International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 04 Issue: 04 | April -2020 ISSN: 2582-3930

The Impact of Blockchain Technology and Machine Learning in Big Data: An Overview

Dr. Mahammad Idrish I. Sandhi[1], Dr. Abdulbasit S. Banga[2], Mr. Tulsidas V. Nakrani [3]

[1,3]Department of Computer Application (MCA), Sankalchand Patel College of Engineering, Visnagar
[2]College of Computing and Informatics, Saudi Electronic University, Jeddah - Male Branch, Kingdom of Saudi Arabia

---

Abstract - The importance of big data in machine learning can not be overdone in recent memory. Through the evolution of big data, most scientific technologies that relied heavily on Brobdingnagian data to unravel complicated issues in human life gained ground; Machine learning is AN instance of those technologies. many machine learning models that supply innovative returns with high potency rates in predicting, detecting, classifying, discovering, and deed in-depth data regarding events that might preferably be terribly troublesome to see, are created doable by big
Data. though huge knowledge has actually helped within the field of machine learning analysis, over the years its mode of acquisition has expose an excellent challenge within the industries, education, and alternative agencies that obtained it for varied functions. this can be as a result of these giant amounts of knowledge can not be keep on personal computers with restricted storage capability, however need the employment of high-capacity servers for effective storage. These servers is also owned by a gaggle of firms or people WHO had the distinctive privilege of modifying the information in their possession as and once deemed relevant, therefore, making a centralized knowledge storage surroundings. Most of them were known as Third Parties (TP) within the knowledge acquisition method. For the services they provided, these sure parties valued the information in their possession in a very pricey means. The adverse impact could be a limitation in varied investigations that would facilitate solve a series of issues in human life. it's value mentioning that the peace of mind that these knowledge ar purchased in a very pricey manner cannot even be warranted by limiting many investigations that thrive on secure knowledge. To curb these events and have higher machine learning models, the incorporation of blockchain technology databases in machine learning. this text appearance at the thought of massive knowledge, machine learning, and blockchains. additionally, it's at however huge knowledge has wedged the machine learning community, the importance of machine learning, and the way BlockChain technology might be used equally to the machine learning community. the aim of this document is to encourage additional analysis to include BlockChain technology into machine learning.

***Key Words*:** Big Data, Machine Learning, Blockchains, Data Preprocessing

1. INTRODUCTION

Data will be outlined as a set of values of a particular variable either qualitative or quantitative [17]. Whereas quantitative data highlights on amount and numbers, qualitative data is a lot of categorical and should be painted by classes like height, color, race, gender, etc. data could be a important resource in each analysis work. the sort of data non inheritable let alone the preprocessing techniques used contribute massively to nice analysis achievements. usually obtained through primary and secondary sources, data is primarily obtained by direct observations and through the conductivity of surveys. Secondarily, data can even be non inheritable through rigorous market studies or info generated electronically or obtained from the worldwide web. Over the years, primary sources of data have provided fastened and comparatively little quantities of data as compared to its secondary sources counterpart. In recent times, the acquisition of data for analysis comes has been created straightforward with the worldwide web. the huge amounts of data being generated per second through numerous social media platforms, on-line marketing platforms, and business websites among others usually defines big data (BD)[22]. These data could also be preprocessed and analyzed upon acquisition to create higher event predictions and data discoveries for the advantage of man. they will even be fed into a machine learning model for machine-controlled series of specific actions. The works of R. Swathi and R. Seshadri, in [18] confirms that a solid relationship exist between machine learning and big data. This relationship is therefore established from the very fact that machine learning models perform relatively higher with big data than with fewer sets of data. the larger the data, the higher the classification rate, potency rate, prediction rate and general system turnout. determination issues which might are rather not possible to contend with [14, 17], Machine learning has wedged greatly in health, industry, transportation, promoting and alternative sectors of human lives through the event of robots to handle activities that square measure nephrotoxic or dangerous to humans, the timely detection of diseases like cancer, eye disease etc., the image of good cars, effective net search, language translations and etc. Over time, the ever-increasing quantity of data from completely different sources couldn't be hold on on personal computers thanks to huge storage capability required and needed several servers for applicable storage These servers may solely be owned by explicit teams of firms or people WHO may afford for each their purchase and maintenance.

These teams conjointly referred to as trusty Parties square measure trusty with voluminous amounts of data, have proprietary data access and unharnessed data resolute people at a fee. baccalaureate being employed to underneath take machine learning comes square measure largely non inheritable from these trusty parties operational under centralized environments. The wave result could be a incapacitating world of inventions because the purchase of data greatly limits the amount and quality of analysis per annum. conjointly the centralized approach greatly limits the irresponsibleness of such data as a result of the singular purpose of failure associated. In machine learning but, unreliable data suggests that lower system turnout thus the requirement for abundant reliable data. The block chain technology might give reliable data for machine learning comes at no charge, through a redistributed access controls approach [15]. variety of nodes square measure connected to every alternative in a very variety of a series and higher cognitive process depends equally on all connected nodes i.e. nobody node takes call for the amount of nodes concerned thus no single purpose of failure [7]. The technology encourages the sharing of information between nodes. Sharing of data between nodes any imply a considerably larger quantity of information among the chain. Such information will then be fed into Machine learning models directly and freely while not the help of a trusty party that may otherwise need pricy quantity of cash i.e. Block chains Databases in Machine learning Models saves cash. In Machine Learning, the larger the data, the higher the accuracy and larger the generalization ability of the model. i.e. Block chain implementation not solely facilitate save cash however conjointly helps in making certain higher machine learning models thanks to its decentralization ability cite100. within the next section, the conception of Blockchain Technology is generally mentioned, Section three discusses Machine learning and its associated technologies. In Section four, the Block chain Technology is mentioned showing however well it might be incorporated into Machine Learning. This paper concludes in section five.

2. BLOCKCHAIN TECHNOLOGY

Blockchain is that the interconnection of suburbanized blocks of data [14]. The technology thrives on peer to look networks so as to attain its decentralization ability. In Blockchains, entries area unit written into a record by every peer. Variety of records of data from a selected peer type a block. Every peer at intervals the network has their own block. These blocks area unit interconnected to create a series of blocks containing info [21].Information flows freely at intervals these enchained blocks. However, entries written into a record by every peer at intervals the network of users needs to be consented to by group [6]. In Blockchain technology, info is formed promptly on the market to all or any peers at intervals a bunch or network. They then use specific protocols to work out whether or not associate info modification or update ought to or not occur. The technology derives its strength from three different technologies. They're Peer to peer Network, Public Key Cryptography and also the Blockchain Protocol [3].

Peer to Peer Network: Peer to peer Technology drives the authorization and decentralization ability of the Blockchain Technology. Peers reach an accord and judge on explicit data updates or amendments. Nobody peer will impact modification to associa te data while not the approval of others [5].

Public Key Cryptography (PuKC): The involvement of PuKC within the blockchain technology ensures a secure digital identity. Victimization the associated non-public and public keys, a digital signature portrayal sturdy sense of possession may be created and thence a secure digital identity. publicly Key cryptography, a user that desires to speak sends a message in conjunction with its public key to a peer. The resceiving peer receives the message and uses their non-public key to decipher and retrieve the message [21, 15]. this kind of securing data provides high authentication access. A feature embedded in Blockchain. The authorization and authentication method concerned in Block chain makes it a force to reckon with in recent times.

Blockchain Protocol: This protocol determines the underlying rules among that blockchain operates i.e. broadcasting a digitally signed data to all or any nodes/peers in an exceedingly network at a given time [1]. The nodes concerned agree on the data update and every node/block gets a replica of the updated information thence no single purpose of failure. the foremost property of blockchain making certain security and overall effectiveness of the technology lies with decentralization /shared controls [21].

2.1 Structure of block

A block is observed a instrumentality that is employed to store knowledge. It's composed of header and body. The Block header consists of:

• *Block version:* determines the set of block validation rules to be followed.

• *Parent block hash:* it's a 256-bit hash worth that point to the previous block.

•*Merkle* tree root hash: it's the hash worth of all the transactions.

• *Timestamp:* indicates this timestamp as seconds.

• *N-Bits*: offers current target of hashing during a compact format.

• *Nonce*: it's a 4-byte field that starts with zero and increments for each hash calculation.

The block body any consists of a group action counter and transactions. the most range of group actions that a block will hold depends on the scale of the block and therefore the size of every transaction. Blockchain makes use of AN uneven cryptography to validate the authentication of transactions [24].

2.2 Mechanism

BlockChain mechanism is sort of straightforward and secure compared to different technologies. The steps concerned during this method are:

• *Triggering a transaction*: this is often the initial stage of the mechanism. throughout this stage one entity begins to form a dealings by causation data i.e. a dealings gets triggered. This dealing is then broadcasted to all or any the peers within the network.

• *Validation and verification of data:* during this stage validation is finished by the miners. The dealings broadcasted encompasses a hash operate hooked up to that that is employed by the miners to induce a correct output. applicable algorithms area unit chosen to get correct results. These results area unit then verified by each peer within the network. once approval from each node within the network it's passed to successive stage.

• *Creation of new block:* once thriving validation and verification, formation of a replacement block takes places. This new block consists of personal Key, hash operate and also the output generated within the previous step.

• *Addition of block to the chain:* The new block is then communicated to all or any the nodes within the network to be later appended to the present chain of blocks within the blockchain digital ledger [24].

2.3 How BlockChain helps in Big Data

Data and its analysis face plenty of challenges. And if the dimensions of the data get larger the matter gets worse. One resolution to the present most typical downside would be to drill BlockChain technology in knowledge analytics. Blockchain incorporates a layer of security to the already secure huge knowledge analytics technique creating it a lot of genuine . This satisfies the 2 main huge knowledge Analysis demands: [24].

• Block Chain has network architecture that produces it nearly not possible to tamper the info by hackers and Trojans. It additionally permits operations to be performed on the info during a additional reliable method.

• Blockchain ensures integrity, higher performance, and additional stability for managing data [24]. BlockChain technologies overcomes the challenges Janus-faced by big data for analysis.

The key enhancements ascertained victimization blockchain for large knowledge analysis would be:

•***Storage:*** data is hold on individual nodes that square measure showing intelligence distributed with no central entity eager to management access to a user's files. This helps in up security and decreasing prices via decentralized file storage [24]. as an example, the Blocks take naming storage system, encompasses a four-tier design that totally utilizes the decentralization characteristics of the blockchain to confirm the high security of the info [24].

•***Security:*** data security is that the primary good thing about victimization blockchain technology. Blockchain mechanism ensures that data is encrypted properly that makes modification of data a troublesome task. Decentralization makes it easier to cross check file signatures across all the ledgers on all the nodes within the network and verify that they haven't been modified. It makes use of agreement protocols across a network of nodes, to validate transactions and record data during a manner that's incorrupt. Hacking of data appears to be not possible as all the info isn't hold on during a single repository. Exhilaration and corruption of data square measure impracticable [24].

•***Tools*:** Accessibility: ton of cash is spent in obtaining tools for analytics. This drawback may well be resolved by blockchain technology because it has managed to expand the supply of the tools with the assistance of reorganization and democratizing the technology. Blockchain has allowed the massive firms to do to create their analytics efforts additional valuable and useful for data scientists [24].

3. MACHINE LEARNING

Machine Learning is a side of computer science that permits computers to perform specific task by learning. Through learning, systems are able to adapt from previous expertise and to perform similar or connected tasks while not being programmed expressly for those tasks. Machine learning makes use of data and numerous algorithms so as to realize the training method. Some machine learning algorithms embody Artificial Neural Networks, Support Vector Machines, and navy bayes etc. Machine learning algorithms need an affordable quantity of data so as to provide an additional generalized and correct conclusion or results [18]. Therefore the link between big data and machine learning. the training processes concerned in machine learning will be supervised, unsupervised , reinforcement [20]

In supervised Learning additionally known as Example Learning, models desired output is already notable. it's solely conferred with associate input example and presupposed to learn to provide the meant output [112 11]. Through numerous value functions like the cross entropy ,Quadratic and Exponential value, the distinction between the output associated meant output is found and an optimizer perform like the Adams Optimizer, random Gradient Descent(SGD) etc accustomed minimize such value. Supervised learning is most frequently employed in applications wherever future predictions bank heavily on historical data. As an example in predicting earthquakes.

In unsupervised Learning, systems area unit expected to be told justified from given inputs; no labels or examples area unit given. The system is meant to explore okay the computer file, determine patterns at intervals associated turn out an output of some kind. This learning method works well on transactional data. as an example, in recommender systems.

Reinforcement Learning is usually employed in game applications wherever rewards or punishments area unit given associate agent supported their actions. Agents' area unit so expected to require actions to maximize their rewards by following the simplest policy. Reinforcement learning consists of three vital options. These embody associate Agent, Actions and therefore the setting. The agent is predicted to perform tasks by taking actions supported their close environments. Reckoning on actions taken, they receive rewards or get tortured. It's so the responsibility of the agent to use best policies thus on increase their rewards.

3.1 Significance of Machine Learning

Machine learning has improved the quality of lives of humans by providing a number of applications to facilitate human living. Among the numerous applications of machine learning in the field of health, science, industries etc. is the timely detection of diseases such as cancer, glaucoma and other diseases which are claiming human lives at a jaw -breaking rate, the visualization of smart cars, effective web search which has made the internet searches more easy, language translations are immensely helping in worldwide communications and limiting the great language barrier among countries, realization of fraud detection and face recognition systems to mention but a few are greatly helping to improve the quality of life of humans. It is in this regard that Machine Learning has remained significant over the years.

4. BLOCKCHAIN IN MACHINE LEARNING

In order to get smart models in Machine learning, great deal of data is needed. This can be as a result of giant data will increase the turnout, helps is creating a a lot of generalized conclusion and produces a lot of economical and reliable system. This can be one amongst the explanations why the importance of big data in machine learning can't be overemphasized. However, incorporating Blockchain databases in Machine learning suggests that having a shared data, having comparatively abundant larger and safer data and having far better machine learning models [3].

(1) Shared Data: The redistributed property of block chains modify for data to be shared among a community of nodes. This provides quick access to data for connected machine learning models implementation. The difficulty of data acquisition has been a serious obstacle to most machine learning researches. Antecedently analyzers went through powerful struggles to urge some fastened quantity of data for his or her research. This issue failed to solely lead to the generation of less reliable and inefficient models, however conjointly served as a serious hindrance to variety of researches. With the introduction of big data, this hurdle may well be crossed; however, a trusty party would be concerned to urge sufficiently great deal of data. These trustees would successively be paid expensively for the info being collected. Blockchain databases but would supply data to researchers for major research comes while not the services of a trusty party owing to its redistributed data sharing ability. [3, 14].

(2) Larger and Safer Data: redistributed data suggests that abundant larger and safer data with data returning from each intrinsic and extrinsic sources. Intrinsic sources of data be classified into native and metropolitan. The info that emanates from a selected place say a selected branch of a corporation will be aforementioned to be native. Combined data from constant company however totally different branches will be termed Metropolitan data. With Blockchain, these data will be shared across and once used as input to a machine learning model, turn out high potency rate as compared to victimization solely regionally non inheritable data. foreign information is also data from connected corporations being shared. Such data once employed in major prognostic machine learning models will in little question build higher predictions. other than deed voluminous quantity of data through such technology at much no expense, the info non inheritable is additionally as safe as heaven [3]

(3) Higher Machine Learning Models: The wavelet result of obtaining great deal of safe data for machine learning researches is that the development of higher and a lot of reliable machine learning models for varied functions as prediction, foretelling, diseases detection, voice and speech recognition, face detection, to say however a number of. [3]

5. CONCLUSION

The paper summarizes briefly the impact of blockchain technology and machine learning in Big Data. The relevance of these technologies and how closely they relate with one another is further discussed citing major applications which

makes use of these technologies together. The aim of this paper is to encourage further research in incorporating BlockChain Technology into Machine Learning.

ACKNOWLEDGEMENT

The paper is basically focused on the basic impact of the blockchain technology and machine learning in the concept of the Big Data. The paper focus on the various field in the blockchain technology to be used and implemented by using the machine learning technology with reference to the Big Data. The main aim of this paper is to encourage the further research in the fields of the blockchain technology and machine learning technology in big Data. This will help the further research steps for the new researcher who is interested to do the research in the areas of Big Data, Blockchain Technology and Machine Learning.

REFERENCES

[1] Francisca Adoma Acheampong, Big Data, Machine Learning and the BlockChain Technology: An Overview, International Journal of Computer Applications (0975 - 8887) Volume 180 - No.20, March 2018.

[2] S. Athmaja, M. Hanumanthappa, and V. Kavitha. A survey of machine learning algorithms for big data analytics. In 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), pages 1–4, March 2017.

[3] Nolan Bauerle. How does blockchain technology work? Available at:[url =https://www.coindesk.com/information/how-doesblockchain- technology-work/,, 2018. Accessed Feb 2018].

[4] Carla E Brodley and Mark A Friedl. Identifying mislabeled training data. Journal of artificial intelligence research, 11:131–167, 1999.

[5] C. Cachin. Blockchains and consensus protocols: Snake oil warning. In 2017 13th European Dependable Computing Conference (EDCC), pages 1–2, Sept 2017.

[6] Michael Crosby, Pradan Pattanayak, Sanjeev Verma, and Vignesh Kalyanaraman. Blockchain technology: Beyond bitcoin. Applied Innovation, 2:6–10, 2016.

[7] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. Journal of the royal statistical society. Series B (methodological), pages 1–38, 1977.

[8] S. Gharatkar, A. Ingle, T. Naik, and A. Save. Review preprocessing using data cleaning and stemming technique. In 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), pages 1–4, March 2017.

[9] Jiawei Han, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. Elsevier, 2011.

[10] T. M. Khoshgoftaar and P. J Rebours. Improving software
quality prediction by noise filtering techniques. Comput Sci Technol, 22:387, 2007.

[11] Sotiris B Kotsiantis, I Zaharakis, and P Pintelas. Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering, 160:3–24, 2007.

[12] Sotiris B Kotsiantis, Ioannis D Zaharakis, and Panayiotis E Pintelas. Machine learning: a review of classification and combining techniques. Artificial Intelligence Review, 26(3):159–190, 2006.

[13] David J Lary, Amir H Alavi, Amir H Gandomi, and Annette L Walker. Machine learning in geosciences and remote sensing. Geoscience Frontiers, 7(1):3–10, 2016.

[14] W. Meng, E. Tischhauser, Q. Wang, Y. Wang, and J. Han. When intrusion detection meets blockchain technology: A review. IEEE Access, PP(99):1–1, 2018.

[15] James Nechvatal. Public-key cryptography. Technical report, NATIONAL COMPUTER SYSTEMS LAB GAITHERSBURG MD, 1991.

[16] M. Ngxande, J. R. Tapamo, and M. Burke. Driver drowsiness detection using behavioral measures and machine learning techniques: A review of state-of-art techniques. In 2017 Pattern Recognition Association of South Africa and Robotics and Mechatronics (PRASA-RobMech), pages 156–161, Nov 2017.

[17] Rod Pierce. What is data? Math Is Fun, Available at:[url = http://www.mathsisfun.com/data/data.html,, 2017. Accessed Feb 2018].

[18] A. Rathor and M. Gyanchandani. A review at machine learning algorithms targeting big data challenges. In 2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT), pages 1–7, Dec 2017.

[19] S. R. Suthar, V. K. Dabhi, and H. B. Prajapati. Machine learning techniques in hadoop environment: A survey. In 2017 Innovations in Power and Advanced Computing Technologies (i-PACT), pages 1–8, April 2017.

[20] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.

[21] Karl Wst and Arthur Gervais. Do you need a blockchain? Cryptology ePrint Archive, Report 2017/375, 2017. https: //eprint.iacr.org/2017/375.

[22] X.Wu, X. Zhu, G. Q.Wu, andW. Ding. Data mining with big data. IEEE Transactions on Knowledge and Data Engineering, 26(1):97–107, Jan 2014.

[23] Li Xiang-wei and Qi Yian-fang. A data preprocessing algorithm for classification model based on rough sets. Physics Procedia, 25:2025–2029, 2012.

[24] Manisha Valera, Parth Patel & Shruti Chettiar, AN AVANT-GARDE APPROACH OF BLOCKCHAIN IN BIG DATA ANALYTICS, International Journal of Computer Engineering & Technology (IJCET), Volume 9, Issue 6, November-December 2018, pp. 115–124, Article ID: IJCET_09_06_014