International Journal of Applied Engineering Research ISSN 0973-4562 Volume 15, Number 8 (2020) pp. 873-883
© Research India Publications. http://www.ripublication.com

Deep learning techniques to diagnose COVID-19 big data based on features extracted from CT and x-ray images

Majd Alhawamdeh[1], Ghaith Jaradat[2], Hamzeh Alhawamdeh[3]

[1]Department of Computer Science, Faculty of Computer Science and Information Technology, Jerash University, P.O.Box: 311-26150, Jerash, Jordan

[2]Department of Computer Science, Faculty of Computer Science and Informatics, Amman Arab University, P.O.Box: 2234-11953, Amman, Jordan

[3]Department of management information, Faculty of Business, Jerash University, P.O.Box: 311-26150, Jerash, Jordan

Abstract

This study aimed at testing the role of Deep learning techniques on predicting COVID-19 big data. The study adopted two tasks to measure Deep learning (classification, clustering), while the big data was measured through three dimensions: volume, variety and velocity. To achieve study aims, the researcher relied on measuring accuracy and parameters settings of classification and clustering techniques, and measuring the features of the covid-19 dataset. First, by presenting questions that reflects the dimensionality of the dataset and the features of the two techniques. Second, by analyzing the outcomes of the artificial neural network and K-means to answer those questions. Also, the results of both techniques, artificial neural network and K-means, proved to be suitable to classify instances into two categories of negative and positive covid-19 cases and some features of both techniques are of no significant impact on accuracy, and the classification has a greatest impact on accuracy contributed to a number of features.

Keywords: Deep learning, big data, covid-19, ANN, K-means.

1. INTRODUCTION

Data is information formed in special formats to be stored, modified, or analyzed. However, when it becomes difficult to store this data because it is large in size or difficult to modify due to its variety and complexity or the difficulty of analyzing it due to the speed of its production, these data are called "big data". Big data is one of the biggest achievements of the age in the field of information technology, because it represents a great challenge for users, companies, and researchers, as traditional tools cannot deal with this type of data. In this article, we will highlight the term "big data", and show the characteristics of the data that gives this name, and we will explain its importance and challenges facing the people who deal with it and the methods used to deal with its problems, and the most important of these methods "in-deep learning". In-deep learning is the type of machine learning that uses a set of sophisticated algorithms to extract highly abstract data from raw data. This is done by building a hierarchy of these algorithms to arrange these data, classify them, and extract useful ones from them, to be stored, modified, or analyzed. Big data contains pieces of information, which is a useful part of this data, and deep learning is the most promising solution for extracting these useful portions of the massive content. Therefore, in-depth learning is a very important tool for making analysis of "big data" easier, by providing the analysis tools with abstract data representation. Finally, the analysis of big data with the help of in-depth learning makes a significant contribution to areas of development and innovation in various sectors, such as health care, banking, education, etc., which makes in-depth learning one of the most sought-after areas of research in the world.

1.1 Research Problem:

The amount of data generated by internet and other technologies are increasingly huge. Chains such as medical records (e.g. Covid-19) are of Terabytes upon Terabytes of data. As the amount of techniques being used in data collection keeps expanding, the amount of observations that could be used as training data also keeps increasing. The problem becomes harder and bigger when each of these data points contains multiple features or attributes. Storing this data results in generating huge datasets in size and dimensions.

When dealing with big data, it is advised by [1] to consider the volume of the dataset, dataset dimensionality, model complexity, and constraints. These settings present the main problems hindering an accurate predictive model from being built properly.

In most cases, deep learning deals with unlabeled big data. In such conditions, it requires clustering techniques to group similar features, where feature selection and extraction takes place in medical images datasets classification. However, the focus of the authors is on classifying the data extracted from the huge collection of Covid-19 images prepared in datasets form by conventional techniques such as ANN and K-means. Hence, this raises the need for examining the accuracy of those techniques and how can be improved in the future in terms of

accuracy in the aspect of big data and specifically the volume dimension.

1.2 Research objectives

1. To apply deep learning techniques (including classification and clustering) for a large amount of medical image data (including Covid-19 datasets) using Weka.
2. To tackle multiple Covid-19 data sources with different data formats from Kaggle and others.

2. LITERATURE REVIEW:

2.1 Big data

Big data is a key concept that cannot be ignored in the world of information technology, due to the prominent increase in data and data-related services, so it is important to explore this field and find ways to improve the provision of data service, especially through cloud computing. Whereas, cloud computing helps address the problem of storage and data service. As big data processing becomes a less expensive task. Big data requires large amounts of data storage, processing, and exchange. This is what traditional platforms cannot do, such as data analysis or data warehouses, or expand easily and without normal costs to meet big data requirements [2].

Big data**:** It the process of flooding digital data which includes texts, sounds, videos, images, and combinations of each and its collected from many resources such as e-mails, social networks, Internet, digitizers, sensors, numerical modeling, scanners, mobile phones, and videos [3]. Big data has 5Vs and it refers to volume of data, Velocity refers to the fast generation and transmission of data across the Internet, Variety refers to the diverse data, Veracity refers to the diversity of quality, Value This refers to the ability to transform a tsunami of data into business [4]. Skourletopoulos *et al*. [5] indicate the big data it's a collection of technologies and technologies that require new forms of integration to reveal large hidden values from the large, complex, and massive scale large data sets. Big data is a term used to refer to an increase in the volume of data that is difficult to store, process and analyze with traditional database technologies. The nature of big data is unclear and includes major processes for identifying and translating data into new visions [6]. Also, IDC (International Classification of Diseases) defined Big Data Technologies as "a new generation of technologies and architectures, designed to extract the most important value from very large amounts of a variety of data, by enabling high speed capture, detection and / or analysis."

During the follow up of traditional data life cycle challenges, digital Big Data poses other technological challenges according to its 5V features which are described by [4] [7] [8] [9]. As follows:

- Data storage: Refers to the size, speed, and variety of big data. Big data storage over traditional physical storage is a problem because it often fails, and traditional data protection mechanisms are not effective with storage. While cloud storage services offer almost unlimited storage with high error tolerance it provides potential solutions to meet the challenges of big data storage.
- Data transmission: The data transfer process continues in different stages of the data life cycle as follows: (1) collecting data from sensors to storage; (2) data integration from multiple data centers; (3) Data management to transfer integrated data to processing platforms (such as cloud platforms) and (4) Data analysis to transfer data from storage to host analysis (such as high-performance computing groups). Transferring large amounts of data presents clear challenges in each of these stages. Therefore, smart pre-processing technologies and data compression algorithms are needed to effectively reduce data. Size before transferring data. Additionally, when transferring large data to cloud platforms from local data centers, how to develop efficient algorithms to automatically recommend appropriate cloud service (location) based on spatial temporal principles to maximize data transfer speed while reducing cost is also difficult.
- Data management: manage, analysis, and visualization of large, disorganized, and efficiently heterogeneous data are difficult things for computers to do. Creating metadata for geospatial data is also difficult due to the intrinsic properties of data with high dimensions (3D space and 1D time) and complexity (for example the correlation between spacetime and dependency). Besides creating metadata, Big Data also poses challenges to database management systems (DBMSs) because of traditional RDBMSs lack the scalability to manage and store large, unstructured data. While non-relational databases (NoSQL) such as MongoDB and HBase are designed for big data, the process of developing ways to customize databases to deal with large geospatial data by developing effective spatial indexing and querying algorithms is still a difficult problem.
- Data integration: Data integration is critical for achieving the 5th V (value) of Big Data through integrative data analysis and cross-domain collaborations. the data integration challenges of schema mapping, record linkage, and data fusion. Metadata is essential for tracking these mappings to make the integrated data sources 'robotically' resolvable and to facilitate large-scale analyses. However efficiently and automatically creating metadata from Big Data is still a challenging task. In the geospatial domain, geo-data integration has sparked new opportunities driven by an ever increasingly collaborative research environment.

In this research we study the three defining properties or dimensions of big data 3Vs which are (volume, variety and velocity); they are discussed as follows.

2.1.1 Volume: The name "Big Data" is related to the big data size. As the size of big data is an enormous amount of data. When determining the value of the data, the size of the data plays a very important role. Where the volume of data is huge, it is actually considered "big data". Consequently, the size of the data determines whether it is possible to consider that this data is large or not, and the classification of big data on this

basis. If we see big data as a pyramid, the Volume of this data is the base for this pyramid [10]. The volume of the data refers to the size of the datasets that must be analyzed and processed, which are now often larger than terabytes and petabytes. The sheer volume of data requires different processing technologies that are different from traditional storage and processing capabilities. In other words, this means that the data sets in Big Data are too large to be processed with a regular laptop or desktop processor [11].

Volume is the bigdata component and is used to determine the amount of big data that is stored and managed by an organization. It assesses the vast amount of data in data stores and concerns about their manageability, accessibility, and scalability [12]. Kuo and Kusiak [13] indicate the volume of big data as the most critical component of big data 5 V's framework, where it determines the capacity of the data infrastructure to store, manage, and deliver data to users and applications. The volume focuses on planning current and future storage capacity - especially as it relates to speed - but also reaping the optimum benefits from efficient use of existing storage infrastructure.

2.1.2 Variety: Refers to the nature of structured and semi-structured and non-structured data. It also indicates heterogeneous sources. Diversity is essentially data access from new sources inside and outside the organization. It can be structured, semi-structured, and non-structured [12]. Whereas Structured Data: These data are basically structured data. Generally, refers to data that has defined the length and formatting of data. And semi-structured data: These data are basically semi-structured data. It is generally a form of data that does not conform to the official data structure. Log files are examples of this type of data [11]. As for non-structured data: This data refers mainly to non-structured data. Generally, refers to data that does not accurately fit into the traditional relational database row and column structure. Text, images, videos, etc. are examples of unstructured data that cannot be stored in rows and columns [14]. The company can get data from many different sources: from internal devices to GPS for smartphones or what people say on social networks. The importance of these information sources varies according to the nature of the work. This data can contain many layers, with different values [13]. Mehta and Pandit [15] indicated the Variety of data is the diversity in the extracted data, which helps users, whether they are researchers or analysts, to choose the appropriate data for their research field and includes structured data in databases and unstructured data such as pictures, clips, audio, and video recordings, SMS, call logs and data GPS maps require time and effort to be properly adapted for processing and analysis.

2.1.3 Velocity: Companies need the information to flow quickly and as soon as possible. So much so, that researchers have indicated that Velocity can be more important than size because it can give organizations a greater competitive advantage. Sometimes it is better to have limited data and a higher Velocity than to have a lot of data and a low Velocity [13]. Velocity: means the speed at which data is produced and extracted to cover the demand for it, as speed is considered A critical component in making decisions based on these data, is the time we spend from The moment this data reaches the moment of the decision based on it [14]. Gupta *et al.* [11] indicated that the data must be available at the appropriate time to make the appropriate business decisions. Velocity refers to the speed of response in the data request and the speed of this large data flow from sources as there is a huge and continuous flow of data such as devices, networks, social media, mobile phones, and so on [15] Using cloud-based sensors and manufacturing, a large amount of data is generated in the manufacture and development of new products and services. Big data enables faster production and services by analyzing them better; Thus, there is more data generation with more flow or velocity. velocity features are categorized (1) Pace of data creation, generation, (2) data processing time, (3) time of information broadcasting [12].

2.2 Deep learning: The principle of deep learning is based on extracting data traits, using a hierarchical learning model. The structure of deep learning consisting of a number of layers of learning with non-linear transformations extracts more abstract attributes from other less abstract attributes from them, and as a result, these abstract attributes represent the sources of variance in the data. Therefore, they are the best structured representation of data [16]. Deep learning also uses big data that has not been subjected to any prior human coordination and this makes them a solution to reduce direct human intervention in the mentoring and education process. As a result, deep learning outcomes are a valuable resource for different applications for example, other learning algorithms, classification and indexing [17]. Berman *et al.* [18] indicate the Machine learning is the most leading solution in "big data analysis", and it is a set of artificial intelligence algorithms, which rely on data analysis and previous patterns, for future decisions. But the nature and structure of data has a major impact on the performance of machine learning algorithms, while simple algorithms will deal well with structured data, unstructured data will reduce the performance of other algorithms of great complexity. Hence, "extracting features" - the process of converting raw data into an appropriate representation based on field requirements - is a difficult task, because it is the most important part of machine learning algorithms, which is the part that entails human intervention. Thus, in-depth learning is one of the most promising solutions for automating the process of feature extraction and handling of disorganized data, using a hierarchical structure and simulation of the human brain. In-deep learning typically uses greedy learning algorithms to train network layers, using massive amounts of data without the need for human assistance. Hence, its network structure, which is based on non-linear hierarchical layers, improves learning outcomes in it. This appears in its applications for computer vision, speech recognition, and natural language processing [19].

Deep learning can be defined as: It's a method (part or field) of machine learning family that is based on learning and improve the computer to do what comes naturally to humans by examining computer algorithms based on artificial neural networks with representation learning. It can be supervised, semi-supervised, or unsupervised [20]. In deep learning methods, the classification tasks performed directly from images, text, or sound. And it's can achieve state-of-the-art accuracy, sometimes exceeding human-level performance [21]. Deep learning Models using a huge and large set of labeled data

and neural network architectures that contain many layers. also, its permitted to be heterogeneous and to deviate widely from biologically informed connectionist models, for the sake of efficiency, trainability and understandability, whence the "structured" part [22].

In this research we use the clustering and classification techniques to tackle the big data. We use the k-means algorithm in the cluster method and the artificial neural network in the classifying method. Where the Classification and clustering are two pattern methods used in identification process in the machine learning. Also, both methods have certain similarities, but the difference between them that classification uses predefined structure and classes in which objects are assigned, while clustering refers to the process of identifying similarities between objects, and integrate it into groups according to the characteristics. And These groups are called "clusters".

2.2.1 Clustering:

Clustering is a technique for organizing a group of data into categories and groups where the objects within a group have high similarity and the two group objects are similar to each other. Here the two clusters can be considered separate. The main goal of grouping is to divide the entire data into multiple clusters. Unlike the classification process, here category labels for things unknown before, and clustering pertains to unsupervised learning [23]. Banerjee & Arora [24] indicate in Clustering, the similarity between two objects is measured by the similarity function, as the distance between these two objects is measured. The shortest distance is the higher similarity, on the contrary, the longest distance is higher the difference. actually, when dealing with a huge amount of data we tend to summarize the huge amount of data to a small number of groups or categories, in order to facilitate the analysis process [25]. Clustering is the process of placing data in similar clusters, and it is a branch of data mining. The Cluster algorithm divides a data set into several clusters, where the similarity between the points within a given cluster is greater than the similarity between two points within two different clusters. The idea of Clustering data is too simple in nature and very close to the human and his thinking way [26]. Clustering algorithms are widely used not only to organize and classify data but to compress data and build the data structure model. Whereas if we can find clusters of data, then a problem model can be built on the basis of those clusters [27]. There are many algorithms used in the data collection process, and we will study the k-means clustering algorithm.

2.2.1.1 K-means algorithm

K-means are one of the most important clustering algorithms it's used to analyze the data by dividing it into clusters of *k* number. Each cluster contains similar elements [28]. K-means is classified as a descriptive algorithm that divides data based on its properties and belongs to unsupervised learning algorithms [29] which create inferences from datasets using only input vectors without referring to known, or labeled outcomes [30]. Xu & Lange [31] indicate that this algorithm is used to collect multiple data depending on its properties to *k* aggregate, and the aggregation process is done by reducing the distances between the data and the cluster centroid. According to Fuente-Tomas *et al.* [32], the K-means algorithm identifies *k* number of centroids and then allocates every data point to the nearest cluster while keeping the centroids as small as possible. The 'means' in the K-means refers to averaging of the data; that is, finding the centroid [33]. Tai *et al.* [34] indicate K-means algorithm works to process the learning data. The K-means algorithm in data mining starts with the first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids It halts creating and optimizing clusters when either: the centroids have stabilized — there is no change in their values because the clustering has been successful, or the defined number of iterations has been achieved. Sharma *et al.* [35] indicate the performance of this algorithm depends on the initial locations of the centers (Centroid), and it is recommended to implement this algorithm several times with different centers each time from the previous times. see for more details.

2.2.2 Classification

Classification is the process of distinguishing things from one another, arranging them, and dividing them according to their similarity into groups, where each category includes a group of common units with each other in certain characteristics or properties, and this concept has many uses in our lives [25]. Banerjee & Arora [24] indicate the classification it's a process in the data based on dividing the knowledge into different subjects while giving a specific code for each topic, by highlighting the topics in a way that helps to link them together so that the general data on that particular progress, and Considerate the relation of all data to the data that follows. There are many benefits to the classification process, most notably: saving time and effort, organizing the place, and making good use of the spaces, as classification and arrangement in matters related to data contribute to obtaining the required material, whether this material is a piece of information, a book, or existing materials In the supermarket, if it were not classified and arranged, it would be difficult for a person to obtain his various needs [23]. It is the classification process actually we grouping similar things together, or next to each other, i.e. arranging things based on their similarities and differences [26]. In addition, Reimers *et al.* [27] indicate the classification is the process of learning a model that shows different categories of data predetermined and its geared with supervised learning. It is a two-step process, consisting of the learning step and the classification step. In the learning step, a categorization model is created and the categorization step is used in the learning model to create class labels for specific data.

2.2.2.1 Artificial neural network

Neural network: It is a circle of neurons, consisting of artificial neurons or nodes. It is a mechanism for processing data by simulating the way the natural neural networks of a person (the human nervous system) are performed [37]. Baroni [38] indicate that a mathematical suggestion and theory that describes how normal human neurons work. By exchanging neural networks signals from one cell to another in the natural networks system (in the natural neural network). The neural network acquires knowledge with learning, and capture and

stores knowledge in neurons linked together by a neural network based on modified weights. Neural networks learn from experience and are used in particular to distinguish patterns and shapes, and this is what distinguishes neural networks from traditional calculation programs that simply implement instructions in a fixed sequential order while the floating inference and reasoning systems are useful for situations where they need to be a human experience which cannot It is translated into a set of equations (integrated into an automated decision-making process) [39]. Owoyele *et al.* [40] and Rahman & Muniyandi [41] indicate that, despite data problems and training difficulties, the network is able not only to make good expectations in the actual time series and in the temporary operations that do not appear in the training phase but also, to produce a series of status variables at earlier times. Moreover, how can the network conclude that there is a role for the status variables when the purpose of that data is unknown and the model's ability to deal with data is separate, so it is more likely to be used in a number of applications.

The method of artificial neural networks (ANN) is perfectly suited with the nonlinear models, because most of the existing systems, including economic systems, do not exist except in the context of dynamic or dynamic models, which makes them all time-changing and the data in them have non-linear relationships [42]. Arunkumar *et al.* [37] indicate that each neural network is arranged in layers of artificial cells: the inner layer and the outer layer and the layers between them or hidden between the income and outer layers. Each cell in one of these layers is connected to all the neurons in the next layer and all the neurons in the layer that precedes it. and the information is processed Through the connection between one neuron and another, where each connection is distinguished by its value it's called Weighting, which constitutes the importance of the link between these two elements. The neuron multiplies each value of the incoming value from the previous layer neurons by the weights of the connections with these neurons, then collects all multiplication products, then subject result to a conversion sequence that varies according to the neuron type, the output of the conversion function is the output of the neuron that is transferred to the neurons of the next layer [43].

3. METHODOLOGY

To measure the variables of the study and its dimensions, the researchers relied on measuring accuracy and parameters settings of classification and clustering techniques, and measuring the features of the covid-19 dataset. First, by presenting questions that reflects the dimensionality of the dataset and the features of the two techniques. Second, by analyzing the outcomes of the ANN and K-means to answer those questions.

3.1 Research questions:

This work addresses the following research questions

• What features/attributes of the Kaggle COVID-19 dataset that have the greatest impact on the classification task?

• What deep learning model has the highest prediction accuracy for the Kaggle COVID-19 dataset with respect to volume dimension using WEKA?

This comprises of testing a number of configurations for both ANN classification and K-means clustering models and then comparing their results.

We first clarify Kaggle's dataset, then the ANN and K-means techniques, and finally the WEKA machine learning workbench.

3.2 Kaggle's dataset

The Covid-19 datasets are abundant in Kaggle in many forms and class attributes. For the purpose of this work we focus on the dataset that handles the readings of x-ray and CT images. The dataset has either two class attributes or three distinguishing between Covid-19 cases and normal cases or other pneumonia cases. The dataset is split into train and test sets. Train set is used to train ANN and k-means models for prediction, while test set is used to feed those models with new data for accuracy evaluation. We have chosen this dataset solely for the purpose of examining the ANN and K-means models over Covid-19 images data in terms of volume dimension of such big data. This comes without the consideration of obtaining a competitive or better accuracy rate compared to other techniques involved in the well-known Kaggle's competition.

3.3 Classification using ANN

Classification is a supervised learning system that uses a labeled dataset representing predictions. It is used as a training set of input-output pairs to find a deterministic function that maps inputs to outputs, predicting future input-output observations while minimizing errors as much as possible [44]. ANNs are relatively raw electronic models that rely on the brain's neural structure. The brain learns from experience and artificial neural networks try to simulate the workings of the brain. Computers do things well but have trouble recognizing simple patterns. While the brain stores information as models. Some of these patterns are very complex and allow us to recognize individual faces from many different angles. The process of storing information in the form of patterns, and the use of those patterns, then solving problems, includes a new field in computing, which does not use traditional programming but involves the creation of vastly parallel networks and training those networks to solve specific problems which are called artificial neural networks so the artificial neural networks are one of the best technique for classifying the data in the data warehouse [23].

3.4 Clustering using K-means

Clustering is an Unsupervised learning system that uses unlabeled datasets to train the system in order to derive structure, from unlabeled data by investigating the similarity between pairs of objects and is usually associated with density estimation or data clustering [44]. K-means is an unsupervised machine learning, it's an eager learner. An eager learner has a model fitting that means a training step but a lazy learner does not have a training phase. The K-means algorithm classifies the Objects into a predetermined number of clusters and each cluster is symbolized by *k*. The selection of cluster centers in this algorithm is done randomly, and it is preferred that these

centers be as far away from each other as possible. The random starting point influences the effectiveness of the assembly process and the results. The complex approach process depends mainly on the values of the initial centers [36].

3.5 Weka

Weka is a free workbench software that contains a collection of visualization tools and algorithms for data analysis and prediction models. It comes with an easy to use graphical interface for all of its functionalities. This software will be used in this work to perform the classification and clustering tasks by implementing the ANN and K-means algorithms. Weka has a variety of parameter settings for each algorithm with many parameters tuning functions. This gives the researchers a rich environment of experiments with less time.

3.6 Study Model:

Figure (1) refers to the study variables, which are big data in its dimensions (volume, variety, velocity) and deep learning in its dimensions (clustering, classification) where it was designed by researchers to help in analyzing and solving the research problem.

4. RESULTS AND DISCUSSIONS

This section discusses experimental settings and experimental results. It also presents the training and testing of ANN and K-means and their obtained results. Datasets that are used in this work are also briefly discussed.

4.1 Experimental settings

This research conducted preliminary experiments to determine the suitable configuration of the ANN and K-means for a proper data analysis. At first, they are implemented with their default parameter settings. Then, their parameter values are carefully tuned to obtain the desired results. Using Weka version 3.8.4, experiments are conducted on a Windows 10 machine with Core i5 processor and 4 GB of RAM.

4.2 Experimental datasets

The datasets that are used in this research are fetched from the Kaggle repository, where a training set of data in the form of *arff* Weka file is predetermined in advanced. In the other hand, testing set of data is randomly generated by the Python-based model provided by the creator of the training dataset. The training dataset consists of 400 samples with 2 attributes (distinct features extracted from medical CT and X-ray images of Covid-19 and nonCovid-19 cases) and 1 class attribute (with 2 values of negative and positive cases of covid-19 cases). Extracted features (namely x and y in the training dataset) are of numeric data type, while the class attribute is in nominal data type. Test datasets have the same form of the training dataset but can be of a huge size, more than 400 samples.

The dataset has the float numerical readings of chest computed tomography (CT) scan images and chest radiology (x-ray) images. The readings present some distinct features of covid-19 cases that are clearly presented in the images and well known to medical experts and radiologist.

4.3 Experimental Results

The following subsections summarizes the obtained results for each model.

4.3.1 Classification results

This work utilizes the ANN for the classification task, where it has a Multilayer Perceptron classifier (MLP). MLP is a class of feedforward ANN where it has a sophisticated architecture, consisting at least 3 layers of nodes including input layer, hidden layer and output layer. Within the MLP, it utilizes a non-linear activation function and a backpropagation for supervised training process. The activation function is used to describe the relations between inputs and outputs in a non-linear way for more flexibility of the model in relaxing arbitrary relations.

In brief, the MLP algorithm is as follows (for more details see [45]):

1. Forward pass (input layer): passing inputs into the model multiplied by weights and add bias at every layer to find the calculated output of the model.
2. Loss calculate (activation function/hidden layer): once a sample data (e.g. a record in a dataset) an output is obtained from the model as a predicted output, which will be labeled with the data that is a real or expected output. Hence, the loss that is to be backpropagated is calculated using the embedded backpropagation algorithm.
3. Backward pass (output layer): finally, and most importantly in training the model, the loss is backpropagated and updates the weights of the model by using gradient. Iteratively, weights will be adjusted according to the gradient flow to a certain direction.

For predicting two categories (negative, positive), the following table shows the classification model (MLP/ANN) for the dataset. The following tables summarizes parameter settings and accuracy of the model.

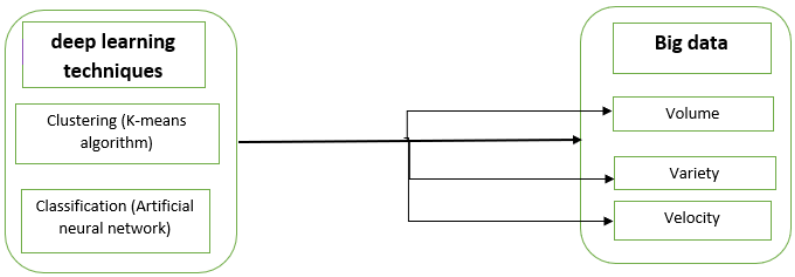

Figure 1: Generic study model

Table 1: Parameter settings for each classification model (MLP)

Model	#Hidden layers	Learning Rate	momentum	Epoch	Test mode
ANN-1	attributes + class	0.3	0.2	500	Training evaluation
ANN-2	attributes + class	0.3	0.2	500	10-fold cross-validation
ANN-3	attributes + class	0.3	0.2	500	Percentage Split 66%
ANN-4	attributes + class	0.3	0.2	10000	Training evaluation
ANN-5	attributes + class	0.3	0.2	10000	10-fold cross-validation
ANN-6	attributes + class	0.3	0.2	10000	Percentage Split 66%
ANN-7	attributes + class	0.3	0.2	100000	Training evaluation

Table 2: Accuracy of the classification models

Model	Correctly classified instances	Incorrectly classified instances	Kappa	Mean absolute error	Relative error	Precision	Recall	F-measure	Time (s)
ANN-1	94%	6%	0.88	0.1002	20.04%	0.941	0.940	0.940	0.1
ANN-2	92%	8%	0.84	0.1082	21.641%	0.921	0.920	0.920	0.1
ANN-3	94.11%	5.88%	0.88	0.098	19.6006%	0.943	0.941	0.941	0.0
ANN-4	94.25%	5.75%	0.885	0.0754	15.081%	0.944	0.943	0.942	0.0
ANN-5	92.75%	7.25%	0.885	0.0969	19.386%	0.928	0.928	0.927	1.92
ANN-6	93.38%	6.61%	0.867	0.09	18.0076%	0.934	0.934	0.934	0.0
ANN-7	94.75%	5.25%	0.895	0.0724	14.4764%	0.949	0.948	0.947	0.0

With a total number of 400 instances in the dataset, it is shown in table 2 that the classifier has obtained highly accurate classifications for all 7 models. It is clear that the seventh model (ANN-7) has the highest accuracy with 94.75% of correctly classified instances, a value of 0.95 for kappa, 0.949 precision, 0.948 recall and 0.947 F-measure. The best experimental configuration of ANN-7 model is shown in table 1, where the number of hidden layers in the MLP is the number of attributes plus the number classes. The ANN has a learning rate of 0.3, momentum of 0.2, 100000 epochs and training evaluation as a testing mode. On the other hand, a relatively lower accuracy of the classification is obtained by ANN-2 with a 92% of the instances are correctly classified.

4.3.2 Clustering results

This work utilizes the K-means for the clustering task to determine the class attributes from the dataset. To achieve this, it calculates squared Euclidean distances to reduce variances within clusters.

In brief, the K-means algorithm is as follows (for more details see [45]):

1. Initialize a set of cluster centers and a set of data points.
2. Calculate the distance between each data point and cluster centers.
3. Assign a data point to the closest cluster center.
4. Recalculate the new cluster center.
5. Recalculate the distance between each data point and new obtained cluster center.
6. Repeat steps 3, 4, and 5 until convergence (if no data point was reassigned).

For predicting two categories (negative, positive), the following table shows the clustering model (K-means) for the dataset. The following tables summarizes parameter settings and accuracy of the model.

Table 3: Parameter settings for each clustering model (K-means)

Model	Distance function	Initialization method	Number of clusters	Seed number	Cluster mode
Kmeans-1	Euclidean distance	Random	2	10	Training evaluation
Kmeans -2	Euclidean distance	Random	2	10	Percentage Split 66%
Kmeans -3	Euclidean distance	Random	2	10	Cluster evaluation via classes
Kmeans -4	Euclidean distance	Random	5	10	Training evaluation
Kmeans -5	Euclidean distance	Random	5	10	Percentage Split 66%
Kmeans-6	Euclidean distance	Random	5	10	Cluster evaluation via classes

Table 4: Accuracy of the clustering models

Model	Sum of squared errors with in clusters	Number of iterations	Number of data objects	Incorrectly clustered instances	*Time (s)*
Kmeans-1	12.966	2	400 (200, 200)	-	0.0
Kmeans -2	9.649	2	264 (69, 67)	-	0.0
Kmeans -3	11.267	9	400 (213, 187)	8.75%	0.0
Kmeans -4	6.783	19	400 (72, 104, 96, 70, 58)	-	0.0
Kmeans -5	4.618	11	264 (28, 28, 21, 18, 41)	-	0.0
Kmeans -6	5.483	20	400 (102, 66, 81, 82, 69)	62.75%	0.0

With a total number of 400 instances in the dataset, it is shown in table 4 that the K-means has obtained highly accurate classifications for 5 out of 6 models. It is clear that the 5th model has obtained the best value for the sum of squared errors within clusters which is about 4.618. As shown in table 3, kmeans-5 model has the configuration of a 66% split cluster mode, and a random initialization for generating 5 clusters out of 264 instances. Three clusters out of five have categorized the instances into the negative class. Despite the percentage of incorrect clustering, all 6 models converge towards the negative class.

4.4 Results discussions

Weka has proven to be flexible in tuning the required parameter settings for a classifier to thoroughly understand attributes in the dataset. Both techniques, MLP and K-means, proved to be suitable to classify instances into two categories of negative and positive covid-19 cases. However, some features of both techniques are of no significant impact on the accuracy. For example, in the case of MLP, increasing epochs is not significant, where the accuracy is not significantly improved. In the case of K-means, the seed number is not significant as well and no explicit accuracy improvement is present. Another example, models with the number of 2 clusters and a dataset with 50% split of its 2 clusters is not statistically relevant where no clear conclusion can be drawn from them.

Meanwhile, the greatest impact on the classification accuracy is contributed to a number of features. They are:

1. In the case of MLP:
    a. Number of hidden layers in relation to the testing mode (number of attributes plus number of classes).
    b. Training evaluation proved better than cross-validation and split 66% training modes. This could be due to the relatively small size of the training dataset compared to a variety of huge randomly generated test datasets (>10,000 instances) for all 7 models.

2. In the case of K-means:

a. Euclidean distance and initializing clusters are the main engine of the algorithm's performance, so they are significant for the accuracy.

b. Number of clusters has a significant impact on the accuracy, where a balanced number of clusters =5 (not too small ≤3 and not too large >5) is significant.

c. Percentage split 66% proved better than training evaluation and split 66% clustering modes. This could be due to the relatively small size of the training dataset compared to a variety of huge randomly generated test datasets (>10,000 instances) for all 6 models. In addition, it could be also due to the small number of categories (2 classes, negative and positive).

The abovementioned features are subjected to directly to answer the first research question (*Q1*), "*What features/attributes of the Kaggle COVID-19 dataset that have the greatest impact on the classification task?*".

Readings of tables 2 and 4 are dedicated to answer the second research question (*Q2*), "*What deep learning model has the highest prediction accuracy for the Kaggle COVID-19 dataset with respect to volume dimension using Weka?*". it is clear that both techniques MLP and K-means are highly accurate. However, MLP proved to be more accurate due to the percentage of incorrectly classified instances (5.25%) and a smallest absolute mean error (0.0724). MLP is clearly powerful when dealing with big datasets at the training evaluation mode. On the other hand, K-means is superior when dealing with big datasets at the percentage split 66% cluster mode. Therefore, MLP is superior for training, while K-means is superior for testing on big datasets. This research work aimed at training and testing the models on big data scheme. It focuses on the volume dimension of a dataset, so the MLP is better and suitable. K-means would be better and suitable for the variety dimension and perhaps the velocity dimension sense clustering is the task of discovering a pattern in the dataset.

5. CONCLUSION

In this study, two deep learning techniques are implemented to perform two tasks, using two techniques ANN for classification and K-means for clustering. They are employed for detecting covid-19 cases from big data of XCR and CT images. Both techniques proved to be accurate, and rapid detectors of covid-19 cases. They handled big test datasets robustly and efficiently. Over a number of big datasets, it has obtained highly accurate classifications. Parameter tuning helped the model to improve the classification accuracy. The most significant enhancement factor is the ANN configuration for the classification task, while the clustering mode and a balanced number of clusters are the significant factors for the clustering task. In future study, a combination between ANN and K-means will be investigated to measure the impact of how the k-means could enhance the classification accuracy in the ANN.

REFERENCES

[1] Bekkerman, R., Bilenko, M., & Langford, J. (2012). *Scaling up machine learning: Parallel and distributed approaches*. Cambridge University Press.

[2] Alyass, A., Turcotte, M., & Meyre, D. (2015). From Big Data Analysis to Personalized Medicine for All: Challenges and Opportunities. *BMC Medical Genomics,* (8): 1–33.

[3] Karch, Marziah (2017). *Google Books Ngram Viewer*
[https://www.lifewire.com/google-books-ngram-viewer-1616701] (Viewed on10/3/2017).

[4] Yang, C., Huang, Q., Li, Z., Liu, K., & Hu, F. (2017). Big Data and cloud computing: innovation opportunities and challenges. *International Journal of Digital Earth,* 10(1), 13-53.

[5] Skourletopoulos, G., Mavromoustakis, C. X., Mastorakis, G., Batalla, J. M., Dobre, C., Panagiotakis, S., & Pallis, E. (2017). Big data and cloud computing: a survey of the state-of-the-art and research challenges. In Advances in mobile cloud computing and big data in the 5G Era (pp. 23-41). Springer, Cham.

[6] Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information systems*, *47*, 98-115.

[7] Li, Z., Yang, C., Liu, K. Fei, H., & Jin, B., (2016). Automatic Scaling Hadoop in the Cloud for Efficient Process of Big Geospatial Data. *ISPRS International Journal of Geo-Information,* 5(10), 173. doi:10.3390/ijgi5100173.

[8] Aghabozorgi, S., Seyed Shirkhorshidi, A., & Ying Wah, T. (2015). Time-series Clustering – A Decade Review. *Information Systems*, 53(C), 16–38.

[9] García, S., Luengo, J., & Herrera, F. (2015). Data Preprocessing in Data Mining. *Intelligent Systems Reference Library* 72. doi:10.1007/978-3-319-10247-4 (Chapter 6).

[10] Hassan, M. M., Gumaei, A., Alsanad, A., Alrubaian, M., & Fortino, G. (2020). A hybrid deep learning model for efficient intrusion detection in big data environment. *Information Sciences*, 513, 386-396.

[11] Gupta, S., Modgil, S., & Gunasekaran, A. (2020). Big data in lean six sigma: a review and further research directions. *International Journal of Production Research*, 58(3), 947-969.

[12] Zwetsloot, I. M., Kuiper, A., Akkerhuis, T. S., & de Koning, H. (2018). Lean Six Sigma Meets Data Science: Integrating Two Approaches Based on Three Case Studies. Quality Engineering, 1–13. doi:10.1080/08982112.2018.1434892.

[13] Kuo, Y. H., & Kusiak, A. (2018). From Data to Big Data in Production Research: the Past and Future Trends. International Journal of Production *Research*: 1–26. doi:10.1080/00207543.2018.1443230.

[14] Zhong, R. Y., C. Xu, C. Chen, & Huang, G. Q. (2017). Big Data Analytics for Physical Internet-Based

Intelligent Manufacturing Shop Floors." *International Journal of Production Research,* 55(9), 2610–2621.

[15] Mehta, N., & Pandit, A. (2018). Concurrence of Big Data Analytics and Healthcare: A Systematic Review. *International Journal of Medical Informatics,* (114), 57–65.

[16] Hossain, M. S., & Muhammad, G. (2019). Emotion recognition using deep learning approach from audio–visual emotional big data. *Information Fusion,* 49, 69-78.

[17] Ardabili, S., Mosavi, A., Dehghani, M., & Várkonyi-Kóczy, A. R. (2019). Deep learning and machine learning in hydrological processes climate change and earth systems a systematic review. *In International Conference on Global Research and Education,* 52-62, Springer, Cham.

[18] Berman, D. S., Buczak, A. L., Chavis, J. S., & Corbett, C. L. (2019). A survey of deep learning methods for cyber security. *Information*, 10(4), 122.

[19] Golmohammadi, M., Harati Nejad Torbati, A. H., Lopez de Diego, S., Obeid, I., & Picone, J. (2019). Automatic analysis of EEGs using big data and hybrid deep learning architectures. *Frontiers in human neuroscience*, 13, 76.

[20] Zhang, Q., Yang, L. T., Chen, Z., & Li, P. (2018). A survey on deep learning for big data. *Information Fusion*, 42, 146-157.

[21] Jan, B., Farman, H., Khan, M., Imran, M., Islam, I. U., Ahmad, A., & Jeon, G. (2019). Deep learning in big data Analytics: A comparative study. *Computers & Electrical Engineering*, 75, 275-287.

[22] Ning, C., & You, F. (2019). Optimization under uncertainty in the era of big data and deep learning: When machine learning meets mathematical programming. *Computers & Chemical Engineering*, (125), 434-448.

[23] Cong, X., Yu, B., Liu, T., Cui, S., Tang, H., & Wang, B. (2020). Inductive Unsupervised Domain Adaptation for Few-Shot Classification via Clustering. *arXiv preprint arXiv*:2006.12816.

[24] Banerjee, A. K., & Arora, N. (2020). Machine learning techniques in biological data classification and clustering: Initiation of a scientific voyage. *Journal of PeerScientist,* 2(1), e1000011.

[25] Scharpf, P., Schubotz, M., Youssef, A., Hamborg, F., Meuschke, N., & Gipp, B. (2020). Classification and Clustering of arXiv Documents, Sections, and Abstracts, Comparing Encodings of Natural and Mathematical Language. *arXiv preprint arXiv*:2005.11021.

[26] Arora, J., Tushir, M., & Kashyap, R. (2020). Improving Semi-Supervised Classification using Clustering. *EAI Endorsed Transactions on Scalable Information Systems*, 7(25).

[27] Reimers, N., Schiller, B., Beck, T., Daxenberger, J., Stab, C., & Gurevych, I. (2019). Classification and clustering of arguments with contextualized word embeddings. *arXiv preprint arXiv*:1906.09821.

[28] Ghezelbash, R., Maghsoudi, A., & Carranza, E. J. M. (2020). Optimization of geochemical anomaly detection using a novel genetic K-means clustering (GKMC) algorithm. *Computers & Geosciences*, 134, 104335.

[29] Stemmer, U. (2020). Locally private k-means clustering. In Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms (pp. 548-559). Society for Industrial and Applied Mathematics.

[30] Xia, C., Hua, J., Tong, W., & Zhong, S. (2020). Distributed K-Means clustering guaranteeing local differential privacy. *Computers & Security*, 90, 101699.

[31] Xu, J., & Lange, K. (2019, May). Power k-means clustering. *In International Conference on Machine Learning*, 6921-6931.

[32] Fuente-Tomas, L. D. L., Arranz, B., Safont, G., Sierra, P., Sanchez-Autet, M., Garcia-Blanco, A., & Garcia-Portilla, M. P. (2019). Classification of patients with bipolar disorder using k-means clustering. *PloS One*, 14(1), e0210314.

[33] Yuan, C., & Yang, H. (2019). Research on K-value selection method of K-means clustering algorithm. *J—Multidisciplinary Scientific Journal*, 2(2), 226-235.

[34] Tai, H., Khairalseed, M., & Hoyt, K. (2020). Adaptive attenuation correction during H-scan ultrasound imaging using K-means clustering. Ultrasonics, (102), 105987.

[35] Sharma, D. K., Dhurandher, S. K., Agarwal, D., & Arora, K. (2019). kROp: k-Means clustering based routing protocol for opportunistic networks. Journal of Ambient Intelligence and Humanized Computing, 10(4), 1289-1306.

[36] Jones, P. J., James, M. K., Davies, M. J., Khunti, K., Catt, M., Yates, T., & Mirkes, E. M. (2020). FilterK: A new outlier detection method for k-means clustering of physical activity. *Journal of Biomedical Informatics*, 103397.

[37] Arunkumar, N., Mohammed, M. A., Mostafa, S. A., Ibrahim, D. A., Rodrigues, J. J., & de Albuquerque, V. H. C. (2020). Fully automatic model-based segmentation and classification approach for MRI brain tumor using artificial neural networks. *Concurrency and Computation: Practice and Experience*, 32(1), e4962.

[38] Baroni, M. (2020). Linguistic generalization and compositionality in modern artificial neural networks. *Philosophical Transactions of the Royal Society B*, 375(1791), 20190307.

[39] Kim, B., Lee, S., & Kim, J. (2020). Inverse design of porous materials using artificial neural networks. *Science advances*, 6(1), eaax9324.

[40] Owoyele, O., Kundu, P., Ameen, M. M., Echekki, T., & Som, S. (2020). Application of deep artificial neural networks to multi-dimensional flamelet libraries and spray flames. *International Journal of Engine Research*, 21(1), 151-168.

[41] Rahman, M. A., & Muniyandi, R. C. (2020). An Enhancement in Cancer Classification Accuracy Using a Two-Step Feature Selection Method Based on Artificial Neural Networks with 15 Neurons. *Symmetry*, 12(2), 271.

[42] Oztekin, Y. B., Taner, A., & Duran, H. (2020). Chestnut (Castanea sativa Mill.) cultivar classification: an artificial neural network approach. *Notulae Botanicae Horti Agrobotanici Cluj-Napoca*, 48(1), 366-377.

[43] Bajželj, B., & Drgan, V. (2020). Hepatotoxicity modeling using counter-propagation artificial neural networks: handling an imbalanced classification problem. *Molecules*, 25(3), 481.

[44] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. Cambridge (Mass).

[45] Agarwal, C.C. (2015). Data Mining: the textbook. Springer Cham Heidelberg, ISBN 978-3-319-14142-8, DOI 10.1007/978-3-319-14142-8.