AL-RYADA JOURNAL FOR COMPUTATIONAL INTELLIGENCE AND TECHNOLOGY (ARJCIT)

VOL. 1 (1) - 2024

Enhancing Big Data Processing Performance using Cutting-Edge Deep Learning Algorithms

Amira Hassan Abed 0000-0001-6617-0596

Business Information systems Department, Faculty of Business Administration, Al RYADA University for science and technology, Cairo, Egypt.

Article history:

Received: 29 - July - 2024

Revised: 09 - Oct - 2024

Accepted: 31 - Oct - 2024

Available online: 01 - Dec - 2024

10.21608/ajcit.2024.307230.1001

Correspondence

Amira Hassan Abed,

Business Informatics Department, Faculty of Business Administration, Al Ryada University for science and technology, Cairo, Egypt
Email:mirahassan61286@gmail.com

ABSTRACT:

The vast quantity of data (Big Data) that is being gathered as a result of the latest developments in networks of sensors and Internet of Things technologies is known as big data (BD). Investigating such large volumes of data requires more efficient methods with excellent analytical precision. Conventional techniques significantly limit the ability to process vast volumes of data in real time. With BD's analytics solutions, deep learning (DL) has begun to take center stage during the last few years. In terms of BD analytics, DL provides more scalable, fast, and accurate outcomes. It has shown previously unheard-of success in disciplines such as speech recognition, computer vision, and natural language processing. Due to its ability to extract complex high-level representations and data scenarios—particularly unsupervised data from large volumes of data—DL is an intriguing and practical tool for BD analytics. Despite this interest, there is not any structured review encompassing DL techniques for BD analytics. The purpose of this survey is to review the BD informatics studies conducted using DL techniques. The possible application of big data based DL is examined in a number of studies that provide extremely accurate analytical results.

KEYWORDS

Machine learning (ML); Big Data analytics (BDA); Deep learning (DL);

1. INTRODUCTION

In recent years, the two of the most active fields in research in science and technology have been DL and analytics for large amounts of data. BD refers to digital data that is hard or impossible to manage and evaluate with conventional tools and technologies [1]. Data analysis and information extraction are essential for the growth of science, national security, healthcare, and workplace decision-making. The demand for real-time analysis of data gave rise to BD analytics. BD analytics is the process of extracting insightful information from vast volumes of data to make the best decisions possible. Data volume has grown dramatically over the last 10 years due to the emergence of new technologies like social networks, cloud computing, and the Internet of Things.

The ever-increasing quantities of data and its potential for all areas of society provide challenges for both data processing and data mining [2]. The traditional machine learning algorithms significantly limit the capacity to handle massive volumes of data in real time and produce conclusions which can be acceptable with high accuracy. Different types of data can be handled with the use of Deep learning (DL), especially given that they can tackle data that has been labelled as well as not. DL is an intriguing field of study within AI. The machine learning methodology known as DL uses both unsupervised and supervised techniques to automatically learn patterns having hierarchies in deep architectures. It has achieved unprecedented success in application in critical sectors such as computer vision and natural language analysis [3]. Due to the fact that it can generate abstractions at high levels from large amounts of data, especially unstructured data; DL is an attractive technique for big data analytics [4]. More specifically, selective analysis, data tags, conceptual indexation, and fast retrieving are among the BD analytics problems that DL may aid with. A wide range of BD analytics problems, such as quickly varying streaming data, widely distributed input sources, poor and inadequate information, large the number of dimensions, methodology scaling, uncategorized data, and variances in the initial data design, must also be solved using DL approaches.

In this review, the authors defined BD and its characteristics in section 2, then the DL and its importance was defined in section 3. After that, the significance of DL techniques for BD analytics was discussed in section 4. The DL architectures were covered in section 5. Finally, the review for some of existing studies that applied DL techniques for improving the performance of BD analytics was reported and discussed in section 6.

2. BIG DATA AND ITS CHARACHTERISTICES

The rise of Big Data has been facilitated by advancements in computational capability, volume of data availability, as well as capacity for storage. Six important topics are focused on by most of the systems currently in use to solve the challenges presented by BD: volume, velocity, variety, veracity, validity, and volatility. The initial term is volume, which suggests the writers have to deal with enormous amounts of information that most conventional methods are unable to handle. For example, every minute, fifteen hours of videos are uploaded to Facebook, resulting in a daily data accumulation of about fifty terabytes. In light of the daily volumes of data generated, the scientists are able to predict the increasing pace of data collection over the next years [5]. The

data is growing at a rate of forty percent per year. Every year, approximately 1 ZB of records are produced. Large companies recently have begun to leverage the benefits of massive amounts of data. High volume is a relative indicator that depends on the current state of the organization and cannot be specified in a predefined method [6]. The authors must deal with an extensive variety of formats for documents, especially unorganized varieties such PDF documents, electronic mail, recordings, and so forth. This is the second challenge. This data should be consolidated for use in subsequent procedures [7]. The startling rate when data keeps accumulating is represented by the third parameter V, velocity, that's got the ability to abruptly suspend the system. It is shown that real-time solutions are necessary. The next two versus, validity and veracity, are closely related: outcome data must be valid and mean data must be as pure, dependable, and valuable as possible for use in further processing stages. The greater number of data streams and styles that are, the harder it is to maintain confidence [8]. The volatility serves as a guide for the optimal duration of data retention in the system. Value was introduced as the eighth V [9] and represents the amount of hidden information in BD.

The six attributes—availability, capacity, credibility, variability, optimizing resources, and velocity—can likewise be applied to the analysis of open research challenges. The authors in [10] mentioned a few problems and unresolved research concerns relevant to the BD management features of variability and manageability. The additional criteria covered in the study [11] are honesty and availability. These parameters are defined as follows:

- **Availability:** This speaks to the notion that users should always have access to information, wherever they are and whenever there is a malfunction. Data analysis methods should be able to process large amounts of data quickly while also offering support for huge quantities of data [12].
- **Scalability**: it is the degree to which a system can effectively handle growing volumes of data. Since 2011, scalability has been a major concern for industrial applications that need to run efficiently on a little amount of memory.
- **Data Integrity:** Indicates the correctness of data. When many users with varying levels of privilege alter data stored in the cloud, the issue gets worse. Cloud is in demand with handling databases. For data integrity, users must so abide by cloud policy [13].
- **Heterogeneity:** highlights the presence for three distinct groups of data: loosely organized, unorganized, and organized [14].
- **Optimization of resources:** it is the efficient use of resources that are already in place. It takes a strong resource efficiency policy to guarantee that BD is accessible to everyone.
- **Velocity:** the speed at which current data is generated and examined. The rate at which data is created is increasing due to the widespread use of digital devices like smartphones. Real-time analyses are thus necessary. These can differ significantly based on the application; therefore they could not be the same for every application. From a stage standpoint, the BD area can also be divided into three main Phases: BD preparation is the procedure of taking specific initial actions regarding records for preparing it, which involves cleaning and various preliminary processes. BD storage is the term for data storage done well. Best practices for managing data to accomplish a variety of objectives, including as classification, grouping, and so on, are referred to as BD organization and processing [15].

3. DEEP LEARNING

One of the most well-liked areas of machine learning (ML) is deep learning (DL), which finds use in almost every sector that handles big data. DL is a possible research direction for higher levels abstractions automated complex feature extraction. DL is the process of learning several levels of representations and forms that help with the understanding of written, recorded, and photographs. DL systems stand out in part because of their ability to train on unlabeled input. We may detect transitional or abstract visualizations by using hierarchical unsupervised learning. More complex properties are determined at each level using the lower qualities as a basis. It can improve the results of classification modeling and has a considerable ability for generalization learning. One application of DL is the extraction of an individual's invariant qualities from an image. It produces more insightful knowledge from unprocessed data and is referred to be our observation variety. In addition, it makes use of a hierarchical, tiered learning architecture that generates more intricate data representation. It layers irregular extractors of features to improve machine learning outcomes, such as an enhanced classification model and a constant property of data representation. Excellent outcomes have been attained in a variety of applications, such as vision for computers, speech recognition, public opinion-based election debate winner prediction, quicker analysis and forecasting of traffic jams in crowded areas, and the identification of a novel mechanism influencing intricate traffic systems. It is challenging for most traditional ML techniques to extract non-linear patterns. DL generates relationships and learning patterns that transcend surrounding relationships. DL not only provides complex data representations but also separates computers from humans [12]. It extracts beneficial details from unstructured information without the help of humans. In short, DL is composed of successive layers that generate local abstracts. Each layer modifies the input in a nonlinear fashion, and the final layer produces a complex abstract picture of the data. greater the number of layers of data we analyze, the more detailed and complicated our representation gets.

4. BIG DATA APPLICATIONS USING DL

The standard activities in the BD procedure for application include the generation of data, handling of data, analysis of data, and data application. BD analytics, or the art of finding patterns in data, is considered to be the most important stage of the process. Due to a variety of problems, BD analytics have grown significantly more difficult and complex than regular-sized data analytics. Large dimensionality, algorithm adaptability, fast-moving data in streams, poor and inadequate data, and other problems are a few of these [20]. In this section, the author examined and evaluated the challenges and potential fixes associated with DL for BD analytics.

a) visualization of data

BD obtained from many domains using a multitude of formats. The densities and representations of each modality differ. Using existing methods, handling such data is almost impossible. The integration of diverse data renders the response to this dilemma viable. DL is appropriate for heterogeneity integration of data since it may identify data variation elements and offer abstract representations for them. DL has demonstrated the ability to successfully incorporate data from several sources [3]. Several deep learning models have been

proposed for integrating heterogeneous data. The research [21] developed a multi-modal deep learning network that learns representations using audio and video data. The authors in ref [22] developed approach for textual and image learning.

b) Noisy and poor-quality data

There are tones of erratic, inaccurate, noisy, and incomplete items in BD. This kind of low quality data is abundant in BD. More than ninety percent of the attribute variables for a physician's diagnosis are lacking in the healthcare industry. It is clear that a large number of traditional learning methods are inapplicable when dealing with data that contains 90% missing values. In recent years, some methods for learning features from low-quality data have been proposed. In[23], a autoencoder architecture was presented in order to obtain reliable features for input that was damaged. The model did a great job at de-noising and restoring pictures. The authors in [24] proposed an incredibly wide completely convolutional auto-encoder model for visual restoration. Due to its reliance on multilayer techniques, the main limitation of this strategy is the local nature of the returned information.

c) Super-high dimensionality

BD is often fairly high dimensional in some sectors. In general, the dimensions of the data expand dramatically with the required time or memory. The problem lays in the fact that existing machine learning and data mining algorithms are not extensible to data with high dimensions (such as photographs) or are computationally wasteful. In reference [25], a brand-new tensor-based representation method for image classification was introduced. By requiring the user to acquire the parameters convolution for image tensors, the technique maintains the spatial details of the image. Furthermore, CNNs are effective in scaling up to high-dimensional input. State-of-the-art performance was attained by CNNs on 256 by 256 RGB images using the ImageNet archive [4].

d) Unsalable computation ability

Many machine learning techniques don't work well on large datasets since they typically have many samples and a large number of features. Several large-scale deep learning models have been developed to acquire attributes and interpretations for vast amounts of data. They can be broadly categorized into three groups: parallel DL designs, improved DL designs, and GPU based implementation [26]. Existing DL systems often make advantage of data or model parallelism; nevertheless, these methods often result in insufficient parallelization performance. The authors proposed FlexFlow, a DL system that automatically finds efficient parallelization alternatives for DNN applications [27]. Six practical DNN experiments were used by the investigators to test FlexFlow across two graphics card clusters. The findings show that FlexFlow outperforms even the most sophisticated parallelization strategies.

e) Fast moving streaming data

One of the hardest things about BD analytics is managing streaming and dynamic input data. The data stream's distribution qualities are dynamic and rapidly changing, making real-time processing necessary because of how quickly they are generated. While approaches capable to handle tremendous amounts of continuous input data are needed, DL is employed to manage streaming data. In recent years, numerous incremental learning methods have been created for high-speed feature learning. The author in ref. [28] proposed a gradual feature learning method to determine the most suitable model complexities for substantial datasets, based on the denoising autoencoder. In a large-scale online setting, the model quickly converges to the optimal number of features. Furthermore, the technique can also be adept at spotting new patterns when the distribution of data in the massive internet data flow varies over time. It was demonstrated in [29] that the Adaptive Deep Belief Network could learn from real-time, non-stationary stream data.

5. DEEP LEARNING ARCHITECTURES

Deep architectures use a set of machine learning techniques called DL to learn several levels of representation. In the last few years, a large number of DL architectures have been successfully developed. The many DL architectures that are commonly used in data analytics are briefly summarized here.

5.1 Autoencoder and Stacked Autoencoders (SAEs)

The most prevalent kind of feed-forward neural networks (Autoencoders), are the building blocks of the SAEs (DL technique) [26]. An autoencoder is a kind of unsupervised learning that consists of three layers: input, hidden, and output. The auto-encoder training procedure consists of two stages: encoding and decoding. An encoder maps the input data into a hidden representation, while a decoder reconstructs the input data from the hidden representation. The two phases of training that SAE goes through are pre-training and fine-tuning. Throughout pre-training, each autoencoder structure proceeds according to unstructured layer-by-layer training procedure from the bottom to the top. This approach is done while all hidden layer properties are learned. Once every hidden layer have been trained, the backward propagation methodology is employed to fine-tune by updating the weights using labelled sets of training data and lowering the cost function [30].

5.2 Deep Belief Network (DBN) and Restricted Boltzmann Machines (RBMs)

The deep belief network is the most widely used and extensively taught architecture in DL [31]. The RBM is the Boltzmann machine type that is most frequently utilized [26]. The RBM is a randomized graphical structure that is a type of stochastic neural network. The network is divided into two parts: the hidden layer and the visible layer. The drawback is that units within the same layer do not interact; connections are only formed between units from other levels. DBNs can be trained on data that is unstructured or structured to represent features. It consists of three layers: output, hidden, and input. RBM builds a two-layer model that includes full communication between layers using DBN. In DBN, supervised fine-tuning and unsupervised pre-training were combined. The unsupervised stages seek for patterns in data distributions without using label information,

whereas the supervised stages conduct local searches for fine tuning [26]. The DBN model is widely used by academics in the literature to efficiently and reliably evaluate vast amounts of data. Part of the solution is a GPU based method that handles enormous amounts of data while cutting down on processing time by using stacked RBM in parallel. What makes DL so powerful is its capacity to train and control thousands of parameters at once. Countless limited Boltzmann systems can be stacked together to form DBN.

5.3 Convolutional Neural Networks (CNNs)

The CNN is a feed-forward, multilayer neural network that employs perceptrons for data analysis and supervised learning. It is mostly used to visual data, such picture categorization. Compared to other neural networks, the architecture of CNN is distinct. Convolutional, sub-sampling, or pooling, and fully linked layers make up CNN's hidden layers. CNNs typically begin with a convolutional layer that receives input layer data. Convolution operations with a small number of identically sized filter maps are handled by the convolutional layer. While sub-sampling is employed to reduce dimension, the convolutional layer performs the convolution process to accomplish weight sharing [32]. To lower the feature map's dimension, a sub-sampling (or pooling) layer is typically applied after the convolutional layer. Usually, a max pooling procedure or an average pooling action might be used to achieve it. CNN employs a fully connected layer and a softmax layer with output classes for recognition and classification after the second stage. In the past few years, CNN has made significant progress in a variety of applications, including text comprehension, speech recognition, picture analysis, and more [26].

5.4 Recurrent Neural Networks (RNNs)

RNNs are thought to be an additional class of deep networks that are particularly effective in modeling sequence data, such as voice or text, for both supervised and unsupervised learning. RNN uses its internal neural network state, which stores a recollection of prior inputs, to learn features for the series data. A guided cycle is used to build the connections between neurons**.** The recurrent neural network integrates the prior hidden representation into the forward pass, capturing the dependence between the current sample and the previous one, in contrast to classic networks where inputs and outputs are independent of one another. Theoretically, arbitrary-length relationships can be captured by recurrent neural networks. Recurrent neural networks, however, have trouble capturing long-term dependencies since the gradient vanishes when they utilize the back-propagation technique to train their parameters. Some models, including LSTM, have been proposed to address this issue by stopping the gradient from diminishing or from exploding [26]. In several applications, including machine translation, speech recognition, and natural language processing, the RNN and its variations have demonstrated exceptional performance.

6. DEEP LEARNING ENHANCING BD PERFORMANCE

In this section the author introduced a number of significant existing academic studies and researches *(as shown in Table 1)* that applied advanced DL techniques and algorithms in many areas (such as medical, finance, smart cities, smart grids....etc) for the propose of improving the results accuracy of BD analysis processes.

In study [33], they suggested a better method for predicting the speed of urban traffic that incorporates DL and input-level data fusion (Transport for Greater Manchester dataset that include 22000 records). The authors suggest a LSTM-NN for speeding traffic prediction that integrates weather and traffic information on a city's roadway system in Greater Manchester, United Kingdom, and is inspired by DL prediction techniques. Comparing the experimental findings to highway-only sources of data for the speed of traffic prediction validates the usefulness of the technique. Linqi Zhu et. al. [34] in order to create a porosity evaluation model, they first offered a novel way to create unlabeled logging huge data. From there, they developed a semi-supervised DL technique appropriate for calculating the porosity of deep-sea sediments that contain gas hydrates. The process of creating huge data logs by expanding 380 initial data samples into 2280 labeled samples and 60050 unlabelled samples lowers the amount of sediment forms in deep sea that require expensive monitoring. The evaluation's findings demonstrate that the model not only outperforms other techniques in the inspection wells that match to the training wells' locations, but it also performs very well in wells that are not included in the modeling. The average relative error of porosity prediction is under four percent when compared to conventional prediction techniques. It offered a fresh concept for the assessment of deep-sea hydrate sediment reserves using intelligent logging.

Mohammed A. [35]: utilizing the NARX neural network framework built around a limited and BD of symmetrical volatility information, he sought to use three DL approaches for daily accuracy improvement prediction (based on Website dataset of Investing.com that include about 18000 records) for the JKII prices. The best DL strategy for forecasting daily accuracy increase in the study is the non-linear autoregressive exogenous (NARX) neural network, which is chosen based on the training and testing criteria with the greatest accuracy score. The results of the experiment show how the LM technique provides the most efficient network solution for the method of prediction together with 24 neurons in the hidden layer throughout a delay parameter of 20. This is in agreement with the MSE and the coefficient of correlation criteria. The most excellent possible prediction accuracy is obtained as a consequence. Lin Wang et al. [36]: in order to obtain intelligent categorization and representation of particular human motions included in video sequences, they merged information linked to BD-based visual analysis with DL. In order to classify sports footage, this article primarily uses an autonomous narrative built around LSTM networks. LSTM structures are employed in the conventional video descriptions model S2VT to learn the mapping connection between sequences of words and video frame sequences. The experimental findings demonstrate that the refined technique presented in this study can raise the categorization accuracy of sports videos. Dabeeruddin S. et al. [37]: the authors of this paper suggested a unique hybrid clustering-based DL technique with improved scalability for STLF at the level of distribution transformers. It looks at the accuracy performance and training time gain while using clustering-based DL modeling for STLF. A subset of a thousand transformer substations from the Spanish distributing power grid data, which includes over twenty-four million load records, is used to evaluate the correctness of the

suggested modeling. The findings show that, when compared to non-clustering models, the suggested model performs better, saving approximately forty-four percent of the training time with retaining accuracy while employing single-core processing.

Mahzad M. et al.'s work **[38]** used a BD-aware DL approach to create an effective Intrusion Detection System (IDS) that could handle these difficulties. They created a particular LSTM architecture, and this model is capable of identifying intricate connections and long-term dependencies between incoming packets of data. By doing this, they may raise the intrusion detection system's accuracy and lower the quantity of false alerts. The suggested approach, dubbed BDL-IDS, performs better than other IDS schemes with respect to of rate of detection (20%), rate of false alarms (60%) accuracy (15%) and training time (70%) compared to other IDS schemes like classical ML along with Artificial Neural Network.

Khadijeh A. at el. [39]: they build Models for estimating yield that aid in lowering energy usage, boosting output, and estimating labor needs for harvesting and storing. This study looked at how well RNN models could forecast tomato and potato harvests employing information about the climate and irrigation intensity. For the purpose of predicting agricultural output, sandy loam soil was used to train the LSTM, GRU, as well as the derived BLSTM and LSTM models. The findings indicate that on the validation set, the application of BLSTM models performed better than the basic models. Furthermore, the BLSTM's yield prediction performance was compared to that of the CNN structure, MLP, and RF. It was discovered that the BLSTM beat the MLP networks, CNN, and RF. Researchers in work [40], developed a classifier that can discriminate among both negative and positive corona-positive X-ray images in order to combat the COVID-19 pandemic. This research applies a Deep Transfer Learning (DTL) approach employing CNN three structures - InceptionV3, ResNet50, and VGG19 - on COVID-19 chest X-ray images using the Apache Spark system as an extensive data framework. With 100% accuracy, the three algorithms are assessed in two classes: COVID-19 and typical X-ray pictures. However, for COVID/Normal/pneumonia, the inceptionV3 model had an accurate detection rate of 97 percent, the ResNet50 model had a detection accuracy of 98.55 %, and the VGG19 model had a detection accuracy of 98.55%.

The proposed Fog BD evaluation framework and BPNN analysis method for IoT sensor deployment applying FDL by the work [41] have created new hurdles for potential machine-to-machine communication strategies. By applying their proposed FBDAM on the most significant Fog applications developed on smart city data—parking, transportation, safety, and IoT data sensing—they have enhanced the results. Their research indicates that in order to benefit and add value for IoT users, FDL's potential have to be fully leveraged. In order to incorporate multisource heterogeneous BD and predict PM2.5 over the vast state of California, which has significant variations in carbon dioxide emissions, the climate, and fire events, the work [42] introduced an ensemble DL technique. they constructed a PM2.5 prediction system with predictions that are uncertain at a high temporally (every week) and geographical (1 kilometer × 1 kilometer) precision for a ten-year timeframe applying DL employing BD integrated from numerous sources. They employed autoencoder-based full RDNs to simulate complex multidimensional interactions involving emissions and transmissions and dispersal parameters, and other relevant factors. Ensemble DL demonstrated its ability to predict PM2.5, with a experimental RMSE of 2.29.

As presented in reference [43], in order to enhance the quality of IoT data gathered concurrently from various sensors, it is advised that LSTM be tailored to the accuracy of each individual sensor. In the experiment using the dataset, the LSTM building method showed a low rate of errors in each sensor. In some data trials, LSTM construction algorithms showed an acceptable degree of inaccuracy in 95 sensors. It is suggested that creating a LSTM is more predictive in both cases that integrating inputs at the same time. Depending on the manner of input, the error rate increases around twenty-nine percent to forty-two percent. This demonstrates that developing and using LSTM with separate input of obtained data is advised achieving better lasting reliance results. Four complementary state-of-the-art technologies—graphics processing units (GPUs), memory-based computation, DL, and BD—were merged by researcher in [44] to produce a novel and all-encompassing approach to large-scale, faster transportation planning. They trained deep networks using over eleven years of Caltrans data, the largest dataset ever employed for research. During learning and estimation, alternative network topologies of networks constructed using DL were tested while working alongside a variety of each of the input attributes. Using the trained CNN model for real-time prediction resulted in higher accuracy for prediction when compared to other approaches. The authors addressed the UNM architecture design based on IoT in [45]. A standard stereophonic microphone can be connected to the Raspberry Pi 4 in order to record background noise for the system. They have used Python to classify a wide variety of sound types. Lastly, adding the observed event details to the cloud-based Firebase database. The CNN model which enhanced and adjusted data for more efficient performance has been used to train the system. With ninety-five classification accuracy, UNM had the highest rating. The authors also attempted the recommended method for instantaneous prediction, and 48 out of 50 experiments were successful.

In addition to unique threats, authors in [46] proposed a model with a higher detection rate and smaller rate of false positives than existing intrusion detection systems for recognizing situational and shared security attacks. To get these kinds of outcomes, they employ an LSTM in their networking Chabot. They built a model using many thousands of particles in their environment, explaining the intangible regularities in the internet and evaluating them in almost real time to identify contextual, point, and collective irregularities. Experiments are conducted using the MAWI dataset, and the findings show a greater rate of recognition versus both sign and traditional anomaly IDS. The authors in [47] trained and evaluated a profoundly DD-ResNet to segment the real clinical volume of cancer treated with breast radiation therapy in a fast and reliable manner utilizing BD. Radiation oncologists with expertise verified the CTV. A fivefold cross-validation was employed to assess the model's efficacy. The precision of segmentation was assessed using a dice-based Similarities Coefficient. In contrast to the other networks, both of which DD-ResNet exhibited higher mean DSC values (0.91 and 0.91), respectively. The LSTM model was utilized by authors in ref. [48] to forecast algal blooms in four of the principal rivers in South Korea. Short-term (one-week) projections were made using regression modeling and DL techniques using a newly produced database on the quantity and quality of water that was obtained through sixteen dam lakes on the rivers. Three different deep learning models (DL) were utilized to predict chlorophyll-a, a known proxy for algal activity: MLP, RNN, and LSTM. The accuracy of the OLS regression evaluation was exceeded by each of the DL models, notably the LSTM one showing the highest predicting rate for harmful algal blooms. Our results indicate that DL and LSTM have promise regarding Bloom of algae predictions.

Table 1: review of Big Data applications using DL for improving accuracy

Reference	Task	Techniques	Dataset	Accuracy	Result
[33]	urban traffic speed prediction	LSTM-NN	Transport for Greater Manchester dataset	98.64%	LSTM-NN and temperature data resulted in improved prediction accuracy.
[34]	prediction for deep-sea gas hydrate-bearing sediment reservoirs	DBM: Deep Boltzmann Machine	Contains about 62,000 record	98%	DBM enhanced the prediction accuracy with large training sets.
[35]	predict the daily accuracy improvement for the Jakarta Islamic Index prices	Levenberg–Marquardt, Bayesian regularization and scaled conjugate gradient	Website dataset of Investing.com	MSE: 30.69891, 40.88045, 45.94484	The best network solution for the prediction process is created using the LM approach.
[36]	classification of human movements in sports videos	long- and short-term memory LSTM, GRU, BPNN	dataset of freestyle gymnastics	14.22% & 14.03%	LSTM & GRU improved the accuracy of sports video classification.
[37]	Short-term Load Forecasting in smart grids	DNNs RNN	Iberdrola dataset & energy consumption dataset.	MAPE: 7.27 & 7.18.	The applied models improved forecasting results.
[38]	Improving Intrusion Detection	LSTM	NSL-KDD Dataset	98%	LSTM optimized the accuracy of Intrusion Detection system.
[39]	estimating Crop yield	LSTM, GRU, CNN, MLP, RF	Agricultural Fadagosa dataset (www.drapc.gov.pt)	99 %, 97%, 98%, 88%, 90%	The season can be predicted with amazing accuracy using the DL.
[40]	Classifying corona-positive X-ray images into positive and negative categories	CNN architectures —InceptionV3, ResNet50 & VGG19	"Coronavirus chest x-ray images" and "Chest X-Ray (Pneumonia)"	97%, 98.55% & 98.55%	Every performance metric showed these models are 100% accurate predictors.
[41]	Improving Fog BD analysis	SVM, SVMG-RBF, BPNN, S3VM, and fusion deep learning	smart city datasets (parking, security transportation)	74.3, 88.2, 91.8, 91.8 & 92.3%	The FDL technique showed the optimal analysis results.
[42]	Calculating PM2.5 levels & the uncertainty of their projections	Full residual deep network (FRDN)	PM2.5 dataset	RMSE : 2.29	enhanced estimation of the spatiotemporal PM2.5 over a wide, diverse area.

Table 1: review of Big Data applications using DL for improving accuracy

References	Task	Techniques	Dataset	Accuracy	Result
[43]	Data quality improvement in IoT environment	Single Dimensional LSTM- multi Dimensional LSTM	100 sensors datasets	MAPE: 0.156 & 0.280	Predictive power of LSTM is higher than the method of simultaneous data input.
[44]	Road traffic prediction	CNN	PeMS vehicles & Traffic flow datasets	98.7%	CNN improved the prediction accuracy.
[45]	Real-time audio classification	CNN	Urbansound8k dataset	95%	CNN gave the best accuracy of approximately 95% for classification.
[46]	Detect security attacks	DRNN & LSTM	MAWI dataset	92%	High detection rate & better point anomalies detection.
[47]	Auto-segmentation of the clinical target volume for breast cancer radiotherapy	deep dilated residual network (DD-ResNet)	early-stage Breast Cancer dataset	95%	DD-ResNet improve the segmentation accuracy of CTV.
[48]	Prediction of Harmful Algal Blooms	ordinary least square OLS, MLP, RNN and LSTM.	dataset from 16 dammed pools on 4 rivers in South Korea	81%, 92% 98,5%, 99%	DL models out-performed the OLS regression analysis.
[49]	Predicting Infectious Disease	DNN, LSTM, ARIMA	the Naver Data Lab dataset	88%, 91%, 72%	the developed model by LSTM was more accurate.
[50]	Enhance health risk prediction	CNN, SVM, KNN, LR	AIA Vitality members dataset (800,000,000)	73.2% , 67.4% , 65.3%, 48.7%	CNN enhanced the accuracy with large training sets.

As presented in [49], this study predicts viral illnesses by varying DL algorithm and accounting for large dataset, incorporating online social networking data. In order to evaluate the effectiveness of DNN versus LSTM learning models, three infectious diseases were predicted within seven days from now using the ARIMA technique. The results show that the LSTM and DNN models can both perform better than ARIMA. Whenever it was about chickenpox prediction, the top-10 DNN and LSTM methods improved the median precision by twenty-four percent and nineteen percent, respectively. In terms of accuracy, the LSTM model fared better than the DNN model while a viral illness was spreading. They believed that the models established in this work could lower expenses to society by eliminating reporting delays from existing monitoring systems. Finally the researchers in [50] created a framework to enhance safety assessments through BD application by utilizing the upgraded fusion module and DL architectures. To generate more detailed and reliable predictions from enormous medical records, DL would be utilized in combination with expertise merging methodologies. Since, it makes it possible to repeatedly draw conclusions about higher levels from minimal data. To show how the framework was applied, an experimental system built on top of the recommended framework was developed. The results assured that using large training data, CNN increased analytical accuracy.

7. CONCLUSION

This review has looked at the application of DL and architectural algorithms towards accuracy related BD analytics problems. In contrast to typical ML techniques, an examination of relevant literature about the application of DL in many domains indicated that DL has the ability to handle a lot of the instructional and informatics difficulties that BD analytics encounters. BD has challenges given dimensions, variation, imprecise categorization, irregular dissemination, and numerous other difficulties, despite offering DL with a large amount of training data. We must find creative ways around these technological challenges and transformative solutions if we are to fully realize the potential of BD. This means that in the upcoming years, more thorough research in the field of DL is required. In the future work, the deep leaning model's performance evaluation strategies, metrics with their equations will be discussed. Also, a new studies that report the time as a performance measure in the big data analytics will be covered in the further study.

REFERENCES

[1] Naglaa S. & Amira H., “Big Data with Column Oriented NOSQL Database to Overcome the Drawbacks of Relational Databases”, International Journal of Advanced Networking and Applications (IJANA), Volume 11 Issue 5, pp. Pages: 4423-4428 (2020).

[2] Amira H. “Recovery and Concurrency Challenging in Big Data and NoSQL Database Systems”, International Journal of Advanced Networking and Applications, Volume 11 Issue 04, pp. Pages: 4321-4329 (2020).

[3] Chen, W. Big Data Deep Learning: Challenges and Perspectives. - Access IEEE, vol. 2, 2014, pp. 514-525.

[4] Amira H. " The Evaluation of Electronic Human Resources (eHR) Management based Internet of Things using Machine Learning Techniques", International Journal of Advanced Networking and Applications, Volume 16 Issue 03, pp.: 6437-6452 (2024) ISSN: 0975-0290.

[5] Amira H. & Mona N., “Business Intelligence (BI) Significant Role in Electronic Health Records - Cancer Surgeries Prediction: Case Study”, International Journal of Advanced Networking and Applications, Vol.: 13 Issue: 06 Pages: 5220-5228 (2022).

[6] Mohamed A. & Amira H. “A comprehensive investigation for Quantifying and Assessing the Advantages of Blockchain Adoption in Banking industry". IEEE. 2024 6th International Conference on Computing and Informatics (ICCI), pp. 322-33.doi: 10.1109/ICCI61671.2024.10485028.

[7] M. Zakir Ullah, Y. Zheng, J. Song et al., “An attention-based convolutional neural network for acute lymphoblastic leukemia classification,” Applied Sciences, vol. 11, no. 22, Article ID 10662, 2021

[8] Amira H. " Deep Learning Techniques for Improving Breast Cancer Detection and Diagnosis", International Journal of Advanced Networking and Applications, Volume 13 Issue 06, pp. : 5197-5214(2022) ISSN: 0975-0290.

[9] Jinchuan Chen, Yueguo Chen, Xiaoyong Du, Cuiping Li, Jiaheng Lu, Suyun Zhao, and Xuan Zhou. Big data challenge: a data management perspective. Frontiers of Computer Science, 7(2):157–164, 2013.

[10] Amira H. & Mona N., “Diabetes Disease Detection through Data Mining Techniques”, International Journal of Advanced Networking and Applications (IJANA), Volume 11 Issue 1, pp. Pages: 4142-4149 (2019)..

[11] Han Hu, & Xuelong Li. Toward scalable systems for big data analytics: A technology tutorial. IEEE Access, 2:652–687, 2024.

[12] Marwa S., Amira H., & Mahmoud A. “A systematic review for the determination and classification of the CRM critical success factors supporting with their metrics”. Future Computing and Informatics Journal. Vol:(3). pp:398-416. (2018)

[13] Amira H. & bahloul, M. (2023) "Authenticated Diagnosing of COVID-19 using Deep Learning-based CT Image Encryption Approach," Future Computing and Informatics Journal: Vol. 8: Iss. 2, Article 4.

[14] Alexandros Labrinidis and Hosagrahar V Jagadish. Challenges and opportunities with big data. Proceedings of the VLDB Endowment, 5(12):2032–2033, 2012.

[15] Amira H. A., Mona N., & Basant S. "The Principle Internet of Things (IoT) Security Techniques Framework Based on Seven Levels IoT's Reference Model" Proceedings of Internet of Things—Applications and Future ITAF 2019. Springer publisher, Part of the Lecture Notes in Networks and Systems book series (LNNS, volume 114).

[16] Katina Michael and Keith W Miller. "Big data: New opportunities and new challenges" [guest editors' introduction]. Computer, 46(6):22–24, 2023.

[17] Amira H. Abed, Faris H. Rizk, Ahmed Mohamed Zaki, Ahmed M. Elshewey. "The Applications of Digital Transformation Towards Achieving Sustainable Development Goals: Practical Case Studies in Different Countries of the World". Journal of Artificial Intelligence & Metaheuristics. Vol. 07, No. 01, PP. 53-66, (2024)

[18] Xue-Wen C. & Xiaotong L. Big data deep learning: challenges & perspectives. IEEE Access, 2:514–525, 2014.

[19] Amira A., Mona N. & Laila A. "A conceptual Framework for Minimizing Peak Load Electricity using Internet of Things", Int. J. of Computer Science and Mobile Computing, Vol. 10. No. 8. pp: 60-71. (2021).

[20] Wang, X. Learning from Uncertainty for Big Data: Future Analytical Challenges and Strategies. - IEEE Systems, Man., & Cybernetics Magazine, April 2016, pp.26-32.

[21] Marwa S., Mahmoud A., Amira H. Abed. "The Success Implementation CRM Model for Examining the Critical Success Factors Using Statistical Data Mining Techniques" International Journal of Computer Science and Information Security (IJCSIS), Vol. 15, No. 1, .p: 455 – 475 (2017).

[22] Amira H. A. " The Applications of Deep Learning Algorithms for Enhancing Big Data Processing Accuracy", International Journal of Advanced Networking and Applications, Vol. 16 Issue 02, pp.: 6332-6341 (2024).

[23] Wang R. Non-local auto-encoder with collaborative stabilization for image restoration. - IEEE Transactions on Image Processing, vol. 25, no. 5, 2016, pp. 2117–2129.

[24] Amira H. A. & Essam M.. "Modeling Deep Neural Networks for Breast Cancer Thermography Classification: A Review Study." International Journal of Advanced Networking and Applications (IJANA), Volume 13 Issue 2, pp.:4939-4946 (2021).

[25] Amira Hassan Abed, Essam M Shaaban, Om Prakash jena & Ahmed A. elngar. "A Comprehensive Survey on Breast Cancer Thermography Classification Using Deep Neural Network ", Machine Learning and Deep Learning in Medical Data Analytics and Healthcare Applications. book. routledge, CRC Press, Taylor and Francis Group Pages: 250-265 (2022).

[26] Zhang, Q. A survey on deep learning for big data. - Information Fusion, vol. 42, 2018. pp. 146–157. (Zhang Q., L.Yang, Z. Chen)

[27] Jia, Z. Beyond data and model parallelism for deep neural networks. - arXiv:1807.05358v1 [cs.DC] 14 Jul 2018, pp.1-15.

[28] Amira A., "Internet of Things (IoT) Technologies for Empowering E-Education in Digital campuses of Smart Cities.", International Journal of Advanced Networking and Applications, Volume 13 Issue 2, pp. Pages: 4925-4930(2021).

[29] Calandra, R, Learning deep belief networks from non-stationary streams. - Artificial Neural Networks and Machine Learning– ICANN 2012. Springer, Berlin Heidelberg. 2012, pp. 379–386.

[30] Amira H., Mona N., & Walaa S. "The Future of Internet of Things for Anomalies Detection using Thermography", International Journal of Advanced Networking and Applications (IJANA), Volume 11 Issue 03 Pages: 4294-4300 (2019) ISSN: 0975-0290.

[31] Amira H. Abed, Mona N., Laila A. & Laila E. " Applications of IoT in Smart Grids Using Demand Respond for Minimizing On-peak Load", International Journal of Computer Science and Information Security (IJCSIS). Vol. 19. No. 8. (2021).

[32] Ahmed M. , Sayed M. , Amel A., Marwa R. & Amira Hassan Abed. Optimized Deep Learning for Potato Blight Detection Using the Waterwheel Plant Algorithm and Sine Cosine Algorithm. Potato Res. (2024). https://doi.org/10.1007/s11540-024-09735-y

[33] Essien, A., Petrounias, I., Sampaio, P., & Sampaio, S. (2019). Improving Urban Traffic Speed Prediction Using Data Source Fusion and Deep Learning. In 2019 IEEE International Conference on Big Data and Smart Computing, BigComp 2019

[34] Linqi Z., & Shiguo W. 2022. Application of unlabelled big data & deep semi-supervised learning to significantly improve the logging interpretation accuracy for deep-sea gas hydrate-bearing sediment reservoirs, Energy Reports, Vol.:8, Pp:2947-2963.

[35] Mohammed A. (2022), "Deep learning with small and big data of symmetric volatility information for predicting daily accuracy improvement of JKII prices", Journal of Capital Markets Studies, Vol. 6 No. 2, pp. 130-147.

[36] Wang, Lin & Zhang, Haiyan & Yuan, Guoliang. (2021). Big Data and Deep Learning-Based Video Classification Model for Sports. Wireless Communications and Mobile Computing. 2021. 1-11.

[37] Syed, Dabeeruddin & Abu-Rub, Haitham & Ghrayeb, Ali & S. Refaat, Shady & Houchati, Mahdi & Bouhali, Othmane & Banales, Santiago. (2021). Deep Learning-Based Short-Term Load Forecasting Approach in Smart Grid With Clustering and Consumption Pattern Recognition. IEEE Access. PP. 1-1.

[38] Mahdavisharif, Mahzad & Jamali, Shahram & Fotohi, Reza. (2021). Big Data-Aware Intrusion Detection System in Communication Networks: a Deep Learning Approach. Journal of Grid Computing.

[39] Alibabaei K. & Lima T.2021. Crop Yield Estimation Using DL Based on Climate Big Data & Irrigation Scheduling. Energies.

[40] Awan M. 2021. Detection of COVID-19 in Chest X-ray Images: A Big Data Enabled Deep Learning Approach. International Journal of Environmental Research and Public Health.

[41] Rajawat A. 2021. Fog Big Data for IoT Sensor Application Using Fusion Deep Learning. Mathematical Problems in Engineering.

[42] Lianfa Li, Mariam Girguis, & Frederick Lurmann, Ensemble-based deep learning for estimating PM2.5 over California with multisource big data including wildfire smoke, Environment International, Vol. (145), 2020, p.p: 106-143. ISSN 0160-4120

[43] Hwang, Chulhyun & Lee, Kyouhwan & Jung, Hoekyung. (2020). Improving data quality using a deep learning network. Indonesian Journal of Electrical Engineering and Computer Science.

[44] Aqib M, & Katib I. Smarter Traffic Prediction Using Big Data, In-Memory Computing, DL & GPUs. Sensors. 2019; 19(9):2206.

[45] Shah S., Tariq Z. & Lee Y., "IoT based Urban Noise Monitoring in Deep Learning using Historical Reports," 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 2019, pp. 4179-4184.

[46] Al Jallad K. (2019). Big data analysis and distributed deep learning for next-generation intrusion detection system optimization. Journal of Big Data. 6. 10.

[47] Men K. & Zhang T. (2018). Fully automatic and robust segmentation of the clinical target volume for radiotherapy of breast cancer using big data and deep learning. Physica Medica. 50. 13-19. 10.1016/j.ejmp.2018.05.006.

[48] Lee S. Improved Prediction of Harmful Algal Blooms in Four Major South Korea's Rivers Using Deep Learning Models. International Journal of Environmental Research and Public Health. 2018; 15(7):1322.

[49] Chae S, Kwon S, & Lee D. Predicting Infectious Disease Using Deep Learning and Big Data. International Journal of Environmental Research and Public Health. 2018; 15(8):1596.

[50] Zhong H. & Xiao J. (2017). Enhancing Health Risk Prediction with Deep Learning on Big Data and Revised Fusion Node Paradigm. Scientific Programming. 2017. 1-18.