ISSN (Print): 0974-6846
ISSN (Online): 0974-5645
Indian Journal of Science and Technology, *Vol 12(37), DOI: 10.17485/ijst/2019/v12i37/148647, October 2019*

SCF: Smart Big Data Classification Framework

Majed Mohaia Alhaisoni[1*], Rabie A. Ramadan[1,2] and Ahmed Y. Khedr[1,3]

[1]Department of Computer Science, University of Ha'il, Ha'il, Saudi Arabia; m.alhasoni@uoh.edu.sa
[2]Computer Engineering Department, Cairo University, Egypt; Rabie@rabieramadan.org
[3]Systems and Computer Engineering Department, Al-Azhar University, Egypt; a.khadr@uoh.edu.sa

Abstract

Background/Objectives: Remote sensing produces huge data to be analyzed for different applications. The aim of this study is to develop smart big data classification platform based on AI techniques. **Methods**: The data differs in its types and format, text, images, audio, and video, and they might be structured or unstructured. Besides, data could be divided into categories, and each category needs to be analyzed by itself. Hence, the first step to handle this massive data is to classify them according to their types. Then, the classification phase is followed by the analysis phase. We proposed to utilize two AI algorithms: Fuzzy KNN and CNN. **Findings**: We proposed a novel and new smart big data classification platform based on AI techniques. It also involves cloud computing as a distributed environment to speed up the classification process of such huge data. The framework proposes a pre-analysis structure with suggested algorithms. The framework is examined against a regular/serial approach, and it proves its efficiency in the big data analysis.

Keywords: Artificial Intelligence, Big Data, Classification, Deep Learning, Neural Networks

1. Introduction

Recently, big data has become a new era for complex data due to its nature and behavior. Such a dominant approach demands careful consideration in terms of data classification and analysis. Therefore, big data has affected the data platform techniques owing to the heterogeneity in data format and size. Moreover, data could be generated from different destinations in various forms. One example on the big data generators is the smart city where it mimics a source of big data[1]. Moreover, Gartner report has released that 20 billion devices will be connected by 2020, which would create huge data[2]. Hence, dealing with massive data is a challenge at the level of classification and analysis[3]. Accordingly, various platform techniques have been developed to overcome such arise issues in big data analysis[4]. At the same time, big data is gaining much attention at level of software and hardware due to the fact compiling such sort of data demands deep inspection and analysis.

Big data can be defined as 5 Vs which refers to velocity, variety, value, veracity, and volume[5]. Various domains could be the source of big data such as internet, social media, short messages, and banking. Therefore, data analysis techniques play a vital role on analyzing these data in order to find very useful information[6]. However, it's very critical to achieve very accurate results of analysis, as such analysis output may lead to decisions to be taken to either act or improve a certain issue[7]. However, the first step to analyze massive data is the classification process. There is a need for efficient distributed classification approach. Therefore, this paper presents a novel smart framework for data classification which combines AI and cloud computing paradigm.

The rest of this paper is organized as follow; related work of data analysis problem is reviewed in section 2. Following that, section 3 presents the proposed framework, sections 4 and 5 show the implementation of the proposed solution and results are discussed thoroughly accordingly.

2. Literature Survey

During the last decades, qualitative research has been benefited from theoretical and mythological research of data analysis. Though, recently big data analysis has gained much attention on the research community[8]. However, data analysis typically falls into one of two

*Author for correspondence

classes which are content and thematic[9]. The content analysis studies the frequency of particular words or phrases. Furthermore, word count and extracting some of the word/phrase semantics such as word place and synonym. Hence, content analysis is valuable in terms of efficiency and easy implementation; however, content analysis faces an issue of its limitation in the richness of the summary data produced[10]. Thematic analysis is more than counting words or extracting data representing phrases. It focuses on identifying and describing both implicit and explicit ideas. Reliability of the thematic analysis could be of greater interest due to generating some codes out of the text that require deep understanding to the raw data.

Analysis of big data goes through multiple phases as shown in Figure 1[1]. These phases are called analysis pipeline. The first phase includes the acquisition and recording where data is collected may be from various sources. Data acquisition could be through sensors and/or data entry, or other sources. The second phase of the analysis is the data extraction/cleaning/annotation. The cleaning process is an important phase in such some of the fields, words, and/or noise could not be needed or important. This phase requires careful attention due to losing data that might affect the overall data analysis process. The third phase includes data classification which leads to the analysis/modeling phase. The last phase is the interpretation. Of course, these phases might not be easy to be implemented[11] due to either the lack of expertise, the unstructured data, and/or the various sources and formats.

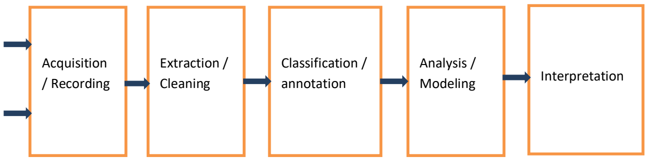


Figure 1. Big data analysis pipeline.

There are many data classification and analysis techniques including statistics methods, intelligent methods, and rough set method. In statistical methods[3], means and standard deviation are the most famous techniques. Both techniques could be done through one sample or two samples. However, this is beneficial to a certain level of confidents. On the other hand, intelligent methods might include fuzzy controllers, neural networks, and neuro fuzzy. However, MapReduce technique[12] could be used to build the fuzzy rules based classification systems. Furthermore, Rough set[13] can be also an efficient method for big data analysis. In addition to that, Machine Learning (ML) is one of the tools used for big data classification and analysis issues since ML acts without any human interventions which is based upon AI techniques[14]. However, ML is classified based as supervised, unsupervised, and reinforcement[15]. Furthermore, one of the main recent techniques used in big data analysis is called deep learning. This technique has attracted more attention recently though it's not human engineered[16], such technique involves supervised and unsupervised methods within the deep architecture of deep learning. The paper utilizes the deep learning in data classification and the next step is to test its efficiency in data analysis as well. Moreover, another study has proposed a novel technique based upon hybrid cluster algorithms[17]. In this study, authors have shown how hybrid cluster algorithm is performing better that existing algorithms in terms of accuracy and quality of produced data.

This paper proposes a new classification framework to serve as a vehicle for big data analysis. The proposed framework is tested against the sequential data classification.

3. Smart Classification Framework (SCF)

Based on our experience, data comes from different sensing devices in various formats. The main problems of analyzing such data are the amount of data and the interrelation between them. Taking into consideration one type of data and building our decision based on its analysis might not be efficient. On the other hand, analyzing all the types of received data is too expensive in terms of resources and time. Therefore, our model in this section is motivated by those two problems. The proposed model tries to handle the incoming data from sensing devices in a distributed manner and tries to classify it accordingly. This leads into distributed analysis to the data later.

Although the proposed approach is a generic approach that can be used with different types of data, certain blocks are suggested for better performance. Some of those blocks are cloud computing, Map-Reduce mechanism, RIS scheduling algorithm. Figure 1 shows the proposed framework diagram. The proposed solution consists of the following blocks:

3.1 Sensing Block

This block encompasses all the sensing devices, including cameras, sensors, software agents, websites, and any other data input. Thus, out of this block, massive data are expected to be received in either a real -time data or offline. Simultaneously, based up on the type of systems and variety of sensing devices, the types and amount of data flowing into the framework are identified.

3.2 Classification Block

One of the main blocks in this framework is the classification block, where Map-Reduce is used to classify different types of data. Map-Reduce is introduced in 2004 by Google[20]. It is a distributed programming model for writing huge, scalable, and fault-tolerant information applications that were created over a cluster of computers to process big information sets. The MapReduce model is based on two main features: the map function and the user-designed reduction function. In particular, the map function generates some intermediate outcomes, processes the input information in the first stage; subsequently, these intermediate outcomes are fed into a second stage in a reduction function that somehow mixes the intermediate outcomes to produce a final output. Figures 2, 3 and 4 show the Map-Reduce programming model.

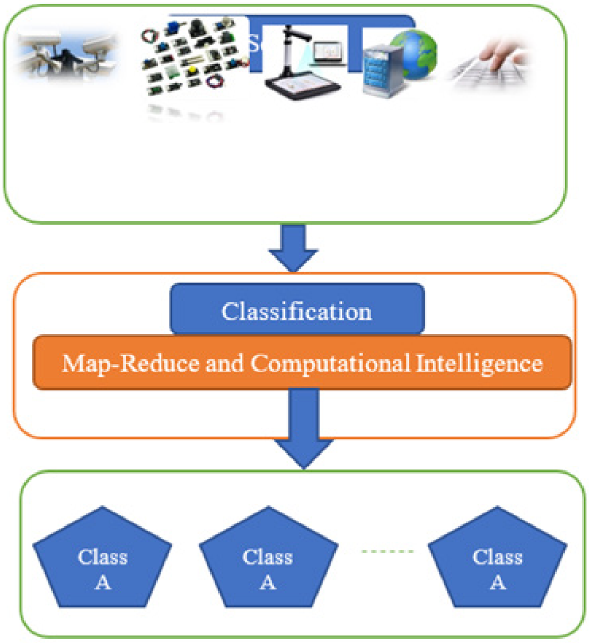


Figure 2. Smart big data classification framework.

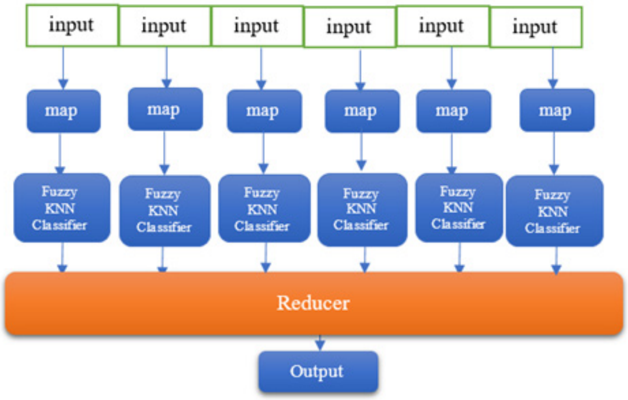


Figure 3. Map-reduce fuzzy based programming model.

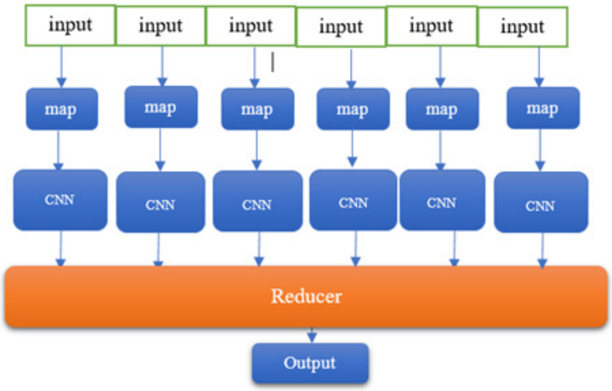


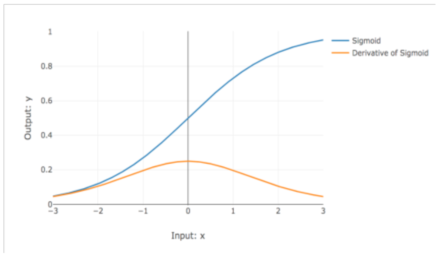
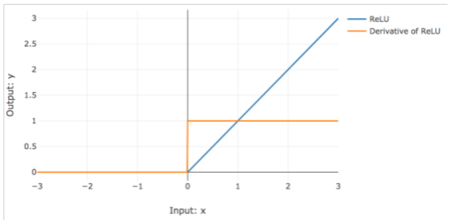
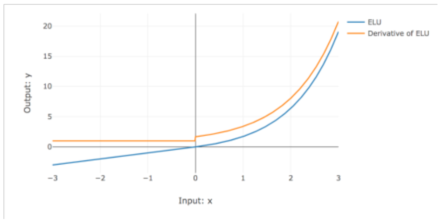
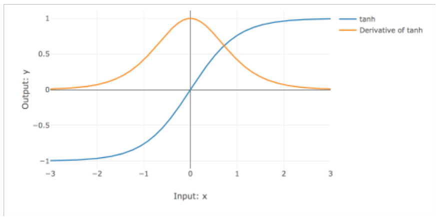
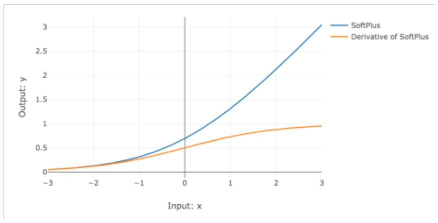
Figure 4. Map-reduce deep learning based programming model.

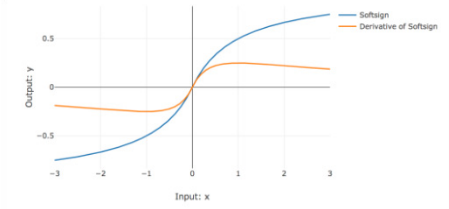
We propose to utilize any computational intelligence technique with Map-Reduce to handle the classification. To prove the efficiency of our proposal, Fuzzy KNN Logic[21] and Deep Neural Networks[22], have been introduced as shown in Figures 2, 3, and 4. These two techniques will be used along with the Map-Reduce for data classification.

The Fuzzy KNN classifier works by assigning to the unlabeled signature a membership value that provides the framework with appropriate data to estimate the decision's certainty. Each one of the groups identified as a fraction of the unlabeled signature determined by the coefficient of Fuzzy membership. This is one of the main advantages of the Fuzzy logic over the smooth logic system when we allocate the record to an unknown category. Two things are classified by the Fuzzy KNN classifier, the Fuzzy information from the training data, and the training data itself. A Fuzzy K-NN Classifier is one of the most popular software techniques due to its simplicity and consistency of the classification decision.

With regards to the deep learning, the classification is done differently where Convolution Neural Network (CNN)[23] is used the best with image classification. Therefore, there is a need for preprocessing for the input data to change it accordingly to image similar format. The input in this research is formed in a 28 × 28 pixels matrix to fit the CNN input. This has been done in the "map" block. In addition, there are different activation functions can be used with CNN. Table 1 shows a summary of the CNN activation functions.

Table 1. CNN activation functions

Function	Equation	Shape
Sigmoid	$f(x)=\alpha(x)=\frac{1}{1+e^{-z}}$ $\frac{d}{dx}f(x)=\frac{d}{dx}\alpha(x)=\frac{e^{-x}}{(1+e\ )}$	
Rectified linear unit (Rlu)	$f(x)=\begin{cases}x, & \text{if } x\geq 0\\ 0, & \text{if } x<0\end{cases}(=\max(x,0))$ $\frac{d}{dx}f(x)=\begin{cases}1 & \text{if } x\geq 0\\ 0 & \text{if } x<0\end{cases}$	
Exponential linear unit (ELU)	$f(x)=\begin{cases}x, & \text{if } x>0\\ \alpha(e^{x}-1), & \text{if } x\leq 0\end{cases}$ $\frac{d}{dx}f(x)=\begin{cases}1 & \text{if } x>0\\ f(x,\alpha)+\alpha & \text{if } x\leq 0\end{cases}$	
TanH	$f(x)=\tan h(x)=\frac{e^{z}-e^{-z}}{e^{z}+e^{-z}}\left(=\frac{\sin h(x)}{\cos h(x)}\right)\left(=\frac{2}{1+e^{-2x}}-1\right)$ $\frac{d}{dx}f(x)=\frac{d}{dx}\tan h(x)=\frac{4}{(e^{-z}+e^{z})^{2}}$	
SoftPlus	$f(x)=\ln(1+e^{z})$ $\frac{d}{dx}f(x)=\frac{1}{1+e^{-z}}$	

Function	Equation	Shape
Softsign	$f(x)=\frac{x}{1+\lvert x\rvert}$ $\frac{d}{dx}f(x)=\frac{x}{(1+\lvert x\rvert)^2}$	

In both cases, we split the data sets into data sets learning and data sets evaluation. Thus, the learning data sets are 75% of all data, and the remaining 25% of all data are the test data sets. MapReduce divides data into different chunks, and the division size is a function of the data size and number of available nodes. As shown in Figures 2 and 3, using the map function, the data sets are split into several mappers. Every mapper has the same sample number. The reduced part takes individual mappers ‘ tests and incorporates them to produce the final outcome. The model's idea is to build on an individual classifier. That classifier is used to identify the test data and send the class label to the reducer function, and then the reducer takes the majority vote to decide for the test data on the final class label.

4. Experimental Results

In this section, the experimental results are shown which examines the efficiency of the proposed classifier. The USTC_SmokeRS dataset[18] is used to evaluate the proposed framework. Therefore, the data set consists of a total of 6225 RGB images from six classes: cloud, dust, haze, land, seaside, and smoke. However, each image was saved in a format of ".tif" with a size of 256 × 256 and a spatial resolution of 1 km. The number of images in each class is shown in Table 2. Figure 5[19] shows a sample of the dataset images.

Table 2. Number of images per class

Class	Cloud	Dust	Haze	Land	Seaside	Smoke
Number of images	1164	1009	1002	1027	1007	1016

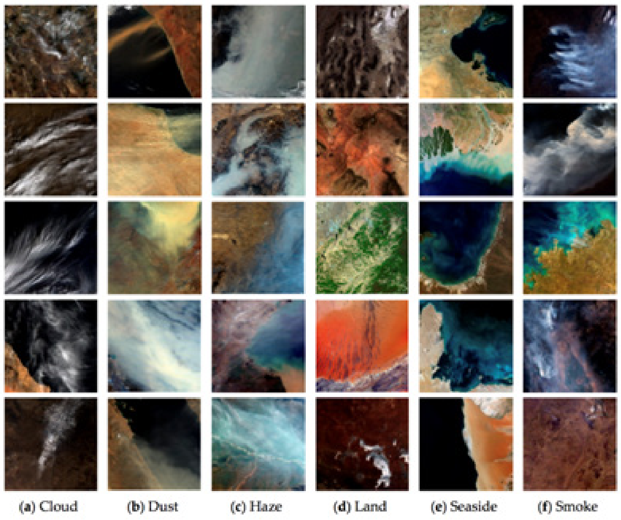
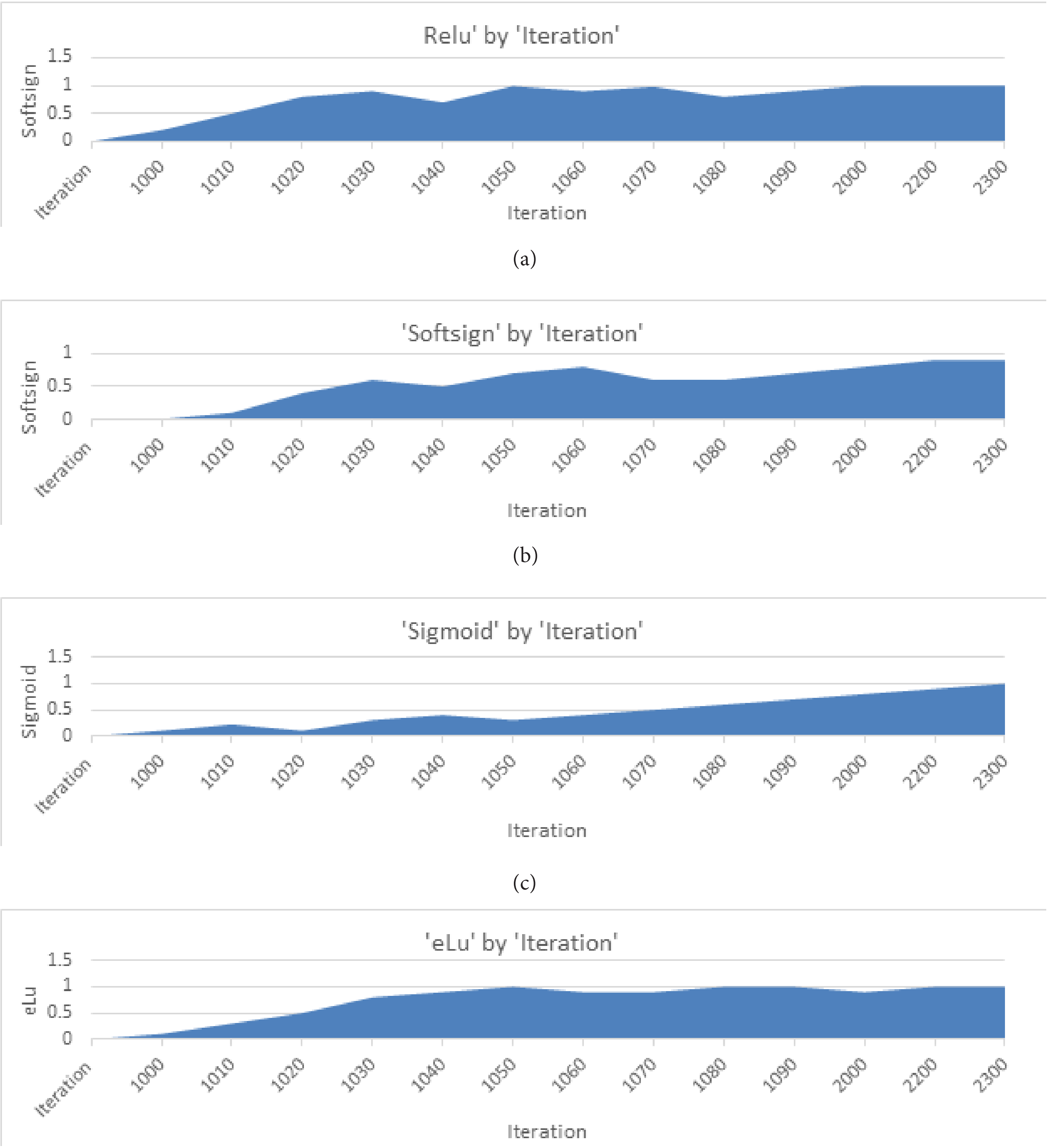

(a) Cloud (b) Dust (c) Haze (d) Land (e) Seaside (f) Smoke

Figure 5. Dataset sample images[19].

After running the training process with different iterations, the accuracy of both algorithms is measured. The performance of CNN with different activation functions is shown in Figure 6. Besides, Table 3 demonstrations the average results of the activation functions. The performance, in this context means the classification percentage. As can be seen, the performance of different activation functions produces similar results; however, Relu, eLu, and SoftPlus are giving the best performance. Also, it is clearly noted that with the increase of the number of iterations, the performance increases which gives a proof of the proposed framework efficiency but few drops in some of the functions such as Softsign, Sigmoid, and SoftPlus. Therefore, those activation functions are recommended for applications with similar classification requirements.

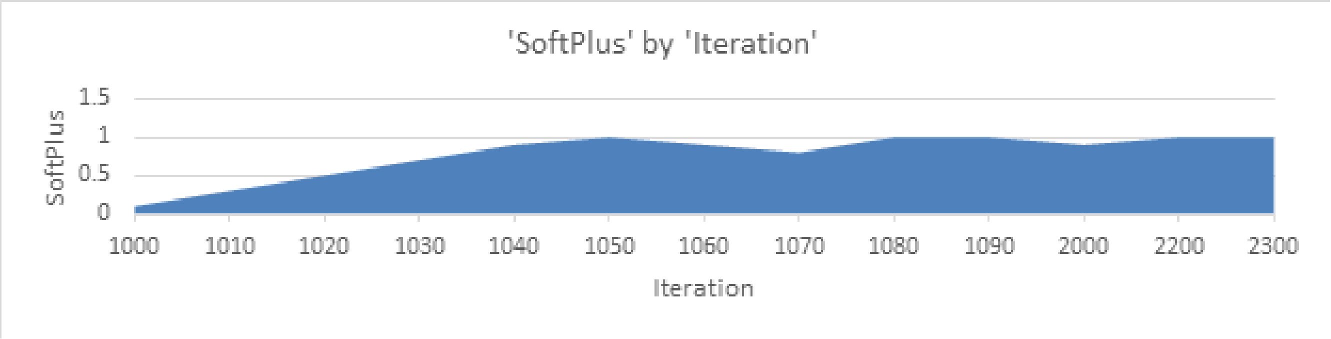

Relu' by 'Iteration'
Softsign
1.5
1
0.5
0
Iteration
1000
1010
1020
1030
1040
1050
1060
1070
1080
1090
2000
2200
2300
Iteration
'Softsign' by 'Iteration'
Softsign
1
0.5
0
Iteration
1000
1010
1020
1030
1040
1050
1060
1070
1080
1090
2000
2200
2300
Iteration
'Sigmoid' by 'Iteration'
Sigmoid
1.5
1
0.5
0
Iteration
1000
1010
1020
1030
1040
1050
1060
1070
1080
1090
2000
2200
2300
Iteration
'eLu' by 'Iteration'
eLu
1.5
1
0.5
0
Iteration
1000
1010
1020
1030
1040
1050
1060
1070
1080
1090
2000
2200
2300
Iteration

(a)

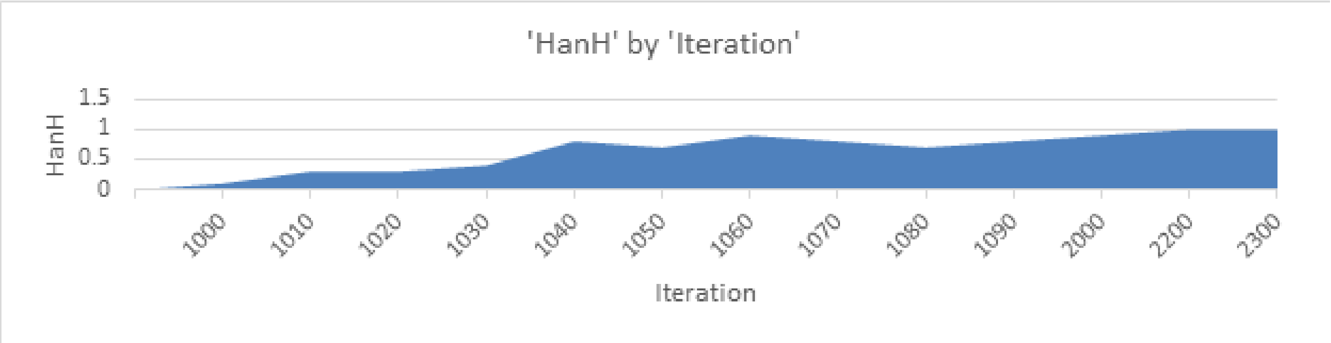
(b)

(c)

(d)



(e)



(f)

Figure 6. CNN performance with different activation functions (a) Relu , (b) Softsign, (c) Sigmoid, (d) eLu, (e) SoftPlus, and (f) HanH.

Table 3. CNN activation functions average performance

Activation Function	Average Performance
Relu	0.82076923
Softsign	0.58538462
Sigmoid	0.48615385
eLu	0.79230769
SoftPlus	0.66923077
HanH	0.77692308

Table 4 shows comparison between the Fuzzy KNN and CNN performances. As can be seen, the performance of Fuzzy KNN over performs the CNN with small percentage.

Table 4. Performance of Fuzzy KNN and CNN

Algorithm	Performance
Fuzzy KNN	0.86
CNN (Relu activation function)	0.82

In summary, we can say that both algorithms work fine in data classification based on our selected dataset. In addition, the speed performance of the proposed classifier reached, on average, 50% faster than the sequential algorithm. Indeed, this performance measurement could be enhanced by increasing the number of used processors as well as the increasing of the different input channels.

5. Conclusion

We proposed a big data classification framework based on AI. We utilized two AI algorithms, Fuzzy KNN and CNN. Both algorithms were examined, and the performance results show that they are suitable for the framework; however, the Fuzzy KNN seems to have a better performance than CNN. The future work of this study is to examine this framework on a large-scale data, and to extend the framework to involve the analysis phase as well.

6. Acknowledgement

The authors would like to thank the Deanship of Scientific Research at University of Hail for their support.

7. Reference

1. Kitchin R. The real-time city? Big data and smart urbanism. GeoJournal. 2014; 79(1):1–14. https://doi.org/10.1007/s10708-013-9516-8.
2. Gartner Says 6.4 Billion Connected 'Things' Will Be in Use in 2016, Up 30 Percent From 2015 [Internet]. [cited 2015 Nov 10]. Available from: https://www.gartner.com/en/newsroom/press-releases/2015-11-10-gartner-says-6-billion-connected-things-will-be-in-use-in-2016-up-30-percent-from-2015.
3. Oussous A, Benjelloun FZ, Ait Lahcen A, Belfkih S. Big data technologies: A survey. Read the latest articles of Journal of King Saud University - Computer and Information Sciences. 2018; 30(4):431–48. https://doi.org/10.1016/j.jksuci.2017.06.001.
4. Londhe A, Rao PP. Platforms for big data analytics: Trend towards hybrid era. 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS); 2017. p. 3235–38. https://doi.org/10.1109/ICECDS.2017.8390056.
5. Assunção MD, Calheiros RN, Bianchi S, Netto MAS, Buyya R. Big data computing and clouds: Trends and future directions. Journal of Parallel and Distributed Computing. 2015; 79-80:3–15. https://doi.org/10.1016/j.jpdc.2014.08.003.
6. Kim JK, Wang Z. Sampling techniques for big data analysis. International Statistical Review. 2019; 87(1):177–91. https://doi.org/10.1111/insr.12290.
7. Chen M, Mao S, Liu Y. Big data: A survey. Mobile Networks and Applications. 2014; 19(2):171–209. https://doi.org/10.1007/s11036-013-0489-0.
8. Northcott R. Big data and prediction: Four case studies. Studies in History and Philosophy of Science Part A; 2019. p. 1–12. https://doi.org/10.1016/j.shpsa.2019.09.002.
9. Brough P. Advanced research methods for applied psychology. Advanced Research Methods for Applied Psychology; 2018. p. 211–23. https://doi.org/10.4324/9781315517971.
10. Vaismoradi M, Turunen H, Bondas T. Content analysis and thematic analysis: Implications for conducting a qualitative descriptive study. Nursing and Health Sciences. 2013; 15(3):398–405. https://doi.org/10.1111/nhs.12048 PMid:23480423.
11. Labrinidis A, Jagadish HV. Challenges and opportunities with big data. Proc. VLDB Endow. 2013; 5(12):2032–33. https://doi.org/10.14778/2367502.2367572.
12. López V, Del Río S, Benítez JM, Herrera F. On the use of MapReduce to build linguistic fuzzy rule based classification systems for big data. IEEE International Conference on Fuzzy Systems; 2014. p. 1905–12. https://doi.org/10.1109/FUZZ-IEEE.2014.6891753.
13. Inhibitory rules in data analysis a rough set approach [Internet]. [cited 2009]. Available from: https://www.springer.com/gp/book/9783540856375.
14. Banka AA, Mir RN. Current big data issues and their solutions via deep learning: An overview. Iraqi Journal for Electrical and Electronic Engineering. 2018; 14(2):127–38.
15. Rudin C, Wagstaff KL. Machine learning for science and society. Machine Learning. 2014; 95(1):1–9. https://doi.org/10.1007/s10994-013-5425-9.
16. Dwivedi K, Biswaranjan K, Sethi A. Drowsy driver detection using representation learning. Souvenir IEEE International Advance Computing Conference (IACC); 2014. p. 995–99. https://doi.org/10.1109/IAdCC.2014.6779459.
17. Kumar S, Singh M. A novel clustering technique for efficient clustering of big data in Hadoop Ecosystem. Big Data Min. Anal. 2019; 2(4):240–7. https://doi.org/10.26599/BDMA.2018.9020037.
18. SmokeNet: Satellite smoke scene detection using convolutional neural network with spatial and remote sensing article [Internet]. [cited 2019 Jul]. Available from: https://www.researchgate.net/publication/334562542_SmokeNet_Satellite_Smoke_Scene_Detection_Using_Convolutional_Neural_Network_with_Spatial_and_Channel-Wise_Attention.
19. SmokeNet: Satellite Smoke Scene Detection Dataset [Internet]. [cited 2015]. Available from: https://webpages.uncc.edu/cchen62/dataset.html.
20. Chu CT, Kim SK, Lin YA, Yu YY, Bradski G, Ng AY, Olukotun K. Map-reduce for machine learning on multicore. Advances in Neural Information Processing Systems; 2007. p. 281–8.
21. Maillo J, Luengo J, García S, Herrera F, Triguero I. Exact fuzzy k-nearest neighbor classification for big datasets. IEEE International Conference on Fuzzy Systems; 2017. p. 1–6. https://doi.org/10.1109/FUZZ-IEEE.2017.8015686.
22. Mills K, Ryczko K, Luchak I, Domurad A, Beeler C, Tamblyn I. Extensive deep neural networks for transferring small scale learning to large scale systems. Chemical Science. 2019; 10(15):4129–40. https://doi.org/10.1039/C8SC04578J. PMid:31015950 PMCid:PMC6460955.
23. Indolia S, Goswami AK, Mishra SP, Asopa P. Conceptual understanding of convolutional neural network- a deep learning approach. Procedia Computer Science. 2018; 132:679–88. https://doi.org/10.1016/j.procs.2018.05.069.