ROMANIAN JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY
Volume **23**, Number S, pp. S117–S129, 2020

Deep Transfer Learning with Apache Spark to Detect COVID-19 in Chest X-ray Images

Houssam BENBRAHIM, Hanaâ HACHIMI, and Aouatif AMINE

BOSS-Team, GS-Laboratory, National School of Applied Sciences,
Ibn Tofail University, Kenitra, Morocco
E-mails: houssam.benbrahim@uit.ac.ma, hanaa.hachimi@univ-ibntofail.ac.ma,
aouatif.amine@uit.ac.ma

Abstract. A chest X-ray test is one of the most important and recurrent medical imaging examinations. It is the first imaging technique that represents a significant role in the diagnosis of SARS-CoV-2 disease. Automatic classification of 2019-nCoV using X-ray images is a major request that can help doctors to make the best decisions. In this paper, we adopted, developed, and validated a Deep Transfer Learning (DTL) method using Convolutional Neural Network (CNN) based models InceptionV3 and ResNet50 with Apache Spark framework for the classification of COVID-19 in chest X-ray images collected from Kaggle repository. High accuracy was obtained by our model in the detection of COVID-19 X-ray images 99.01% by the pre-trained InceptionV3 model and 98.03% for the ResNet50 model.

Keywords: COVID-19, SARS-CoV-2, 2019-nCoV, chest X-ray images, deep transfer learning, convolutional neural network, CNN, apache spark, InceptionV3, ResNet50.

1. Introduction

On March 11, 2020, the World Health Organization (WHO) has declared SARS-CoV-2, recently named COVID-19 [1], as a global pandemic. The WHO on May 4, 2020, pointing to the over 3,442,200 confirmed cases globally, leading to at least 239,700 deaths [2]. Morocco is among the countries affected by this pandemic. Moreover, since the first case was detected on March 2, 2020, Morocco has marked on May 4, 2020, 5,053 confirmed cases among them, 1,653 recovered and 179 death, as well as 40,249 cases have been excluded after a negative analysis in the laboratory [3]. The percentage of recovered persons is 32.71% compared to the percentage of deaths represents 3.54%. Morocco has only carried out a 45,302 analysis test until now. This means that the Ministry of Health has done about 13 tests for every 10,000 people (knowing that the population of Morocco is currently 35,895,089 on May 4, 2020 [4]).

Given the epidemiological situation in Morocco, the specificity of the tests, and the recommendations of the WHO, only the following laboratories are authorized to carry out the examinations in March 2020: The Institute of Hygiene in Rabat, the Pasteur Institute of Morocco in Casablanca, and The laboratory of the Mohammed V Military Training Hospital in Rabat. The Ministry of Health revealed in April 2020, that laboratory analyzes are also available in six university hospitals in the cities of Rabat, Casablanca, Marrakech, Oujda, Fez, and Agadir [3]. Morocco uses the PCR technique or the blood test (the presence or absence of antibodies) to detect COVID-19 in suspected cases. With its two methods, the results take about 5 to 24 hours based on where the hospital is located in which the samples were taken. But these results can be accelerated depending on both the symptoms of pneumonia and chest X-ray tests.

Chest X-ray is the first imaging method that plays an important role in the examination of SARS-CoV-2 disease. Several studies have been proved that the use of CNN for the detection of COVID-19 in chest X-ray or Computed Tomography (CT) images, gives relevant results and it should be surrounded by more attention and priority [5, 6, 7, 8].

Using Big Data technologies with deep learning can overcome many challenges and help to achieve encouraging results. The combination between the two technologies Apache Spark and Transfer Learning allows data analysts to classify images with very hay efficiency and they permit to create models that can be running on clusters [9].

The general objective of our paper is to develop a system based on Deep Transfer Learning (DTL) using Convolutional Neural Networks (CNN) based on pre-trained models InceptionV3 and ResNet50 with the framework of big data Apache Spark. This architecture will be able to detect COVID-19 in chest X-ray images obtained from two databases of Kaggle repository. This model will be an intelligent platform dedicated to helping doctors make better decisions about coronavirus and to encourage specialists in the field of medical imaging in Morocco to use its advanced techniques as a diagnostic tool.

2. Related Work

2.1. Big Data

Big data is a blanket term for the non-traditional strategies and technologies needed to collect, organize, and process insights from large datasets [10]. There are multiple definitions of Big Data, it is sometimes complicated to approve on a single definition, each theme focuses on a particular aspect of this concept. For example, SAS Company, the world leader in business analytics software defines Big Data as [11]: "Big data is a term that describes the large volume of data – both structured and unstructured – that inundates a business on a day-to-day basis. But it’s not the amount of data that’s important. Its what organizations do with the data that matters. Big data can be analyzed for insights that lead to better decisions and strategic business moves".

These massive data are characterized by 10 Vs: Volume (a vast amount of data), Velocity (speed of data), Variety (multiple types and forms of data), Variability (data in change), Veracity (data accuracy), Validity (level of quality, governance), Vulnerability (new security concerns), Volatility (the time required for storage and validity of data), Visualization (data view mode), and Value (potential of big data to create a change) [12, 13].

The amount of data is growing significantly over the past few years, therefore, the need for data analytics frameworks is growing. Different platforms for the treatment of Big Data have been built for different purposes. The most famous platform for Big Data analytics is the Apache Spark framework [14]. Spark has the potential to process information with various structures. It is very fast, it supports several programming languages, it combines the functionality of machine learning, and it integrates with several platforms [15].

Big Data in healthcare is evolving to provide insight from terribly massive knowledge sets, and it producing important results whereas reducing prices as well as ensuring very good quality of care for patients [16]. Apache Spark can play a very important role in the analysis of healthcare, it can help doctors for better diagnosis diseases especially for image classification [17].

2.2. Big Data Analytics

Big data analytics refers to the action of analyzing huge volumes of structured and unstructured data, from different sources, and in different sizes, using advanced analytic techniques. Big data Analysis permit to researchers, business and analysts to obtain better and faster decisions. They can use advanced analytics techniques such as machine learning, deep learning, predictive analytics, etc., in the context of the big data, to gain new ideas, new approaches, and new insights [18]. The aim is to discover patterns and connections that can be invisible, and that furnish precious insights about the users who created it. Big Data analytics allows the health care sector to produce information from data to create decisions, improve diagnoses and treatments, develop the quality of care at a lower cost, and to present better results [19].

2.3. Apache Spark

Apache Spark is an open-source data-processing framework characterized by speed, ease of use, and sophisticated analytics. Spark runs in-memory, on clusters, it isn't tied to Hadoop's MapReduce two-stage paradigm, and it has the lightning-fast performance [20]. Spark can run as a standalone, on a top of Hadoop YARN, Mesos, or on the Cloud, where it can read data directly from HDFS, Cassandra, Hbase, Hive, and Tachyon. In addition to its in-memory processing, graph processing, and machine learning, Spark can also handle streaming [21]. Spark uses the Resilient Distributed Datasets (RDDs) to store data in memory (RAM). Users can read data from disk and write in the RDDs to create a job with the iterative operation as well as they can execute several queries on the same subset of data continuously by keeping it in-memory with interactive mode [22]. Spark batch processing offers incredible speed advantages, trading off high memory usage. Spark Streaming is a good stream processing solution for workloads that value throughput over latency [23]. We can summarize the characteristics of the Apache Spark in Table 1.

Table 1. Apache Spark characteristics

Spark characteristics		
Design	Source	UC Berkeley
	Implementation (core)	Scala
	Current version	2.4.5
	Development	Complete project Apache Software foundation (ASF).
	Use by the company	Amazon,Baidu, eBay Inc., NASA JPL,Yahoo, etc.
Languages	Languages support	Java, Scala, Python, and R.
	Library	Machine Learning (Mllib), SQL (Spark SQL), Graph and Parellel Graph (GraphX), and Streaming (Spark streaming).
	API	Spark Scala API, Spark Java API, Spark Python API, Spark R API, and Spark SQL.
Execution	Mode of execution	Standalone, Hadoop, Mesos, and Kubernetes.
	Processing model	Micro-Batch
	Management	Sparkcontext
	Processing	Executors
Model	Transformation	Transformation into peer collections Key/value.
	Pradigm	RDD
	Iterations	Interactive programs are divided into several Indepent jobs.
Security	Spark RPC	Spark supports authentication for a Communication protocol between Spark processes channels using a shared secret.
	Local Storage Encryption	Spark uses the encryption of temporary data written to local disks.
	Web UI	Allowing authentication for the Web UIs is done using javax servlet filters.

Data	Access data	HDFS, Cassandra, Hbase, Hive, Tachyon, and Hadoop.
	Data streaming	HDFS, Flume, Kafka, etc.
Streaming	Streaming	Streaming processing over the batch system.
	Streaming latency	High
	Streaming throughput	High

2.4. Deep Learning

Deep learning or deep structured learning represents a class of advanced machine learning techniques. It relies on algorithms using mathematical operations based essentially on Artificial Neural Networks [24]. This method exploits many hidden layers for extracting and transforming features. Each layer takes as input the output of the previous one, they accepting data to be processed and they intended to deliver the result of the calculation [25]. With deep learning techniques, the machine becomes capable of learning without needing to be programmed. Deep Learning was applied to multiple problems for example in automatic speech recognition, image recognition, natural language processing, drug discovery and toxicology, customer relationship management, recommendation systems, and bioinformatics [26].

2.5. Convolutional Neural Network for Image Classification

Convolutional Neural Network (CNN) named ConvNet is a type of artificial neural network in which the connection between neurons is inspired by the visual cortex of animals. [27]. It is a Deep Learning algorithm that can be able to classify input images, chest X-ray images for example, as either infected by COVID-19 or not [5]. ConvNet can itself extracts the features, which makes the classification faster, and more accurate. There are four main layers in the CNN: The Convolution layer, Non-Linearity (ReLU) layer, Pooling or Sub Sampling layer, and the Classification (Fully Connected Layer) [28]. The first one is used to extract features from the input image. The second layer is an element-wise operation, it is integrated to replace all negative pixel values in the feature map by zero. The third one, the Pooling layer, is applied to decrease the dimensionality of each feature map but to conserve the most significant information. Finally, the last one means that every neuron in the precedent layer is attached to every neuron on the next layer [29].

2.6. Transfer Learning

Transfer Learning (TL) is a technique that enables the knowledge acquired to be transferred to a “source” dataset to better process a new called “target” dataset. It is a deep learning and machine learning method that permits to resolve the basic problem of insufficient training data [30]. TL can be explained where a model created for a task is reused as the beginning step for a model on a second task. Image recognition is one of the main areas of application for TL [31]. We can summarize the transfer learning concept of building and training machine learning models in Fig. 1.

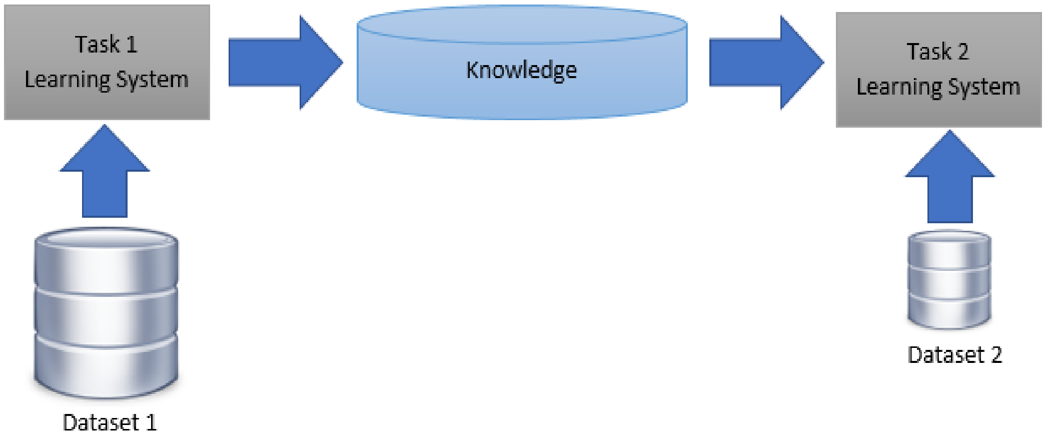


Fig. 1. The Transfer Learning approach

In transfer learning, we can use the knowledge of already trained models to build new models and even solve problems such as having less data for the new task, which in contrast to traditional learning that is isolated and takes place only based on specific tasks.

3. Material and methods

In this paper, we have employed Deep Learning Pipelines on Apache Spark to permit fast transfer learning. We have used the pre-trained CNN InceptionV3 and ResNet50 architectures and logistic regression to classify the chest X-ray images. Pipelines is an open-source project for deep learning. It is a high-level framework that supports deep learning workflows via the Apache Spark. Deep learning Pipelines library is included in Databricks Runtime ML [32]. Logistic regression is a statistical method employed on machine learning to analyze independent features that define an outcome, in our case two kinds of images (COVID-19 and Normal) [33]. InceptionV3 and ResNet50 models with weights are two pre-trained architecture on ImageNet. Inception-v3 [34] network model is the third edition of Google's Inception CNN. It is a deep neural network based on the TensorFlow using for image analysis and object detection. ResNet50 [35] is a Residual Network that is 50 layers deep. It’s a subclass of CNNs, with ResNet most popularly used in the field of image recognition and classification.

3.1. COVID-19 X-ray Images dataset

In this study, we used a chest radiograph or chest X-ray (CXR) images and we selected two datasets. We have worked with 320 images in total, classed in two categorize: 160 images for patients affected by COVID-19 and 160 Normal images. These images have been obtained from Kaggle repository exactly from two datasets respectively called "COVID-19 chest xray" [36] and "Chest X-Ray Images (Pneumonia)" [37]. Our objective is to create a dataset with chest X-ray images as a frontal projection that assembled 160 normal and 160 COVID-19 patients. In Fig. 2 and Fig. 3, there are chest X-ray images of COVID-19 and normal patients, respectively.

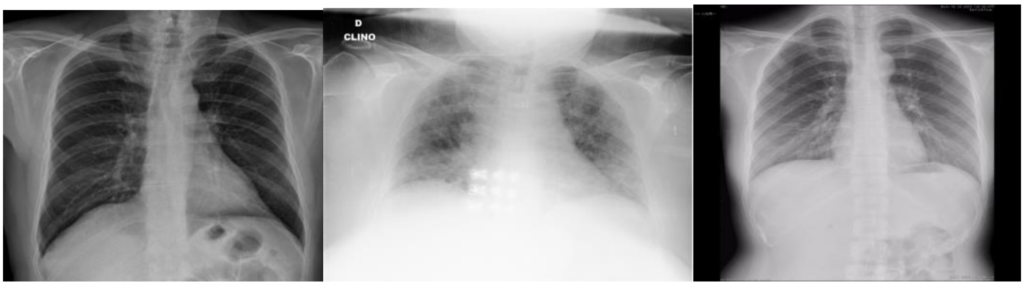


Fig. 2. COVID-19 chest X-ray images

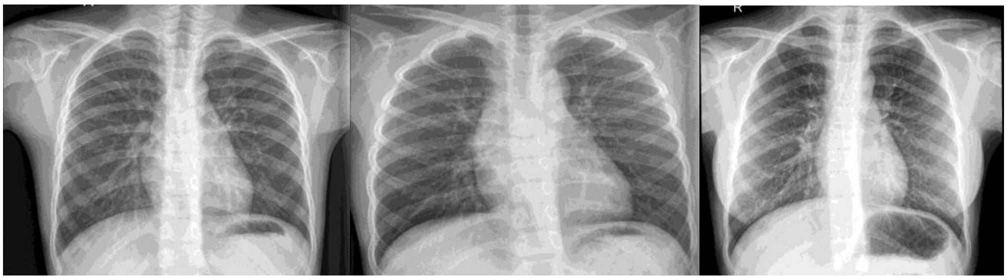


Fig. 3. Normal chest X-ray images

3.2. Apparatus

For our experience, we used the Databricks Workspace [38]. It is an Apache Spark-based analytics platform. Databricks is a collaborative environment that permits users to implement all of their analytical processes in a single space and to manage machine learning models throughout their life cycle. In this platform, we created a cluster to execute our model as a set of commands. Table 2 gives a detailed description of our cluster.

Table 2. The general description of the cluster

Cluster Name	Databricks Runtime Version	Instance	Spark Environment Variables
Deep Learning withe Apache Spark	Runtime: 6.4 (Scala 2.11, Spark 2.4.5).	1 Driver: 15.3 GB Memory, 2 Cores, 1 DBU	PYSPARK version supports only Python 3

We have stored the chest X-ray images in Databricks File System (DBFS). It is a distributed file system that allows us to store the data for queries inside Databricks and it accessible on Databricks clusters. We have created two paths: dbfs/ml/database/covid19 and dbfs/ml/database/normal to store respectively, the chest X-ray images affected by COVID-19 and the normal images.

4. Results

In this study, two different convolutional neural network-based models (InceptionV3 and ResNet50) with Apache Spark trained and tested on chest X-ray images to detect the coronavirus pneumonia infected patients. For this reason, we used the Deep transfer Learning technique with a combination of Deep Learning Pipelines on Apache Spark (using Databricks workspace) and logistic regression. The general process of our work is summarized in Fig. 4.

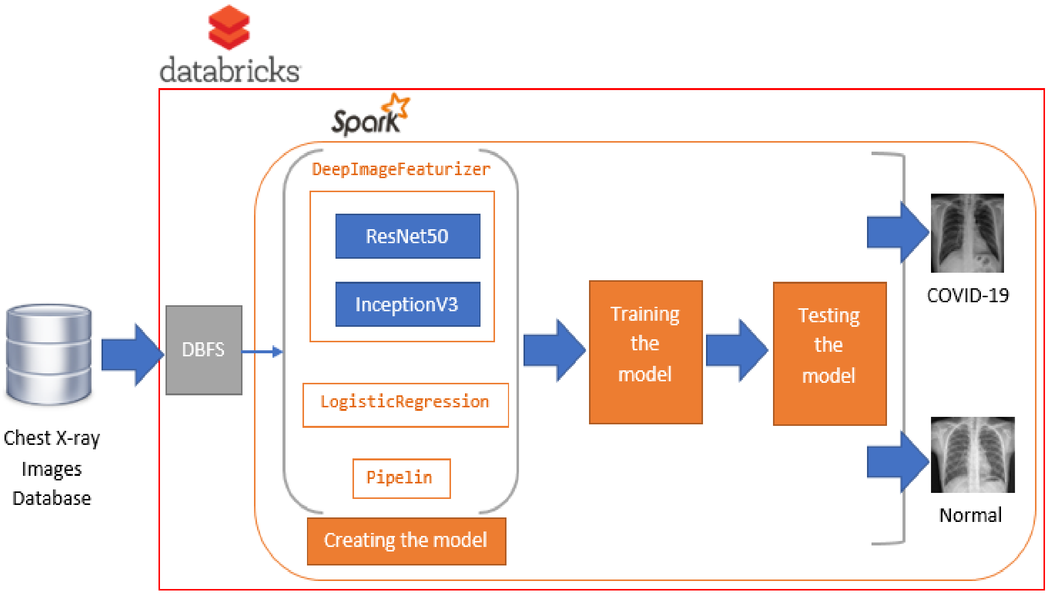
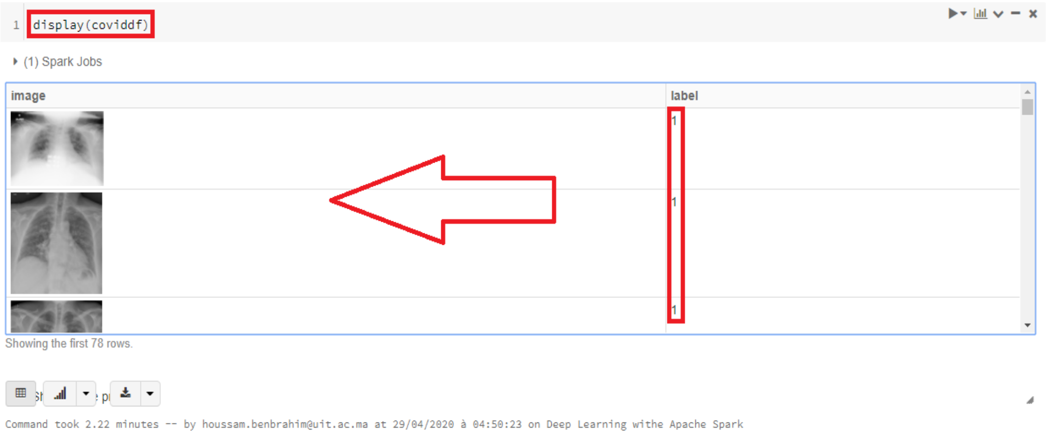


Fig. 4. Deep transfer learning withe Apache Spark architecture of our work model

Before the implementation of our model, we loaded the images. For this reason, we used the Deep Learning Pipelines to load images into a Spark DataFrame, and we assigned the values "1" and "0" as labels for images affected by COVID-19 and normal images, respectively. Fig. 5 and Fig. 6 illustrate this operation.



```
display(coviddf)
```

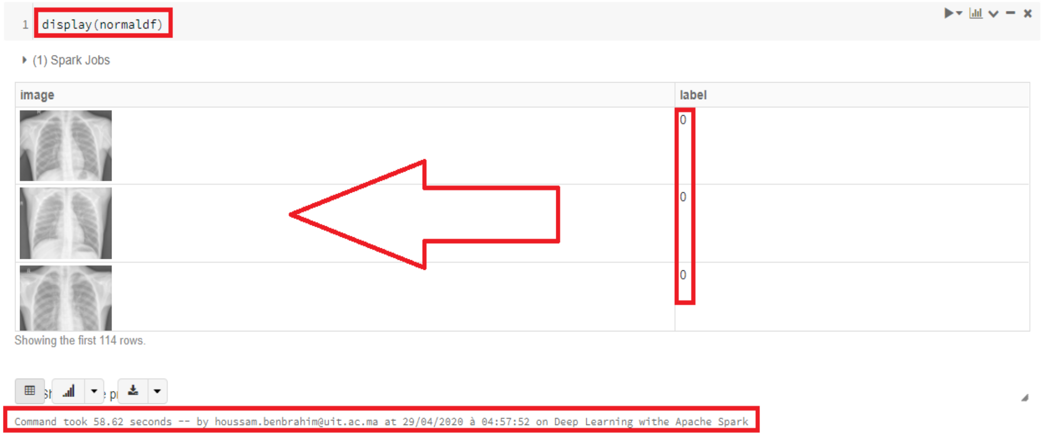
▸ (1) Spark Jobs

image	label
	1
	1
	1

Showing the first 78 rows.

```
Command took 2.22 minutes -- by houssam.benbrahim@uit.ac.ma at 29/04/2020 à 04:50:23 on Deep Learning withe Apache Spark
```

Fig. 5. COVID-19 images with label 1



```
display(normaldf)
```

▸ (1) Spark Jobs

image	label
	0
	0
	0

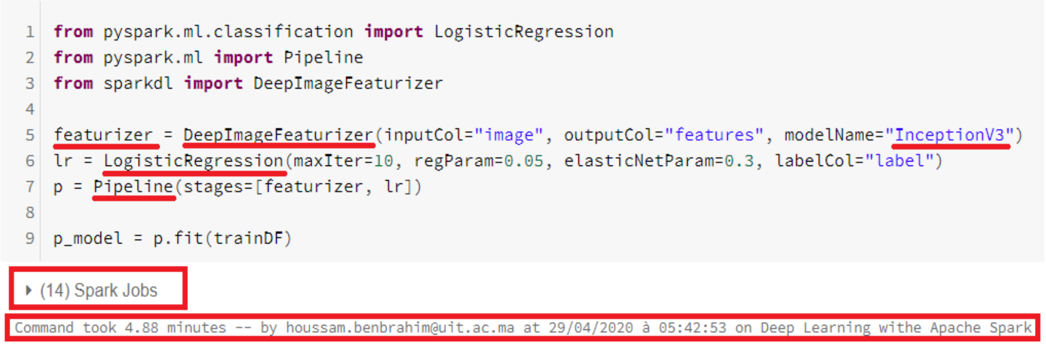
Showing the first 114 rows.

```
Command took 58.62 seconds -- by houssam.benbrahim@uit.ac.ma at 29/04/2020 à 04:57:52 on Deep Learning withe Apache Spark
```

Fig. 6. Normal images with label 0

The dataset used was split into two independent datasets with 70% and 30% for training and testing respectively. The reason to opt for this decision is that this partition helps our model to be able to learn the general principles in the training phase, this means that our model sees more examples and, therefore, finds a better solution. And if we don't have enough test data, our final assessment of the generalizability of the model may not be exact.

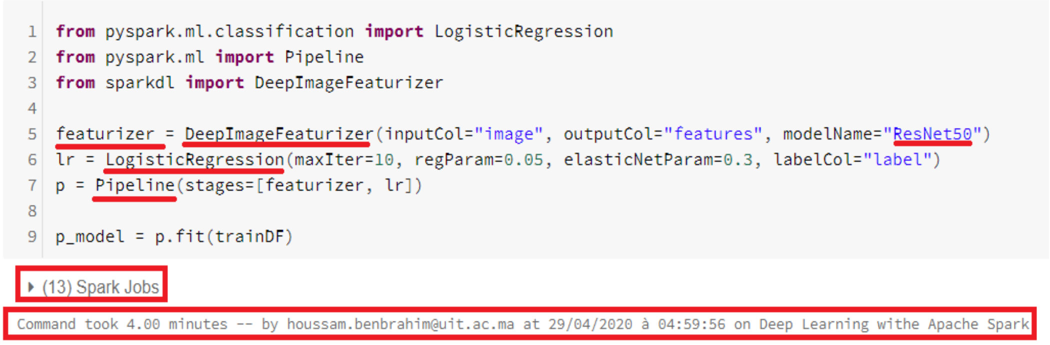
Deep Learning Pipelines allow fast transfer learning on Apache Spark cluster with the notion of a Featurizer. For our experiment, we have worked on the InceptionV3 and ResNet50 models and we have used a DeepImageFeaturizer. This last is considered an important function of Deep Learning Pipelines, which automatically peels off the last layer of a pre-formed CNN model by introducing the output of all previous layers as features for the logistic regression algorithm. Fig. 7 and Fig. 8 illustrate an extract from the program which exploits the techniques of DeepImageFeaturizer, Logistic regression, and Pipeline.



```
from pyspark.ml.classification import LogisticRegression
from pyspark.ml import Pipeline
from sparkdl import DeepImageFeaturizer

featurizer = DeepImageFeaturizer(inputCol="image", outputCol="features", modelName="InceptionV3")
lr = LogisticRegression(maxIter=10, regParam=0.05, elasticNetParam=0.3, labelCol="label")
p = Pipeline(stages=[featurizer, lr])

p_model = p.fit(trainDF)
```

```
(14) Spark Jobs
Command took 4.88 minutes -- by houssam.benbrahim@uit.ac.ma at 29/04/2020 à 05:42:53 on Deep Learning withe Apache Spark
```

Fig. 7. An extract from the program for InceptionV3

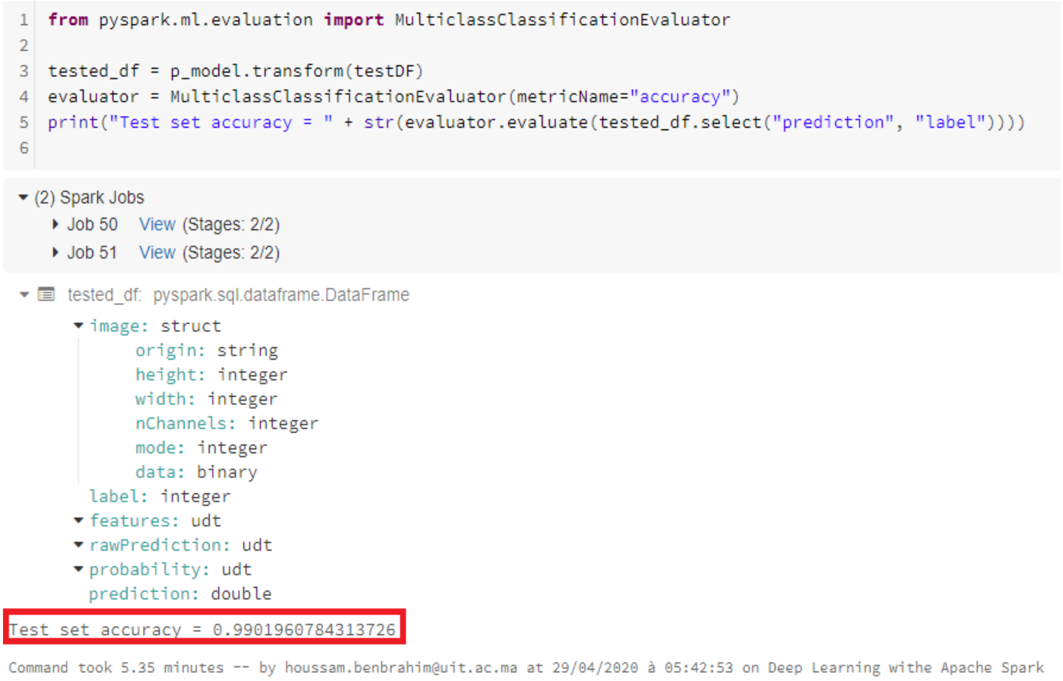


```
from pyspark.ml.classification import LogisticRegression
from pyspark.ml import Pipeline
from sparkdl import DeepImageFeaturizer

featurizer = DeepImageFeaturizer(inputCol="image", outputCol="features", modelName="ResNet50")
lr = LogisticRegression(maxIter=10, regParam=0.05, elasticNetParam=0.3, labelCol="label")
p = Pipeline(stages=[featurizer, lr])

p_model = p.fit(trainDF)
```

```
(13) Spark Jobs
Command took 4.00 minutes -- by houssam.benbrahim@uit.ac.ma at 29/04/2020 à 04:59:56 on Deep Learning withe Apache Spark
```

Fig. 8. An extract from the program for ResNet50

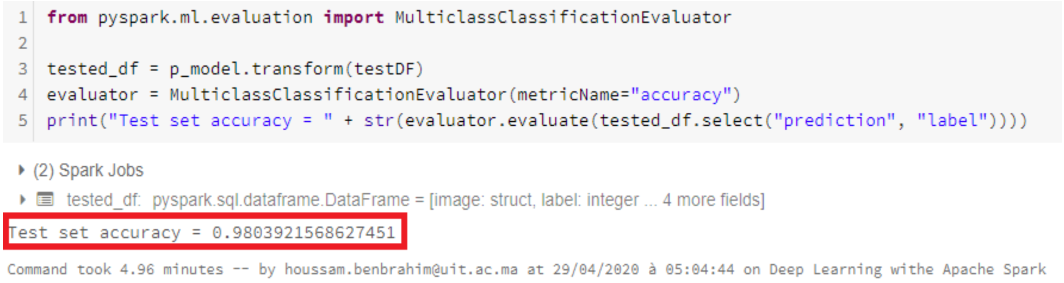
In this study, we have used four criteria to test the performance of our model, which are: Accuracy, F1-Score, weighted Precision, and weighted Recall. The first performance index tested is accuracy. It refers to the proximity of the measurements to a determined value. With InceptionV3 our model achieves the accuracy of 99.01% which is recorded in Fig. 9, on the other hand with ResNet50 our architecture achieves the accuracy of 98.03% which is displayed in Fig. 10.



```python
from pyspark.ml.evaluation import MulticlassClassificationEvaluator

tested_df = p_model.transform(testDF)
evaluator = MulticlassClassificationEvaluator(metricName="accuracy")
print("Test set accuracy = " + str(evaluator.evaluate(tested_df.select("prediction", "label"))))

```

```
▾ (2) Spark Jobs
    ▸ Job 50   View (Stages: 2/2)
    ▸ Job 51   View (Stages: 2/2)
▾ ▤ tested_df:  pyspark.sql.dataframe.DataFrame
      ▾ image: struct
            origin: string
            height: integer
            width: integer
            nChannels: integer
            mode: integer
            data: binary
        label: integer
      ▾ features: udt
      ▾ rawPrediction: udt
      ▾ probability: udt
        prediction: double
Test set accuracy = 0.9901960784313726
Command took 5.35 minutes -- by houssam.benbrahim@uit.ac.ma at 29/04/2020 à 05:42:53 on Deep Learning withe Apache Spark
```

Fig. 9. Training accuracy for InceptionV3



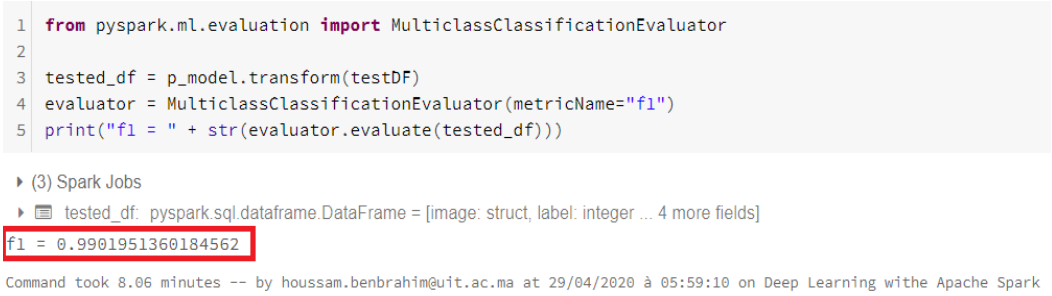
```python
from pyspark.ml.evaluation import MulticlassClassificationEvaluator

tested_df = p_model.transform(testDF)
evaluator = MulticlassClassificationEvaluator(metricName="accuracy")
print("Test set accuracy = " + str(evaluator.evaluate(tested_df.select("prediction", "label"))))
```

```
▸ (2) Spark Jobs
▸ ▤ tested_df:  pyspark.sql.dataframe.DataFrame = [image: struct, label: integer ... 4 more fields]
Test set accuracy = 0.9803921568627451
Command took 4.96 minutes -- by houssam.benbrahim@uit.ac.ma at 29/04/2020 à 05:04:44 on Deep Learning withe Apache Spark
```

Fig. 10. Training accuracy for ResNet50

Even if the highest training accuracy is obtained with the InceptionV3 model, we can notice that our model gives important values whatever the pre-trained CNN model used.

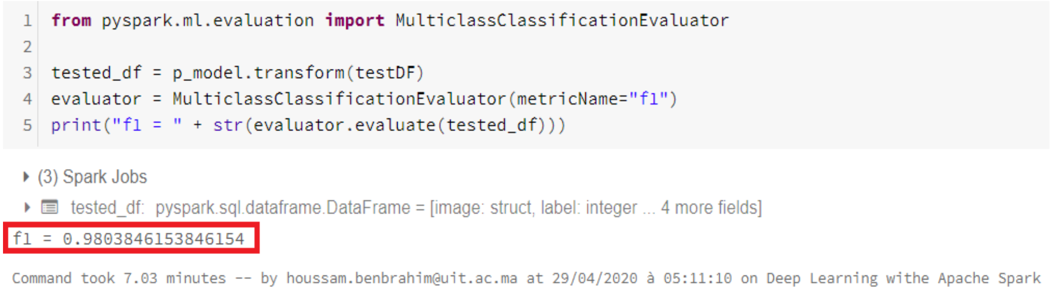
The second index tested is the F1-Score, it is a measure of a test's accuracy. Our model with InceptionV3 attains a value of 0.9901 and the ResNet50 attains 0.9803. Fig. 11 and Fig. 12 respectively represent the results of F1-Score for the model proposed with InceptionV3 and ResNet50.



```python
from pyspark.ml.evaluation import MulticlassClassificationEvaluator

tested_df = p_model.transform(testDF)
evaluator = MulticlassClassificationEvaluator(metricName="f1")
print("f1 = " + str(evaluator.evaluate(tested_df)))
```

```
▸ (3) Spark Jobs
▸ ▤ tested_df:  pyspark.sql.dataframe.DataFrame = [image: struct, label: integer ... 4 more fields]
f1 = 0.9901951360184562
Command took 8.06 minutes -- by houssam.benbrahim@uit.ac.ma at 29/04/2020 à 05:59:10 on Deep Learning withe Apache Spark
```

Fig. 11. F1-Score for InceptionV3



```
from pyspark.ml.evaluation import MulticlassClassificationEvaluator

tested_df = p_model.transform(testDF)
evaluator = MulticlassClassificationEvaluator(metricName="f1")
print("f1 = " + str(evaluator.evaluate(tested_df)))
```

```
▸ (3) Spark Jobs
▸ tested_df:  pyspark.sql.dataframe.DataFrame = [image: struct, label: integer ... 4 more fields]
f1 = 0.9803846153846154

Command took 7.03 minutes -- by houssam.benbrahim@uit.ac.ma at 29/04/2020 à 05:11:10 on Deep Learning withe Apache Spark
```

Fig. 12. F1-Score for ResNet50

The last two indices tested are weighted Precision and weighted Recall. The first one affords the weighted mean of precision with weights equal to class probability and the second one gives the weighted mean of recall. ResNet50 attains 98.03% (weighted Precision) and 98.11% (weighted Recall), such as InceptionV3 achieves 99.01% (weighted Precision) and 99.03% (weighted Recall). We can summarize the results of its last two indexes in Fig. 13.

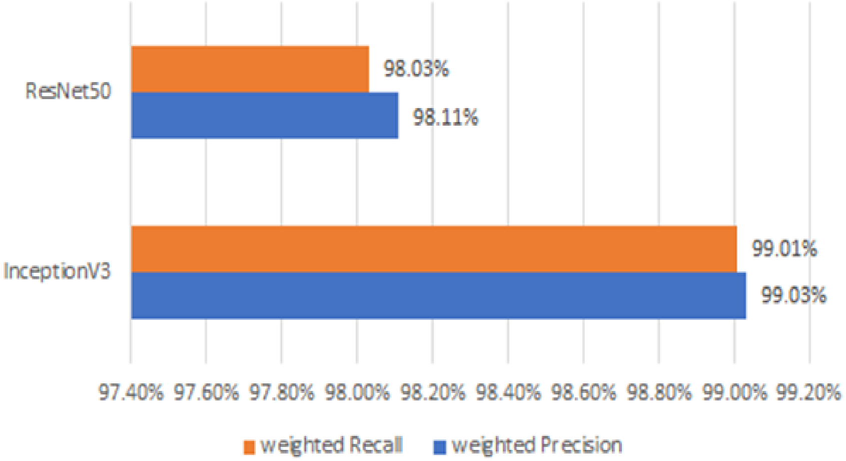


Fig. 13. Weighted Recall and weighted Precision for ResNet50 and InceptionV3

Our model showed very high values for the four indices tested. Even if we used two pre-trained models of CNN (ResNet50 and InceptionV3) the results were very important and especially for InceptionV3. The combination of the deep transfer learning method and the Apache Spark framework has shown great performance, efficient analysis, and advanced results. This fusion between its two techniques has proven that our model able to easily detecting people with COVID-19 and without COVID-19 in X-ray images.

5. Discussion

In this paper, a method founded on deep transfer learning using CNN based InceptionV3 and ResNet50 with Apache Spark was proposed for the detection of COVID-19 in chest X-ray images. The experimental results proved that our model achieved a very high accuracy of 99.01% for InceptionV3 and 98.03 for the ResNet50 model. We also tested other performance measurements as F1-Score, weighted Precision, and weighted Recall. All the latest indices have validated very high values ranging from 98% up to 99%.

In a similar study [5], the authors have created a classification of COVID-19 in chest X-ray images using DeTraC deep CNN. Their experimental results showed that the model obtained an accuracy of 95.12%. In another work [6], the authors created the

COVID-Net model (a deep convolutional neural network) for COVID19 detection. It achieved a 92.6% test accuracy. In [39], the authors developed an automatic detection of COVID-19 using a raw chest X-ray. Their proposed model was created to furnish accurate diagnostics for binary classification (COVID vs. No-Findings) and multi-class classification (COVID vs. No-Findings vs. Pneumonia). Their proposed Darknet-19 model produced a classification accuracy of 98.08% for binary classes and 87.02% for multi-class cases images. In [40], a new COVIDX-Net framework has been proposed to automatically identify COVID-19 in X-ray images based on seven deep learning classifiers: VGG19, DenseNet201, ResNetV2, InceptionV3, InceptionResNetV2, Xception, and MobileNetV2. Their proposed COVIDX-Net achieved a good and similar performance score for the VGG19 and DenseNet201models with f1-scores of 0.89 and 0.91for normal and COVID-19, respectively. In [41], the authors developed an automatic detection of COVID-19 using X-ray images and Deep CNNs, the pre-trained ResNet50 provides 98% of the accuracy, 97% accuracy for InceptionV3, and 87% accuracy for Inception-ResNetV2. In another paper [42], the authors suggested a classification model based on ResNet50 plus SVM to detect COVID-19. Their model reported 95.38% accuracy. Through the foregoing, we can remark that:

- None of the compared works has used the Apache Spark framework as an analysis tool for the detection of COVID-19 in chest X-ray images.
- Most of the articles cited have worked with CNN pre-trained models.
- The overall size of the database of all the works indicated above was varied between 50 and 13800 chest X-ray images.
- There is a remarkable difference in the performance of the models proposed in the cited articles and our suggested model.

After all that, we can see that the combination of deep transfer learning with Apache Spark using advanced techniques as pre-trained CNN InceptionV3 and ResNet50 models, Pipelines, and logistic regression algorithm gave important results for detecting COVID-19 in chest X-ray images. Apache Spark proved an advanced performance in image classification techniques. This framework offers a fast, efficient, and intelligent environment for data analysis.

6. Conclusion

SARS-CoV-2 is a very dangerous virus, as soon as it arrived in Morocco, several measures have been taken by the state to limit its spread, but it has reached progressive figures, neither in terms of the number of persons affected nor in terms of the number of deaths.

Through this work, we have developed an architecture on Deep Transfer Learning with Apache Spark to detect COVID-19 in chest X-ray Images. In this study, we have used two convolutional neural network-based models (InceptionV3 and ResNet50), as well as the image classification process for our model was founded on deep learning Pipelines and logistic regression. Our model was implemented and tested on two databases drawn from the Kaggle repository. We have elaborated our experience in Databricks workspace, in which we have stored the images as well we exploited the advanced performance of Apache Spark the big data analysis framework. In this work, we tested four performance indices for deep transfer learning: Accuracy, F1-Score, weighted Precision, and weighted Recall. The results obtained were very high for both InceptionV3 and ResNet50. The values marked for InceptionV3 exceed 99%, and for ResNet50 the values mentioned going up to 98%. With its performance obtained by our model, we can claim that our architecture is capable to detect Covid-19 disease in chest X-ray images with high accuracy.

These results can convince the health specialists in Morocco to use these advanced techniques against 2019-nCoV disease. Our model will simplify, speed up, and reduce the costs of coronavirus detection. Our future work aims to attach TensorFlow in Apache Spark and creates a novel model to detect COVID-19 in chest CT images. The combination of its last two technologies makes the work environment more interesting and we can build a model running on large computing clusters.

Acknowledgements. The authors are grateful to the “anonymous” reviewers for evaluation of the manuscript.

Funding Statement. The authors received no specific funding for this study.

Conflicts of Interest. The authors declare that they have no conflicts of interest to report regarding the present study.

Authors ’contributions. Conceived the model: HB. Further processed and analyzed the data: HB. Wrote the paper: HB. Corrected and agreed with the paper: HH, AA.

References

[1] CHAN J. F. W., YUAN S, KOK K. *et al*, A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *The Lancet*, 2020, vol. 395, no 10223, pp. 514-523.

[2] World Health Organization, Coronavirus disease (COVID-19), https://covid19.who.int/, accessed May 4, 2020.

[3] Ministry of Health of Morocco**,** The official portal of Coronavirus in Morocco**,** http://www.covidmaroc.ma/Pages/AccueilAR.aspx, accessed May 4, 2020.

[4] High Commission for Planning of Morocco, https://www.hcp.ma/Demographie-population_r142.html, accessed May 4, 2020.

[5] ABBAS A., ABDELSAMEA M. M., GABER, M. M., Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network. *arXiv preprint arXiv:2003.13815*, 2020.

[6] WANG L., WONG A., COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest radiography images. *arXiv preprint arXiv:2003.09871*, 2020.

[7] GOZES O., FRID-ADAR M., GREENSPAN H. *et al*, 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 arXiv:2003.05037*, 2020.

[8] WANG S., KANG B., MA J. *et al*, A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19). *MedRxiv*, 2020.

[9] KHUMOYUN A., CUI Y., HANKU L., Spark based distributed deep learning framework for big data applications. In: *2016 International Conference on Information Science and Communications Technologies (ICISCT)*. IEEE, 2016. pp. 1-5.

[10] CHEN M., MAO S., LIU, Y., Big data: A survey. *Mobile networks and applications*, 2014, vol. 19, no 2, pp. 171-209.

[11] SAS Company, Big Data What it is and why it matters, https://www.sas.com/en_us/insights/big-data/what-is-big-data.html, accessed 16 April 2020.

[12] FIRICAN G., The 10 Vs of big data, https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx, 2017.

[13] SUWINSKI P., ONG C., LING M. H., POH Y. M., KHAN A. M., ONG H. S., Advancing personalized medicine through the application of whole exome sequencing and big data analytics. *Frontiers in genetics*, 2019, vol. 10, pp. 49.

[14] SHORO A. G., SOOMRO T. R., Big data analysis: Apache spark perspective. *Global Journal of Computer Science and Technology*, 2015.

[15] SALLOUM S., DAUTOV R., CHEN X., PENG P. X., HUANG J. Z., Big data analytics on Apache Spark. *International Journal of Data Science and Analytics*, 2016, vol. 1, no 3-4, pp. 145-164.

[16] BURGHARD, C. Big data and analytics key to accountable care success. IDC health insights, 2012, pp. 1-9.

[17] 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’)*. Springer, Cham, 2016. pp. 265-272.

[18] SATYANARAYANA L. V., A Survey on challenges and advantages in big data. *IJCST*, 2015, vol. 6, no 2, pp. 115-119.

[19] RAGHUPATHI W., RAGHUPATHI V., Big data analytics in healthcare: promise and potential. *Health information science and systems*, 2014, vol. 2, no 1, pp. 3.

[20] WANG K., KHAN M. M. H., Performance prediction for apache spark platform. In: *2015 IEEE 17th International Conference on High Performance Computing and Communications, 2015 IEEE 7th International Symposium on Cyberspace Safety and Security, and 2015 IEEE 12th International Conference on Embedded Software and Systems*. IEEE, 2015. pp. 166-173.

[21] FRAMPTON M., Mastering apache spark. Packt Publishing Ltd, 2015.

[22] ZAHARIA M., CHOWDHURY M., DAS T. *et al.*, Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In: *Presented as part of the 9th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 12)*. 2012. pp. 15-28.

[23] MARCU O. C., COSTAN A., ANTONIU G., PÉREZ-HERNÁNDEZ, M. S., Spark versus flink: Understanding performance in big data analytics frameworks. In: *2016 IEEE International Conference on Cluster Computing (CLUSTER)*. IEEE, 2016. pp. 433-442.

[24] BENGIO Y., COURVILLE A., VINCENT P., Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 2013, vol. 35, no 8, pp. 1798-1828.

[25] 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 journal of radiology*, 2017, vol. 18, no 4, pp. 570-584.

[26] Hordri, N. F., Yuhaniz, S. S., & Shamsuddin, S. M., Deep learning and its applications: a review. In: *Conference on Postgraduate Annual Research on Informatics Seminar*, 2016.

[27] KIM, Y., Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*. 2014.

[28] HIDAKA, A., KURITA, T. Consecutive dimensionality reduction by canonical correlation analysis for visualization of convolutional neural networks. In: *Proceedings of the ISCIE International Symposium on Stochastic Systems Theory and its Applications*. The ISCIE Symposium on Stochastic Systems Theory and Its Applications, 2017, pp. 160-167.

[29] YAMASHITA R., NISHIO M., DO, R. K. G., TOGASHI, K., Convolutional neural networks: an overview and application in radiology. *Insights into imaging*, 2018, vol. 9, no 4, pp. 611-629.

[30] TAN C., SUN F., KONG T., ZHANG, W., YANG, C., LIU, C., A survey on deep transfer learning. In: *International conference on artificial neural networks*. Springer, Cham, 2018. pp. 270-279.

[31] ALO M., BALOGLU U. B., YILDIRIM Ö., ACHARYA, U. R., Application of deep transfer learning for automated brain abnormality classification using MR images. *Cognitive Systems Research*, 2019, vol. 54, pp. 176-188.

[32] EL-AMIR H., HAMDY M., Deep Learning Pipeline.2020.

[33] DREISEITL S., OHNO-MACHADO L., Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics*, 2002, vol. 35, no 5-6, pp. 352-359.

[34] SZEGEDY C., VANHOUCKE V., IOFFE S., SHLENS J., WOJNA, Z., Rethinking the inception architecture for computer vision. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016. pp. 2818-2826.

[35] HE K., ZHANG X., REN S., SUN, J., Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016. pp. 770-778.

[36] Kaggle**,** COVID-19 chest xray, https://www.kaggle.com/bachrr/covid-chest-xray, accessed April 13, 2020.

[37] Kaggle, Chest X-Ray Images (pneumonia), https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia, accessed April 16, 2020.

[38] ETAATI, L., Azure Databricks. In: *Machine Learning with Microsoft Technologies*. Apress, Berkeley, CA, 2019, pp. 159-171.

[39] OZTURK T., TALO M., YILDIRIM E. A., BALOGLU U. B., YILDIRIM O., ACHARYA U. R, Automated detection of COVID-19 cases using deep neural networks with X-ray images. *Computers in Biology and Medicine*, 2020, p. 103792.

[40] HEMDAN E. E. D., SHOUMAN M. A., KARAR M. E, Covidx-net: A framework of deep learning classifiers to diagnose covid-19 in x-ray images. *arXiv preprint arXiv:2003.11055*, 2020.

[41] NARIN A., KAYA C., PAMUK Z., Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks. *arXiv preprint arXiv:2003.10849*, 2020.

[42] SETHY P. K., BEHERA S. K, Detection of coronavirus disease (covid-19) based on deep features. *Preprints*, 2020, vol. 2020030300, p. 2020.