Machine Learning Algorithms for Big Data Mining Processing: A review

DJAFRI Laouni[1] and GAFOUR Yacine[2]

[1] Department of Computer Science, Ibn Khaldoun University, Tiaret, Algeria, EEDIS laboratory, Djillali Liabes University, Sidi Bel Abbes, Algeria
djaafri29tp@gmail.com,

[2] Department of Computer Science, Ibn Khaldoun University, Tiaret, Algeria, EEDIS laboratory, Djillali Liabes University, Sidi Bel Abbes, Algeria
gayacine1@gmail.com,

Abstract. Big data mining is an excellent source of information and knowledge from systems to end users. However, managing such amounts of data or knowledge requires automation, which leads to serious consideration of the use of machine learning algorithms. Machine learning helps us make decisions if there is no right way to solve a problem identified in previous knowledge bases, and that is, too, one of the most widely used analysis and modeling tools for this purpose. In this work, we present an in-depth study that helps us to choose the best machine learning algorithms in order to process big data and extract knowledge from it, so that, this treatment can be very flexible, either in a simple system with sequential computing, or in a distributed system with parallel computing. To achieve this, we will, first and foremost, test the accuracy of the results provided by the classifiers; here we mean the strength and flexibility of a classifier when it comes to dealing with big data mining. Second, we will also test the execution speed for each classifier in complex cases; that is, when the classifier will not be sufficient to solve a particular problem in the context of big data mining, especially if all cases are dealt with quickly and efficiently. The results obtained in this paper demonstrated the superiority of certain classifiers over others in certain cases, and demonstrated their failure in other cases, the reason being due to the nature of the dataset, in particular the number of instances, the number of attributes , and the number of classes.

Keywords: Artificial Intelligence; Big Data Mining; Supervised classification; Binary Classification; Multi-class Classification.

1. Introduction

Machine learning and data mining are not the same, but cousins. Machine learning is a branch of artificial intelligence that provides systems that can learn from data. Machine learning is often used to classify data or make predictions, based on known properties in the data learned from historical data that's used for training. Data mining is sorting through data to identify patterns and establish

relationships. Generally, data mining (sometimes called knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining is the analysis of data for relationships that have not previously been discovered. It is an interdisciplinary subfield of computer science, the computational process of discovering patterns in large data sets ("Big Data") involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. So, data mining works to provide insights and discovery of unknown properties in the data. Machine learning can be carried out through either supervised learning or unsupervised learning methods. The unsupervised learning uses algorithms that operate on unlabeled data, namely, the data input where the desired output is unknown. The goal is to discover structure in the data but not to generalize a mapping between inputs to outputs. The supervised learning (It is the subject of our concern) use labeled data for training. Labeled data are datasets where the input and outputs are known. The supervised learning method works to generalize a relationship or mapping between inputs to outputs. There is an overlap between the two. Often data mining uses machine learning methods and vice versa, where machine learning can use data mining techniques.

2. Literature Survey

2.1 Big data analytics and Machine Learning

Today, in our world, whoever has more information has more power; this information is extracted from a large amount of data. Big data is generated terribly daily, unlike what we were living in at the end of the last century as the amount of data produced at that time is very small as the data is only generated when certain types of events occur, and you can live weeks and months without producing a single piece of data. But, today we can never do that because data is everywhere; it is produced by individuals, groups, companies, and even things that depend on the Internet, etc. Data analysis is extremely important, especially for companies for public and private companies of all types and services. Companies use this analytics to make informed decisions about self-strategies, including recruitment, marketing, and branding. In general, these analyzes can be used to predict unknowns or what we call extrapolation, what makes the big data concept even more important is actually the concept of artificial intelligence. We especially mention machine learning; thanks to its advantages such as being fast, automaticity, having no acquisition costs and saving on labor, it increases companies to be superior in competition.

2.1.1 Big data analytics

Since the advent of the Internet to this day, we have seen explosive growth in the volume, velocity and variety of data created daily [1]; this amount of data is generated by a variety of methods such as click stream data, financial transaction data, log files generated by web or mobile applications, sensor data from Internet

of Things (IoT), in-game player activity and telemetry from connected devices, and many other methods [2][3].This data is commonly referred to as "Big Data" because of its volume, the velocity with which it arrives and the variety of forms it takes. In 2001, Gartner proposed a three-dimensional or 3 Vs (Volume, Variety and Velocity) view of the challenges and opportunities associated with data growth [4]. In 2012, Gartner updated this report as follows: Big data is high volume, high speed, and / or wide variety of information resources that require new forms of processing to improve decision making [5]. Often times, these Vs are supplemented by a fourth V, is Veracity: How accurate is the data? [6][7]. We can extend this model to the Big Data dimensions over ten Vs: volume, variety, velocity, veracity, value, variability, validity, volatility, viability and viscosity[3][8][9][10][11][12][13][14]. Accordingly, the increasing digitization of our activities, the ever-increasing ability to store digital data, and the accumulation of information of all kinds, is generating a new sector of activity aimed at analyzing these large amounts of data.

Analytics is a broad term that encompasses the processes, technologies, frameworks and algorithms to extract meaningful insights from data. Raw data in itself does not have a meaning until it is contextualized and processed into useful information. Analytics is this process of extracting and creating information from raw data by filtering, processing, categorizing, condensing and contextualizing the data. This information obtained is then organized and structured to infer knowledge about the system and/or its users, its environment, and its operations and progress towards its objectives, this is known as big data mining [15][16] . Its main purpose is to extract and retrieve desired information or patterns from a large amount of data [17]. It is usually performed on a large amount of structured or unstructured data using a combination of techniques that make it possible to explore these large amounts of data, automatically or semi-automatically [18][19].

2.1.2 Machine Learning

Big Data Mining is a great source of information and knowledge from systems to other end users. However, managing such a large amount of data or knowledge requires automation, which leads to serious thinking about the use of machine learning techniques. Machine learning consists of many powerful algorithms for learning patterns, knowledge acquisition, and predicts future events. Specifically, these algorithms work by searching a group of possible predictive models to capture the best relationship between descriptive features and target functions in the dataset. Based on this, the machine learning algorithm makes the selection during the training process. The clear criterion for driving this choice is the search for data-compatible models [5][20] . We can then use this model to make predictions for new cases (instances) [21]. Therefore, Machine learning, which is one of the sub domains of artificial intelligence, aims to automatically extract and exploit the information present in the dataset, that is, equipping machines with human intelligence, so that they are able to make predictions based on a huge amount of data, which is an almost impossible task for a human being [22]. For

example, machine learning plays a key role in better understanding and coping with the COVID-19 crisis, in which machine learning algorithms allow computers to mimic human intelligence and ingest large volumes of data to quickly identify models and information; these models are used to predict new observed values. After that, smart decisions can be taken to help us out of the crisis [23][24].

Machine learning algorithms are broadly classified into three categories: supervised, unsupervised and reinforcement learning [25]. In our work, we have relied on supervised algorithms in order to build predictive models; so that, it connects past and current datasets with the help of labeled data to predict future events [26]. We can simply say that supervised learning refers to known labels (predicted classes are known beforehand) as a set of samples to predict future events [27][28]. It is divided into three phases: the learning phase, the validation phase and the test phase. Supervised learning is also divided into two broad categories [29]: classification and regression. Classification algorithms are suitable for the system that produces discrete responses [30]. In other words, responses are categorical variables, whereas regression algorithms are al- gorithms that develop a model that relies on equations or mathematical operations based on the values taken from input attributes to produce a continuous value representing the output [29]. This means that, the input of these algorithms can take continuous and discrete values depending on the algorithm, whereas the output is a continuous value [30]. Supervised learning algorithms in the context of big data are more complex. Nowadays, there is a growing interest in social, economic, health, safety, and other issues that need to be solved using big data analysis and machine learning algorithms. These two concepts are starting to gain attention in many scientific researches. For example, but not limited to, in the business world, most decisions would be much easier if we can anticipate the likelihood, or propensity of customers to take different actions using machine learning algorithms. Successful applications of propensity modeling include predicting the likelihood of customers moving from one mobile operator to another, responding to particular marketing efforts, or purchasing different products [31]. Also, organizations can use the machine learning algorithms to better control and manage the situation in the event of risks [32]. In the healthcare world, these algorithms can help professionals make better diagnoses by tapping into large collections of historical examples on a scale beyond anything an individual might see in their career. For example, predicting optimal doses based on past dose data and associated outcomes [33]. In a similar study conducted by D.Nguyen et.al [34] in order to find out the optimal distribution of prostate cancer radiotherapy a patient will receive. Currently, if we are talking about fighting epidemic diseases and how to prevent them, we are talking more specifically about the Corona virus pandemic. In early 2020, coinciding with the emergence of this pandemic in China, December 2019 [35], and to this day, the machine learning algorithms is used terribly in most, if not we say, in all scientific research related to fighting this virus [23][24][27][28]. Therefore, big data mining and machine learning are two promising technologies used by many healthcare providers use to help medical

experts in order to solve real problems. But, most of the works in this regard have centered on predictions for prevention and saving lives. Predictions are mainly based on supervised algorithms. For example, A.Ardakani et.al [35] they adopted the Deep Convolutional Neural Network method to build predictive models. Also, T.Ozturk et.al [36] they adopted Convolutional Neural Network method for prediction. A similar work done by L.Sun et.al [37] in which they used the SVM method. Another work presented by J.Wu et.al [38] in the same context, no less important than the other works, which is based on the random forests algorithm. Also, in the context of data mining, there is a comparative study for a better precision in the prediction of cardiovascular diseases carried out by R.Sharma and S.N.Singh [39], where several classifiers have been used including Naive Bayes, C-PLS, KNN and decision tree.

3. Summary of the papers reviewed

In this part, we have selected the most recent works that compare classifiers. Thus, we can choose the best and optimal classifier. But, before starting the summary of the papers reviewed and the experiences, we ask the following question: Why did we choose the six classifiers in our work?

We have selected the six classifiers because they are the most widely used and are highly suitable for big data analytics. To confirm our choices, we quote the following:

SVM (Support Vector Machine) over the past decade, SVM has been gradually integrated into the Big Data field. It solves big data classification problems. In particular, it can help multi-domain applications in a big data environment [40].

ANN (Artificial neural networks) constitute a realistic criterion in the Big Data field , thus knowledge of this field is of paramount importance for those who wish to extract significant information from the big data available to date [41].

KNN (K-Nearest Neighbors) is widely used in big data analytics, especially if it is developed more and more in order to give satisfactory classification results [42][43].

RF (Random Forests) seem insensitive to over-fitting, this method generally does not require a lot of parameter optimization efforts. Random forests therefore avoid one of the main pitfalls of Big Data approaches in machine learning [44].

LR (Logistic Regression) gives better result for analyzing the big data [45].

BN (Bayesian Network) or (Naïve Bayes) can also be used in the Big Data field, it is very useful for generating synthetic data when the actual data is insufficient [46].

We have done a thorough research by investigating some of the related work to find out which of the classifiers works best to give an excellent classification result compared to other classifiers. For reference only, these classifiers used the same datasets and default hyper-parameters in each independent work. Table 1 summarizes the results obtained as follows:

References	YEAR	Classifiers used	Performance metrics	The best classifier
[47]	2020	ANN, SVM	**Accuracy** **ANN=0.982** SVM=0.941	ANN with a difference of (0.041) compared to the next classifier (SVM).
[48]	2020	ANN,SVM,RF	**Accuracy** **ANN=0.8218** **SVM=0.8229** RF=0.7839	ANN and SVM with a very slight increase compared to ANN (difference=0.0011)
[49]	2020	SVM,RF, NB,KNN	**Accuracy** SVM=0.582609 **RF=0.721739** NB=0.553623 KNN=0.628986	RF with a difference of (0.092753) compared to the next classifier (KNN).
[50]	2020	RF,LR,KNN	**Accuracy** **RF=0.9321** LR=0.9096 KNN=0.9096	RF with a difference of (0.0225) compared to the next classifier (LR and KNN).
[51]	2020	LR, RF	**Accuracy** **LR=0.80** RF=0.79	LR with a difference of (0.01) compared to the next classifier (RF).
[52]	2020	LR,NB,ANN, RF,SVM	**Precision** LR=0.530 NB=0.515 **ANN=0.700** RF=0.450 SVM=0.451	ANN is much better than other classifiers, with a difference of (0.17) compared to the next classifier (LR).
[53]	2020	KNN,LR,RF, SVM	**Accuracy** KNN=0.6462 LR=0.8232 **RF=0.8461** SVM=0.8050	RF with a difference of (0.0229) compared to the next classifier (LR).
[54]	2020	RF,KNN,SVM	**Accuracy RF=0.9984** KNN=0.9983 SVM=0.9976	RF with a difference of (0.0001) compared to the next classifier (KNN).
[55]	2021	RF,KNN,LR, NB	**Accuracy RF=0.8677** KNN=0.6629 LR=0.6192 NB=0.6056	RF with a difference of (0.2048) compared to the next classifier (KNN).
[56]	2020	KNN,RF,SVM, NB,LR,ANN	**Accuracy** KNN=0.8916 RF=0.9460 SVM=0.8418 NB=0.8233 LR=0.8408 **ANN=0.9580**	ANN With a difference of (0.012) compared to the next classifier (RF).

Table 1. A survey to choose the best classifier.

From Table 1, we note that in most cases the RF classifier gives a good classification result than SVM and LR [49][50][53][54][55] , and in some cases the SVM classifier gives a good classification result than RF [48][52]. In other cases, the LR classifier gives us good results than RF and SVM [52][51]. We also see that in most cases these classifiers (RF, SVM, LR) give good classification results than KNN and NB [53][50].

Through all the works previously presented in Table 1, we note that when an ANN classifier exists among these classifiers it gives good classification results compared to other classifiers, despite the diversity of the nature and size of the dataset [47][48][52][56].

From here, we conclude that the ANN classifier is the best and first classifier for dealing with big data mining, followed by the RF classifier in the second place, followed to a lesser extent by SVM and LR, then NB in the last place.

4. Practical proofs and discussion

From Table 1, we note that some classifiers sometimes outperform others, and fail at other times.

In order to objectively discuss the work presented in Table 1 and to see the stability or volatility (fluctuation) of the classification results, we conducted two experiments to test these classifiers in the context of big data mining.

We developed our software with the tools and APIs under the Linux Ubuntu operating system 20.04.1, Pycharm CE 2020.2, Java development kit 11.0.8, Apache Spark 3.0.1, Python 3.8, PyQt5 desinger 5.

- The first experiment: In this experiment, we use the KDD Cup 2012 dataset[3], knowing that this dataset has two classes (binary classification). KDD Cup 2012 saved in the LIBSVM format, the size of this dataset is detailed in the following Table:

Data-set	**Instance-training**	**Instances- validation**	**Features**	**Classes**
KDD 2012 (2 Go)	119705032 (1.60 Go)	29934073 (458 Mo)	54686452	2

Table 2. Characteristics of the KDD Cup 2012 dataset.

- The second experiment: In this experiment, we use the Mnist8m dataset, knowing that this dataset has ten classes (multi-class classification). To see more information on this database, visit this site[4]

[3] http://www.csie.ntu.edu.tw/∼ cjlin/libsvmtools/datasets/binary.html
[4] https://www.csie.ntu.edu.tw/∼ cjlin/libsvmtools/datasets/multiclass.html

Data-set	Instances training	Features	Classes
Mnist8m (2.75 Go)	8100000	784	10

Table 3. Characteristics of the Mnist8m dataset.

The results (classifier performance metrics) obtained are shown below:

Metrics (%) / Classifier	LR	RF	SVM	ANN
Precision	0.9099952241540803	0.8345043151705323	0.8488354118601478	1.0
Accuracy	0.9167988783529498	0.8121272365805169	0.8319844609906119	1.0
Recall	0.9107581787524190	0.8121272365805168	0.8319844609906117	1.0
F-measure	0.9133063016519345	0.8160949483720379	0.8366175902829752	1.0
RMSE	0.1088647507021327	0.3935667105928753	0.4098969858505704	0.0
MAE	0.2049216107001812	0.4099316501604129	0.1680155390093881	0.0
MSE	0.1702721106852340	0.9420281768726457	0.1775217779419465	0.0
ROC	0.6164675070485738	0.5060675170482035	0.5710370150412737	1.0
AUC	0.5777808410445922	0.3160580181014294	0.4462580781142544	1.0
Time (s)	94	9	6	5

Table 4: Big Data Mining binary classification.

From Table.4 which represents the big data mining binary classification, where we obtained a precision for binary classification equal to 100% using the ANN classifier, it is followed by LR in second place with an precision equal to 90.99%, then comes in third place the SVM classifier with an precision equal to 84.88%, and finally comes in fourth place the RF classifier with an precision equal to 83.45%.

Metrics(%)/Classifier	LR	RF	SVM	ANN
Precision	0.88557010181677	0.77996042809628	1.0	0.91642815472205
Accuracy	0.88581392104896	0.73163912885459	0.94646349909959	0.91624279416751
Recall	0.88581392104896	0.73163912885459	0.94646349909959	0.91624279416751
F-easure	0.88553253087226	0.74477511561745	0.97249550226594	0.91620695452747
RMSE	0.13935667105928	0.21644947572987	0.23137956024766	0.21564348435074
MAE	0.40993165016041	0.48797841706505	0.05353650090040	0.31129196337741
MSE	0.19420281768726	0.24675495508314	0.05353650090047	0.14777890810444
ROC	0.59993165016041	0.51074125016131	0.0	0.73230124012973
AUC	0.48794202817687	0.37974402511608	0.0	0.68707544271661
Time (s)	370	32	52	71

Table 5: Big Data Mining multi-class classification

In addition, from Table.5 which represents the big data mining multi-class classification, where we obtained a precision for multi-class classification equal to 100% using the SVM classifier. On the other hand, we got the two metrics ROC and AUC equal to 0. This indicates an over-fitting problem. So, we conclude that SVM does not perform well in big data mining classification (multi-class classification). Thus, the ANN classifier comes in first place with an precision

equal to 91.64% with a processing time equal to 71 seconds, then it is followed by LR with an precision equal to 88.55%, but the processing time long enough (370s). After that, the RF classifier comes in third place with an precision equal to 77.99% and a somewhat acceptable processing time (32s).

5. Conclusion and future scope

In conclusion, our study shows that the majority of machine learning algorithms are influenced in terms of performance metrics and execution time. The ranking of these algorithms changes in terms of efficiency and effectiveness depending on the nature of the datasets used, particularly in the context of big data analytics. But the ANN classifier was less affected than other classifiers, as it retained its first place despite the diversity of data sets. Based on this, we consider the ANN classifier to be the best for dealing with big data mining. As future work, we can change the hyper-parameters and increase the number of hidden layers and/or increase the number of neurons per hidden layer to improve classification accuracy more and more, which we simply call the deep learning.

References

1. R. Sathyaraj, L. Ramanathan, K. Lavanya, V. Balasubramanian, J. Saira Banu, Chicken swarm foraging algorithm for big data classification using the deep belief network classifier, Data Technologies and Applications (2020). doi:https://doi.org/10.1108/DTA-08-2019-0146.
2. P. O'Donovan, K. Leahy, K. Bruton, T. J. O'Sullivan, Big data in manufacturing: a systematic mapping study, Journal of Big Data 20 (2) (2015). doi:DOI 10.1186/s40537-015-0028-x.
3. R. Hariri, E. Fredericks, K. Bowers, Uncertainty in big data analytics: survey, opportunities, and challenges, J Big Data 44 (6) (2019). doi:https://doi.org/10.1186/s40537-019-0206-3.
4. Chen, Mao, Liu, Big data : a survey, mobile networks and application 19 (2) (2014) 171–209.
5. Erl, Khattak, Buhler, Big data fundamentals : Concepts, drivers and techniques, Prentice Hall Press (2016).
6. Chan, An architecture for big data analytics, Communications of the IIMA 13 (2) (2013) 1–13.
7. Roos, Deutsch, Corrigan, Zikopoulos, Parasuraman, Giles, Harness the power of big data : The ibm big data platform, New York : McGraw-Hill (2013).
8. N. Khan, H. Shah, G. Badsha, A. A. Abbasi, M. Alsaqer, S. Salehian., 10 vs, issues and challenges of big data, In International Conference on Big Data and Education ICBDE '18 (2018) 203–210.
9. Kayyali, D. Knott, S. V. Kuiken, The big-data revolution in us health care : Accelerating value and innovation, Mc Kinsey Company 2 (8) (2013) 1–13.
10. A. Katal, M. Wazid, R. Goudar, Big data : issues, challenges, tools and good practices, In Contemporary Computing (IC3) Sixth International Conference.IEEE (2013) 404–409.
11. M. Ferguson, Enterprise information protection- the impact of big data, IBM (2013).

12. Ripon, Arif, Big data : The v's of the game changer paradigm, In IEEE 18th International Conference on High Performance Computing and Communications ; IEEE 14th International Conference on Smart City ; IEEE 2nd International Conference on Data Science and Systems (2016). doi:DOI 10.1109/HPCC-SmartCity-DSS.2016.8.
13. IBM, The top five ways to get started with big data (2014).
14. N. Elgendy, A. Elragal, Big data analytics: A literature review paper. in: Perner p. (eds) advances in data mining. applications and theoretical aspects, ICDM 8557 (2014).
15. T. Cen, Q. Chu, R. He, Big data mining for investor sentiment, Journal of Physics: Conference Series 1187 (5) 2019.
16. C. Dunren, S. Mejdl, P. Zhiyong, From big data to big data mining: Challenges, issues,and opportunities, DASFAA Workshops LNCS 7827 (2013) 1–15.
17. A. Oussous, F.-Z. Benjelloun, A. Lahcen, S. Belfkih, Big data technologies: A survey, Journal of King Saud University - Computer and Information Sciences (2017). doi:doi: http://dx.doi.org/ 10.1016/j.jksuci.2017.06.001.
18. W. Xindong, Z. Xingquan, W. Gong-Qing, D. Wei, Data mining with big data, IEEE Transactions on Knowledge and Data Engineering 26 (1) (2014) 97–107. doi:DOI: 10.1109/TKDE.2013.109.
19. Z. Xingquan, D. Ian, Knowledge discovery and data mining: Challenges and realities, Hershey, New York ISBN 978-1-59904-252 (2007).
20. S. Bailly, G. Meyfroidt, J. Timsit, What's new in icu in 2050: big data and machine learning, Intensive Care Med 44 (2018) 1524–1527. doi:https://doi.org/10.1007/s00134-017-5034-3.
21. P. V. Klaine, M. A. Imran, O. Onireti, R. D. Souza, A survey of machine learning techniques applied to self-organizing cellular networks, IEEE Communications Surveys and Tutorials 19 (4) (2017) 2392–2431. doi:10.1109/COMST.2017.2727878.
22. K. Burhan, Rashidah, Hunain, Critical insight for mapreduce optimization in hadoop, Inter- national J of Computer Science and Control Engineering 2 (1) (2014) 1–7.
23. C. An, H. Lim, D. Kim, Machine learning prediction for mortality of patients di- agnosed with covid-19: a nationwide korean cohort study, Sci Rep 10 (2020). doi:https://doi.org/10.1038/s41598-020-75767-2.
24. D. Goodman-Meza, A. Rudas, J. Chiang, P. Adamson, J. Ebinger, N. Sun, A machine learn- ing algorithm to increase covid-19 inpatient diagnostic capacity, PLoS ONE 15 (9) (2020). doi:https://doi.org/10.1371/journal.pone.0239474.
25. A. Dasgupta, A. Nath, Classification of machine learning algorithms, International Journal of Innovative Research in Advanced Engineering 3 (3) (2016).
26. N. Mathkunti, S. Rangaswamy, Machine learning techniques to identify dementia, SN Comput Sci 1 (118) (2020). doi:https://doi. org/10.1007/s42979-020-0099-4.
27. L. Muhammad, E. Algehyne, S. Usman, A. Ahmad, C. Chakraborty, I. Mohammed, Super- vised machine learning models for prediction of covid-19 infection using epidemiology dataset, SN Computer Science 2 (1) (2020). doi:10.1007/s42979-020-00394-7.
28. Y. Li, Z. Hai-Tao, G. Jorge, A machine learning-based model for survival prediction in patients with severe covid19 infection, medRxiv (2020). doi:https://doi.org/https://doi.org/10.1101/2020.02.27.20028027.
29. G. James, D. Witten, T. Hastie, R. Tibshirani, Statistical learning. in: An introduction to statistical learning, Springer Texts in Statistics, Springer, New York, NY 103 (2013) 15–57.

30. P. Siirtola, J. Roning, Comparison of regression and classification models for user independent and personal stress detection, Sensors (2020).
31. A. Coulet, M. Chawki, N. Jay, N. Shah, M. Wack, M. Dumontier, Predicting the need for a reduced drug dose, at first prescription, Scientific Reports 8 (1) (2018). doi:10.1038/s41598- 018-33980-0.
32. D. Nguyen, T. Long, X. Jia, W. Lu, X. Gu, Z. Iqbal, S. Jiang, A feasibility study for predicting optimal radiation therapy dose distributions of prostate cancer patients from patient anatomy using deep learning, Scientific reports 9 (1) (2019). doi:https://doi.org/10.1038/s41598-018- 37741-x.
33. S. Lalmuanawma, J. Hussain, L. Chhakchhuak, Applications of machine learning and artificial intelligence for covid-19 (sars-cov-2) pandemic: A review, Chaos, Solitons and Fractals 139 (C) (2020). doi:10.1016/j.chaos.2020.110059.
34. Q. Pham, D. C. Nguyen, T. Huynh-The, W. Hwang, P. N. Pathirana, Artificial intelligence (ai) and big data for coronavirus (covid-19) pandemic: A survey on the state-of-the-arts, IEEE Access 8 (2020) 130820–130839. doi:10.1109/ACCESS.2020.3009328.
35. A. A. Ardakani, A. Kanafi, U. R. Acharya, N. Khadem, A. Mohammadi, Application of deep learning technique to manage covid-19 in routine clinical practice using ct images: Re- sults of 10 convolutional neural networks, Computers in Biology and Medicine 121 (2020). doi:https://doi.org/10.1016/j.compbiomed.2020.103795.
36. T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim, U. Rajendra Acharya, Auto- mated detection of covid-19 cases using deep neural networks with x-ray images, Computers in Biology and Medicine (2020). doi:10.1016/j.compbiomed.2020.103792.
37. L. Sun, G. Liu, F. Song, N. Shi, F. Liu, S. Li, P. Li, W. Zhang, X. Jiang, Y. Zhang, L. Sun, X. Chen, Y. Shi, Combination of four clinical indicators predicts the severe/critical symptom of patients infected covid-19, Journal of Clinical Virology (2020). doi:10.1016/j.jcv.2020.104431.
38. J. Wu, P. Zhang, L. Zhang, W. Meng, J. Li, C. Tong, Y. Li, J. Cai, Z. Yang, J. JZhu, M. Zhao, H. Huang, X. Xie, S. Li, Rapid and accurate identification of covid-19 infection through machine learning based on clinical available blood test results, medRxiv (2020). doi:https://doi.org/10.1101/2020.04.02.20051136.
39. R. Sharma, S. N. Singh, Data mining classification techniques – comparison for better accuracy in prediction of cardiovascular disease, Int. J. Data Analysis Techniques and Strategies 11 (4) (2019).
40. E. Sadrfaridpour, T. Razzaghi, I. Safro, Engineering fast multilevel support vector machines, Machine Learning 108 (2019). doi:https://doi.org/10.1007/s10994-019-05800-7.
41. H. Chiroma, U. A. Abdullahi, S. M. Abdulhamid, A. A. AlArood, L. A. Gabralla, N. Rana,T. Herawan, Progress on artificial neural networks for big data analytics: A survey, IEEE Access 7 (2019). doi:10.1109/access.2018.2880694.
42. Z. Deng, X. Zhu, D. Cheng, M. Zong, S. Zhang, Efficient knn classification algorithm for big data, Neurocomputing 195 (2016) 143–148. doi:10.1016/j.neucom.2015.08.112.
43. W. Xing, Y. Bei, Medical health big data classification based on knn classification algorithm, IEEE Access 8 (2020) 28808–28819. doi:10.1109/ACCESS.2019.2955754.
44. L. Djafri, D. Amar-Bensaber, R. Adjoudj, Big data analytics for prediction : parallel process- ing of the big learning base with the possibility of improving the final result of the prediction, Information discovery and delivery 46 (3) (2018) 147–160. doi:https://doi.org/10.1108/IDD- 02-2018-0002.
45. S. Dhamodharavadhani, R. Rathipriya, Enhanced-logistic-regression-(elr)-model-for-big-data, IGI Global (2019). doi:10.4018/978-1-7998-0106-1.ch008.

46. M. Scutari, C. Vitolo, A. Tucker, Learning bayesian networks from big data with greedy search: computational complexity and efficient implementation, Stat Comput 29 (2019) 1095–1108. doi:https://doi.org/10.1007/s11222-019-09857-1.
47. M. Fengying , J. Zhang, W. Liang , and J. Xue, Automated Classification of Atrial Fibrillation Using Artificial Neural Network for Wearable Devices, Mathematical Problems in Engineering (2020), Article ID 9159158, https://doi.org/10.1155/2020/9159158.
48. J. Miao, W. Zhu, PRECISION-RECALL CURVE (PRC) CLASSIFICATION TREES, arXiv:2011.07640v1 [stat.ML] (2020)
49. R. Naseem , B. Khan, M. Arif Shah, K. Wakil, A. Khan , W. Alosaimi, M. Irfan Uddin , B. Alouffi, Performance Assessment of Classification Algorithms on Early Detection of Liver Syndrome, Journal of Healthcare Engineering (2020), Article ID 6680002, https://doi.org/10.1155/2020/6680002.
50. H. Eedi, M. Kolla, MACHINE LEARNING APPROACHES FOR HEALTHCARE DATA ANALYSIS, Journal of Critical Reviews ISSN- 2394-5125 7(4)(2020).
51. F. RUSTAM , A. MEHMOOD , M. AHMAD, S. ULLAH, D. M. KHAN , G. SANG CHOI , Classification of Shopify App User Reviews Using Novel Multi Text Features,IEEE Access,(2020). doi: 10.1109/ACCESS.2020.2972632.
52. A. Lamurias, S. Jesus, V. Neveu, R. M. Salek, F. M. Couto, Information Retrieval using Machine Learning for Biomarker Curation in the Exposome-Explorer, bioRxiv , (2020). doi: https://doi.org/10.1101/2020.12.20.423685.
53. X. Zhang, H. Saleh , E. M. G. Younis, R. Sahal , A. A. Ali, Predicting Coronavirus Pandemic in Real-Time Using Machine Learning and Big Data Streaming System, Complexity , Article ID 6688912, (2020).https://doi.org/10.1155/2020/6688912.
54. K. M. Ghori, M. Imran, A. Nawaz, R. A. Abbasi, A. Ullah, L. Szathmary, Performance analysis of machine learning classifiers for non-technical loss detection, Journal of Ambient Intelligence and Humanized Computing, (2020).https://doi.org/10.1007/s12652-019-01649-9.
55. M. Hanafy, R. Ming , Machine Learning Approaches for Auto Insurance Big Data, Risks 42(9), (2021). https://doi.org/10.3390/risks9020042.
56. Y. Muhammad, M. Tahir, M. Hayat, K. Chong, Early and accurate detection and diagnosis of heart disease using intelligent computational Model, Scientific Reports,10:19747 (2020).https://doi.org/10.1038/s41598-020-76635-9.