Volume: 16 Issue: 02 Pages: 6332– 6341 (2024) ISSN: 0975-0290

The Applications of Deep Learning Algorithms for Enhancing Big Data Processing Accuracy

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

----------**ABSTRACT**----------

Big Data (BD) is the massive amount of data that has been collected as a result of recent developments in sensor networks and IoT technology. More effective techniques with high analytical accuracy are required for the investigation of such vast amounts of data. The ability to analyze large amounts of data in real time is severely limited by the standard neural network and artificial intelligence algorithms. In the past several years, DL has started to take center stage in BD's analytics solutions. When it comes to BD analytics, DL can produce results that are more accurate, quicker, and scalable. In domains including natural language processing, speech recognition, and computer vision, it has achieved before unseen success. DL is an interesting and useful technique for BD analytics because of its capacity to extract high-level complicated representations as well as data scenarios, particularly unsupervised data from big volume data. To the best of our knowledge, no comprehensive survey covering all DL approaches for BD analytics exists, despite this interest. The current survey's goal is to examine the BD analytics research that has been done with DL methods. Several studies that offer very accurate analytical findings explore the potential use of DL with BD analytics.

Keywords - Big Data; Deep learning (DL); Convolutional Neural Network; Autoencoder and Stacked Autoencoder; Deep belief network; Recurrent Neural Network.

Date of Submission: July 25, 2024 Date of Acceptance: Aug 11, 2024

I. INTRODUCTION

Deep learning and big data analytics have emerged as the two most active areas of scientific and technical study in recent years. Digital data that is difficult or impossible to handle and analyze using traditional tools and technologies is referred to as BD [1]. For the purpose of national security, healthcare, scientific advancement, and making informed decisions in the workplace, data analysis and knowledge extraction are critical. BD analytics was made possible by the need for real-time data analysis. The process of gleaning valuable insights from massive amounts of data in order to arrive at the best possible judgments is known as BD analytics. In the past ten years, as new technologies such as cloud computing, IoT, and social networks have emerged, the amount of data has increased significantly. Data mining as well as data processing face challenges due to the rapidly growing volume of data and the opportunity it presents for all societal sectors [2]. The ability to handle large amounts of data in real time and provide findings that are acceptable with high accuracy is severely limited by the standard neural network and artificial intelligence algorithms.

DL skills, particularly its capacity to handle both labeled and unlabeled data, which are frequently gathered copiously in BD, can help handle various kinds of data. Within artificial intelligence (AI), DL is a compelling area of study. DL is the name given to ML approaches that automatically learn structures with hierarchy in deep architectures using both supervised and unsupervised methods. In crucial domains including computer vision, voice and audio processing, and human language processing, it has attained unparalleled triumph in application [3]. DL is a desirable technique for BD analytics because of its capacity to extract sophisticated, high-level abstractions and data visualizations from massive amounts of data, particularly unstructured data [4]. More precisely, DL may help solve BD analytics issues including discriminative modeling, data tagging, semantic indexing, and quick information retrieval. DL techniques must also be used to solve a variety of BD analytics challenges, including rapidly moving streaming data, widely dispersed input sources, noisy and low-quality data, high dimensionality, algorithm scalability, unsupervised and un-categorized data, restricted supervised / labeled data, and raw data format variations.

II. BIG DATA

Increases in computing power, data volume accessibility, and data storage capacity have all contributed to the emergence of BD. The majority of the technologies in use today to address the problems posed by BD are concentrated on six key issues: volume, velocity, variety, validity, validity, and volatility. The first is volume, which indicates that we are dealing with enormous data sets that most conventional algorithms are unable to handle. For instance, 15 hours of movies are uploaded to Facebook every minute, accumulating almost 50 terabytes of data per

Volume: 16 Issue: 02 Pages: 6332– 6341 (2024) ISSN: 0975-0290

day. We are able to forecast the rising rate of data in the upcoming years based on the daily amounts of data that are generated [5]. The growth rate of the data is 40% annually. Approximately 1.2 ZB of data are created annually. Big businesses like Twitter, Facebook, and Yahoo have just started to take use of the advantages of huge volume data. High volume is not defined in a predetermined way; rather, it is a relative metric that is dependent on the enterprise's present circumstances [6]. The second difficulty is variety, which simply means we have to deal with a wide range of file formats, including unstructured ones like PDFs, emails, audio files, and so on. For use in later procedures, this data ought to be consolidated [7]. The third V, velocity, denotes the striking rate at which data are arriving, which has the potential to quickly hang the system. The requirement for real-time algorithms is demonstrated. The following two Vs—Veracity and Validity—have a lot in common: for usage in subsequent processing stages, mean data has to be as pristine, reliable, and valuable as feasible, and result data needs to be valid. Maintaining confidence becomes increasingly challenging the more data sources and types there are [8]. The last V, volatility, indicates how long data should be kept in the system in order for it to be useful. Value, which stands for the quantity of hidden information contained in BD, was introduced by McKinsey as the seventh V [9]. Six characteristics can also be used to analyze open research problems: availability, scalability, integrity, heterogeneity, resource optimization, and velocity (which is connected to stream processing). The authors in [10] discussed a few difficulties and unresolved research issues pertaining to the BD management elements of heterogeneity and scalability. The study[11] discusses other factors including availability and honesty. The definition of these parameters is as follows:

- **Availability:** This refers to the idea that information ought to be available to users at all times and locations, even in the event of a failure. Large volumes of data should be supported by data analysis techniques, which should also provide a fast stream processing 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:** describes the existence of three distinct categories of data: semi-structured, unstructured, and structured [14].
- **Resource optimization:** it is the effective use of already-existing resources. Robust resource optimization policy is required to ensure widespread access to BD.
- **Velocity:** the rate at which new data is created and analyzed. The proliferation of digital devices such as smartphones and tablets has led to an acceleration in the rate of data creation. Analysis in real time are therefore required. These vary greatly depending on the application, therefore they may be different for every application. Additionally, the BD region may be split into three primary Phases from a stages perspective: BD preprocessing is the process of taking certain initial steps toward data in order to prepare it, such as cleaning and other preparatory measures. BD storage refers to the proper way to store data. BD organization as well as processing refers to the best practices for handling data to achieve various goals, like categorization, grouping, and so forth [15].

III. DEEP LEARNING SIGNIFICANCE

The branch of ML known as "deep learning" is one of the most popular subjects and is being used in practically every industry that uses large data. A potential line of inquiry for high-level abstraction automating complicated feature extraction is DL. Learning several tiers of abstractions and representations that aid in the interpretation of text, audio, and image data is known as DL. The capacity of DL systems to use unlabeled input for training is one of their distinctive features. Through hierarchical unsupervised learning, we are able to find intermediate or abstract representations. At each level, higher-level features are defined based on lower-level characteristics. It has a significant capacity for learning generalization and can enhance the outcomes of classification modeling. Extraction of a person's invariant traits from a picture is one use of DL. In layman's terms, it is known as our observation variety and it generates more meaningful knowledge from raw data. Furthermore, it employs a tiered, hierarchical learning architecture that produces more complex data representation. In order to improve ML outcomes, such as an improved classification model and invariant characteristic of data representation, it stacks non-linear feature extractors. Outstanding results have been obtained in a range of applications, including speech recognition, computer vision, election debate winner prediction based on public opinion, faster analysis and prediction of traffic jams in congested areas, and the discovery of a new mechanism affecting complex traffic systems. Non-linear patterns are difficult for most conventional ML methods to extract. Learning patterns and relationships that go beyond neighboring relationships are produced by DL. In addition to offering sophisticated data representations, DL also decouples robots from humans [12]. Without human assistance, it derives valuable information (features, representation) from unsupervised data. To put it simply, DL is made up of successive layers, each of which produces a local abstract. Every layer applies a nonlinear alteration to its input, and the output of the last layer is an intricate abstract representation of the data. Our representation becomes more intricate and complex the more layers of data we process. Data is transformed in a very non-linear way to get the final representation. Rather than attempting to extract predetermined representations, Dl seeks to identify invariant patterns by disentangling the variables causing variance in the data. DL models outperform shallow learning methods for learning compact representations. Due to their reduced computational requirements, the compact representations are efficient. Large volumes of data may be learned in nonlinear representations thanks to it [11].

Volume: 16 Issue: 02 Pages: 6332– 6341 (2024) ISSN: 0975-0290

IV. DEEP LEARNING IN BD ANALYTICS

Data production, data management, data analytics, and data application are typical steps in the BD application process. Finding patterns in data is known as BD analytics, and it's regarded as the most crucial step in the entire process. BD analytics are now far more challenging and complex than normal-sized data analytics due to a number of issues (such as large dimensionality, scalability of algorithms, rapid moving streaming data, noisy and low quality data, and so on) [20]. We analyzed and discussed the difficulties and potential solutions related to DL for BD analytics in this part.

a) Intricate data visualization

BD is often gathered from several domains with a variety of modalities. The representation, distribution, and density of each modality vary. It is nearly hard to handle such data using current approaches. The integration of heterogeneous data makes the answer to this challenge conceivable. Because DL can learn data variation elements and provide abstract representations for them, it is more suitable for heterogeneous data integration. It has been shown that DL is highly successful in incorporating data from many sources [3]. A few multi-model DL models have been suggested for the integration of heterogeneous data. For instance, the study [21] used audio and video data to create a multi-modal DL model that learns representations. A multimodal Deep Boltzmann Machine (DBM) for text data and picture object feature learning was created by the authors in ref. [22]

b) Noisy and poor-quality data

BD contains an enormous amount of noisy, imprecise, erroneous, and incomplete items. BD is full of this kind of poor quality data. In the clinic and health sectors, for instance, more than 90% of the attribute values for a doctor's diagnosis are missing. It is evident that many conventional learning techniques are invalid for handling data that has 90% missing values. A few techniques for learning features for low-quality data have been suggested in the last several years. A non-local auto-encoder model was introduced in [23] to acquire trustworthy features for damaged input. In terms of picture restoration and de-noising, the model performed admirably. For picture restoration, authors in [24] suggested an extremely deep fully convolutional auto-encoder network. The primary constraint of this approach is the local character of the features that are retrieved because it relies on convolutional procedures.

c) Super-high dimensionality

In certain fields, BD is frequently quite high dimensional. Generally speaking, the amount of time or memory needed increases exponentially with the size of the data. The issue is that current data mining and ML techniques are either computationally inefficient or poorly scalable to high-dimensional data (like photos). A novel tensor-based representation approach for image categorization was presented in ref. [25]. The method preserves the spatial information of the picture by having the user learn the parameter tensor for image tensors. Additionally, CNNs scale up to high-dimensional input with effectiveness. CNNs achieved state-of-the-art performance on 256x256 RGB pictures from the ImageNet dataset [4].

d) Unsalable computation ability

Several frequently employed methods for ML and data mining do not perform well in huge datasets because they usually contain a high number of characteristics and several class kinds of samples. Various enormous scale DL models have been designed to learn features and representations for massive volumes of data. They may be roughly divided into three groups: GPU-based implementation, parallel DL models, and enhanced DL models [26]. Data or model parallelism are widely used in existing DL systems; yet, these approaches frequently lead to inadequate parallelization performance. FlexFlow, a system for DL that automatically identifies effective parallelization options for DNN applications, was suggested by authors [27]. The authors tested FlexFlow on two GPU clusters using six real-world DNN benchmarks, and the results demonstrate that FlexFlow performs noticeably better than the most advanced parallelization techniques.

e) Fast moving streaming data

The handling of streaming and rapidly changing input data is one of the most difficult parts of BD analytics. The distribution properties of the data stream are dynamic and very quickly, necessitating real-time processing due to their rapid generation. DL is used to manage streaming data because algorithms that can handle massive volumes of ongoing input data are required. Many incremental learning techniques for high-speed feature learning have been introduced in recent years. Relying on the denoising autoencoder, Zhou et al. [28] suggested a progressive feature learning technique to find the ideal model complexity for large-scale datasets. In an extensive online scenario, the model rapidly converges to the ideal amount of characteristics. Furthermore, as the data distribution in the enormous online data stream changes over time, the algorithm is also good at identifying new patterns. Adaptive Deep Belief Network was shown in [29] to be able to learn from live, non-stationary stream data.

V. DEEP LEARNING ARCHITECTURES

A group of ML methods known as DL are used to learn several layers of representation in deep architectures. Many DL structures have already been created in the past few years. Below is a quick summary of the various DL structures that are frequently used in analytics of data.

Autoencoder and Stacked Autoencoders (SAEs)

SAEs are a popular DL approach that are created by stacking several autoencoders, which are the most common type of feed-forward neural networks [26]. The input layer, hidden layer, and output layer are the three layers that make up an autoencoder, a type of unsupervised learning structure

Volume: 16 Issue: 02 Pages: 6332– 6341 (2024) ISSN: 0975-0290

(Fig. 1). The encoding stage and the decoding stage are the two phases of the autoencoder training process. The input data is mapped into a hidden representation using an encoder, and the input data is reconstructed from the hidden representation using a decoder. Pre-training and fine-tuning are the two steps that SAE is normally trained through. Every auto-encoder model goes through an unsupervised layer-by-layer training process from the bottom to the top during the pre-training phase. Until all hidden layer parameters are taught, this process is repeated. The back-propagation technique is used to update the weights using labeled training sets and decrease the cost function after all hidden layers have been trained in order to accomplish fine-tuning [30].

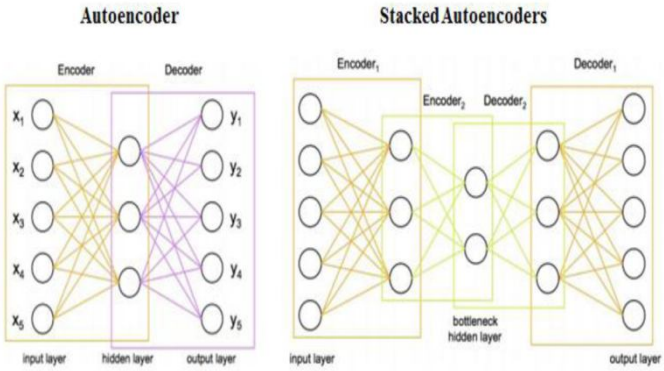

Fig. 1: Architecture of autoencoders

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

The most often used and well taught architecture in DL is the deep belief network [31]. As seen in Figure 3, many restricted Boltzmann machines stack DBN. The Boltzmann machine variant that is most widely used is the RBM [26]. A kind of stochastic neural network, the RBM is a probabilistic graphical model. There are two levels in the network: the visible layer and the hidden layer (**Fig. 2**). The limitation is that connections are only made between units from different levels; there is no interaction between units in the same layer. Structured and unstructured data may be used to teach deep belief networks how to represent features. There are three layers in it: input, concealed, and output. DBN is used by RBM to build a two-layer model with complete connectivity between them. Unsupervised pre-training and supervised fine-tuning were merged in DBN. While the supervised stages carry out local search for fine tuning, the unsupervised stages aim to learn data distributions without utilizing label information [26]. Many researchers in the literature utilize the DBN model to analyze large amounts of data reliably and effectively. A GPU-based approach that uses stacked RBM in parallel to handle massive volumes of data while minimizing processing time is used in part. DL's ability to train and manage millions of parameters at once is what gives it its power. A deep belief network can be created by stacking many limited Boltzmann machines together.

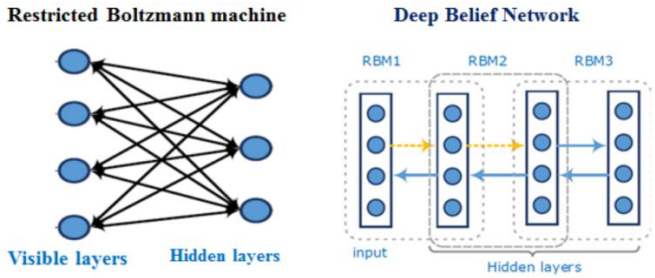

Fig. 2: Deep Belief Network Architecture

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 (Fig. 3). 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].

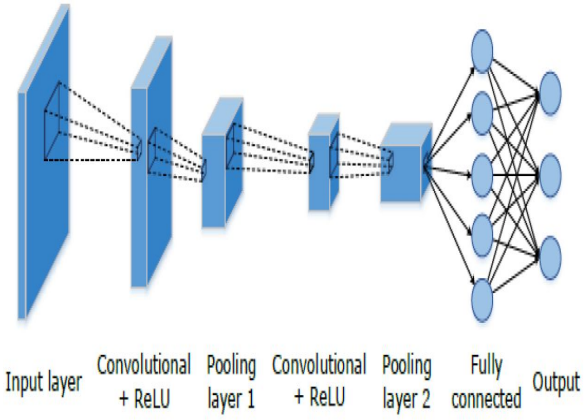

Fig. 3: Architecture of a CNN

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 **(Fig. 4.).** The

Volume: 16 Issue: 02 Pages: 6332– 6341 (2024) ISSN: 0975-0290

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.

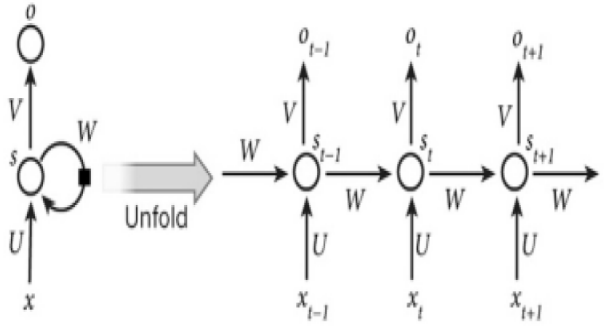

Fig. 4: Architecture of a RNN

VI. DEEP LEARNING IMPROVED BD ACCURACY

In this section we 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.

Aniekan E. et al. [33]: they suggested a better method for predicting the speed of urban traffic that incorporates DL and input-level data fusion. We 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 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.

Volume: 16 Issue: 02 Pages: 6332– 6341 (2024) ISSN: 0975-0290

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.

Javed Mazhar at el. [40]: Researchers 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%.

Anand S. et al. [41]: new challenges for possible machine-to-machine communication techniques have been raised by their suggested Fog BD analysis model (FBDA) and BPNN analytical approach for IoT sensor deployment employing fusion deep learning (FDL). They have improved the outcomes by using their suggested FBDAM to the most important Fog applications created on smart city data (places to park, transport, safeguards, and sensing IoT data). According to their research, FDL's capabilities should be fully utilized to benefit and provide value to Internet of Things consumers.

Lianfa L. at al. [42]: presented an ensemble DL technique to integrate multisource diverse BD and estimate PM2.5 throughout California, a huge region with considerable variability in emissions, terrain, weather, and wildfire events. they created a PM2.5 prediction algorithm with estimates of uncertainty at a high temporal (weekly) and spatial (1 km × 1 km) resolution for a 10-year period (2008–2017) employing ensemble-based DL using BD integrated from multiple sources. To simulate intricate nonlinear interactions between PM2.5 emission, transmission and dispersion parameters, and other important aspects, they used autoencoder-based complete residual deep networks. With an overall mean training RMSE of 1.54 μg/m3 (R2: 0.94) and a test RMSE of 2.29 μg/m3 (R2: 0.87), Ensemble DL was able to predict PM2.5.

According to **Chulhyun H. et al. [43],** when using LSTM to improve data quality in an Internet of Things (IoT) setting in which data has been collected simultaneously from multiple sensors, it is recommended that LSTM be constructed specifically for each sensor's accuracy. The individual LSTM method of construction demonstrated a low error rate in every sensor in the experiment utilizing the entire set of data. 95 LSTM building techniques demonstrated an acceptable level of error in 95 sensors in certain data trials. It is proposed that in both scenarios, building an individual LSTM has a better predictive capability than importing data all at once. In example, contingent upon the input method, the error rate rises from 29% to 42%. This shows that for better long-term dependence results, building and employing LSTM with distinct input of acquired data is recommended.

Muhammad A. [44] combined four complimentary cutting-edge technologies—BD, DL, memory-based computation, and graphics processing units (GPUs)—to create an innovative and comprehensive strategy toward large-scale, quicker, and real-time traffic prediction. They used the largest dataset yet used in deep learning research—more than 11 years' worth of data from Caltrans—to train deep networks. For training and prediction, different network designs of models based on DL were examined in conjunction with multiple combinations of the data's input properties. When compared to alternative methods, the prediction accuracy increased when the CNN model that had been trained was used for real-time prediction.

Urban Noise Monitoring (UNM) structure prototype based on Internet of Things was addressed by **Sayed Kh. et al. [45]**. The Raspberry Pi 4 can be linked to a normal stereophonic microphone to record ambient noises for the system. They have categorized many sound kinds using TensorFlow on the Raspberry Pi. Lastly, updating the Firebase cloud-based database with the reported event details. They trained our system using a 2D CNN model, which augments and normalizes data for more effective performance training. The UNM recorded the highest categorization accuracy, which was almost 95%. Additionally, we tried the UNM approach to immediate prediction, and 48/50 of the tests passed with success.

Khloud A. et al. [46]: suggested a model to identify contextual and collective security attacks in addition to novel threats with a lower false positive rate and greater detection rate than currently in use IDS. they use a DNN called Long Short Term Memory (LSTM) in Networking Chatbot to achieve such results. From a series of hundreds of thousands of packets throughout their environment, they created a model that explains the abstract typical behavior of the network and evaluates them in nearly real time to discover point, collective, and contextual anomalies. The MAWI dataset is used for experiments,

Volume: 16 Issue: 02 Pages: 6332– 6341 (2024) ISSN: 0975-0290

TABLE 1: Applications of DL in Big Data

References	Task	Techniques	Dataset	Accuracy	Result
Aniekan E. et. al. [33]	urban traffic speed prediction	LSTM-NN	Transport for Greater Manchester dataset	98.64%	LSTM-NN and temperature data resulted in improved prediction accuracy.
Linqi Zhu et. al. [34]	Predict 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.
Mohammed A. [35]	predict the daily accuracy improvement for the Jakarta Islamic Index prices	Levenberg–Marquardt, Bayesian regularization & 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.
Lin Wang et. al. [36]	classification of human movements in sports videos	long- and short-term memory LSTM, GRU, BPNN	dataset of freestyle gymnastics	14.22% & 14.03%	Improved the accuracy of sports video classification.
Dabeeruddin S. et. al. 2021 [37]	Short-term Load Forecasting (STLF) in smart grids	DNNs RNN	Iberdrola dataset & energy consumption dataset.	MAPE: 7.27, 7.18.	The applied models improved forecasting results.
Mahzad M. et. al. [38]	Improving Intrusion Detection	LSTM	NSL-KDD Dataset	98%	LSTM optimized the accuracy of Intrusion Detection system.
Khadijeh A. et al. [39]	estimating Crop yield	LSTM, GRU, CNN, MLP, RF	Agricultural Fadagosa dataset	99 %, 97%, 98%, 88%, 90%	the season can be predicted with amazing accuracy using the DL model.
Mazhar J. et. al. [40]	Classifying corona-positive X-ray images into positive & negative	CNN architectures —InceptionV3, ResNet50, & VGG19	“Coronavirus chest x-ray images” & “Chest X-Ray” datasets	97% 98.55% 98.55%	Every performance metric showed these models are 100% accurate predictors.
Anand S. at. el. [41]	Improving Fog BD analysis	SVM, SVMG-RBF, BPNN, S3VM, and fusion deep learning	(parking,security transportation & sensor IoT dataset)	74.3, 88.2, 91.8, 91.8 & 92.3%	The FDL technique showed the optimal analysis results.
Lianfa L. et. al. [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.
Chulhyun H. et. al. [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.
Muhammad A. et. al. [44]	road traffic prediction	CNN	PeMS vehicles & Traffic flow datasets	98.7%	CNN improved the prediction accuracy.
Sayed Kh. et. al. [45]	real-time audio classification	CNN	Urbansound8k dataset	95%	CNN gave the best accuracy of approximately 95%.
Khloud A. et. al. [46]	Detect security attacks	DRNN & LSTM	MAWI dataset	92%	High detection rate & better point anomalies detection.
Kuo M. et. al. [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.
Sangmok L. and Donghyun L. [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.
Sangwon Ch. et. al. 2018 [49]	Predicting Infectious Disease	DNN, LSTM, ARIMA	the Naver Data Lab dataset	88%, 91%, 72%	the developed model by LSTM was more accurate.
Hongye Z.& Jitian X. [50]	Enhance health risk prediction	CNN, SVM, KNN, LR	AIA Vitality members dataset (800,000,000)	73.2% , 67.41% , 65.32% , 48.86%	CNN enhanced the accuracy with large training sets.

Volume: 16 Issue: 02 Pages: 6332– 6341 (2024) ISSN: 0975-0290

and the results demonstrate a higher detection rate than both signature and conventional anomaly IDS.

A deeply dilation residual network (DD-ResNet) was trained and tested by **Kuo M. et al. [47]** in order to quickly and reliably auto-segment the actual clinical volume of cancer of the breast radiation therapy using BD. Radiation oncologists with experience verified the CTV. They used a fivefold cross-validation to evaluate the model's performance. The Dice similarity coefficient (DSC) was used to measure the accuracy of segmentation. Compared to each of the other two networks (DDCNN: 0.85 and 0.85; DDNN: 0.88 and 0.87), the mean DSC values of DD-ResNet (0.91 and 0.91) were greater.

Sangmok L. and Donghyun L. [48] used the LSTM model to predict algal blooms in four of South Korea's major rivers. Regression analysis and deep learning approaches were used to make short-term (one-week) forecasts using a freshly created dataset on water quality and quantity that was taken from 16 dammed pools on the rivers. Chlorophyll-a, a known surrogate for algal activity, was predicted using three DL models: MLP, RNN, and LSTM. All of the DL models outperformed the OLS regression analysis, with the LSTM model exhibiting the greatest prediction rate for dangerous algal blooms. Our findings show that LSTM and DL have the potential to be used for algal bloom prediction.

Sangwon, Ch., and others [49] By adjusting the settings of DL algorithms and taking into account BD, including social media data, this study forecasts infectious illnesses. While predicting three infectious illnesses one week from now, the ARIMA was used to compare the performance of DNN against LSTM learning models. The outcomes demonstrate that both the DNN as well as LSTM models are able to outperform ARIMA. The highest-10 DNN and LSTM algorithms increased average accuracy by 24% and 19%, respectively, when it came to chickenpox prediction. When an infectious disease was spreading, the LSTM model outperformed the DNN model in terms of accuracy. They thought that by removing reporting delays from current monitoring systems, the models developed in this work could reduce costs to society.

Hongye Zhong & Jitian Xiao [50] developed the framework to improve health prediction using the updated fusion node and DL frameworks to improve health risk predictions from BD. DL would be used in conjunction with knowledge fusion approaches to produce predictions from massive health data that are more thorough and trustworthy. because it enables the repeated inference of high-level data from low-level data. An experimental system was created based on the suggested framework to demonstrate how the framework was implemented. The outcome guaranteed shown the convolutional neural network improved the analysis accuracy with big training sets.

VII. CONCLUSION

The use of algorithms for DL and architectures to accuracy-related BD analytics challenges has been examined in this review. In contrast to standard ML techniques, an analysis of important literature about the use of DL in many domains demonstrated that DL offers the ability to address many of the learning and analytics problems that BD analytics faces. However, although providing ample training data for DL, BD also poses issues with size, heterogeneity, noisy labeling, non-stationary distribution, and many other issues. To fully utilize BD, we must overcome these technological obstacles with innovative ideas and game-changing solutions. This calls for more, in-depth research in the area of DL in the years to come.

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] Abed A. H. “Recovery and Concurrency Challenging in Big Data and NoSQL Database Systems”, International Journal of Advanced Networking and Applications (IJANA), 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, (Chen W., X. Lin)

[4] Najafabadi, M. Deep Learning applications and challenges in Big Data analytics. - Journal of Big Data, vol.2, no.1, 2015, pp.2- 21. (Najafabadi M., F. Villanustre, T. Khoshgoftaar, N. Seliya,)

[5] Abed A. & 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] Chen, W. Big Data Deep Learning: Challenges and Perspectives. - Access IEEE, vol. 2, 2014, pp. 514-525, (Chen W., X. Lin)

[8] Abed A. " Deep Learning Techniques for Improving Breast Cancer Detection and Diagnosis", International Journal of Advanced Networking and Applications (IJANA), 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”,

Volume: 16 Issue: 02 Pages: 6332– 6341 (2024) ISSN: 0975-0290

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] Abed A. & 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 and Metaheuristics (JAIM). Vol. 07, No. 01, PP. 53-66, (2024)

[18] Xue-Wen Chen and Xiaotong Lin. Big data deep learning: challenges and perspectives. IEEE Access, 2:514–525, 2014.

[19] Amira A., Mona N. & Laila A. “A concepual Framework for Minimizing Peak Load Electricity using Internet of Things”, International Journal 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 (Wang, X., Y. He)

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

[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 H., Essam M., Om Prakash jena & Ahmed A.. "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.

[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, Volume 11 Issue 03 Pages: 4294-4300 (2019).

[31] Amira H., 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. Vol. 19. No. 8. (2021).

[32] Ahmed M. , Sayed M. , Amel A., Marwa R. & Amira H. Abed. Optimized Deep Learning for

Volume: 16 Issue: 02 Pages: 6332– 6341 (2024) ISSN: 0975-0290

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.