BiLSTM_SAE:A Hybrid Deep Learning Framework for Predictive Data Analytics System in Traffic Modeling

Shubhashish Goswami (subh.goswami@gmail.com)
National Institute of Technology Uttarakhand
Abhimanyu Kumar
National Institute of Technology Uttarakhand

Research Article

Keywords: Big Data Analytics, Bi-directional LSTM, SAE, Random forest, Machine Learning

Posted Date: January 4th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2422617/v1

Additional Declarations: No competing interests reported.

BiLSTM_SAE:A Hybrid Deep Learning Framework for Predictive Data Analytics in Traffic Modeling

Shubhashish Goswami[1,3][0000-0002-6129-9822] and Abhimanyu Kumar[2]

[1, 2] National Institute of Technology Uttarakhand
subh.goswami@gmail.com
[3] Dev Bhoomi Uttarakhand University Dehradun

Abstract

Big data has been utilized and attracted various researchers due to the phenomenal increase in computational application which has developed an overwhelming flow of data. Further, with an expeditious blooming of emerging applications such as social media applications, semantic Web, and bioinformatics applications, data heterogeneity is increasing swiftly. Accordingly, a variety of data needs to be executed with less high accuracy and less. However, effective data analysis and processing of large-scale data are compelling which is considered a critical challenge in the current scenario. To overcome these issues, various techniques have been developed and executed but still, it is significant to improve in accuracy. The current study proposed a hybrid technique of BiLSTM-SAE has been proposed for business big data analytics. Bidirectional LSTM is considered as an advanced version of the conventional LSTM approach. The performance comparison of the proposed method BiLSTM-SAE with existing Random forest-RF has been processed. The final result reported that the proposed method BiLSTM-SAE had been procured with better accuracy of 0.836. Moreover, the training and validation accuracy and loss on different performance metrics have been studied and conducted in the research.

Keywords: Big Data Analytics, Bi-directional LSTM, SAE, Random forest, Machine Learning.

1. Introduction

Big data is considered an emerging technology that enables the potential of managing a massive amount of data in an effective way, and it acquires the analytical expertise to overcome certain constraints prevailing in conventional data processing methods [6].

Recently, due to the increase of data volume amount in digital network systems, focusing on emerging techniques is significant in order to provide efficient results. Therefore, various SM - Social Media applications including Facebook and Twitter have recently captivated the attention of users in recent years that significantly eventuated the data expansion and growth of valuable data may be unrestrained in the coming years [20].

A considerable amount of data that have been consistently produced for various purposes in a data deluge epoch and sometimes unprecedented and increases sales. Further, massive datasets are gathered and examined in various domains that include security, engineering sciences, biomolecular research, social networks, transportation [18], business analytics [19], and e-commerce. Especially, digital data has been produced and taken from different digital devices and it has been rapidly growing at impressive rates. Based on the recent information collected regarding the digital data, it has grown 9 times in 2011 and it has been increasing rapidly in volume the total amount can be raised and reach 35 GB in 2020. Hence, Big Data has become an important term in digital technology. Various surveys have been conducted on Big Data to check the effectiveness of other applications and every perspective on this technology is different that includes research status, challenges, analytics background, and opportunities [1].

Moreover, the digital world has rapidly grown, and DL - Deep Learning [14] [15] and Big Data has been acquired great attention, especially in data science. The main purpose of Big Data - BD is that gather a massive collection of data that is completely raw and complex to manage and examine utilizing traditional tools. The data can be found digitally in different sizes, formats, and shapes, hence it is significant to handle and analyze massive amounts of data based on the requirements of the firms. The top companies use Big Data technology such as Google, Amazon, Yahoo, and Microsoft for storing and managing data. Certainly, companies like Facebook, Twitter, and YouTube which are considered popular social media management a huge amount of data produced by massive users. Nevertheless, the collected huge amount of data cannot be properly controlled by conventional tools. Hence, various firms have developed different products by utilizing the concept of Big Data Analytics for simulation, demonstration, monitoring,

and data analysis which helps to fulfill most of the business requirements that make it a significant subject in data science. Further, the main fundamental mission of Big Data - BD analytics is to identify a useful pattern by extracting from the massive amount of information or data that can be utilized in the prediction process and decision-making concepts. Although, there are still limitations that exist in BD Analytics, especially in DA - data analysis and Machine Learning - ML includes input size, formats, data quality, reliability, data storage, data tagging, data streaming, and much more [2].

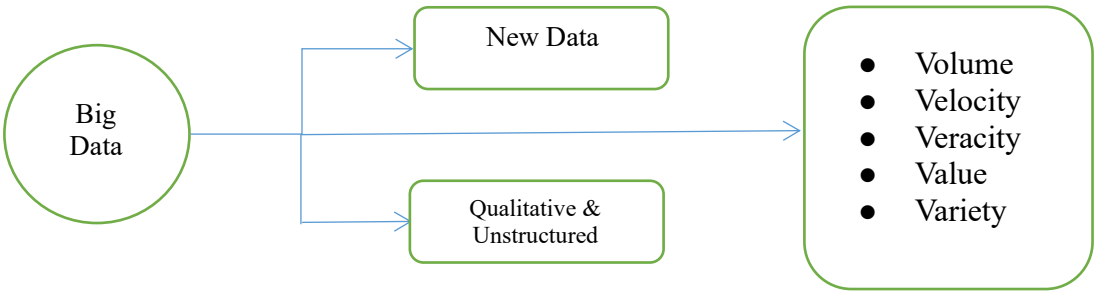


Figure 1.1 - Features of Big Data

Most conventional data processing systems do not have the capability to manage large datasets. Besides, processing and extracting including heterogeneous and large data is complex. In such instances, big data plays an important role in processing the large-scale data collected from various network sources. The important Bid Data features include volume, variety, veracity, velocity, and values. Certainly, Big Data has been classified into two types: Qualitative and Unstructured data. Further, the information gathered in the databases of applications including information technologies, healthcare domains [14], [16], and financial institutions are in the form of unstructured data. Whereas qualitative data is the data that estimates and characterizes big data. Qualitative data can be managed, observed and recorded and the features of the big data are illustrated in figure 1.1.

Moreover, the big data indicate the large data sets that are broadly being utilized among researchers around the world. Therefore, traditional relational databases have been utilized in many researches and have the capability to manage Big Data. Various datasets [13] are collected from different sources such as social media, transaction applications, sensors, web services, and much more. The details of the Big Data features are mentioned below:

Volume - This term indicates the massive amount of data generated every second has oscillated amidst terabytes (TB) to zeta-bytes (ZB). These kinds of Big Data sets can be managed to utilize distributed systems [3].
Velocity - This term indicates the generated data ad processed to assemble the demands.
Variety - This defines the extensive range of data that can be utilized.
Veracity - This denotes the data quality and defines the abnormality, noise, biases, and much more in the data.
Value - This denotes the valuable knowledge extracted from the data.

The technologies and techniques of Big Data Science have the opportunity to penetrate entire facets of the research domains and business. In modern industries or business enterprises in the digital era, the insights of utilizing BD analytics are compelling variations and enhancements in every field. Different Big Data surveys have been conducted in existing research [4]. In this paper, we have focused on a different perspective of existing research and system frameworks conducted for the different layers of the ML and DL techniques which is the data analytics models. The main contribution of this research study is as follows and the main focus of the proposed methodology is to develop an efficient approach for business Big Data analytics using an enhanced DL algorithm. In the current research, a hybrid approach of BiLSTM-SAE algorithm (Bidirectional Long short term memory [17] - stacked autoencoder) is employed for performing business data analytics. The main objectives of the current research are as follows:

- To outline recent trends and the advances in the field of Business Analytics utilizing various DL - Dee Learning algorithms from an operational standpoint.
- To identify a significant use cases utilizing DL algorithms for enhancing the decision making capability in the field of business operations.
- To develop an effective technique for adequately tuning the operational parameters of the Big Data processing model without compromising efficiency and accuracy.

- To resolve or overcome the certain complex problems persist due to data dimensionality by utilizing a DL-based technique with feature extraction that process on larger feature sets without impacting the computational system speed.
- To validate the efficiency of the developed Big Data processing model by testing it for larger datasets.

2. Literature Survey

In [5], the different deep learning (DL) techniques has been reviewed with the Big Data advancements that review the advancements of DL techniques and easy-to-understand initiated from shallow NN - Neural Networks, to prominent CNN - Convolutional Neural Network, legendary RNN - recurrent neural network, GAN - generative adversarial neural network, GNN - graph neural network, and variational AE - autoencoder. Further, the DL methods are examined with implementations in developing and adopting their own techniques that are based on DL. Eventually, the common difficulties of utilizing DL have been examined and certain aspects including the reinforcement learning algorithm have been explained.

In [6], the present state-of-the-art of Big Data has been reviewed in various aspects on representation, reduction & cleaning, processing, and security, and integration to provide Big Data-as-a-service which is a high-quality has also been examined. Therefore, the framework includes three plane application plane, cloud plane, and sensing plane that systemically handle all limitations of the above-mentioned aspects. Moreover, it indicates the working process of the architecture, a tensor-based multiple clustering on returning data is demonstrated, which enables various research on the system.

In [8], a real-time violence detection system has been utilized in the research with the massive input streaming information that manages the violation with simulation processed on human intelligence. Basically, the input represented in the system consists of a large number of video streams in real time that are executed in the Spark framework. Therefore, the Spark framework has been utilized in various existing research and the

frames are divided and the details on the individual frames are processed and extracted utilizing the function of HOG - Histogram of Oriented Gradients. These divided frames are tagged in accordance with the violence method features, human part and negative model that is utilized to train the BDLSTM - Bidirectional Long-Short-Term Memory network for identification of violent scenes. Certainly, the bidirectional LSTM has the possibility to access the relevant data in reverse and forward directions. Eventually, the performance of the technique has been validated and procured an accuracy of 94.5%.

In [9], the accuracy and performance have been evaluated by the procured results based on experiments conducted on the training of LSTM, BiLSTM, and unidirectional LSTM, and these techniques are analyzed in the existing research. Here, the main focus is on data training processed from right to left (opposite direction), but normally the regular data training that is right to left had a significant and positive effect on enhancing time series precision in forecasting. Finally, the procured results on the utilization of an additional layer for the training process helped in enhancing the forecasting accuracy by 37.78% and it was advantageous for modeling. However, it has been observed that training occurred on BiLSTM was slower and it consumes more time during the fetching of data to reach a certain equilibrium state.

In [10], BiLSTM - Bidirectional LSTM has been utilized in the research and was an advanced version of conventional LSTMs. Bidirectional LSTM improves the system architecture performance with the problems persist due to sequence classification, where the entire input sequence provided are accessible. It can train 2 LSTMs instead of 1-LSTM on the sequence (input). Here, the large samples have been considered in high dimension in order to monitor the data with the integration of BLSTM and autoencoder that eventually developed to enhance the prediction accuracy. Further, an autoencoder has been utilized as a feature extractor in order to monitor and compress the condition of the data. Therefore, this method was designed to capture the features based on bidirectional long-range dependencies. With the comparison of other methods including LSTM, CNN, SVR - Support Vector Machine, and MLP - multi-layer perceptron during experimentation, this hybrid model - autoencoder-BLSTM procured better results.

Eventually, this method enables strong support in the maintenance strategy and health management development of turbofan engines.

In [11], the research represents a comparative analysis of deep learning methods that includes RNN - Recurrent Neural Network, BiLSTM, GRUs - Gated recurrent units, LSTM, and VAE - Variational AutoEncoder to forecast both recovered and new cases. Therefore, this study mainly focused on the accordance of details gathered from six countries. Eventually, the results demonstrated promising results in forecasting COVID-19 cases when compared to the other algorithms.

In [21], reviews the application of ML-algorithms in BD- analytics and the challenges associated with it. ML in terms of selecting an optimal path without any predefined knowledge. These algorithms enable big data processing systems to automatically detect the right path for a specific process based on previous data and to predict the next sequence. The study focuses on the application of ML algorithms with respect to big data and its computing platforms. It deals with the issues of ML in data analytics and investigates new opportunities for ML. From the existing literary works, it can be observed that big data poses significant challenges to ML algorithms for extracting an appropriate pattern from the test data.

In [22], discusses the application of deep learning algorithms for processing big data in the IoT environment. The data obtained from various sources is facilitated by performing data analytics in IoT systems. Initially, the data characteristics are articulated by identifying two prominent solutions for data in IoT systems with a ML perspective such as IoT big data and streaming data analysis. The study also discusses the efficacy of deep learning algorithms in achieving the desired analytics in the IoT environment. The capability of the DL algorithms in processing big data is explored and a comprehensive analysis of various DL algorithms was discussed. The prominent observations are analyzed and summarized which validated the potential of DL algorithms in data analytics. Among various deep learning algorithms, recurrent neural networks (RNNs) are gaining prominence because of their superior prediction capability.

In [23], discussed the integration of BD-analytics with DL algorithms. The proposed research consisted of a comprehensive review analysis wherein it was inferred that the conventional algorithms of NN- neural networks and AI-artificial intelligence possess various limitations, and these limitations introduce challenges for processing big data in real-time applications. Hence, the proposed research focussed on introducing the mechanism of deep learning algorithms for addressing the limitations associated with AI and neural networks. However, it was observed that big data analytics demanded a process that is incorporated with multiple iterations and calculations where each calculation required an algorithm or multiple algorithms for computing. The analysis discussed the influence of machine learning and deep learning in BD processing to satisfy different applications and demands of the users in real-time scenarios. Consequently, other techniques of deep learning algorithms were also discussed to validate their efficacy in addressing various complexities and challenges related to big data analytics. Additionally, other related techniques such as TL - transfer learning were also discussed and analyzed properly in order to support the research of DL-algorithms.

In [24], proposed a Meticulous-Fuzzy-Convolution-C-Means (MFCCM) algorithm by transforming the behavior of a Convolutional-Neural-Network (CNN) in order to incorporate the prominent feature analysis of deep learning algorithms. The preliminary aim of the proposed research is to data processing by using an optimized BD algorithm by incorporating the mechanism of efficient feature selection. In this research, the mechanism includes the implementation of the Deep CNN approach with the FCM for selecting prominent features. The proposed MFCCM algorithm exhibited superior results by providing accurate segmentation under uncertainties such as the presence of variance noisy data. The performance of the proposed MFCCM algorithm was tested for various image processing applications. Results showed that the proposed approach was more suitable for Big Data Analytics algorithms with a drawback of time-related complexities.

Various ML algorithms are incorporated in PBA for evaluating the previously stored data and to infer prominent observations for future implementation. However, despite the

effectiveness of the PBA technique, certain limitations degrade the performance of PBA - Predictive Big Data Analytics in the extraction of big data: Firstly, PBA suffers from the issue of data dimensionality and the second one is the huge sample size. These issues increase the computational cost of the predictive model and cause algorithmic instability in PBA systems. These limitations can be resolved by employing efficient ML or DL algorithms such as Scalable Random Forest (SRF) for high dimensional big data [25].

The current research has employed an enhanced version of deep learning algorithms or hybrid approaches for enhancing the prediction accuracy, computational time, and effectiveness for performing business data analysis in big data systems. However, there are certain challenges with respect to deep learning techniques which need to be addressed for the successful implementation of deep learning algorithms for big data analytics.

- Due to the data dimensionality issues in big data, it is challenging to recognize relevant features. Though existing deep learning techniques are efficient to handle complex data structure, they are not much effective for multi feature selection. More advanced hybrid approaches are required for extracting relevant features from complex data structures [26].
- Efficient design of the deep neural networks is still a big challenge. Training the DL algorithms to improve their efficiency is an additional task which increases the design complexities and computational cost. Hence it is essential to reduce the computational complexity.
- Deep learning algorithms are implemented for learning representative attributes from high dimensional manufacturing big data for obtaining accurate prediction accuracy. A wide range of models such as SVM, Decision Trees, Random Forest etc are commonly deployed for obtaining effective solutions. Though these models yield better results, there is still a large scope for improving the overall prediction accuracy [27].
- Application of DL-methods to industrial big data applications introduces unique challenges such as highly skewed CD - class distribution, restricted data, and

existence of rare classes such as failures, the need for explainable decisions, and decision making to optimize the process effectively [28].

3. Research Methodology

In the current research, a hybrid technique of BiLSTM-SAE has been proposed for business big data analytics. Bidirectional LSTM is considered as an advanced version of the conventional LSTM approach. Though Conventional LSTM shows superior performance with a desired level of accuracy, there are certain limitations associated with it such as LSTM possesses a limited amount of data set and the performance of LSTM was tested for smaller datasets. Hence in this study, a Bidirectional LSTM is employed for achieving better accuracy and the BiLSTM approach is integrated with an SAE algorithm for strengthening the potential of the big data processing system. The basic illustration of the business big data analytics is illustrated in figure 3a.

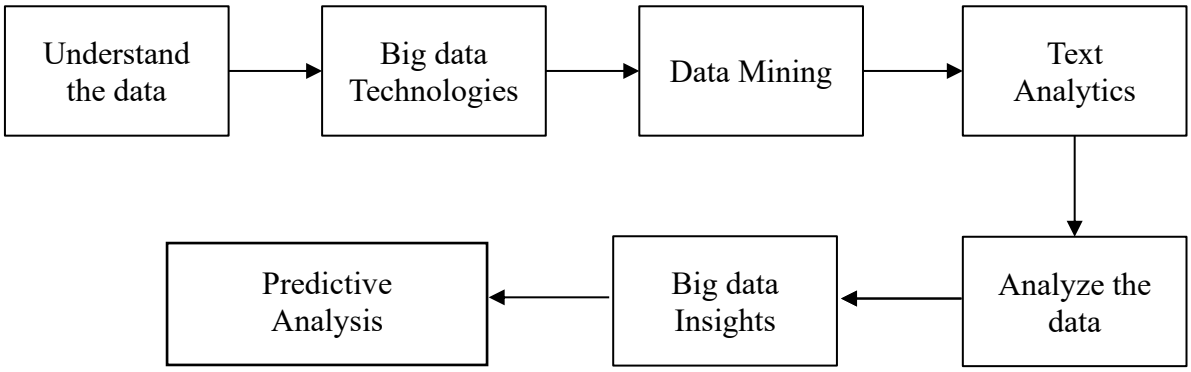


Figure 3a Illustration of business big data analytics

3.1 Dataset Extraction and Preparation

For experimentation, two types of the dataset include the weather dataset defines traffic volume and the sensor dataset indicates the incident has been gathered from the Kaggle and utilized for the research. Firstly, the exploratory data analysis that consists of testing and training data based on the Indian Metro Data has been used which includes the attributes. The testing dataset contain the attributes such as date_time (local IST), is_holiday, air_population_index (10 - 300), humidity (Celsius), wind_speed (miles per

hour), wind_direction (0 - 360 degree), visibility_in_miles (miles), dew_point (Celsius), temperature (Kelvin), and rain_p_h (mm) with values and graphical representation.

Secondly, the Baton Rouge traffic incidents consist of the attributes such as incident number, crash date, street address, city, state, zip code, district, zone, subzone, total vehicles, road class, hit & run, train, fatality, injury, pedestrian, intersection, nearest street, manner of collision, roadway surface, roadway condition, roadway design, alignment, primary factor, secondary factor, weather, location type, roadway relation, access control, lighting, longitude, latitude, and geolocation. Eventually, these two datasets are joined by employing k-means clustering for grouping the data based on clustering with the clustering values 0,1,2 that represent the severity of the traffic based on traffic, weather, and incident. Eventually, the two datasets - exploratory data analysis and baton rouge traffic incidents are combined based on the severity in order to move to the next preprocessing level.

3.2 Preprocessing

In the preprocessing stage, the combined datasets were read and employed preprocessing techniques in order to remove redundant data from the dataset and fill the empty values into 0 (Zero). Normally, the preprocessing technique is considered a significant process in every experiment that eradicates duplicate values from the dataset. Further, applying a label encoder in the process helps in the conversion part where all the strings presented in the dataset will be converted to integers. Finally, the correlated values increases more than 0.90, and the feature value was dropped in order to find the correlation features.

3.3 Applying Smote

In this step, the SMOTE - Synthetic Minority Oversampling Technique was employed for inconsistent data handling this method help to identify imbalanced datasets and this is also known as data augmentation used for the minority class. In the current research, oversampling has been applied along with the SMOTE. However, the main issue that persists with imbalanced classification is due to a few instances of the minority class occurring in the model that helps to learn the certain decision boundary in an effective

way. This problem can be solved utilizing an oversample of the instances presented in the minority class. Therefore, it can be obtained by duplicating instances from the minority class found in the above-mentioned training dataset that assists in fitting a model.

Moreover, this indicates counterbalancing the class distribution, however, it was not capable to provide certain additional data to the presented method. The enhancement of duplicating instances presented from the minority class helps to synthesize instances that come new from the minority class. Certainly, this kind of process is known as data augmentation in order to create effective and tabular data. SMOTE has been used in most of the existing research that assists in different ways by choosing instances that are nearly close in the specific feature space. Particularly, random instances presented from the minority class were initiated by choosing it. In-Addition, a randomly chosen neighbor has been used in the research and a synthetic instance was developed at a randomly picked point amidst two instances in feature space.

3.4 Classification Model - Proposed Hybrid Deep learning (Bi-directional LSTM with Stacked Auto Encoder)

In the classification model step, the bi-directional LSTM stacked autoencoder has been utilized in the research with the model architecture represented below in table I. Bidirectional LSTM is considered as an advanced version of conventional LSTMs. Bidirectional LSTM improves the system architecture performance with the problems persist due to sequence classification, where the entire input sequence provided are accessible. In the current research, It trains two LSTMs instead of one LSTM on the input sequence. The architecture of the bidirectional LSTM is given in figure 3b.

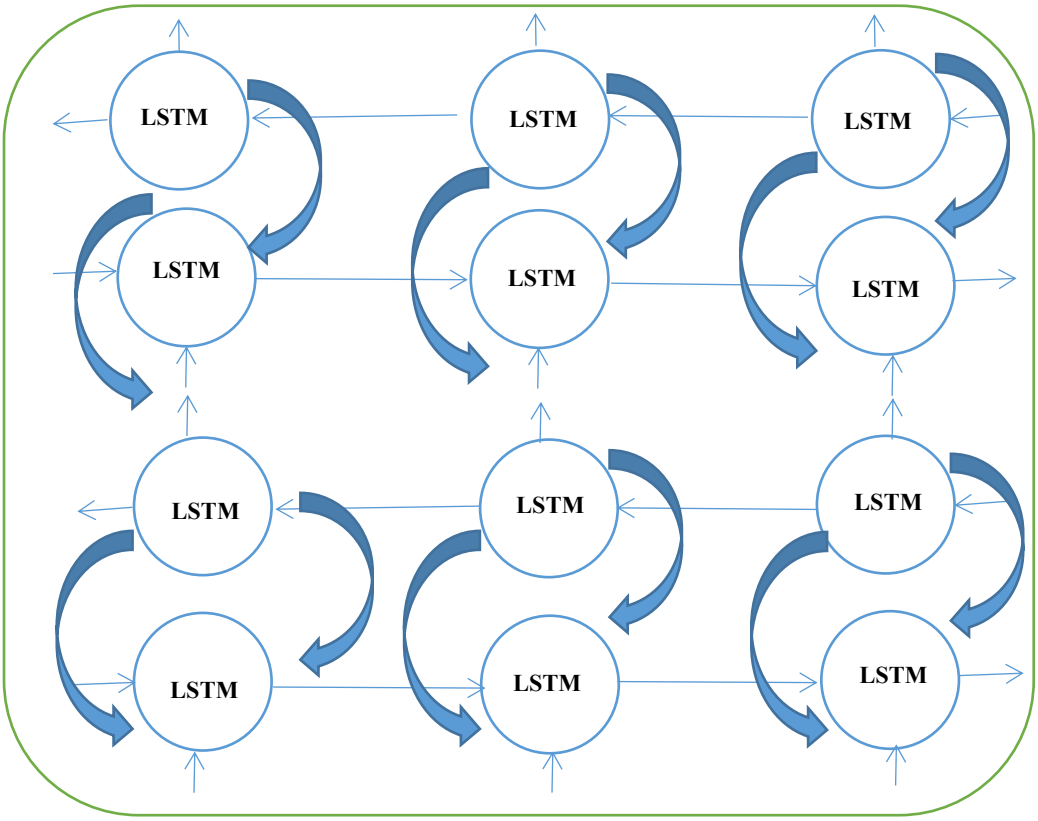


Figure 3b - Architecture of BiLSTM

BLSTM duplicates the first recurrent layer in the network in such a way that two layers are formed side-by-side. It allows the input sequence to the 1stlayer and allows a reversed copy of the represented input sequence to the 2nd layer and it regulates the data flow across the cell. LSTM effectively resolves the issue of gradient exploding and vanishing issues, which commonly occurs while training conventional RNN. At every time step represented as t, the output of the LSTM is updated as shown in below equations:

$$i_t = \sigma\left(W_i x_t + U_i h_{t-1} + b_i \right) \ldots\ldots (1)$$

$$f_t = \sigma\left(W_f x_t + U_f h_{t-1} + b_f \right) \ldots\ldots (2)$$

$$o_t = \sigma\left(W_o x_t + U_o h_{t-1} + b_o \right) \ldots\ldots (3)$$

$$\overline{c_t} = \tanh\left(W_c x_t + U_c h_{t-1} + b_c \right) \ldots\ldots (4)$$

$$c_t = f_t o\, c_{t-1} + i_t o \overline{c_t} \ldots\ldots\ldots (5)$$

$$h_t = o_t o \tanh c_t \ldots\ldots\ldots (6)$$

Where x_t is defined as the input data at time represented as t and $\acute{C}_t$ is the present state of the memory cell and C_t is defined as the updated state (memory cell) and h_t is the final output of the LSTM cell.

Table I - LSTM Stack Auto Encoder

Model: "sequential"

Layer (type)	Output Shape	Param #
lstm (LSTM)	(None, 47, 228)	209760
dropout (Dropout)	(None, 47, 228)	0
dense (Dense)	(None, 47, 32)	7328
batch_normalization (BatchNormalization)	(None, 47, 32)	128
activation (Activation)	(None, 47, 32)	0
dropout_1 (Dropout)	(None, 47, 32)	0
flatten (Flatten)	(None, 1504)	0
repeat_vector (RepeatVector)	(None, 47, 1504)	0
lstm_1 (LSTM)	(None, 47, 32)	196736
dense_1 (Dense)	(None, 47, 64)	2112
flatten_1 (Flatten)	(None, 3008)	0
dense_2 (Dense)	(None, 3)	9027
activation_1 (Activation)	(None, 3)	0
flatten_2 (Flatten)	(None, 3)	0

Total params: 425,091
Trainable params: 425,027
Non-trainable params: 64

Above the table, I define the model “sequential” with the Layer type, Output shape, and Parameters. Therefore, the model was initiated with an Encoder which is known as an input layer which is considered an LSTM layer it is followed by batch_normalization, dense layer, and flatten with a smaller size, and the return sequences are taken from layer 1. In the next step, the acquired data will be fed to a repeat vector that takes the single vector and reshapes it in an efficient way that helps to feed to the Decoder network process that is symmetrical to the encoder. In the execution, the activation function RELU has been utilized and the accuracy has been evaluated in accordance with the original input. The total parameters taken for the experimentation process are 425,091, trainable parameters 425,027, and non-trainable parameters 64.

Moreover, Auto encoders are the neural units that learn the encoding of the inputs in order to retrieve the original input from the encodings as well as possible. In autoencoders, when the number of nodes in the hidden layers is increased then the network will risk itself to learn the “Identity function” wherein the output of the encoder will be the same as the input making the autoencoder less productive. In order to overcome this limitation, the study employs stacked autoencoders by corrupting the data purposefully by tuning the input values randomly to zero. In this study, the collected raw data will be combined and pre-processed to identify the missing values within the dataset.

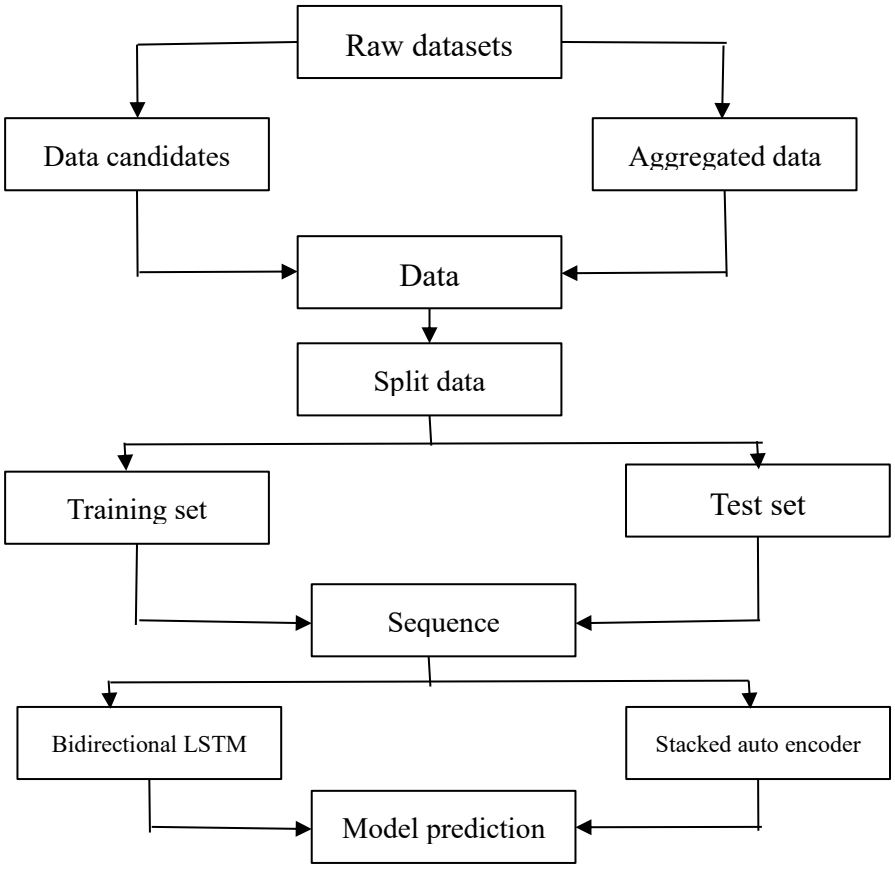


Figure 6.3a - The workflow of the BiLSTM and SAE

Further, the data will be split into two separate datasets such as training and testing datasets. Lastly, BiLSTM and SAE models will be incorporated for data prediction. The workflow of the BiLSTM and SAE is illustrated in figure 6.3a.

4. Experimental Results and Analysis

The current research focused on a hybrid techniques of BiLSTM-SAE for business big data analytics. For experimentation, two types of dataset includes weather dataset defines traffic volume and sensor dataset indicates incident has been gathered from the Kaggle has been utilized for the research. The traffic flow prediction has been mentioned based on Traffic Severity and time (minutes) based on the original and predicted traffic flow in the graphical representation in figure 4.1a.

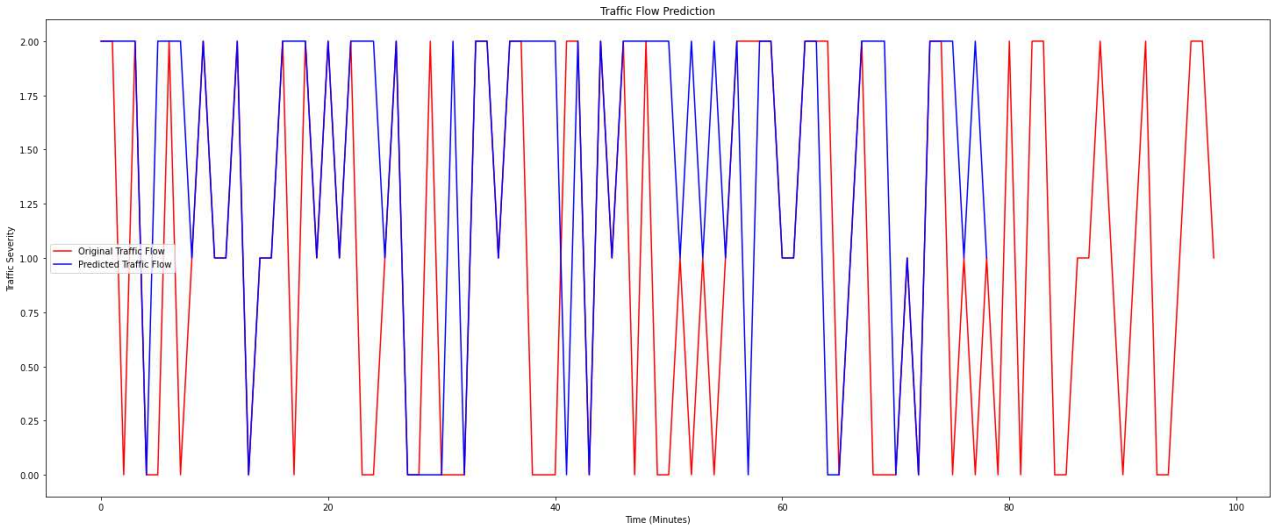
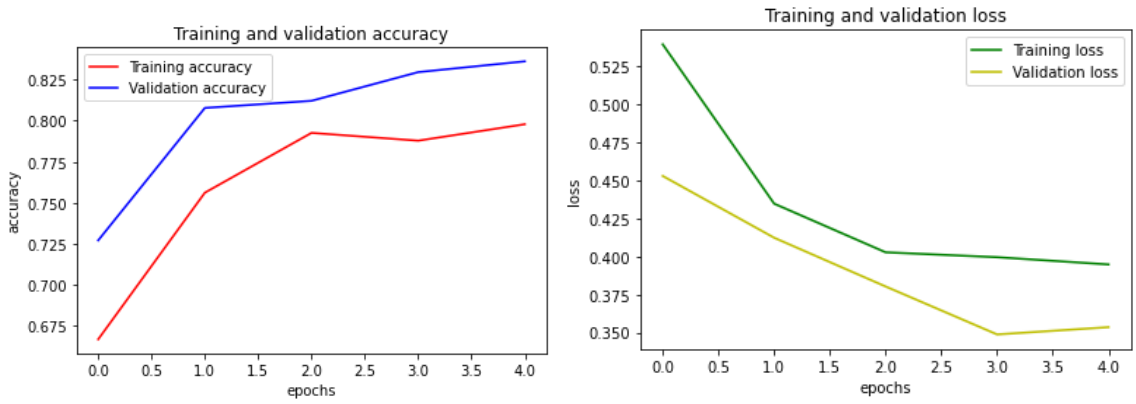


Figure 4.1a - Traffic Flow Prediction

4.1 Performance Evaluation

The performance comparison of the proposed method BiLSTM-SAE with existing Random forest - RF has been processed. The final result reported that the proposed method BiLSTM-SAE had been procured with better accuracy of 0.836. Moreover, the training and validation accuracy and loss has been evaluated and represented in the graphical form in figure 4.1 a & b.



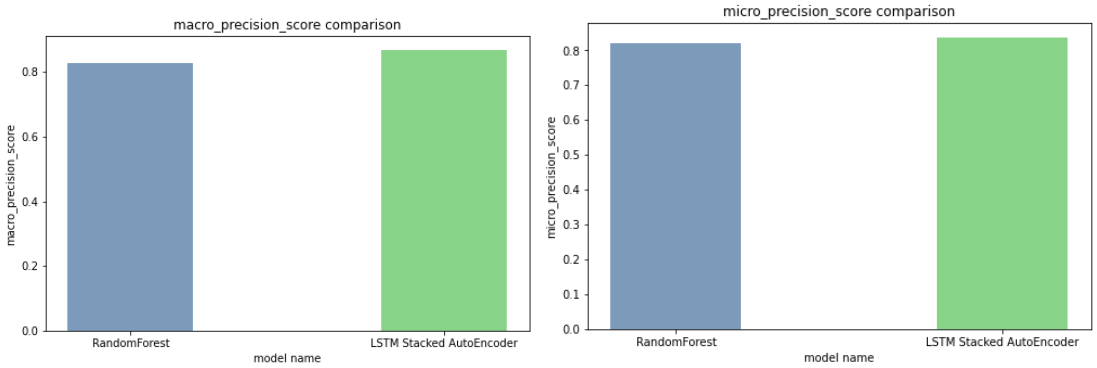
(a) **(b)**

Figure 4.1 a & b - Training and Validation based on accuracy (a) and loss (b)

Here, we have utilized various performance metrics of the multi-class classification model with the classification accuracy, precision, sensitivity, recall, and F-measure. We have considered both micro and macro evaluation for precision, sensitivity, and recall.

Precision

$$\textbf{Precision} = \frac{TP}{TP + FP} \text{ --------- (7)}$$

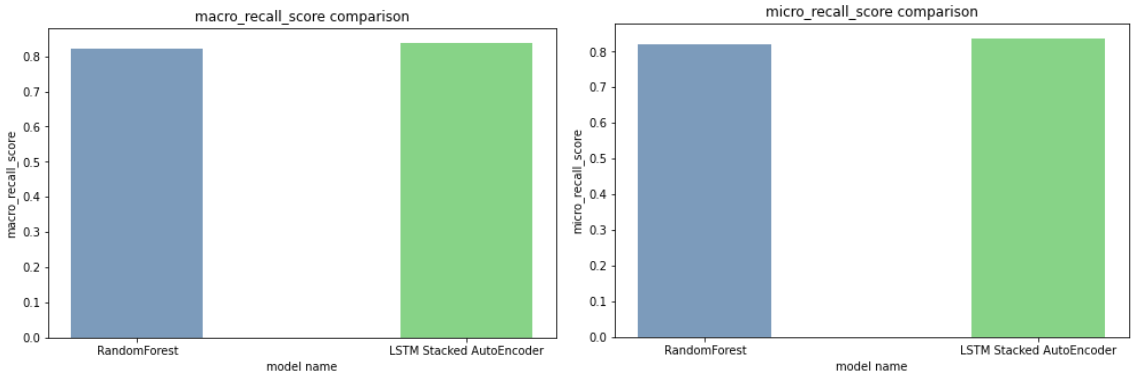


(a) (b)

Figure 4.2 a & b - Comparison of macro precision score & micro precision score based on Random Forest (RF) and LSTM stacked AutoEncoder

Recall

$$\textbf{Recall} = \frac{TP}{TP + FN} \text{ ---------- (8)}$$



(a) **(b)**

Figure 4.3 a & b - Comparison of macro recall score & micro recall score based on Random Forest (RF) and LSTM stacked AutoEncoder

Sensitivity

$$\textbf{Sensitivity} = \frac{TP}{TP + FN} \text{ ---------- (9)}$$

Accuracy

$$\textbf{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \text{ ----------(10)}$$

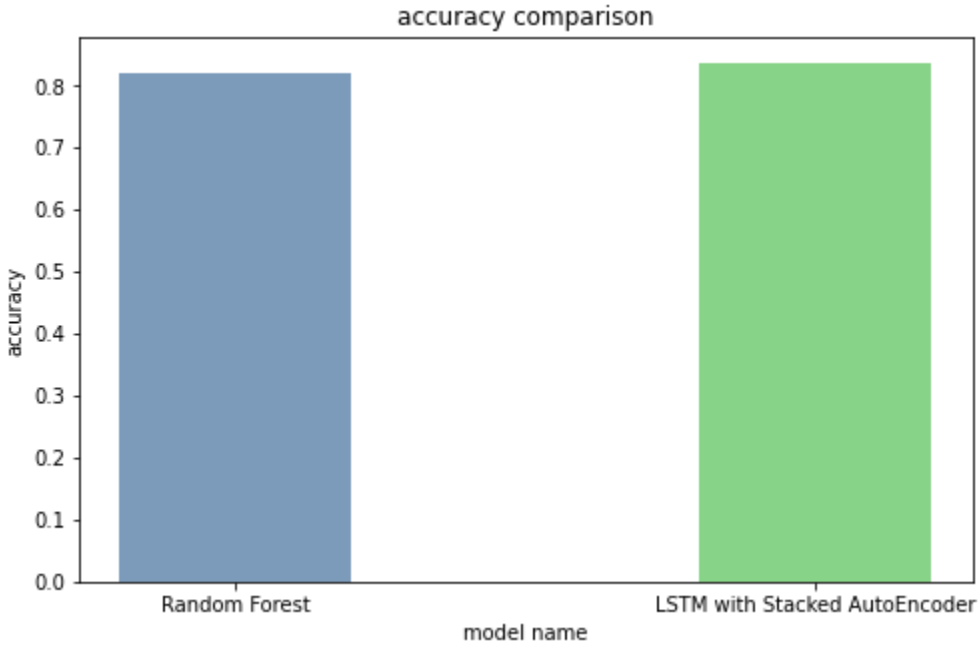


Figure 4.4 a - Accuracy comparison based on Random Forest (RF) and LSTM stacked AutoEncoder

F-Measure

$$\textbf{F-Measure} = \left(\frac{2}{recall^{-1} + precision^{-1}}\right) = 2.\frac{precision.recall}{precision + recall} \text{---------- (11)}$$

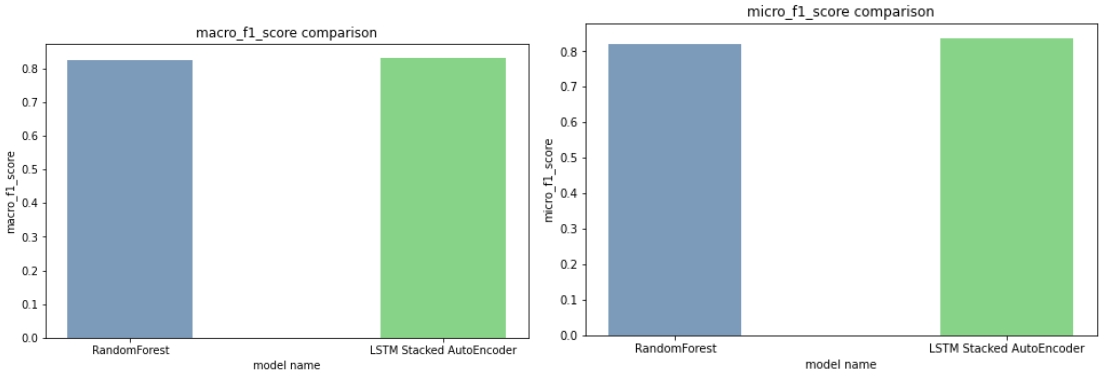


Figure 4.5 a & b - Comparison of macro F1 score & micro F1 score based on Random Forest (RF) and LSTM stacked AutoEncoder

TP - True Positive, TN - True Negative, FP - False Positive, FN - False Negative;

True Negative - TN - The actual value mentioned was False; and the method predicted False;

False Positive - FP - The actual value mentioned was False; and the method predicted True;

False Negative - FN - The actual value mentioned was True, and the method predicted False;

True Positive TP - The actual value mentioned was True, and the method predicted True.

Furthermore, tabular II explains the comparison table on Proposed model - LSTM stacked AutoEncoder and Existing model - Random Forest in accordance with the performance evaluation metrics such as Macro_precision_score, Micro_precision_score, Macro_recall_score, Micro_recall_score, Macro_f1_score, and Micro_f1_score with the procured results 83.6%, 86.7%, 83.6%, 83.7%, 83.1%, and 83.6%. Compared to the existing model based on certain performance evaluation metrics, the proposed model achieved better results.

Tabular II - Comparison table on Proposed model and Existing model

Model Name	Accuracy	Macro_precision_score	Micro_precision_score	Macro_recall_score	Micro_recall_score	Macro_f1_score	Micro_f1_score
Proposed Model(LSTM stacked AutoEncoder)	0.836	0.867	0.836	0.837	0.836	0.831	0.836
Existing Model(Random Forest)	0.821	0.828	0.821	0.821	0.821	0.823	0.821

5. Conclusion

In the current research, the proposed method with the integration of the BiLSTM-SAE algorithm. Here, this study integrates two efficient data processing methods in order to extract the positive attributes from these techniques. Further, BiLSTM has the capability to learn automatically and is able to sequence predictions based on previous data. Whereas Stacked autoencoders have the capability to obtain superior accuracy with less computation time. The proposed hybrid method utilized in the study focuses on improving the high-speed performance of BD - processing systems with high scalability and accuracy in business operations in DA. The experimentation process has been executed based on the provided merged datasets and procured 83.6% accuracy using BiLSTM-SAE and compared to the existing methods (Random Forest), the currently proposed method achieved better results.

Declarations

- No funding was received for conducting this study.
- The authors have no financial or proprietary interests in any material discussed in this article.
- Data will be made available on reasonable request.
- Code will be made available on reasonable request.

- The study conceptualization and design, data collection, analysis and result interpretation, as well as the production of the manuscript are all solely the responsibility of the first author with the guidance of second author.

References:

[1] Goswami, S., Kumar, A(2022). Survey of Deep-Learning Techniques in Big-Data Analytics. Wireless Personal Communications 126, 1321–1343

[2] Jan, B., Farman, H., Khan, M., Imran, M., Islam, I. U., Ahmad, A., ... & Jeon, G. (2019). Deep learning in big data analytics: a comparative study. *Computers & Electrical Engineering*, *75*, 275-287.

[3] Athmaja, S., Hanumanthappa, M., & Kavitha, V. (2017, March). A survey of machine learning algorithms for big data analytics. In *2017 International conference on innovations in information, embedded and communication systems (ICIIECS)* (pp. 1-4). IEEE.

[4] Elshawi, R., Sakr, S., Talia, D., & Trunfio, P. (2018). Big data systems meet machine learning challenges: towards big data science as a service. *Big data research*, *14*, 1-11.

[5] Li, Y., Huang, C., Ding, L., Li, Z., Pan, Y., & Gao, X. (2019). Deep learning in bioinformatics: Introduction, application, and perspective in the big data era. *Methods*, *166*, 4-21.

[6] Wang, X., Yang, L. T., Liu, H., & Deen, M. J. (2017). A big data-as-a-service framework: State-of-the-art and perspectives. *IEEE Transactions on Big Data*, *4*(3), 325-340.

[7] Subbu, K. P., & Vasilakos, A. V. (2017). Big data for context aware computing–perspectives and challenges. *Big Data Research*, *10*, 33-43.

[8] Fenil, E., Manogaran, G., Vivekananda, G. N., Thanjaivadivel, T., Jeeva, S., & Ahilan, A. J. C. N. (2019). Real time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional LSTM. *Computer Networks*, *151*, 191-200.

[9] Siami-Namini, S., Tavakoli, N., & Namin, A. S. (2019, December). The performance of LSTM and BiLSTM in forecasting time series. In *2019 IEEE International Conference on Big Data (Big Data)* (pp. 3285-3292). IEEE.

[10] Song, Y., Shi, G., Chen, L., Huang, X., & Xia, T. (2018). Remaining useful life prediction of turbofan engine using hybrid model based on autoencoder and bidirectional long short-term memory. *Journal of Shanghai Jiaotong University (Science)*, *23*(1), 85-94.

[11] Zeroual, A., Harrou, F., Dairi, A., & Sun, Y. (2020). Deep learning methods for forecasting COVID-19 time-Series data: A Comparative study. *Chaos, Solitons & Fractals*, *140*, 110121.

[12] Kim, J., & Moon, N. (2019). BiLSTM model based on multivariate time series data in multiple field for forecasting trading area. *Journal of Ambient Intelligence and Humanized Computing*, 1-10.

[13] Sun, T., Yang, C., Han, K., Ma, W., & Zhang, F. (2020). Bidirectional spatial–temporal network for traffic prediction with multisource data. *Transportation research record*, *2674*(8), 78-89.

[14] Mengara Mengara, A. G., Park, E., Jang, J., & Yoo, Y. (2022). Attention-Based Distributed Deep Learning Model for Air Quality Forecasting. *Sustainability*, *14*(6), 3269.

[15] Chou, C. H., Huang, Y., Huang, C. Y., & Tseng, V. S. (2019, April). Long-term traffic time prediction using deep learning with integration of weather effect. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 123-135). Springer, Cham.

[16] Zhang, L., Liu, P., Zhao, L., Wang, G., Zhang, W., & Liu, J. (2021). Air quality predictions with a semi-supervised bidirectional LSTM neural network. *Atmospheric Pollution Research*, *12*(1), 328-339.

[17] Abduljabbar, R. L., Dia, H., & Tsai, P. W. (2021). Development and evaluation of bidirectional LSTM freeway traffic forecasting models using simulation data. *Scientific reports*, *11*(1), 1-16.

[18] Li, T., Ni, A., Zhang, C., Xiao, G., & Gao, L. (2020). Short-term traffic congestion prediction with Conv–BiLSTM considering spatio-temporal features. *IET Intelligent Transport Systems*, *14*(14), 1978-1986.

[19] Kraus, M., Feuerriegel, S., & Oztekin, A. (2020). Deep learning in business analytics and operations research: Models, applications and managerial implications. *European Journal of Operational Research*, *281*(3), 628-641.

[20] Huang, C. (2020). Special issue on deep learning-based neural information processing for big data analytics.

[21] Goswami, S., Kumar, A. (2022). Traffic Flow Prediction Using Deep Learning Techniques. In: Chaubey, N., Thampi, S.M., Jhanjhi, N.Z. (eds) Computing Science, Communication and Security. COMS2 2022. Communications in Computer and Information Science, vol 1604. Springer, Cham, 198-213.

[22] Mohammadi, M., Al-Fuqaha, A., Sorour, S., & Guizani, M. (2018). Deep learning for IoT big data and streaming analytics: A survey. *IEEE Communications Surveys & Tutorials*, *20*(4), 2923-2960.

[23] Talha, M., Ali, S., Shah, S., Khan, F. G., & Iqbal, J. (2019). Integration of Big Data and Deep Learning. In *Deep Learning: Convergence to Big Data Analytics* (pp. 43-52). Springer, Singapore.\

[24] Balakrishnan, N., Rajendran, A., & Palanivel, K. (2019). Meticulous fuzzy convolution C means for optimized big data analytics: adaptation towards deep learning. *International Journal of Machine Learning and Cybernetics*, *10*(12), 3575-3586.

[25] Oo, M. C. M., & Thein, T. (2019). An efficient predictive analytics system for high dimensional big data. *Journal of King Saud University-Computer and Information Sciences*.

[26] Amanullah, M. A., Habeeb, R. A. A., Nasaruddin, F. H., Gani, A., Ahmed, E., Nainar, A. S. M., ... & Imran, M. (2020). Deep learning and big data technologies for IoT security. *Computer Communications*, *151*, 495-517.

[27] Huang, C. (2020). Special issue on deep learning-based neural information processing for big data analytics.

[28] Gupta, C., & Farahat, A. (2020, August). Deep Learning for Industrial AI: Challenges, New Methods and Best Practices. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 3571-3572).