International Journal of
Environmental Research and Public Health

MDPI

Article

Detection of COVID-19 in Chest X-ray Images: A Big Data Enabled Deep Learning Approach

Mazhar Javed Awan [1,*], Muhammad Haseeb Bilal [1], Awais Yasin [2], Haitham Nobanee [3,4,5,*], Nabeel Sabir Khan [6] and Azlan Mohd Zain [7]

1 Department of Software Engineering, University of Management and Technology, Lahore 54770, Pakistan; s2020114006@umt.edu.pk
2 Department of Computer Engineering, National University of Technology, Islamabad 44000, Pakistan; awaisyasin@nutech.edu.pk
3 College of Business, Abu Dhabi University, Abu Dhabi 59911, United Arab Emirates
4 Oxford Centre for Islamic Studies, University of Oxford, Marston Rd, Headington, Oxford OX3 0EE, UK
5 Faculty of Humanities & Social Sciences, University of Liverpool, 12 Abercromby Square, Liverpool L69 7WZ, UK
6 Department of Computer Science, University of Management and Technology, Lahore 54770, Pakistan; nabeel.bloch@umt.edu.pk
7 UTM Big Data Centre, School of Computing, Universiti Teknologi Malaysia, Skudai 81310, Malaysia; azlanmz@utm.my
* Correspondence: mazhar.awan@umt.edu.pk (M.J.A.); nobanee@gmail.com (H.N.)

Abstract: Coronavirus disease (COVID-19) spreads from one person to another rapidly. A recently discovered coronavirus causes it. COVID-19 has proven to be challenging to detect and cure at an early stage all over the world. Patients showing symptoms of COVID-19 are resulting in hospitals becoming overcrowded, which is becoming a significant challenge. Deep learning's contribution to big data medical research has been enormously beneficial, offering new avenues and possibilities for illness diagnosis techniques. To counteract the COVID-19 outbreak, researchers must create a classifier distinguishing between positive and negative corona-positive X-ray pictures. In this paper, the Apache Spark system has been utilized as an extensive data framework and applied a Deep Transfer Learning (DTL) method using Convolutional Neural Network (CNN) three architectures —InceptionV3, ResNet50, and VGG19—on COVID-19 chest X-ray images. The three models are evaluated in two classes, COVID-19 and normal X-ray images, with 100 percent accuracy. But in COVID/Normal/pneumonia, detection accuracy was 97 percent for the inceptionV3 model, 98.55 percent for the ResNet50 Model, and 98.55 percent for the VGG19 model, respectively.

Keywords: COVID-19; corona virus; pneumonia; chest X-ray; CNN; transfer learning; big data; public health; data bricks; Apache Spark; ResNet50; InceptionV3; VGG19; SparkDL; machine learning; deep learning

Citation: Awan, M.J.; Bilal, M.H.; Yasin, A.; Nobanee, H.; Khan, N.S.; Zain, A.M. Detection of COVID-19 in Chest X-ray Images: A Big Data Enabled Deep Learning Approach. *Int. J. Environ. Res. Public Health* **2021**, *18*, 10147. https://doi.org/10.3390/ijerph181910147

Academic Editors: Quynh Nguyen and Paul B. Tchounwou

Received: 22 August 2021
Accepted: 21 September 2021
Published: 27 September 2021

1. Introduction

The primary instance of COVID-19 was accounted for in Wuhan, China [1]. The virus marked a pandemic on 11 March 2020 [2]. On 14 September 2021, the WHO recorded 4,636,153 deaths and 225,024,781 affirmed cases. As of 12 September 2021, an aggregate of 5,534,977,637 vaccine doses had been distributed [3,4]. This pandemic affected almost all countries.

There is pressure on testing laboratories due to the global epidemiological condition. COVID-19 tests are accessible to check for present or past diseases. A viral test decides if you are contaminated right now. Antigen testing and nucleic acid amplification tests (NAATs) are two types of viral tests that can be utilized. Neutralizer testing ought not to be used to analyze contamination that is present [5].

A chest X-ray is the most widely recognized imaging procedure used to analyze SARS-CoV-2 infection. Utilizing a Convolutional Neural Network to perceive COVID-19 radiology images gives extensive results, as per a couple of studies [6–8], and it merits further attention and significance.

The blend of the two innovations Apache Spark and Transfer Learning, permits information investigators to recognize photographs rapidly and effectively and encourages models that can run on clusters [9,10].

Big data analytics could be helpful for researchers as well as businesses. Its techniques have been implemented for fake news prevention, as mentioned in [11]. Similarly, in [12], social media are analyzed through big data and are used for stock-market predictions. It also has been used for review processing [13]. Moreover, in [14], the big data approach is used for sales analytics, such as predicting black Friday sales based on historical data.

"Big Data" is an all-encompassing label that encompasses non-traditional tactics and innovation for acquiring, sorting, and producing experiences from massive datasets [15]. The volume of information has expanded drastically lately, requiring the improvement of data analytics frameworks. The Apache Spark framework [16] is the most notable stage for big data analytics. Spark can process information utilizing an assortment of structures. It is lightning quick, upholds various programming languages, incorporates machine-learning usefulness, and associates with multiple platforms [17]. It has characteristics described as 10 V's in Figure 1.

Figure 1. 10'Vs of big data characteristics.

Big data in medical care is developing to bring understanding from substantial information sets, and it is yielding tremendous results while bringing down costs and guaranteeing great patient care [18]. Apache Spark can assume an essential part in medical services investigation, helping clinicians better diagnose problems, especially picture classification [19]. Information-retrieval models need increasing detail as mentioned in [20], and information-extraction techniques were mentioned in [21].

Big data analytics is the practice of utilizing advanced analytical techniques to examine massive volumes of information. Scientists, organizations, and examiners can use extensive data analysis to settle on better and quicker choices.

They can utilize modern analytics procedures, such as machine learning, deep learning, prescient examination, and others to create novel thoughts, approaches, and experiences

regarding big data [22]. The goal is to uncover hidden patterns and associations that disclose crucial information about the clients who developed it ahead of time. Big data analytics can improve findings and therapies, enhance administration quality at a lower cost, and deliver better outcomes [23].

Apache Spark is an open-source information handling system that spotlights speed, effortlessness, and progressed investigation. Spark runs in-memory on clusters, is not constrained by Hadoop's two-stage MapReduce model, and is extremely fast [24]. Spark may run on its own, on top of tools, and it can read data directly from the file systems of these tools. Spark can deal with spilling notwithstanding in-memory preparing, diagram handling, and machine learning [25]. The Spark stores data in memory utilizing Resilient Distributed Datasets (RDDs). Clients can utilize the iterative activity to peruse information from the disc and compose it to RDDs to assemble a task. Alternatively, they can use intuitive mode to run numerous questions on a similar subset of data [26]. The Spark Streaming is a decent stream-preparing alternative [27]. Figure 2 throws light on Apache Spark and its relation to the BigDL library for deep learning.

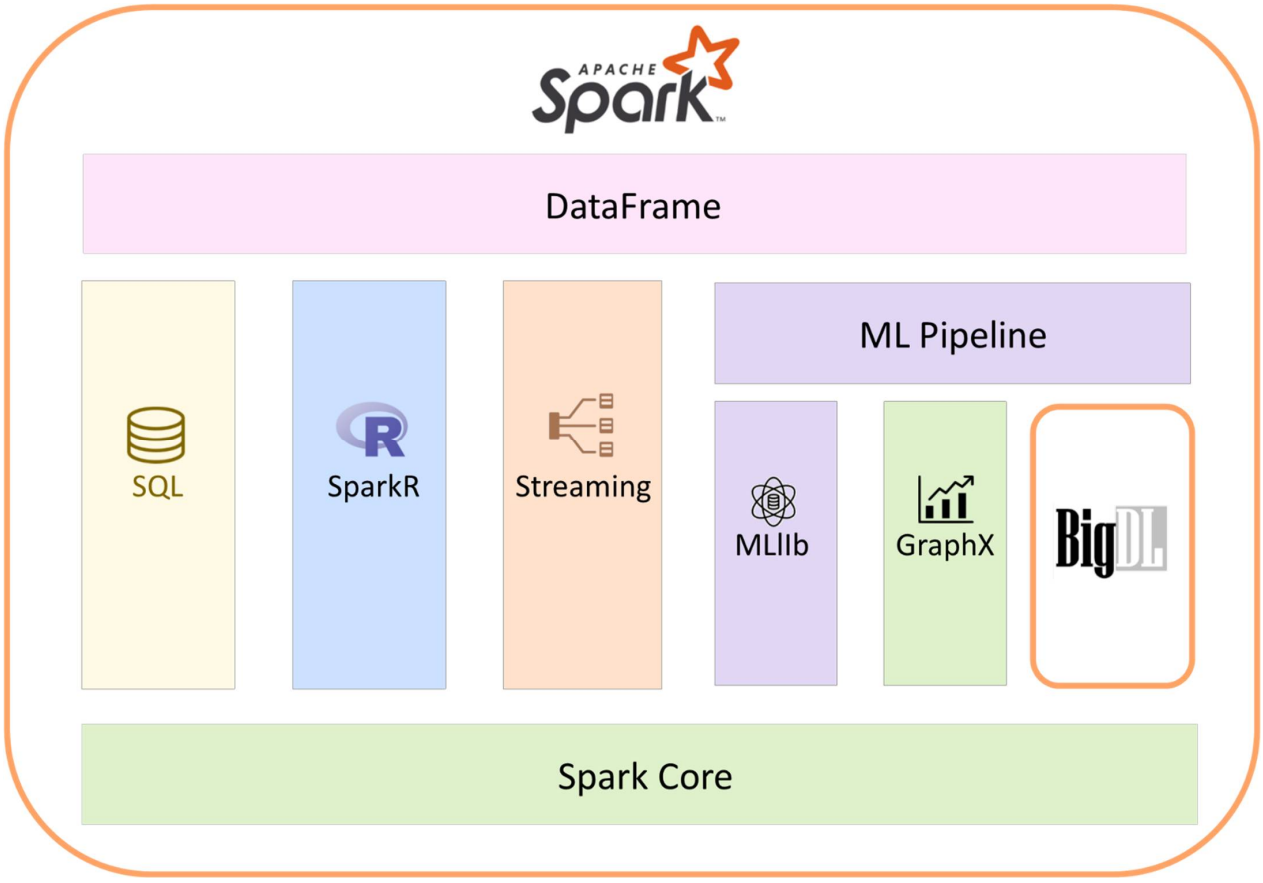


Figure 2. Apache Spark and its relation to BigDL library for deep learning.

Deep learning is a subset of state-of-the-art artificial intelligence calculations, frequently known as deep organized learning. It is subject to mathematical action procedures and Artificial Neural Networks [28]. This methodology extricates and changes attributes by using various secret layers. Each layer gets the yield of the first one as info, acknowledges contributions to be prepared, and means to deliver the estimation's outcome [29].

Deep-learning approaches empower machines to learn without the requirement for programming. Automatic speech acknowledgment, natural language handling, medical prescription turn of events and toxicology, client relationship management, recommenda-

tion frameworks, and bioinformatics are only a couple of the issues that deep learning has been utilized to address [30]. Spark with deep understanding is also used in [31].

The CNN (Convolutional Neural Network) is an artificial neural network powered by an animal's visual cortex [32]. A deep learning algorithm can order input pictures, for example, chest X-ray pictures as tainted or not through COVID-19 [33]. The CNN will separate features, making grouping quicker and more exact. The CNN's four fundamental layers are the Convolution, Non-Linearity (ReLU), Pooling, and Classification layers [34,35].

Transfer Learning (TL) is a procedure for moving acquired data to an original dataset to all the more likely process another destination dataset. TL happens when a model is reused as the beginning stage for another undertaking. One of the critical spaces of utilization for TL [36] is picture recognition. Rather than old-style learning, which is secluded and centred around singular tasks, transfer learning permits us to apply the knowledge on recently prepared models to foster new models and even tackle difficulties, for example, having less information for the new assignment.

The main intention of this investigation is to make a Deep Transfer Learning (DTL) structure utilizing Convolutional Neural Network (CNN) with the Apache Spark big data platform, in light of pre-arranged models InceptionV3, ResNet50, and VGG19. This work was inspired by deep-learning work in [37–40].

Our contribution is significant, which has been described in bullet points:

- We propose a novel approach with logistic regression for features extraction and CNN-based architectures of pre-trained VGG19, InceptionV3, and ResNet50 to detect COVID-19 from chest X-rays;
- To the best of our knowledge, our architectures with state-of-art architectures are effective and accurate after using the big-data framework, the Apache Spark in the pipeline;
- We appraise our architectures, classifying 100% in cases of COVID-19 and healthy patient chest X-ray images.
- Finally, our datasets are considerable, consisting of 1063 images for a 3-class classifier and 708 total images for a 2-class classifier.

The following is how the rest of the article is organized. Section 2 is about related work. Materials and methodology are explained in Section 3. The approach and outcomes of the experiments are detailed in full in Section 4. Section 5 consists of a prolonged discussion. Finally, Section 5 brings the conclusion.

2. Related Work

Deep learning approaches are mostly without big data. But some approaches to deep learning have been used without Spark. Their advantages and disadvantages have been discussed here. The study [41] used the CoroDet model has two class-based classifications of COVID-19 positive and negative patients. There are three types of classifiers: two-class (COVID and healthy patients), three-class (COVID positive patients, healthy person, and non-COVID pneumonia infected person), and four-class (COVID positive patients, healthy person, non-COVID pneumonia infected person, and non-COVID bacterial pneumonia). For low-quality X-ray pictures, the model performed poorly. According to the author, it has the most extensive dataset, and it outsmarted all previously used models in terms of accuracy. In the future, authors will want to use a larger dataset.

This study presents an automated, low-cost, rapid, and high-performance method for detecting COVID-19 disease. In-depth features from CT scan pictures are recuperated utilizing a CNN with a pre-trained CNN-based DenseNet201 design and transfer learning in [42]. The execution was evaluated using an Extreme Learning Machine (ELM) classifier. It outperformed the pre-trained models. The utilization of different ELM classifier initiation functions is dependent on deep features. In the future, authors will want to make web and mobile applications for the help of doctors using this technology.

A deep-learning algorithm named "COVIDetection-Net" was proposed in [43] to recognize coronavirus from chest radiography pictures. The proposed system utilizes the

ShuffleNet and SqueezeNet designs for deeply learned attributes, just as it uses Multiclass Support Vector Machines (MSVM) for recognition and characterization.

The algorithm in [44] can group the information into three classifications: corona-positive patients, other pneumonia-infected patients, and no contaminations. Three particular learning algorithms are utilized to make the model: CNN, VGG-16, and ResNet-50. The chest X-ray is from Kaggle's library. VGG16 beat CNN and ResNet-50 in each of the three learning algorithms to assess the model's performance. Just two current COVID-19 mechanized screening strategies are contrasted with the version of the suggested TLCoV model. Data augmentation is additionally utilized. As a result of its predominant dynamic directing technique, the creator's model uses the Capsule Network.

Three deep CNN models, AlexNet, GoogleNet, and ResNet, were pre-trained utilizing transfer learning to establish model parameters [45]. Softmax was used as the fully connected layer's classification technique. Relative majority voting was used to obtain the ensemble classifier EDL-COVID. According to the findings, the ensemble model outperformed the component classifier in terms of overall classification performance.

The creators of [46] made two sickness diagnosis algorithms: a deep neural network (DNN) because of picture fractal properties and a CNN strategy that examines lung pictures straightforwardly. The segmentation method is also used to locate sick tissue through the CNN model.

The authors have presented a quick detection method based on X-ray image processing that they believe would benefit society. In [47], the authors proposed using X-ray scans to discover COVID-19 patients using the nCOVnet algorithm. In AI, they employed deep learning. In under 5 s, the proposed model could identify a COVID-19 positive patient. With the bit of data accessible, the creators could accomplish a valid positive rate of 97.62 percent. Many of the previous studies that claimed accuracy of up to 98 percent did not account for the possibility of data leaking, which the authors addressed while training nCOVnet, resulting in impartial results.

In [48], machine learning models were used to predict COVID-19 spread. Similarly, big data analytics is being used extensively to detect other diseases and drugs, as mentioned in [49–52].

3. Materials and Methods

This section presents the dataset description and methods. Section 3.1 belongs to the description of the dataset. Finally, the proposed big-data approach is explained through pre-trained CNN models ResNet50, InceptionV3, and VGG19 models.

3.1. Coronavirus X-ray Images Dataset

Chest radiograph or chest X-ray pictures were used in this investigation through two datasets. These images were taken from the Kaggle repository and came from two datasets: "Coronavirus chest x-ray images" [53] and "Chest X-Ray images (Pneumonia)" [54].

3.1.1. Two-Class Classifier Dataset Description

In a two-class classifier, 708 X-ray images are used altogether, separated into two classifications: 354 COVID-19 infected patients' X-ray images and 354 normal X-ray images. A dataset of 354 typical and 354 COVID-19 patients was assembled utilizing front-facing projections of chest X-ray pictures. The images used for testing were unknown for models. COVID-19 infected patients' X-ray images were labeled with 1, while normal X-ray images were marked with 0. Table 1 describes the dataset sample of two classes.

Table 1. Numerical description of the prepared dataset for a binary classifier.

The Classes	Number of Images
COVID-19	354
Normal	354

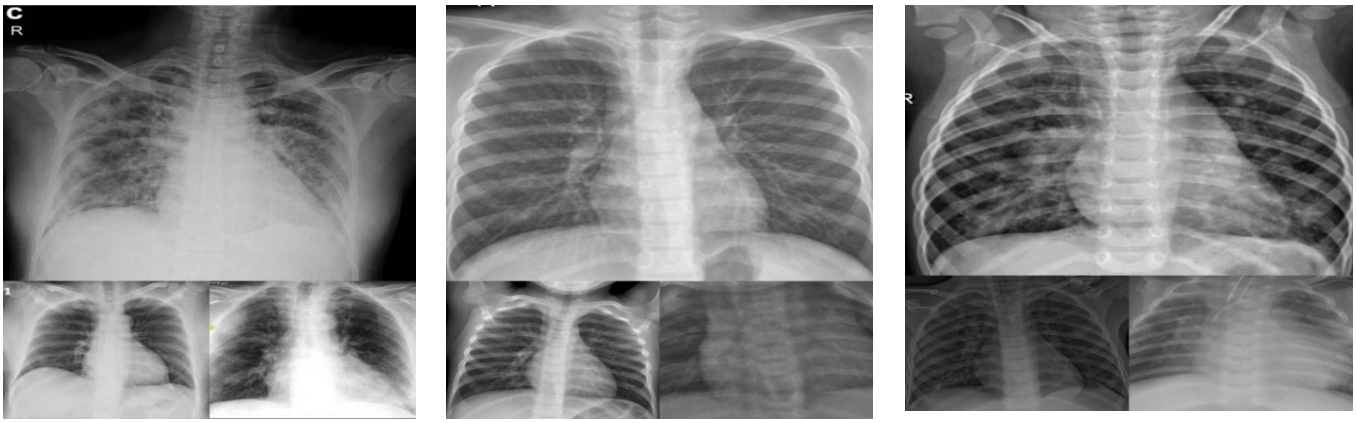
3.1.2. Three-Class Classifier Dataset Description

In the three-class classifier, there were 1063 images from the two databases mentioned above. It had three classifications: 354 pictures for COVID-19 patients, 355 for pneumonia-infected patients, and 354 for healthy individuals. COVID-19 infected patients' X-ray images were labeled with 1. Normal X-ray images were tagged with 0. At the same time, pneumonia-infected chest X-ray images were marked with 2. Table 2 describes the number of samples of the three classes.

Table 2. Numerical description of the prepared dataset for three-class classifier.

The Classes	Number of Images
COVID-19	354
Normal	354
Pneumonia	355

Figure 3a–c are the samples of X-rays of COVID-19, healthy, and pneumonia chest images, respectively.



(**a**) (**b**) (**c**)

Figure 3. Sample X-ray images (**a**) COVID-19 lung X-ray images; (**b**) Healthy chest X-ray images; (**c**) Pneumonia chest X-ray images.

3.2. Our Approach

The transfer learning pipelines were used on Apache Spark in this exploration. The pre-arranged CNN architecture InceptionV3, Residual Net(ResNet50), and Visual Geometry group (VGG19) models were utilized along with logistic regression to sort out the chest X-ray pictures on Data Bricks File System(DBFS) of three classes. Figure 4 shows the architecture of our approach.

For deep learning at a more profound level, Databricks Runtime ML incorporates the Pipelines library.The framework architecture of the Pipelines is a deep-learning open-source project. It is a significant level system that utilizes Apache Spark through machine lerning and deep learning as well [55–57].

Logistic regression is a machine-learning-based statistical method for dissecting autonomous highlights that characterize a result. This model used three types of pictures (COVID-19 patients, pneumonia patients, and non-infected healthy persons)[58]. On ImageNet, there are three pre-trained designs: InceptionV3, ResNet50, and VGG19 models with loads.

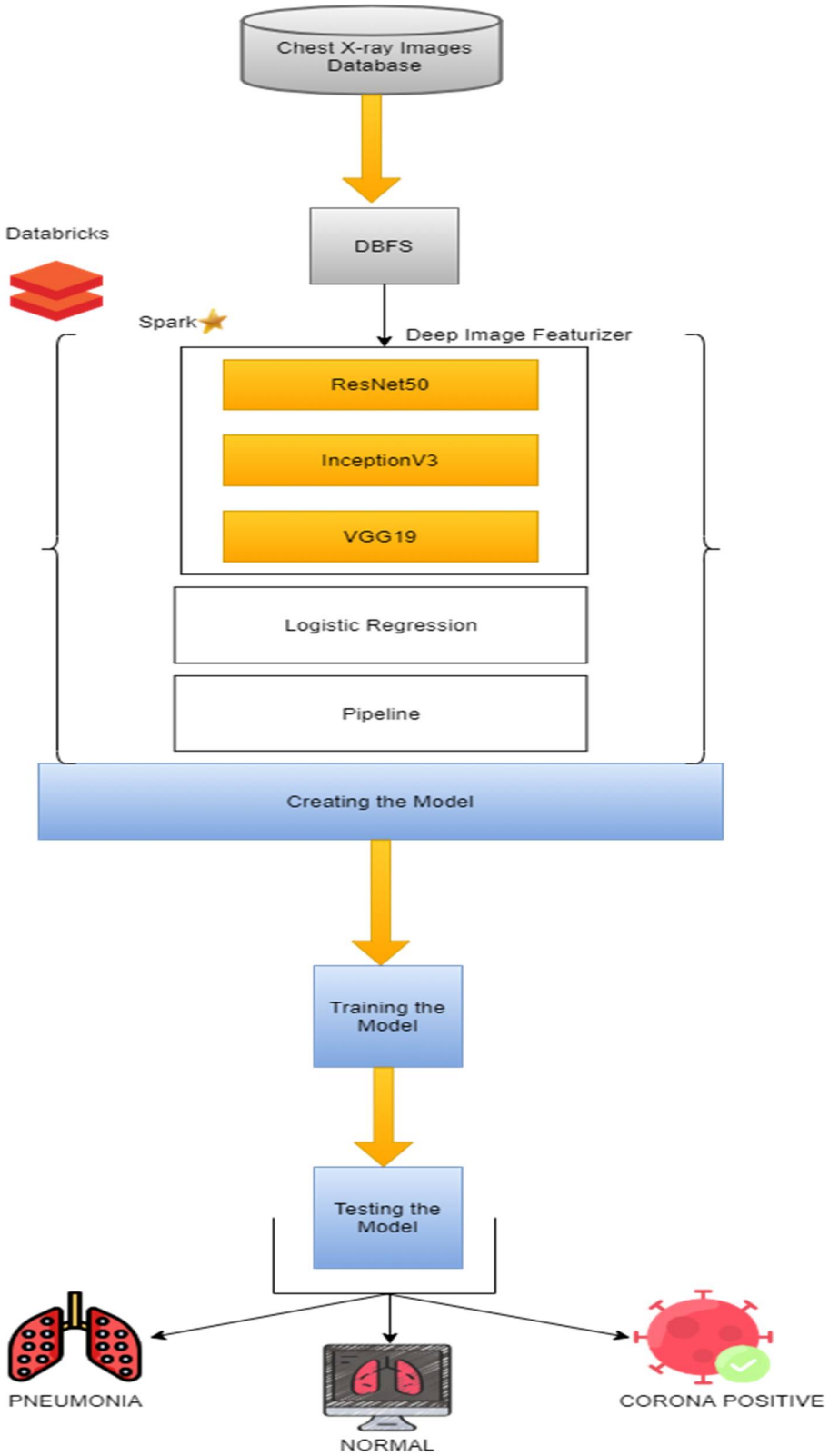


Figure 4. The working model incorporates deep-transfer learning with the Apache Spark architecture InceptionV3, Residual Net (ResNet50), and Visual Geometry group (VGG19) on Data Bricks File System(DBFS).

The 3rd cycle of Google's Inception CNN network model is Origin V3 [59] a deep neural network for image analysis and object recognition based on TensorFlow. A Residual Network with 50 layers is known as ResNet50 [60]. ResNet is a subclass of CNN that is most often utilized in picture acknowledgement and grouping. At the same time, VGG19 evolved from the VGG model, which is also the CNN evolution.

The InceptionV3 has been created for the identification of images. The Inception network was an essential milestone in the development of CNN classifiers. Before its inception, most popular CNNs stacked convolution layers deeper and deeper, hoping to get better performance. The Inception network, on the other hand, was complex [61]. Its constant evolution leads to the creation of several versions of the network. InceptionV3 incorporated needed upgrades stated for InceptionV2 [62].ResNet, short for Residual Networks, is a classic neural network used as a backbone for many computer vision tasks [63]. ResNet50 is a combination of 48 convolution layers along with one MaxPool and one Average Pool layer [64]. The main innovation of ResNet is the skip connection [65]. The VGG19 is the modified form of the VGG [66]. All three models were improved from conventional CNN models. It can be easily used along with transfer learning for image classification. It has 16 layers. This model will be smartly devoted to specialists' help settling on better COVID choices and urging medical imaging experts to receive new methodologies as an asymptomatic instrument.

4. Experiment and Results

4.1. Experiment Setup

The Databricks Workspace was utilized for testing [67]. It is an analytics stage based on Apache Spark. It is a cloud-based platform similar to the Google Colab cloud environment [68]. Databricks is a synergistic platform that permits clients to concentrate the entirety of their logical tasks and oversee machine-learning models over as long as they can remember the cycle. A group has been developed in this platform to run the model according to the assortment of guidelines.

The Databricks File System was utilized to save the chest X-rays pictures (DBFS). It is a dispersed record framework that allows storing of information for inquiries inside Databricks and makes it reachable across clusters. A cluster was created using Apache Spark.

In a two-class classifier, meaning it has two paths to store. First is the path of the typical images, and the second path is chest X-rays about COVID-19. A three-class classifier has three paths to store. The first two paths are similar, and the third path is about pneumonia chest X-ray images. The dataset was used 80% for training and 20% for testing. Therefore, the model was trained on 566 X-ray images, and 142 images were used for testing. Moreover, 74 were normal X-ray images and 68 were COVID-19 infected patients' X-ray images.

In the case of the three-class classifier, the dataset was used 80% for training and 20% for testing. Therefore, the model was trained on 856 X-ray images, and 207 images were used for testing. The images used for testing were unknown for models. Moreover, testing data was divided. There were 74 normal X-ray images, 66 COVID-19 infected patients' X-ray images, and 67 pneumonia infected patients' X-ray images.

4.2. Evaluation Metrics

We calculated our results using mean accuracy, precision, recall, confusion matrix, area under curves, and training losses.

The deep-learning approach is applied through logistic regression to extract features. In this examination, three CNN-based models were prepared and tried on chest X-ray pictures to identify COVID-19 or pneumonia-infected people utilizing Apache Spark. The deep-learning pipelines were used to stack photos into a Spark DataFrame, and views influenced by COVID-19 and ordinary pictures were named with the qualities "1" and

"0" individually. Infected chest X-ray images were labelled with the quality "2" in the three-class classifier pneumonia.

The idea of a featurizer in deep learning pipelines empowers quick transfer learning on an Apache Spark cluster. A DeepImageFeaturizer was utilized, and the InceptionV3, ResNet50, and VGG19 models were used for this examination. Four measurements were used to assess the model's exhibition in this investigation: accuracy, weighted recall, and weighted precision. The primary exhibition metric that was analyzed was accuracy. It alludes to how close the estimations are to a foreordained value.

The two-class classifier model performed with 100 percent accuracy with InceptionV3, ResNet50, and VGG19. The precision and recall also showed 100 percent results. This model correctly predicted samples for confirmation. In terms of the three-class classifier, InceptionV3, ResNet50, and VGG19 were 97, 98.55, and 98.55 percent, respectively. Table 3 shows the performance comparison of InceptionV3, ResNet50, and VGG19 for both binary and three-class classifiers. Table 3 shows the result of three models with two and three classes.

Table 3. Overall performance comparison of 3-class and binary classifier InceptionV3, ResNet50, and VGG19 models.

Model	Classes	Mean Accuracy	Precision	Recall	Mean AUC
Inception V3	COVID-19, Normal COVID-19, Normal, Pneumonia	1 97.10%	1 0.9713	1 0.9710	1 0.9784
ResNet50	COVID-19, Normal COVID-19, Normal, Pneumonia	1 98.55%	1 0.9855	1 0.9855	1 0.9890
VGG19	COVID-19, Normal COVID-19, Normal, Pneumonia	1 98.55%	1 0.9855	1 0.9855	1 0.9893

The binary classifier training-loss graph from iteration 0 to 10 for VGG19, ResNet50, and InceptionV3 is shown in Figure 5.

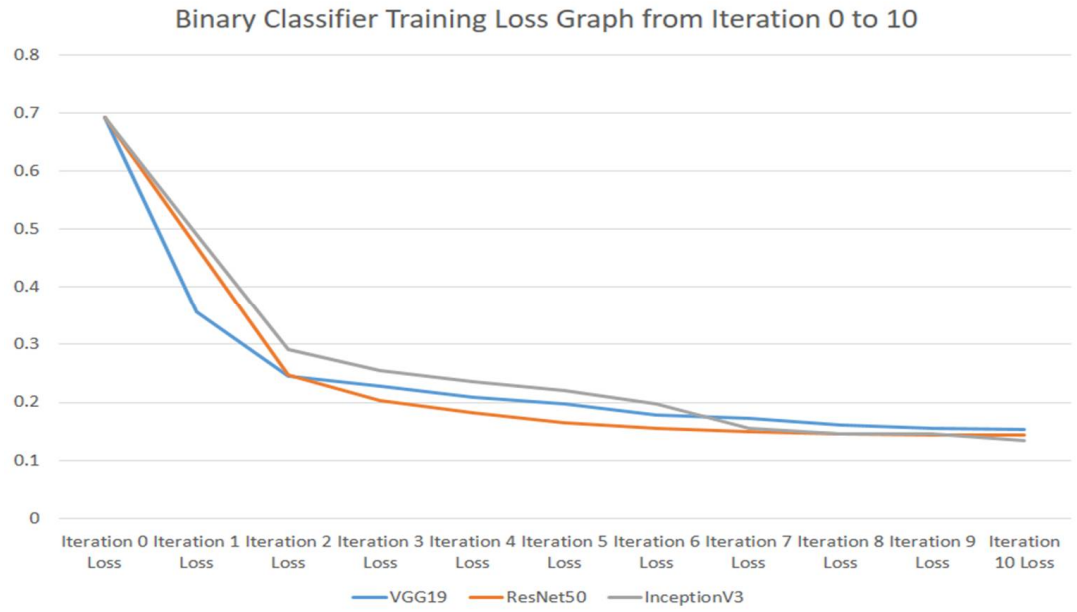


Figure 5. Binary classifier training-loss graph from iteration 0 to 10 for VGG19, ResNet50, and InceptionV3 Models.

The three-class classifier training-loss graph from iteration 0 to 10 for VGG19, ResNet50, and InceptionV3 is shown in Figure 6.

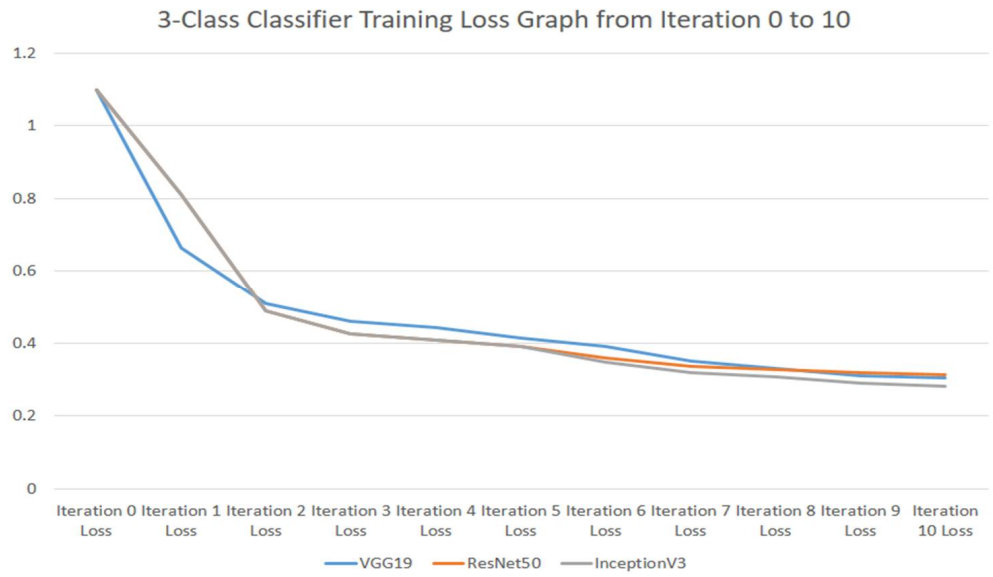


Figure 6. Three-class classifier training-loss graph from iteration 0 to 10 for VGG19, ResNet50, and InceptionV3 models.

5. Discussion

In this work, an architecture for detecting COVID-19 in chest X-ray pictures was proposed based on deep transfer learning through the Apache Spark framework. According to the exploratory findings, our model attained an accuracy of 100% for InceptionV3, the ResNet50 Model, and VGG19. All of the most recent indices have confirmed 100% correct results. For the three-class classifier, the accuracy obtained for InceptionV3, ResNet50, and VGG19 was 97%, 98.55%, and 98.55%, respectively. The confusion matrices for two-class classifiers obtained for Inception V3, ResNet50, and VGG19 models are shown in Figure 7.

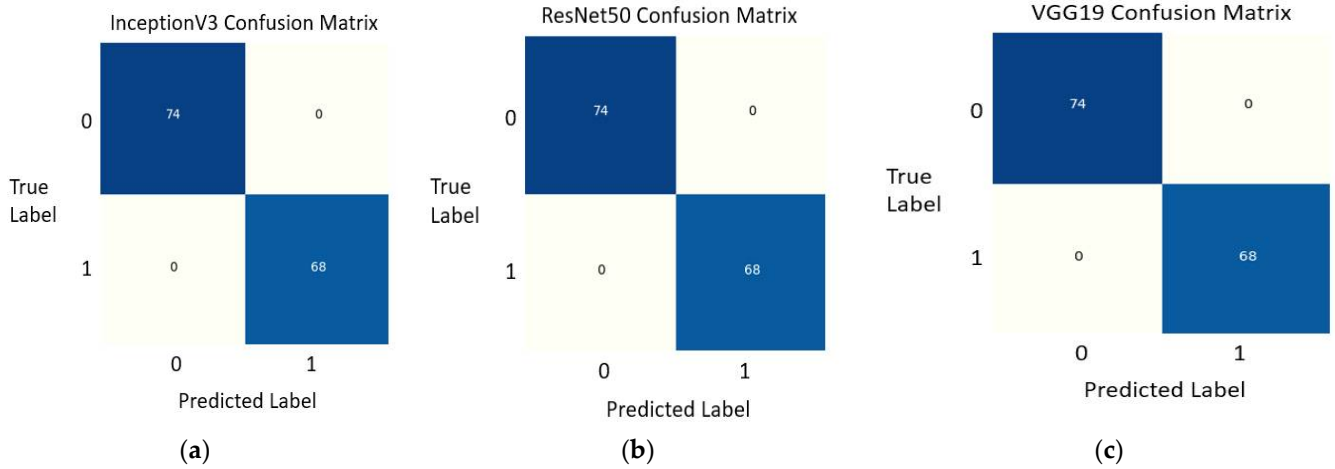
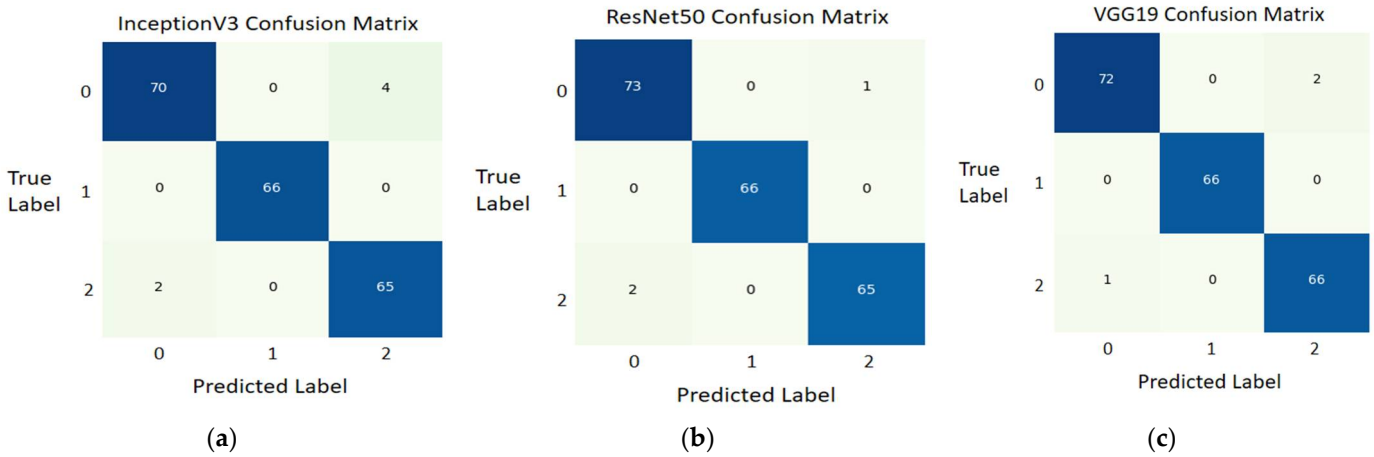


Figure 7. Two-class classifier confusion matrix; (**a**) Confusion matrix of Inception V3 model; (**b**) Confusion matrix of ResNet50 model; (**c**) Confusion matrix of VGG19 model.

These confusion matrices show that 74 values were predicted correctly. Therefore, 74 values are True Positive (TP). It means that these X-ray images of COVID-19 patients were detected correctly. Similarly, 68 normal photos are not affected with COVID-19 and were also classified and detected correctly. This category lies in True Negative (TN). At the same time, no value lies in False Positive (FP) and False Negative (FN). The area under the curve is also 1.

The confusion matrices for the three-class classifier obtained for Inception V3, ResNet50, and VGG19 models are shown in Figure 8.



(**a**) (**b**) (**c**)

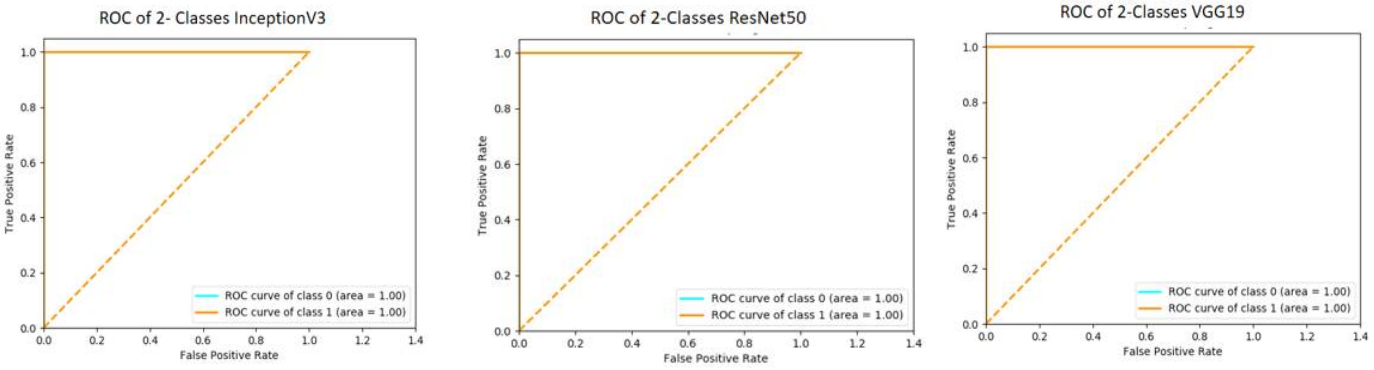
Figure 8. Three-class classifier confusion matrix; (**a**) Confusion matrix of InceptionV3 model; (**b**) Confusion matrix of ResNet50 model; (**c**) Confusion matrix of VGG19 model.

The confusion matrix of InceptionV3 shows that 70 typical images' values were predicted correctly, while four normal images were not predicted correctly. Similarly, 66 COVID-19 X-ray images and 65 pneumonia X-ray images were correctly predicted. Two pneumonia X-ray images were not predicted correctly.

The confusion matrix of ResNet50 shows that 73 normal images' values were predicted correctly, while 1 normal image was not predicted correctly. Similarly, 66 COVID-19 X-ray images and 65 pneumonia X-ray images were predicted correctly. The two pneumonia X-ray images were not predicted correctly.

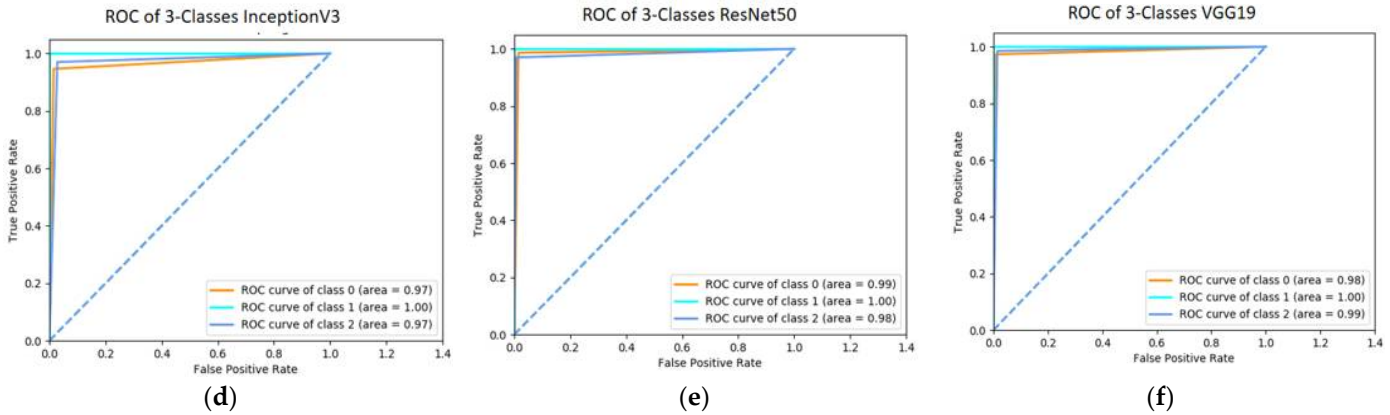
The confusion matrix of VGG19 shows that 72 normal images values were predicted correctly, while 2 normal images were not predicted correctly. Similarly, 66 COVID-19 X-ray images and 66 pneumonia X-ray images were predicted correctly. One pneumonia X-ray image was not predicted correctly.

Figure 9 shows the binary-class classifier area under curve ROC plot and the 3-class classifier area under curve ROC plot for InceptionV3, ResNet50 and VGG19.



(**a**) (**b**) (**c**)

Figure 9. *Cont.*



(**d**) (**e**) (**f**)

Figure 9. ROC AUC for two and three classes: (**a**) AUC two-class classifier InceptionV3 model; (**b**) AUC two-class classifier ResNet50 model; (**c**) AUC of two-class classifier VGG19 model; (**d**) AUC of three-class classifier InceptionV3 model; (**e**) AUC of three classifier ResNet50 model; (**f**) AUC of three-class classifier VGG19 model.

The Apache Spark system, joined with the deep-transfer learning strategy, created excellent execution, proficient examination, and progressed discoveries. Because of the combination of these three methodologies, our model can undoubtedly distinguish between corona-affected images and healthy chest X-ray images. We also compared our result with previous related studies in Table 4

Table 4. Comparing the state of the artworks with our proposed model.

Author, Year	Architecture	2 Class	3 Class	4 Class
Hussain et al., 2021 [51]	Novel CNN Model CoroDet	99.1%	94.2%	91.2%
M. Turkoglu, 2021 [52]	ELM and Deep Neural Network	-	98.36%	-
Das et al., 2021 [54]	CNN, VGG-16 ad ResNet-50	-	VGG = 97.67% ResNet-50 = 96.41% CNN = 93.67%	-
Zhou et al., 2021 [55]	AlexNet, GoogleNet, ResNet and SoftMax for Classification	-	GoogleNet = 98.25% ResNet = 98.56% SoftMax = 98.56% The ensemble model outperformed the component classifier.	-
Hassantabar et al., 2020 [56]	Deep Neural Network (DNN) and Convolutional Neural network (CNN)	CNN = 93.2% DNN = 83.4%	-	-
Panwar et al., 2020 [57]	Deep learning neural network model using nCOVnet algorithm	97.97%	-	-
Proposed Model	InceptionV3, ResNet50, VGG19	Inception V3 = 100% ResNet50 = 100% VGG19 = 100%	Inception V3 = 97% ResNet50 = 98.55% VGG19 = 98.55%	-

The table, as mentioned above, compared the proposed model with some other similar works. This proposed model performed better as compared to other models in terms of performance and innovation.

Aside from the excellent result, however, there are some limitations of our study. Firstly, our dataset is biased in the case of two classes. Secondly, if the noise in data increased, then the performance would also be decreased. Thirdly, our model is significantly

slower due to an operation of logistic regression and transfer learning of various CNN architectures in the pipeline. Due to several layers, the training process takes a lot of time if the computer does not have a good GPU. Fourthly, the decisions of multiple radiologists are not considered in the final prediction. Lastly, our architectures have excellent performance while classifying images that are very similar to the dataset. However, if the images contain tilt or rotation, our architectures usually have difficulty organizing the image.

6. Conclusions

COVID-19 is a highly hazardous virus. World governments have adopted various methods to halt the spread of the virus, depending on their resources. Several countries expect new waves of this virus due to new deadly virus variants, and some impose lockdowns to control the spread. Countries are hastening their vaccination process according to their resources and availability of vaccine supply. Quick identification of positive corona patients and their isolation is effective in reducing spread, according to WHO. Because of this fact, a deep transfer-learning approach with Apache Spark architecture was created to recognize the coronavirus in chest X-ray images. Our image classification system relied on deep-learning pipelines and logistic regression, and three CNN-based models were used in this study, namely InceptionV3, ResNet50, and VGG19. Two databases from the Kaggle repository were utilized to create and test the model. A binary-class classifier and a three-class classifier were proposed in our architecture.

Databricks workspace has been used as an enormous data analytics platform to process X-ray images, and Apache Spark's framework played a significant role in this process. In this investigation, weighted precision, weighted recall, and accuracy were examined as execution measurements for deep transfer learning. For InceptionV3, ResNet50, and VGG19, the outcomes were phenomenal. These three models named InceptionV3, ResNet50, and VGG19 gave 100% accuracy for binary-class classification. All performance measurements proved that these three models are predicting 100% correctly. InceptionV3, ResNet50, and VGG19 gave 97%, 98.55%, and 98.55% accuracy, respectively, when the 3-classes were classified. It has been ensured that architectural design can accurately detect coronavirus infection in chest X-ray images based on the outcomes generated by our model.

These findings may persuade health professionals worldwide to utilize these cutting-edge strategies to combat the coronavirus pandemic. A model will be developed for sudden spikes in demand for gigantic PC clusters utilizing the blend of its last two advancements, which makes the workplace intriguing. This model will make coronavirus detection simpler, quicker, and more affordable.

In future work, the integration of TensorFlow into Apache Spark will foster a new model for detecting coronavirus in chest X-rays, magnetic resonance imaging, and CT scans on a 4-class classifier.

Author Contributions: All authors have contributed equally to this manuscript. 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: This dataset available online and anyone can be used.

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

References

1. World Health Organization. Archived: WHO Timeline—COVID-19. 2021. Available online: https://www.who.int/news/item/27-04-2020-who-timeline---covid-19 (accessed on 1 August 2021).
2. Cucinotta, D.; Vanelli, M. WHO declares COVID-19 a pandemic. *Acta Bio Med. Atenei Parm.* **2020**, *91*, 157.
3. Nguyen, T.T.; Criss, S.; Dwivedi, P.; Huang, D.; Keralis, J.; Hsu, E.; Phan, L.; Nguyen, L.H.; Yardi, I.; Glymour, M.M. Exploring US shifts in anti-Asian sentiment with the emergence of COVID-19. *Int. J. Environ. Res. Public Health* **2020**, *17*, 7032. [CrossRef]

4. World Health Organization. Coronavirus (Covid-19) Dashboard. 2021. Available online: https://covid19.who.int/.
5. Liu, X.-J.; Mesch, G.S. The adoption of preventive behaviors during the COVID-19 pandemic in China and Israel. *Int. J. Environ. Res. Public Health* **2020**, *17*, 7170. [CrossRef] [PubMed]
6. Abbas, A.; Abdelsamea, M.M.; Gaber, M.M.J. Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network. *Appl. Intell.* **2021**, *51*, 854–864. [CrossRef]
7. Wang, L.; Lin, Z.Q.; Wong, A. Covid-net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest x-ray images. *Sci. Rep.* **2020**, *10*, 1–12.
8. Gozes, O.; Frid-Adar, M.; Greenspan, H.; Browning, P.D.; Zhang, H.; Ji, W.; Bernheim, A.; Siegel, E. Rapid AI development cycle for the coronavirus (COVID-19) pandemic: Initial results for automated detection & patient monitoring using deep learning CT image analysis. *arXiv preprint* **2020**, arXiv:2003.05037.
9. Wang, S.; Kang, B.; Ma, J.; Zeng, X.; Xiao, M.; Guo, J.; Cai, M.; Yang, J.; Li, Y.; Meng, X.; et al. A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19). *Eur. Radiol.* **2021**, *31*, 1–9. [CrossRef]
10. Khumoyun, A.; Cui, Y.; Hanku, L. Spark based distributed deep learning framework for big data applications. In Proceedings of the 2016 International Conference on Information Science and Communications Technologies (ICISCT), Tashkent, Uzbekistan, 2–4 November 2016; pp. 1–5.
11. Abdullah, A.; Awan, M.; Shehzad, M.; Ashraf, M. Fake News Classification Bimodal using Convolutional Neural Network and Long Short-Term Memory. *Int. J. Emerg. Technol. Learn.* **2020**, *11*, 209–212.
12. Awan, M.J.; Rahim, M.S.M.; Nobanee, H.; Munawar, A.; Yasin, A.; Zain, A.M. Social Media and Stock Market Prediction: A Big Data Approach. *Comput. Mater. Contin.* **2021**, *67*, 2569–2583. [CrossRef]
13. Ahmed, H.M.; Awan, M.J.; Khan, N.S.; Yasin, A.; Shehzad, H.M.F. Sentiment Analysis of Online Food Reviews using Big Data Analytics. *Ilkogr. Online* **2021**, *20*, 827–836.
14. Awan, M.J.; Rahim, M.S.M.; Nobanee, H.; Yasin, A.; Khalaf, O.I.; Ishfaq, U. A big data approach to black friday sales. *Intell. Autom. Soft Comput.* **2021**, *27*, 785–797. [CrossRef]
15. Chen, M.; Mao, S.; Liu, Y. Big data: A survey. *Mob. Netw. Appl.* **2014**, *19*, 171–209. [CrossRef]
16. Shoro, A.G.; Soomro, T.R. Big data analysis: Apache spark perspective. *Glob. J. Comput. Sci. Technol.* **2015**, *15*.
17. Salloum, S.; Dautov, R.; Chen, X.; Peng, P.X.; Huang, J.Z. Big data analytics on Apache Spark. *Int. J. Data Sci. Anal.* **2016**, *1*, 145–164. [CrossRef]
18. Burghard, C. Big data and analytics key to accountable care success. *IDC Health Insights* **2012**, *1*, 1–9.
19. Archenaa, J.; Anita, E.M. Interactive big data management in healthcare using Spark. In Proceedings of the 3rd International Symposium on Big Data and Cloud Computing Challenges (ISBCC–16′), Chennai, India, 10–11 March 2016; pp. 265–272.
20. Rehma, A.A.; Awan, M.J.; Butt, I. Comparison and Evaluation of Information Retrieval Models. *VFAST Trans. Softw.* **2018**, *6*, 7–14.
21. Alam, T.M.; Awan, M.J. Domain analysis of information extraction techniques. *Int. J. Multidiscip. Sci. Eng.* **2018**, *9*, 1–9.
22. Satyanarayana, L.V. A Survey on challenges and advantages in big data. *Int. J. Comput. Sci. Technol.* **2015**, *6*, 115–119.
23. Raghupathi, W.; Raghupathi, V. Big data analytics in healthcare: Promise and potential. *Health Inf. Sci. Syst.* **2014**, *2*, 1–10. [CrossRef]
24. Wang, K.; Khan, M.M.H. Performance prediction for apache spark platform. In Proceedings of the 2015 IEEE 17th International Conference on High Performance Computing and Communications, 2015 IEEE 7th International Symposium on Cyberspace Safety and Security, 2015 IEEE 12th International Conference on Embedded Software and Systems, New York, NY, USA, 24–26 August 2015; pp. 166–173.
25. Frampton, M. *Mastering Apache Spark*; Packt Publishing Ltd.: Birmingham, UK, 2015.
26. Zaharia, M.; Chowdhury, M.; Das, T.; Dave, A.; Ma, J.; McCauly, M.; Franklin, M.J.; Shenker, S.; Stoica, I. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In Proceedings of the 9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12), San Jose, CA, USA, 25–27 April 2012; pp. 15–28.
27. Marcu, O.-C.; Costan, A.; Antoniu, G.; Pérez-Hernández, M.S. Spark versus flink: Understanding performance in big data analytics frameworks. In Proceedings of the 2016 IEEE International Conference on Cluster Computing (CLUSTER), Taipei, Taiwan, 12–16 September 2016; pp. 433–442.
28. Bengio, Y.; Courville, A.; Vincent, P. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828. [CrossRef]
29. Lee, J.-G.; Jun, S.; Cho, Y.-W.; Lee, H.; Kim, G.B.; Seo, J.B.; Kim, N. Deep Learning in Medical Imaging: General Overview. *Korean J. Radiol.* **2017**, *18*, 570–584. [CrossRef] [PubMed]
30. Benbrahim, H.; Hachimi, H.; Amine, A. Deep transfer learning with apache spark to detect COVID-19 in chest x-ray images. *Rom. J. Inf. Sci. Technol.* **2020**, *23*, S117–S129.
31. Aftab, M.O.; Awan, M.J.; Khalid, S.; Javed, R.; Shabir, H. Executing Spark BigDL for Leukemia Detection from Microscopic Images using Transfer Learning. In Proceedings of the 2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA), Riyadh, Saudi Arabia, 6–7 April 2021; pp. 216–220.
32. Zhang, Y.; Wallace, B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv* **2015**, arXiv:1510.03820.
33. Ouchicha, C.; Ammor, O.; Meknassi, M. CVDNet: A novel deep learning architecture for detection of coronavirus (Covid-19) from chest x-ray images. *Chaos Solitons Fractals* **2020**, *140*, 110245. [CrossRef] [PubMed]

34. Hidaka, A.; Kurita, T. Consecutive dimensionality reduction by canonical correlation analysis for visualization of convolutional neural networks. In Proceedings of the 48th ISCIE International Symposium on Stochastic Systems Theory and its Applications, Fukuoka, Japan, 4–5 November 2016; pp. 160–167.
35. Yamashita, R.; Nishio, M.; Do, R.K.G.; Togashi, K. Convolutional neural networks: An overview and application in radiology. *Insights Imaging* **2018**, *9*, 611–629. [CrossRef]
36. Talo, M.; Baloglu, U.B.; Yıldırım, Ö.; Acharya, U.R. Application of deep transfer learning for automated brain abnormality classification using MR images. *Cogn. Syst. Res.* **2019**, *54*, 176–188. [CrossRef]
37. Awan, M.J.; Rahim, M.M.S.; Salim, N.; Mohammed, M.A.; Garcia-Zapirain, B.; Abdulkareem, K.H. Efficient Detection of Knee Anterior Cruciate Ligament from Magnetic Resonance Imaging Using Deep Learning Approach. *Diagnostics* **2021**, *11*, 105. [CrossRef] [PubMed]
38. Awan, M.J.; Raza, A.; Yasin, A.; Shehzad, H.M.F.; Butt, I. The Customized Convolutional Neural Network of Face Emotion Expression Classification. *Ann. Rom. Soc. Cell Biol.* **2021**, *25*, 5296–5304.
39. Mujahid, A.; Awan, M.J.; Yasin, A.; Mohammed, M.A.; Damaševičius, R.; Maskeliūnas, R.; Abdulkareem, K.H. Real-Time Hand Gesture Recognition Based on Deep Learning YOLOv3 Model. *Appl. Sci.* **2021**, *11*, 4164. [CrossRef]
40. Anam, M.; Ponnusamy, V.A.; Hussain, M.; Nadeem, M.W.; Javed, M.; Goh, H.G.; Qadeer, S. Osteoporosis Prediction for Trabecular Bone using Machine Learning: A Review. *Comput. Mater. Contin.* **2021**, *67*, 89–105. [CrossRef]
41. Hussain, E.; Hasan, M.; Rahman, A.; Lee, I.; Tamanna, T.; Parvez, M.Z. CoroDet: A deep learning based classification for COVID-19 detection using chest X-ray images. *Chaos Solitons Fractals* **2021**, *142*, 110495. [CrossRef] [PubMed]
42. Turkoglu, M. COVID-19 Detection System Using Chest CT Images and Multiple Kernels-Extreme Learning Machine Based on Deep Neural Network. *IRBM* **2021**, *42*, 207–214. [CrossRef]
43. Elkorany, A.S.; Elsharkawy, Z.F. COVIDetection-Net: A tailored COVID-19 detection from chest radiography images using deep learning. *Optik* **2021**, *231*, 166405. [CrossRef] [PubMed]
44. Das, A.K.; Kalam, S.; Kumar, C.; Sinha, D. TLCoV- An automated Covid-19 screening model using transfer Learning from chest X-ray images. *Chaos Solitons Fractals* **2021**, *144*, 110713. [CrossRef] [PubMed]
45. Zhou, T.; Lu, H.; Yang, Z.; Qiu, S.; Huo, B.; Dong, Y. The ensemble deep learning model for novel COVID-19 on CT images. *Appl. Soft Comput.* **2021**, *98*, 106885. [CrossRef] [PubMed]
46. Hassantabar, S.; Ahmadi, M.; Sharifi, A. Diagnosis and detection of infected tissue of COVID-19 patients based on lung x-ray image using convolutional neural network approaches. *Chaos Solitons Fractals* **2020**, *140*, 110170. [CrossRef]
47. Panwar, H.; Gupta, P.; Siddiqui, M.K.; Morales-Menendez, R.; Singh, V. Application of deep learning for fast detection of COVID-19 in X-Rays using nCOVnet. *Chaos Solitons Fractals* **2020**, *138*, 109944. [CrossRef]
48. Gupta, M.; Jain, R.; Arora, S.; Gupta, A.; Awan, M.J.; Chaudhary, G.; Nobanee, H. AI-enabled COVID-19 outbreak analysis and prediction: Indian States vs. Union Territories. *Comput. Mater. Contin.* **2021**, *67*, 933–950. [CrossRef]
49. Ali, Y.; Farooq, A.; Alam, T.M.; Farooq, M.S.; Awan, M.J.; Baig, T.I. Detection of schistosomiasis factors using association rule mining. *IEEE Access* **2019**, *7*, 186108–186114. [CrossRef]
50. Javed, R.; Saba, T.; Humdullah, S.; Jamail, N.S.M.; Awan, M.J. An Efficient Pattern Recognition Based Method for Drug-Drug Interaction Diagnosis. In Proceedings of the 2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA), Riyadh, Saudi Arabia, 6–7 April 2021; pp. 221–226.
51. Nagi, A.T.; Awan, M.J.; Javed, R.; Ayesha, N. A Comparison of Two-Stage Classifier Algorithm with Ensemble Techniques on Detection of Diabetic Retinopathy. In Proceedings of the 2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA), Riyadh, Saudi Arabia, 6–7 April 2021; pp. 212–215.
52. Awan, M.J.; Yasin, A.; Nobanee, H.; Ali, A.A.; Shahzad, Z.; Nabeel, M.; Zain, A.M.; Shahzad, H.M.F. Fake News Data Exploration and Analytics. *Electronics* **2021**, *10*, 2326. [CrossRef]
53. Kaggle. COVID-19 Chest X-ray Challenge. 2021. Available online: https://www.kaggle.com/bachrr/covid-chest-xray (accessed on 15 January 2020).
54. Kaggle. Chest X-ray Images (Pneumonia). 2021. Available online: https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia (accessed on 15 January 2018).
55. Khalil, A.; Awan, M.J.; Yasin, A.; Singh, V.P.; Shehzad, H.M.F. Flight Web Searches Analytics through Big Data. *Int. J. Comput. Appl. Technol.* **2021**, in press.
56. Awan, M.; Khan, R.; Nobanee, H.; Yasin, A.; Anwar, S.; Naseem, U.; Singh, V. A Recommendation Engine for Predicting Movie Ratings Using a Big Data Approach. *Electronics* **2021**, *10*, 1215. [CrossRef]
57. Awan, M.J.; Khan, M.A.; Ansari, Z.K.; Yasin, A.; Shehzad, H.M.F. Fake Profile Recognition using Big Data Analytics in Social Media Platforms. *Int. J. Comput. Appl. Technol.* **2021**, in press.
58. Dreiseitl, S.; Ohno-Machado, L. Logistic regression and artificial neural network classification models: A methodology review. *J. Biomed. Inform.* **2002**, *35*, 352–359. [CrossRef]
59. Wang, S.-H.; Muhammad, K.; Hong, J.; Sangaiah, A.K.; Zhang, Y.-D. Alcoholism identification via convolutional neural network based on parametric ReLU, dropout, and batch normalization. *Neural Comput. Appl.* **2020**, *32*, 665–680. [CrossRef]
60. 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.

61. Mahdianpari, M.; Salehi, B.; Rezaee, M.; Mohammadimanesh, F.; Zhang, Y. Very deep convolutional neural networks for complex land cover mapping using multispectral remote sensing imagery. *Remote Sens.* **2018**, *10*, 1119. [CrossRef]
62. Chu, G.; Potetz, B.; Wang, W.; Howard, A.; Song, Y.; Brucher, F.; Leung, T.; Adam, H. Geo-aware networks for fine-grained recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Korea, 27 October–2 November 2019; pp. 247–254.
63. Pan, X.; Luo, P.; Shi, J.; Tang, X. Two at once: Enhancing learning and generalization capacities via ibn-net. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 464–479.
64. Tahir, H.; Iftikhar, A.; Mumraiz, M. Forecasting COVID-19 via Registration Slips of Patients using ResNet-101 and Performance Analysis and Comparison of Prediction for COVID-19 using Faster R-CNN, Mask R-CNN, and ResNet-50. In Proceedings of the 2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), Bhilai, India, 19–20 February 2021; pp. 1–6.
65. Wang, M.; Gong, X. Metastatic cancer image binary classification based on resnet model. In Proceedings of the 2020 IEEE 20th International Conference on Communication Technology (ICCT), Nanning, China, 28–31 October 2020; pp. 1356–1359.
66. Mateen, M.; Wen, J.; Song, S.; Huang, Z. Fundus image classification using VGG-19 architecture with PCA and SVD. *Symmetry* **2019**, *11*, 1. [CrossRef]
67. Ilijason, R. Getting Started with Databricks. In *Beginning Apache Spark Using Azure Databricks*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 27–38.
68. Awan, M.; Rahim, M.; Salim, N.; Ismail, A.; Shabbir, H.J. Acceleration of knee MRI cancellous bone classification on Google colaboratory using convolutional neural network. *Int. J. Adv. Trends Comput. Sci.* **2019**, *8*, 83–88. [CrossRef]