International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

An Improvised Mechanism for Optimizing Fault Detection for Big Data Analytics Environment

Dr. M. Sudha Paulin[1], Dr. K. Subramani[2], Dr. P. Muralidharan[3], Dr. Maruthamuthu[4], R. Senthil Kumar[5]

Submitted: 12/10/2023 **Revised**: 13/12/2023 **Accepted**: 22/12/2023

Abstract: In the applications of fault detection, the inputs are the data reflected from health state of the observed system. A major challenge to finding errors is the nonlinear relationship between the data. Big data has other drawbacks, and the volume and speed with which it is generated are reflected in the data streams themselves. In this paper, we develop a deep learning model that aims to provide fault detection in big data analytics engine. This investigation develops an approach for fault detection in large datasets using unsupervised learning. In this research, an unsupervised method of learning is developed specifically for the task of classifying large datasets. To discover regular textual patterns in large datasets, this research use data visualization methods. In this virtual environment, we employ an unsupervised learning method of machine learning that does not require human oversight. Instead, the system should be allowed some leeway to work and find things on its own. The unsupervised learning approach utilizes data that has not been tagged. In contrast to supervised learning, this approach can handle complex tasks.

***Keywords*:** *Fault Detection, Optimisation, Big Data*

1. Introduction

When the benefits of big data are exploited to their full potential, a new era of economic growth and transformation is ushered in. Beyond the realm of big data, the fundamental strategy of modern enterprises is the monetization of usable data. New rivals have the ability to entice workers who possess crucial skills for managing large amounts of data. Integrating a company big data can result in a variety of positive outcomes for the company, including the creation of new markets, products, and customers; improvements in customer service; informed strategic direction; and increased operational efficiency [1]. People that use big data have access to a variety of advantages and opportunities, but they also have to deal with a variety of challenges [2]. There are problems with the capturing of data, storing of data, retrieval of data, sharing of data, and analysis of data. If big data is going to realize its full potential, the problems that have been outlined above need to be solved. Despite this, the data volume is too great for the analytic methods that are now in use. For many years, the hallmarks of computer architecture have been a low input/high output ratio as well as a huge central processor unit that consumes a lot of power [3]. As a consequence of the imbalanced nature of the system, we can only think about big data to a limited extent. According to Moore law, which was developed by [4], the performance of central processing units (CPUs) and disk drives roughly doubles every 18 months. However, throughout the previous ten years, there has been an increase in the disk rotational velocity. While the input and output speeds for consecutive inputs have increased significantly with density, the input and output speeds for random inputs have increased only modestly. This is because random inputs are more difficult to process than consecutive inputs.

Even while progress in information processing technology is slow and steady, the amount of data that can be accessed is expanding at an exponential rate. In the relatively recent past, the field of big data analysis has only seen the introduction of a few handful of technologies that can adequately handle the challenges. In a number of big data applications, such as Cassandra, HBase, and Hadoop, even the most cutting-edge methods have been unable to solve real-time problems such as data analysis, visualization, sharing, searching, and storing. Some of these real-time problems include data sharing, searching, and storing. While MapReduce and Hadoop are good at managing and processing data, their capacity to conduct strategic query processing is restricted, and the necessary infrastructure only provides them with minimal support. Software for statistical analysis such as MATLAB, R, and SAS does not scale well enough to be used on huge datasets. Although it offers a framework for selecting algorithms in a graph-

[1,2,3]*Assistant Professor ,Business and Management-Kengeri campus,Christ (Deemed to be University),Bangalore .*
[4]*Assistant Professor/MCA, Madanapalle Institute of Technology & Science, Angallu, Madanapalle,India*
[5]*Associate Professor, Sathak College of Engineering & Technology, Ramnad,Tamilnadu*
Email sudha.paulin@christuniversity.in[1], Ksubramani12@gmail.com[2], muralidharan.p@christuniversity.in[3], drmaruthamuthur@mits.ac.in[4] & srisenthil2011@gmail.com[5]

based context, Graph Lab data management skills leave a lot to be desired, despite the fact that the framework is provided by the software. As a direct result of this, the resources that are required to fully use big data have not yet been established.

Big data analytics is plagued by a number of problems, some of which include lack of data security, timeliness and consistency, as well as scalability and incompletion [5]. Before beginning the investigation, it is essential to correctly assemble the massive amount of data. Even when taking into account the many various kinds of datasets, there are still some challenging difficulties to overcome, such as how to analyze semi- or totally unstructured data, how to access and display data in an effective manner, and so on and so forth. It is absolutely necessary to have a solid understanding of the pre-processing method that is being applied in order to achieve better results in terms of both the analysis and the data outputs. Datasets are often very large in size (expressed in terabytes), and they are compiled from a wide number of different sources. Because of this, contemporary real-time databases are more susceptible to problems such as noise in the data, incompleteness in the data, and inconsistency in the data. As a result, a wide variety of pre-processing methods are utilized so that inconsistencies in the data can be corrected and noise can be removed [6].

Every component process faces its own one-of-a-kind challenges in terms of how it relates to the many data-driven applications. As a result, the remaining challenges associated with maintaining the secrecy of the data ought to be solved in the investigations that will follow. Using several encryption techniques on content that is fundamentally incompatible offers a huge challenge, as does the encrypting of a vast amount of data, which presents a considerable challenge in and of itself [10]-[13]. A significant worry is the maintenance of the data confidentiality during the outsourcing process. The formulation of privacy policies that take into account the myriad of different reasons why consumers might be concerned about the safety of their personal information is an essential step. In addition, the information of users needs to be shielded from misuse and leaks, and those who break the rules need to be called out. Customers are unable to carry out a physical review of data collected directly from cloud services as a result of the practice of data outsourcing. Due to the fact that this is the case, the integrity of the data is being undermined [14]. Therefore, additional studies are needed to overcome the challenges, improve the analysis, and make the presentation and storage of data more efficient.

The fact that the hashing models that have been developed up until this point are not sufficient for the enormous amount of data that already exists is one of the most significant challenges to the integrity of the data. Verifying the integrity of the memory is a difficult task since there is very little information that can be found about the internal memory, and there is also no aid that can be found for obtaining data from the outside. The questions that require responses can be found further down.

2. Related works

In order to collect vast amounts of data for the sake of scientific research, sectors such as oceanology, environmental studies, astrophysics, and genomics make use of equipment and sensors with maximum throughput. The National Science Foundation (NSF) in the United States has just made an announcement on the launch of a brand new program called BIGDATA. The objective of this program is to enhance our capacity to derive actionable insights from vast stores of complex digital data. Only a handful few scientific domains have been able to successfully exploit big data platforms to generate important findings. [Case in point:] [Case in point:] [7], is an example of a project that encourages students, researchers, and teachers to pursue careers in plant sciences by utilizing inter-operative analysis software, a coordination environment, and a data service. This project was created by SA Goff and his colleagues. Experiment data, reference data, observation data, model or analog data, and derived data were all incorporated into this iPlant.

Over the course of the past few years, the importance of large amounts of data has grown across a variety of industries. Because it contains large datasets, conventional databases are unable to store and process it because it is too big. The big data process has a high capacity to manage such enormous and extensive data sets as a result of the gigantic dataset that is processed during the processing phase in the time that is given for that phase. Text analysis was shown to have a significant amount of potential, despite the fact that it is still in its infant phases. This presents a challenge as the majority of companies obtain their data in an unstructured format. On the other hand, the majority of analytic techniques can only operate effectively on structured data.

A cross-disciplinary overview of the big data field in their comprehensive review of visualization approaches. This overview covered the research challenges, achievements, and visualization tools and techniques that are associated with the big data field. This method not only provides a creative solution to the difficulties that are based on the current condition of big data visualization, but it also provides a comprehensive summary of the key concerns that have been gained in state-of-the-art visualization approaches to big data. In

other words, not only does this method provide a creative solution to the difficulties that are based on the current condition of big data visualization, but it also provides a solution to An overview of the evolution of the visualization method over the past few years includes categorization of the data types, analytical approaches, visualization techniques, and tools. This overview provides an overview of the evolution of the visualization method. The findings of this method highlight the shortcomings of the conventional methods used to display data. As a direct result of this, an enormous amount of data was analyzed, together with the constraints that come with human vision.

These approaches made more effective use of data analytics, which contributed to a rise in both their value and their efficacy. The huge amount of raw data that has been collected over the past few years is what is meant when people talk about big data. It has been noted that some conditions lead to an increase in the volume and volatility of big data, which makes it challenging to both access and manage the data. It was discovered that the technology used for analyzing big data are more complicated, and the majority of firms do not have the personnel necessary to successfully apply the data analytics process. Therefore, systems that make use of data visualization make this challenge easier to manage and present an opportunity for more effective analysis and administration of data.

It has come to everyone attention that the process of text visualization is a significant and quickly developing area within the field of information visualization. It was difficult to identify other works that were comparable to this one, particularly ones that made use of visual metaphors or had certain goals in mind for this process of visualization. Due to the fact that this was the circumstance, a visual survey was executed in an interactive manner utilizing text visualization strategies. This survey is helpful for recent researchers who are doing work in the field of visualization because it includes background information on the subject. When comparing the various survey approaches, a text visualization taxonomy was utilized as the criterion. Text visualization had a significant role in the development of the taxonomy. At long last, a presentation was made of the findings of the evaluations that were performed on the entering data.

In a study that utilized a combination of research approaches, a data mining technique and information visualization were utilized in order to convert qualitative data into quantitative information. A structure for the digital mixed-methods research process was developed by combining the quantitative methods of information visualization and data mining across several levels with the qualitative method of multimodal discourse exploration. This resulted in the creation of a framework for the process. After that, an appropriate framework was built using technology that was integrated, theoretically sound, and empirically proven. This was done so that the huge datasets of multimodal texts could be analyzed. The methodology that was proposed has been used extensively in a variety of settings to derive vital information from geotagged public data. For example, public narratives of extreme occurrences can be used as a substantial data source for crisis management. This is particularly useful in situations involving large numbers of people.

3. Proposed Method

The speed at which big data moves is one of its most beneficial characteristics, evaluating it in real time is one of the most important elements in every research project using big data. One of the challenges that real-time processing has in the setting of big data is the incredible rate at which data is accumulated. This is only one of the many challenges that real-time processing faces. Unfortunately, in order to carry out the feature learning process in big data, a large number of parameters need to be included in this method. This results in a high level of computational complexity in a variety of deep learning techniques, particularly in large-scale DNN. Due to the complexity of this task with massive data when utilizing traditional deep learning approaches, it has been prioritized for rapid attention. A straightforward and effective strategy for learning features has been developed, and it can be used to a number of other incremental learning methodologies.

Approaches to online learning that make use of incremental learning techniques for keeping tabs on the parameters may benefit substantially from using a stable network as the starting point from which to receive fresh information if they are implemented online.

The information of the existing network is carried over to the newly introduced objects, but the weights are not changed in order to prevent the network from becoming overtrained. The process of determining the bound in advance is regarded as difficult because of the degree to which this tactic is reliant on previous information. Only neural networks with two layers can reap the benefits of the bound weight modification operation; nevertheless, this strategy can be difficult to implement in incremental deep learning systems.

Deep learning for low-quality data

The process of learning features now faces a new challenge in the form of the dependability of massive amounts of data. Because of factors such as noise, imprecise objects, redundant objects, a significant

quantity of incomplete data, and inaccurate objects, the only data that can be acquired with big data is of poor quality. In this setting, the production of data of a poor quality is necessary for a variety of reasons.

A number of different origins could be the cause of inadequate data. For example, sensors have the ability to capture a great amount of data; nevertheless, there are occasions when they are unable to accurately record all of the features of an object. Errors in data transmission can also contribute to the overall level of noise in large data sets when they occur in the underlying network.

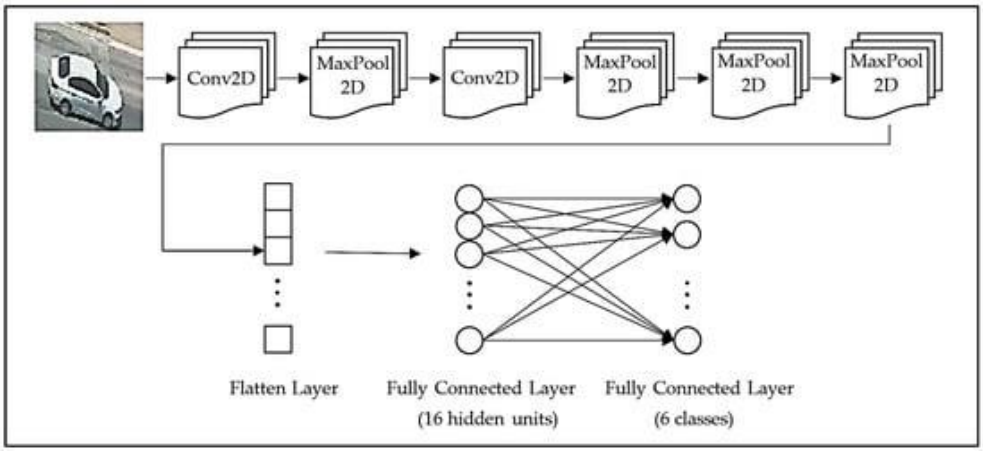

Fig 1. Architecture of the proposed model.

The design of the proposed show is appeared in Figure 1. As appeared within the figure, within the to begin with arrange of the engineering, which compares to the include extraction arrange, two convolutional layers (Conv2D) and four max pooling layers (MaxPool2D) are utilized. Each of the Conv2d layers has 16 channels with 5 × 5 channel measure, and both layers utilize Rectifier Straight Unit (ReLU) as the enactment work. Each of the MaxPool2D layers, on the other hand, has 2 × 2 channel estimate and a walk esteem of 2.

The possibility that some data may be of low quality is disregarded by a number of deep learning methods. To put it another way, the techniques of deep learning were developed solely for the purpose of analyzing the qualities that are associated with high-quality data. The author [8] and colleagues introduced a stacked denoising auto-encoder method as a means of accomplishing feature learning with data that had been damaged by noise.

The gradient descent method is utilized in order to train the parameters of the objective function. To get x, we add some isotropic pepper noise or Gaussian noise. They came up with the idea of using stacked denoising auto-encoders as a method for carrying out the practice of feature learning with data that had been damaged by noise. An imputation auto-encoder model was utilized in order to learn the characteristics of incomplete objects [9]. By making some of object x attribute values zero, we are able to generate a simulated incomplete object denoted by the symbol x'.

This network was used to learn the features of incomplete objects. The results of this experiment demonstrate that the non-local auto-encoder method produces superior performance in the picture restoration process as well as the denoising procedure.

Next we will supplied the entire architecture of the Softmax weighted regression that was built on the Deep Recurrent Neural Network and was used to accomplish a higher level of feature description. This design was utilized to achieve a higher level of feature description. We were able to locate these concealed features automatically in the enormous data set without any assistance from the monitoring signal by using a priori knowledge of these features, and we used these features to build examples that would enhance our subsequent, monitored learning (e.g., object classification). In the second step of the process, a vector that makes use of the recently acquired filter bank was utilized in order to encrypt the various image patches that were chosen.

3.1 System overview

Any one of DL numerous levels has the potential to produce a nonlinear data response emanating from the input layer. Because the performance of DL may be replicated by the human brain and the neurons in the signal processing system, researchers were more interested in the results from DL than in the results from the earlier machines. The work that is being suggested will be carried out in two stages. In the beginning, a fed-forward neural network is used in conjunction with a recurrent neural network to rapidly learn features from a variety of different data sets without the assistance of a human. The weighted regression comes next. It is

recommended that the softmax algorithm be utilized when classifying the dataset on the basis of the features that have already been determined. To validate our hypothesis, we placed an unlisted dataset into the first layer and then read it out of the last layer. This allowed us to see whether or not our prediction was correct. The weighted Softmax regression is then applied to the acquired output, which is the result of a series of layers between the first and fifth (final) layers that distill the functions of the prior part and feed them to the subsequent section. The weighted Softmax regression is then applied to the acquired output.

3.2 Pre-processing

Altering the accommodation can be accomplished with the use of pre-processing techniques such as standardization and fading.

The procedure of normalization, which is a standardized method employed in image processing, was utilized to bring about changes in the intensity of individual pixels. One example of when the 3D RGB is modified is shown in photographs with a lower contrast that have been normalized for glare.

This method of lightening utilizes a zero-phase component analysis on photographs that have already been processed. The ZCA was a correlation approach that was used to throw out old and new order information, and the training phase was affected by the acquisition of new order information in advance. ZCA whitening will typically identify the set of uniformly specified filters that transform w to m, as well as the outcomes of the image processing. This is done so so that the whitening can be performed.

3.3 Classification

After acquiring a collection of qualities, one can utilize those features again and over again to classify a wide variety of different subclasses. The method was entirely predicated on feature learning, and throughout each and every one of the tests, a solitary classification algorithm was utilised in conjunction with this algorithm-weighted softmax regression classifier. We began by calculating the network weight by progressively pretraining the network layers. After that, we raised the available restrictions in an effort to locate the most effective network model combination of hidden nodes and layers. In addition to that, GD was put to use in order to train the best parameter of the SR model, and SR was then put to use in order to classify aspects of facial expressions. In the end, the back-propagation, or BP, method was employed to fine-tune the RNN overall weight in order to improve both the performance and the robustness of the network.

The Softmax algorithm, which was established through the process of logical classification, is utilized in the multi-class classification that is performed by the approach. In a greater degree, the Logic Classifier sidestepped the challenges of nonlinear classification and performed exceptionally well when it came to resolving issues involving binary classification. During this phase of the sorting process, the output probability is calculated, and a threshold is applied in order to determine the final class. This results in the task becoming a binary classification problem, in which the probability that was attained is compared to the threshold.

4. Results and Discussions

The classification results on a number of different large datasets are presented to assist in making a comparison between the suggested method and other common learning strategies. Several different amounts of aggregated data sets are utilized here in order to validate the results. Because it contains a vast library with 100,000 unlabeled examples, unsupervised learning works exceptionally well with the STL. It was decided to use the unlabeled STL-10 subset as the replacement for the initial training data.

We raised the image size from 32 by 32 pixels per pixel to 64 by 64 pixels per pixel so that we could acquire a number of shown items that was equivalent to the number of shown items in other datasets that are being studied on CIFAR-10. One more Caltech-101 dataset contains 102 classes, consisting of 101 object classes and one background class. In addition to it, a second dataset known as Caltech-256 has been collected. The MATLAB software functions as a copy of the work that was meant to be done. It is possible to find all of the input data as well as the output data in a single Mat file, which makes it an excellent choice for processing massive volumes of data. During the course of the simulation, a steady stream of relatively insignificant amounts of data will be saved and reported as being loaded into the memory of the system.

When evaluating the effectiveness of various machine learning methodologies, some of the approaches that are employed include false negative, true negative, false positive, and true positive.The classification of the procedure proposed is not less than the CNN technique and the MCCNN. The evaluation shows that the combination of unsupervised feature learning and controlled fine tuning can significantly improve execution.

Performance Metrics	Training	Testing	Validation
Accuracy %	99.2224	97.8488	97.5054
Precision	0.9898	0.9494	0.9191
Recall	0.9898	0.9797	0.9393
F-measure	0.9999	0.9696	0.9595

Table 1: Analysis of accuracy

This method is not ranked lower than the CNN method or the MCCNN method in the classification hierarchy. The findings of the study indicate that performance can be significantly improved by combining unsupervised feature learning with monitored fine tuning. This was demonstrated by the fact that the performance was much improved.

5. Conclusions

The DNN technique provides conclusive evidence that the high-quality findings that were reached through the examination of the four datasets may be trusted. We have demonstrated that the components of heightened attention and thick extraction can be as fundamental as the unsupervised research computation itself. In addition, we have shown that the central conclusion that heightened attention and thick extraction are significant is supported by our findings. The proper response to this question is that the highlights are no longer visible as the layer rises. This is because the highlights are a component of the heterogeneous information.

Deep learning has the potential advantage over manufacturing algorithms and other traditional forms of machine learning in that it can potentially find a solution to data assessment and learning challenges in the massive volume of input data. This is in comparison to other traditional forms of machine learning. In particular, it makes it easier to distinguish between complex data representations and the vast amounts of unsupervised data. The BDA toolbox is built on the foundation of information research that is derived from huge raw data collections that are often unregulated and unlabeled. This model makes use of a technique known as deep learning so that it can learn from extremely large amounts of data.

Accuracy and misclassification are two variables that are utilized in the process of comparing the executions. The weighted Softmax regression model achieves superior accuracy in classification for a deep recurrent neural network than either the multimodal deep study technique or the deep calculation strategy. This is the case for both of these methods together. This work has been expanded to incorporate the execution of the deep learning framework so that the grouping and error detection activities can be carried out simultaneously.

References

[1] Sarangi, S., Sahu, B. K., & Rout, P. K. (2022). An Advanced Fault Detection Technique for DG Integrated Microgrid Using Fast Fourier Discrete Orthonormal Stockwell Transform-Based Hybrid Optimized Kernel Extreme Learning Machine. *Iranian Journal of Science and Technology, Transactions of Electrical Engineering*, *46*(2), 329-351.

[2] Alissa, K., H. Elkamchouchi, D., Tarmissi, K., Yafoz, A., Alsini, R., Alghushairy, O., ... & Al Duhayyim, M. (2022). Dwarf Mongoose Optimization with Machine-Learning-Driven Ransomware Detection in Internet of Things Environment. *Applied Sciences*, *12*(19), 9513.

[3] Zhang, J., Qu, Z., Chen, C., Wang, H., Zhan, Y., Ye, B., & Guo, S. (2021). Edge learning: The enabling technology for distributed big data analytics in the edge. *ACM Computing Surveys (CSUR)*, *54*(7), 1-36.

[4] Uppal, M., Gupta, D., Juneja, S., Dhiman, G., & Kautish, S. (2021). Cloud-based fault prediction using IoT in office automation for improvisation of health of employees. *Journal of Healthcare Engineering*, *2021*.

[5] Danthuluri, S., & Chitnis, S. (2021). Energy and cost optimization mechanism for workflow scheduling in the cloud. *Materials Today: Proceedings*.

[6] Ning, P., Zhou, Z., Cao, Y., Tang, S., & Wang, J. (2021). A Concurrent Fault Diagnosis Model via

the Evidential Reasoning Rule. *IEEE Transactions on Instrumentation and Measurement*, *71*, 1-16.

[7] Thilagaraj, M., Dwarakanath, B., Pandimurugan, V., Naveen, P., Hema, M. S., Hariharasitaraman, S., ... & Govindan, P. (2022). A Novel Intelligent Hybrid Optimized Analytics and Streaming Engine for Medical Big Data. *Computational and Mathematical Methods in Medicine*, *2022*.

[8] Sreedevi, A. G., Harshitha, T. N., Sugumaran, V., & Shankar, P. (2022). Application of cognitive computing in healthcare, cybersecurity, big data and IoT: A literature review. *Information Processing & Management*, *59*(2), 102888.

[9] Sahoo, S. (2021). Big data analytics in manufacturing: a bibliometric analysis of research in the field of business management. *International Journal of Production Research*, 1-29.

[10] Zhang, Z., Huang, W., Liao, Y., Song, Z., Shi, J., Jiang, X., ... & Zhu, Z. (2022). Bearing fault diagnosis via generalized logarithm sparse regularization. *Mechanical Systems and Signal Processing*, *167*, 108576.

[11] Rao, G. S., Armstrong Joseph, J., Dhiman, G., Mohammed, H. S., Degadwala, S., & Bhavani, R. (2022). Novel big data networking framework using multihoming optimization for distributed stream computing. *Wireless Communications and Mobile Computing*, *2022*.

[12] Sharma, S., Gahlawat, V. K., Rahul, K., Mor, R. S., & Malik, M. (2021). Sustainable innovations in the food industry through artificial intelligence and big data analytics. *Logistics*, *5*(4), 66.

[13] Yao, Y., & Shekhar, D. K. (2021). State of the art review on model predictive control (MPC) in Heating Ventilation and Air-conditioning (HVAC) field. *Building and Environment*, *200*, 107952.

[14] Zhou, X., Xu, X., Liang, W., Zeng, Z., & Yan, Z. (2021). Deep-Learning-Enhanced Multitarget Detection for End–Edge–Cloud Surveillance in Smart IoT. *IEEE Internet of Things Journal*, *8*(16), 12588-12596.