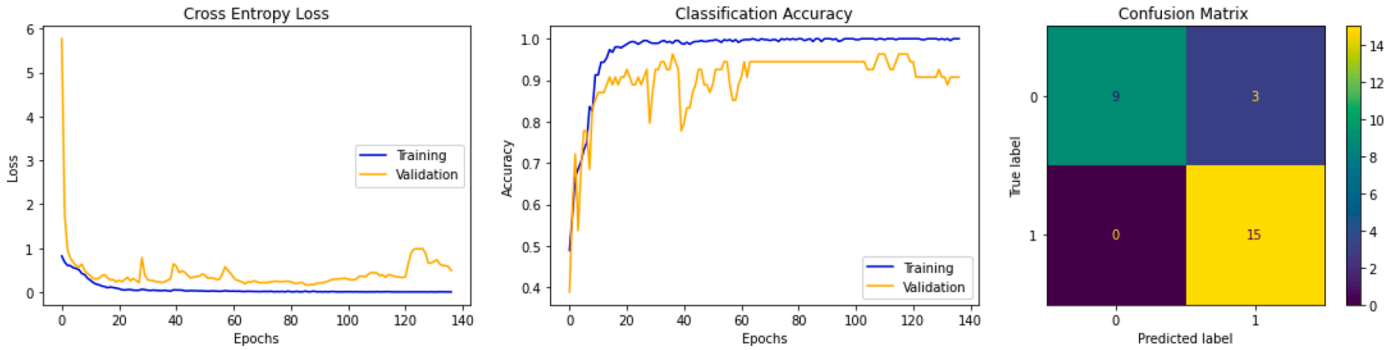


Fig. 20. Confusion matrix and diagnostic curves of LSTM + image model.

employs two simple global hyperparameters to efficiently trade off accuracy versus latency. MobileNet could be used for object detection, fine-grain classification, face recognition, large-scale geolocation, and other applications.

The following are the benefits of using MobileNet over other cutting-edge deep learning models. It reduced network size to 17MB and parameter count to 4.2 million. It is more performant and useful for mobile applications. It has a convolutional neural network with a low latency. Advantages always have some drawbacks, and with MobileNet, it's the accuracy. Even though MobileNet is smaller, has fewer parameters, and performs faster, it is less accurate than other cutting-edge networks. ResNet models reduce training time while increasing accuracy by not activating all neurons in every epoch. Furthermore, the model employs a clever strategy for improving model training performance by learning the feature once and then not attempting to learn it again; instead, it focuses on learning additional features. While VGG significantly improved speed and accuracy by introducing pretrained models and increasing model depth. The model's nonlinearity increased as the number of layers with smaller kernels increased. Unlike Inception v1 to v3, the Inception-ResNet-v2 model makes use of residual networks to improve the accuracy and convergence speed of the original model.

The LSTM based text model performs poorly as compared to the BoW approach text model with reduction in accuracy. This can be attributed to the fact that usually, attention mechanism don't work very well with clinical data, as also shown by the study conducted by researchers in Korea (Kim et al., 2020). Keywords play a more important role and hence, a more simple text model in this case performs better.

6. Conclusion and future outlook

A multi-modal approach is presented in this paper to classify a patient as covid positive or negative using the image of the chest X-ray/CT scan and the clinical notes provided with the scan. Data augmentation techniques are used to overcome the problem of small data sets, and they have been shown to improve model performance. The multi-modal is also compared to previously trained models. The final multi-modal results in a 97.8 percent accuracy on the testing data, with only one data point misclassified. The study takes a unique approach to identifying COVID-19 cases by relying solely on scan images and corresponding notes. This research can benefit all researchers working in this field around the world. The limitation of the study is the size of the dataset, in future a big data (comprises of text and image data both) can be generated and used. This study's future scope cannot be limited to hardware implementation, hybrid classification, etc. Applications can be expanded to include other types of medical data with additional classifiers, neural networks, and other AI and data techniques (Nasir et al., 2022a).

CRediT authorship contribution statement

Nida Nasir: Conceptualization, Methodology, Software, Writing – original draft. **Afreen Kansal:** Conceptualization, Methodology, Software, Writing – original draft. **Feras Barneih:** Investigation, Methodology, Validation, Writing – original draft. **Omar Al-Shaltone:** Investigation, Methodology, Validation, Writing – original draft. **Talal Bonny:** Supervision, Writing – review & editing. **Mohammad Al-Shabi:** Supervision, Writing – review & editing. **Ahmed Al Shammaa:** Funding acquisition, Project administration.

Declaration of Competing Interest

Authors declare that they have no conflict of interest.

Data Availability

Data will be made available on request.

Acknowledgment

We would like to extend sincere thanks to the University of Sharjah and its Research Institute of Science and Engineering (RISE) especially to the Bio-Sensing Research Group for supporting this work.

References

Apostolopoulos, I. D., & Mpesiana, T. A. (2020). COVID-19: Automatic detection from X-ray images utilizing transfer learning with convolutional neural networks. *Physical and Engineering Sciences in Medicine, 43*(2), 635–640.

Barneih, F., Nasir, N., Alshaltone, O., Qatmah, M., Bonny, T., Al Shabi, M., et al. (2022). Artificial neural network model using short-term Fourier transform for epilepsy seizure detection. *2022 advances in science and engineering technology international conferences (ASET)* (pp. 1–5). IEEE.

Cohen, J. P., Morrison, P., & Dao, L. (2020a). COVID-19 image data collection. arXiv:2003.11597https://github.com/ieee8023/covid-chestxray-dataset.

Cohen, J. P., Morrison, P., & Dao, L. (2020b). COVID-19 image data collection. 10.48550/ARXIV.2003.11597.

Cohen, J. P., Morrison, P., Dao, L., Roth, K., Duong, T. Q., & Ghassemi, M. (2020c). COVID-19 image data collection: Prospective predictions are the future. arXiv:2006.11988https://github.com/ieee8023/covid-chestxray-dataset.

Cohen, J. P., Morrison, P., Dao, L., Roth, K., Duong, T. Q., & Ghassemi, M. (2020d). COVID-19 image data collection: Prospective predictions are the future. arXiv:2006.11988https://github.com/ieee8023/covid-chestxray-dataset.

Dash, S., Verma, S., Bevinakoppa, S., Wozniak, M., Shafi, J., & Ijaz, M. F. (2022). Guidance image-based enhanced matched filter with modified thresholding for blood vessel extraction. *Symmetry, 14*(2), 194.

De Miranda, A. S., & Teixeira, A. L. (2020). Coronavirus disease-2019 conundrum: Ras blockade and geriatric-associated neuropsychiatric disorders. *Frontiers in Medicine, 7*, 515.

El Asnaoui, K., & Chawki, Y. (2021). Using X-ray images and deep learning for automated detection of coronavirus disease. *Journal of Biomolecular Structure and Dynamics, 39*(10), 3615–3626.

Hall, L. O., Paul, R., Goldgof, D. B., & Goldgof, G. M. (2020). Finding COVID-19 from chest X-rays using deep learning on a small dataset. arXiv preprint arXiv:2004.02060.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Hemdan, E. E.-D., Shouman, M. A., & Karar, M. E. (2020). COVIDX-Net: A framework of deep learning classifiers to diagnose COVID-19 in X-ray images. arXiv preprint arXiv:2003.11055.

Horry, M. J., Chakraborty, S., Paul, M., Ulhaq, A., Pradhan, B., Saha, M., et al. (2020). COVID-19 detection through transfer learning using multimodal imaging data. *IEEE Access, 8*, 149808–149824.

Ismael, A. M., & Şengür, A. (2021). Deep learning approaches for COVID-19 detection based on chest X-ray images. *Expert Systems with Applications, 164*, 114054.

Ji, Q., Huang, J., He, W., & Sun, Y. (2019). Optimized deep convolutional neural networks for identification of macular diseases from optical coherence tomography images. *Algorithms, 12*(3), 51.

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C., Liang, H., Baxter, S. L., et al. (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell, 172*(5), 1122–1131.

Kim, J., Lee, S., Hwang, E., Ryu, K. S., Jeong, H., Lee, J. W., et al. (2020). Limitations of deep learning attention mechanisms in clinical research: Empirical case study based on the korean diabetic disease setting. *Journal of Medical Internet Research, 22*(12), e18418.

Kucirka, L. M., Lauer, S. A., Laeyendecker, O., Boon, D., & Lessler, J. (2020). Variation in false-negative rate of reverse transcriptase polymerase chain reaction–based SARS-CoV-2 tests by time since exposure. *Annals of Internal Medicine, 173*(4), 262–267.

Maghdid, H. S., Asaad, A. T., Ghafoor, K. Z., Sadiq, A. S., Mirjalili, S., & Khan, M. K. (2021). Diagnosing COVID-19 pneumonia from X-ray and CTimages using deep learning and transfer learning algorithms*, vol. 11734. Multimodal image exploitation and learning 2021* (pp. 99–110). SPIE.

Mahdianpari, M., Salehi, B., Rezaee, M., Mohammadimanesh, F., & Zhang, Y. (2018). Very deep convolutional neural networks for complex land cover mapping using multispectral remote sensing imagery. *Remote Sensing, 10*(7), 1119.

Nasir, N., Alshaltone, O., Barneih, F., Al-Shabi, M., Bonny, T., & Al-Shammaa, A. (2021). Hypertension classification using machine learning-part I. *2021 14th international conference on developments in esystems engineering (DESE)* (pp. 464–468). IEEE.

Nasir, N., Kansal, A., Alshaltone, O., Barneih, F., Sameer, M., Shanableh, A., et al. (2022a). Water quality classification using machine learning algorithms. *Journal of Water Process Engineering, 48*, 102920.

Nasir, N., Oswald, P., Alshaltone, O., Barneih, F., Al Shabi, M., & Al-Shammaa, A. (2022b). Deep DR: Detection of diabetic retinopathy using a convolutional neural network. *2022 advances in science and engineering technology international conferences (ASET)* (pp. 1–5). IEEE.

Ohata, E. F., Bezerra, G. M., das Chagas, J. V. S., Neto, A. V. L., Albuquerque, A. B., de Albuquerque, V. H. C., et al. (2020). Automatic detection of COVID-19 infection using chest X-ray images through transfer learning. *IEEE/CAA Journal of Automatica Sinica, 8*(1), 239–248.

Phan, T. (2020). Novel coronavirus: From discovery to clinical diagnostics. *Infection, Genetics and Evolution, 79*, 104211.

Phankokkruad, M. (2020). COVID-19 pneumonia detection in chest X-ray images using transfer learning of convolutional neural networks. *Proceedings of the 3rd international conference on data science and information technology* (pp. 147–152).

Qatmh, M., Bonny, T., Barneih, F., Alshaltone, O., Nasir, N., Al-Shabi, M., et al. (2022). Sleep apnea detection based on ECG signals using discrete wavelet transform and artificial neural network. *2022 advances in science and engineering technology international conferences (ASET)* (pp. 1–5). IEEE.

Sahinbas, K., & Catak, F. O. (2021). Transfer learning-based convolutional neural network for COVID-19 detection with X-ray images. *Data science for COVID-19* (pp. 451–466). Elsevier.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 4510–4520).

Satia, I., Bashagha, S., Bibi, A., Ahmed, R., Mellor, S., & Zaman, F. (2013). Assessing the accuracy and certainty in interpreting chest X-rays in the medical division. *Clinical Medicine, 13*(4), 349.

Seidaliyeva, U., Akhmetov, D., Ilipbayeva, L., & Matson, E. T. (2020). Real-time and accurate drone detection in a video with a static background. *Sensors, 20*(14), 3856.

Shaik, N. S., & Cherukuri, T. K. (2022). Transfer learning based novel ensemble classifier for COVID-19 detection from chest CT-scans. *Computers in Biology and Medicine, 141*, 105127.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Soares, E., Angelov, P., Biaso, S., Froes, M. H., & Abe, D. K. (2020). SARS-CoV-2 CT-scan dataset: A large dataset of real patients CT scans for SARS-CoV-2 identification. MedRxiv.

Wang, L., Lin, Z. Q., & Wong, A. (2020a). COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. *Scientific Reports, 10*(1), 1–12.

Wang, N., Liu, H., & Xu, C. (2020b). Deep learning for the detection of COVID-19 using transfer learning and model integration. *2020 IEEE 10th international conference on electronics information and emergency communication (ICEIEC)* (pp. 281–284). IEEE.

Wong, H. Y. F., Lam, H. Y. S., Fong, A. H.-T., Leung, S. T., Chin, T. W.-Y., Lo, C. S. Y., et al. (2020). Frequency and distribution of chest radiographic findings in COVID-19 positive patients. *Radiology, 296*(2), E72–E78.

Woźniak, M., Siłka, J., & Wieczorek, M. (2021). Deep neural network correlation learning mechanism for CT brain tumor detection. *Neural Computing and Applications, 05841*, 1–16.

Zhang, R., Guo, Z., Sun, Y., Lu, Q., Xu, Z., Yao, Z., et al. (2020). COVID19XrayNet: A two-step transfer learning model for the COVID-19 detecting problem based on a limited number of chest X-ray images. *Interdisciplinary Sciences: Computational Life Sciences, 12*(4), 555–565.

Zhao, J., Zhang, Y., He, X., & Xie, P. (2020). Covid-ct-dataset: a ct scan dataset about covid-19. *arXiv preprint arXiv:2003.13865, 490*.