Iraqi Journal of Science

ISSN: 0067-2904

Leveraging Hadoop and Hybrid Deep Learning on Home Datasets for Business Intelligence

Asaad R. Kareem , Hasanen S. Abdullah*
Computer Science Department, the University of Technology, Baghdad, Iraq

Received: 4/2/2024 Accepted: 14/8/2024 Published: 30/7/2025

Abstract

In today's data-driven business environment, organizations' dependence on cutting-edge technologies is constantly increasing in order to gain meaningful information from big data. Raw data's size and complexity make it unusable for decision-making. To address this, BI systems transform raw data into clear, insightful information by leveraging the Hadoop framework to process big data. This study is concentrated on the creation of a business intelligence (BI) system depending upon deep learning (DL) approaches, especially one-dimensional convolutional neural networks (1-D CNN) and long short-term memory (LSTM). The proposed approach takes advantage of the DL algorithms for examining and picking up intricate patterns in sequential data, which is helpful in accurately anticipating the results and offering relevant insights. The predictive capabilities of the proposed system are enhanced through a combination of the 1D CNN and LSTM models, enabling it to grasp spatial and temporal data dependencies. Parallel computing and distributed processing made possible by the Hadoop model have increased the efficiency of big data management, which ensures performance and scalability while working with large datasets. This study aims to show how well the proposed business intelligence system is based on hybrid deep learning to make predictions by utilizing big data analyses. The results show the superiority of the integrated CNN-LSTM model, which operates on a data block size of 512 MB, over a data block size of 64 MB for home data sets, in addition to its superiority over two machine learning models (decision tree and booster regression) at the same block size.

Keywords: business intelligence, 1D CNN, LSTM, deep learning, Hadoop framework, predictive analytics, big data.

الاستفادة من Hadoop والتعلم العميق الهجين على مجموعات البيانات المنزلية لذكاء الأعمال

اسعد رحيم كريم, حسنين سمير عبد ألله*
قسم علوم الحاسوب، الجامعة التكنولوجية، بغداد، العراق

الخلاصة

يتزايد اعتماد المؤسسات على التقنيات المتطورة باستمرار في بيئة الأعمال الحديثة المعتمدة على البيانات للحصول على معلومات مفيدة من البيانات الضخمة. إن حجم البيانات الأولية وتعقيدها يجعلها غير قابلة للاستعمال في اتخاذ القرار. تعالج أنظمة ذكاء الأعمال هذه المشكلة من خلال تحويل البيانات الأولية إلى

*Email: hasanen.s.abdullah@uotechnology.edu.iq

معلومات واضحة وثاقبة من خلال الاستفادة من إطار عمل Hadoop لمعالجة البيانات الضخمة. تركز هذه الدراسة على إنشاء نظام ذكاء الأعمال (BI) يعتمد على أساليب التعلم العميق (DL)، وخاصة الشبكات العصبية التلافيفية أحادية البعد (1D CNN) والذاكرة الطويلة قصيرة المدى (LSTM). يمكن الاستفادة من النهج المقترح لخوارزميات (DL) لفحص الأنماط المعقدة في البيانات التسلسلية والتقاطها، وهو ما يساعد في توقع النتائج بدقة وتقديم الأفكار ذات الصلة. يتم تعزيز القدرات التنبؤية للنظام المقترح من خلال دمج نماذج CNN 1D وLSTM ، مما يمكّنه من فهم تبعيات البيانات المكانية والزمانية. أدت الحوسبة المتوازية والمعالجة الموزعة التي أصبحت ممكنة بفضل نموذج Hadoop إلى زيادة كفاءة إدارة البيانات الضخمة، مما يضمن الأداء وقابلية التوسع أثناء العمل مع مجموعات البيانات الكبيرة. تهدف هذه الدراسة إلى إظهار مدى قدرة نظام ذكاء الأعمال المقترح القائم على التعلم العميق الهجين في عمل التنبؤات من خلال الاستفادة من تحليلات البيانات الضخمة. تشير النتائج تفوق نموذج CNN–LSTM المتكامل الذي يعمل على كتلة بيانات بحجم 512 ميجابايت على كتلة بيانات بحجم 64 ميجابايت لمجموعات البيانات المنزلية، بالإضافة إلى تفوقه على نموذجين للتعلم الآلي (شجرة القرار والانحدار الداعم) لنفس حجم الكتلة.

1. Introduction

Business intelligence is the process of converting raw data into useful and relevant information for decision-making [1]. One of the most popular approaches to BI implementation is to use the Hadoop model, a distributed computing model that facilitates the processing, analysis, and storage of large data amounts [2, 3]. Hadoop provides the ability to distribute data over a cluster of machines, which has resulted in faster analysis and processing. It provides various libraries and tools, like Hive and MapReduce, which are used for facilitating data processing and analysis. Those tools present efficient capabilities for querying and analysis and are designed specifically for large datasets [4]. Organizations can gain important information by effectively storing, processing, and analyzing massive data amounts using the Hadoop platform for BI [5, 6]. ANNs represent an algorithm type inspired by the human brain’s functions and structure; they are utilized in ML’s sub-field of deep learning [7]. It has been utilized in general for improving qualitative as well as quantitative prediction analyses in several disciplines [8]. Computer model learning, which is usually referred to as deep learning, directly completes the tasks of classification using text, audio, or images [9]. The DL models can achieve the highest accuracy levels and occasionally outperform humans in their performance [10]. The 1D CNN-LSTM model has been trained with the use of a multi-layered NN model, in addition to massive amounts of labeled data. An ML approach that is referred to as DL extracts features and tasks from data directly. Combining a DL method based upon LSTMs and CNNs with the Hadoop model can result in providing advanced capabilities for data analyses, facilitating accurate predictions, and enabling scalable processing of large household datasets. Those insights can assist businesses in optimizing their operations, improving sales strategies, and increasing customer satisfaction [11]. The system that has been suggested consists of 4 phases, which are: pre-processing and data collection, Hadoop framework DL, and BI. Incorporating the BI framework with DL techniques can enhance their decision-making processes and achieve more robust analytical outcomes by enhancing predictive capabilities, error rates, and overall accuracy. The comprehensive development of the proposed system involves integrating BI with big data and the Hadoop framework and enhancing it with deep learning techniques, as flexibility in data visualization plays a crucial role. This flexibility enables a clear and precise representation of results and predictions while explicitly quantifying accuracy and error rates.

2. Related Work

Hind et al., 2020 [12] utilized the whole solutions cycle of BI, which includes planning, development, and design. They evaluated the suggested business intelligence solution through several iteration tests. Their study aimed to discuss BI's role and impact in the optimization of business decisions and strategies. However, the specific results of those tests have not been stated.

The work of Zeel et al. [13] combines machine learning (ML) methods like decision tree (DT), logistic regression, stochastic gradient descent, SVMs, and multinomial naive Bayesian with NLP methods like deep learning LSTM and BERT in the year 2021. The study utilized Microsoft's Power BI for Business Intelligence (BI) analysis. The objective of the study was to improve efficiency and enhance customer satisfaction in companies. However, the study failed to specify the precise outcomes and their influence.

In 2021, Zhi-Xiong et al. [14] utilized a descriptive survey approach to study the effect of business intelligence on networked learning, innovation, and financial performance in startups. The results showed that business intelligence increased innovation by 0.99 and financial performance by 0.311. Startup intelligence positively influenced network learning (0.537), which enhanced innovativeness (0.632) and financial performance (0.397).

In 2023, Tarek et al. [15] utilized ML and DL methods to detect customer satisfaction tones from big data on social media. The Random Forest classifier achieved 99.1% F1-measure, while the Support Vector Machine performed well without preprocessing at 93.4%. The DNN algorithm showed superior performance.

Vaishali et al. [16] presented an optimized HPMR (Hadoop MapReduce) model in 2019. His work maximizes memory utilization, balances CPU and I/O system performance, and achieves about 30% better optimization than previous models when tested on Wikipedia data using the Word-Count application.

T. Lakshmi Siva Rama Krishna et al. [17] evaluated HDFS performance in 2014 for write and read operations on large and small files using a five-node Hadoop cluster. According to the findings, HDFS works well for files larger than the default block size but poorly for files smaller.

In 2022, Omar et al. presented PABIDDL, a prediction method that relies on big data analysis and DL for large-scale data [18]. It employs CNN DL for text classification, GloVe for data initialization, and MapReduce for Hadoop-based big data reduction. The method's higher performance was demonstrated empirically using IMDB and MR datasets, with a recall of 0.90%, accuracy of 0.93%, and F1-score of 0.92%.

Thaseen et al. presented a Hadoop framework in 2021 [19] that uses automatic tuning and adjusted under-sampling to identify malicious IoT traffic, minimizes computation, and leverages big data platforms. ML techniques use the immune network to optimize parameters. For the BoT_IoT and ToN_IoT datasets, an accuracy of 99% and 90% was attained. 19% more accuracy and 3–4 hours less time spent with HDFS's MapReduce feature.

Muhammad Tawfiqul Islam et al. [20] in 2022 presented a novel RL model together with job scheduling on a cloud-deployed Spark cluster to achieve SLA objectives. TF-Agent's framework implements two DRL-based schedulers that maximize executor placement and take advantage of cloud VM pricing. To shorten job durations and lower cluster VM use costs, DRL-based agents learn the characteristics of the jobs. The suggested DRL-based approach reduces VM utilization costs by up to 30%, according to the results.

3. Theoretical Background

3.1 Dataset

The big data includes all transactions made by a group of 2,500 households that frequently shop at one specific retailer. It depicts the shopping habits of these households over the course of two years. This includes complete information about every purchase made by each household, without any restrictions on particular product types or categories. The dataset also includes demographic details and historical records of direct marketing interactions for selected households [21], [22].

3.2 Pre-processing

- **Missing values**

Missing values refer to data values that are not recorded for a specific variable in an observation [23]. The presence of missing data is prevalent across various research domains and can have a substantial impact on the interpretability of data findings [24], Utilizing the provided formula restricts values within every one of the columns to a range between 0 and 1.

$$\mathrm{me} = \frac{s}{n} \ \mathrm{Eq.\,1}$$

Where:
me: Mean
s: Sum of all data points
n: Number of data points

- **Ordinal encoder**

Ordinal encoding is an ML method that gives each category a distinct integer value to transform categorical data into numerical form. This encoding enables the use of categorical variables as input for neural networks. The benefit of ordinal encoding lies in its capability to preserve the order of categories [25].

- **Normalization**

A data transformation method called min-max scaling replaces each value in a column with a new value based on a specific formula, much like z-score normalization [26]. The formula for min-max scaling is as follows:

$$\mathrm{n} = \frac{(\mathrm{V} - \mathrm{vmin})}{(\mathrm{vmax} \ - \mathrm{vmin})} \ \mathrm{Eq.\,2}$$

Where: n is our new value
V represents the original cell value
vmin denotes the minimum value of the column, while the vmax represents the maximum value of the same column
Applying this formula ensures that the values of each column will be transformed to fall within the range of 0 to 1.

3.3 Hadoop Framework

Hadoop is a powerful framework that enables node-based systems to store massive amounts of data [27]. Hadoop design makes use of multiple components to enable parallel data processing:

1- Using Hadoop HDFS for storing data across slave machines will guarantee high availability and fault tolerance.

2- Hadoop cluster uses Hadoop YARN for resource management, effectively allocating resources for optimal performance.

3- Data can be processed in a distributed manner using Hadoop MapReduce, which enables scalable and parallel data processing.

4- To guarantee synchronization throughout a cluster and promote reliable and seamless operations, use Zookeeper.

HDFS, or Hadoop Distributed File System, comprises three key components:

- NameNode: Acts as the central authority, storing and managing file system metadata in RAM and on disk.
- Secondary NameNode: Maintains a copy of NameNode metadata on disk, providing additional reliability.
- DataNode: Stores data as blocks, forming a distributed storage infrastructure.

Together, these components provide distributed storage, reliability, and effective data management inside HDFS. Figure 1 illustrates the structure of Hadoop.

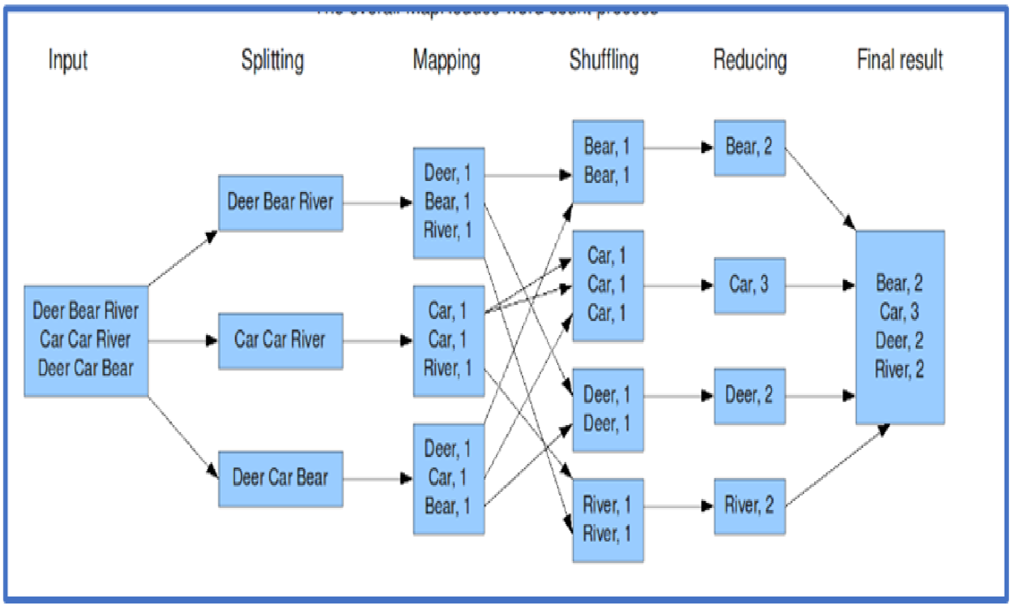


Figure 1: Hadoop MapReduce deep diving and tuning [28]

3.4 Convolutional Neural Network

A DL algorithm called a CNN is used for tasks involving image processing and recognition. Pooling, convolutional, and fully connected layers of FCL make up this structure. Convolutional layers use filters to extract features, like textures, shapes and edges, from input images. After that, the result is run through a series of pooling layers to down-sample feature maps while maintaining all relevant data. Using the processed features, FCLs then classify or generate predictions about the image. Figure 2 depicts the 1D CNN structure [29], [30]. The convolution layer, connecting input x_i^k and output y_i^k, performs matrix multiplication between the previous layer's input and a filter bank w_{ij}^k,it also includes the addition of a regularization term (bias) b_j^k. Eq. 3 can be used to compute the convolution layer [31]:

$$y_i^k = \sum\nolimits_i \left(x_i^k * w_{ij}^k\right) + b_j^k \text{ Eq. 3}$$

The activation function of the nonlinear layer, which addresses the Dying Relu problem and prevents the exclusion of negative inputs, is a Leaky rectified linear unit (LeakyRelu). The mathematical expression for LeakyRelu can be found in Eq. 4.

$$f(x)LeakyRelu = \begin{Bmatrix} x, if\ x\ >\ 0 \\ mx, x\ \leq\ 0 \end{Bmatrix} \text{ Eq. 4}$$

The leak factor *mx* has a small value (e.g., 0.001), introducing a slight negative slope in the activation function for improved learning, allowing information flow even for negative inputs. The maximum pooling layer selects the highest value in non-overlapping feature matrices. The fully connected layer can use a nonlinear activation function like Leaky ReLU. For classification tasks, the output layer employs SoftMax activation, while linear regression is used for regression tasks. Equation (5) can be used to compute both SoftMax and linear regression functions.

$$\sigma\ (\text{v})\text{i} = \frac{e^{vi}}{\sum_{j=1}^{K} e^{vj}}\ for\ i = 1, \ldots\ldots\ldots, k\ and\ v = (v1, \ldots\ldots\ldots.vk)\ \in$$

$$R^K \text{ Eq. 5}$$

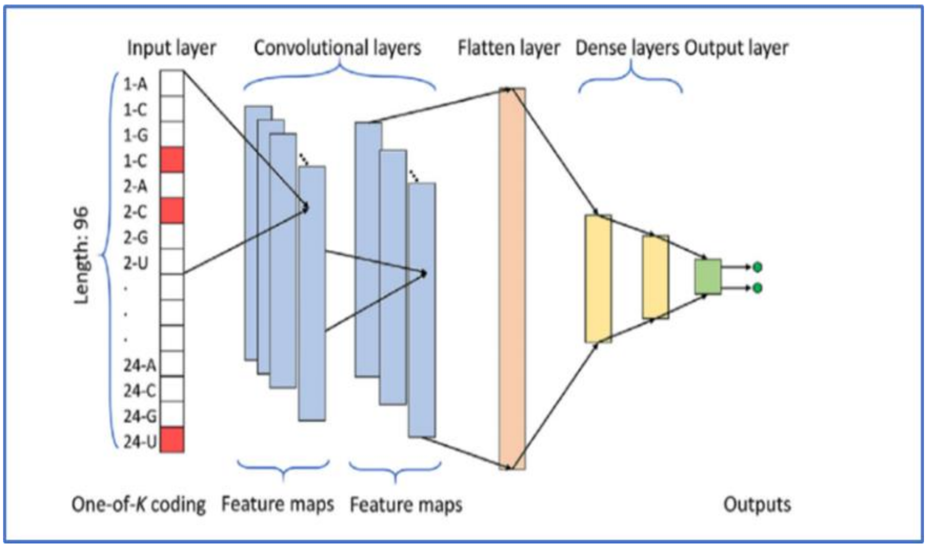


Figure 2: 1D-CNN model architecture. It consists of two convolutional layers, one flattening layer, two dense layers and one output layer [32]

3.4 Long Short-Term Memory

Recurrent neural networks RNNs of the Long Short-Term Memory (LSTM) type are frequently employed in AI and DL [33]. Because LSTM has feedback connections built in, unlike feed-forward NNs, it can analyze complete data sequences instead of simply single data points like video, audio, or image [34]. When it comes to tasks like classification and prediction with time series data, LSTM networks are a great fit. They deal with the vanishing gradient issue that comes up frequently when training regular RNNs. LSTMs frequently outperform RNNs, hidden Markov models, and other sequence-based learning techniques in terms of accuracy and performance because of their capacity to accommodate different sequence gap lengths [35].

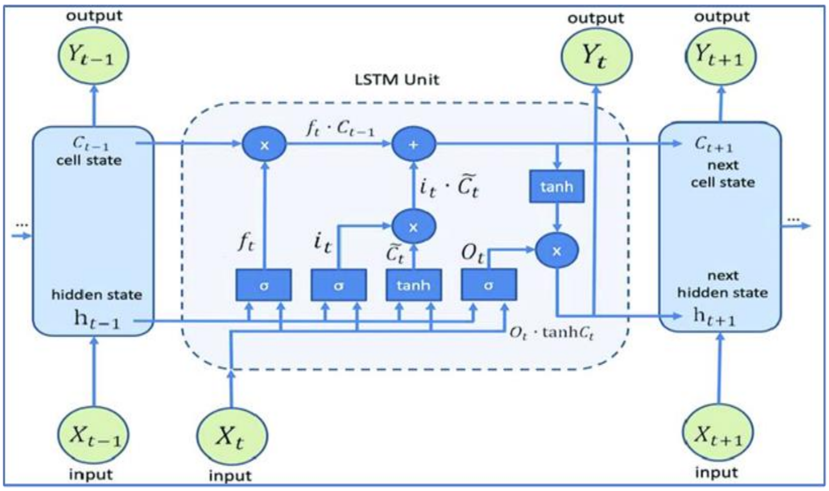


Figure 3: The Structure of the LSTM Model [36]

The sigmoid function σ, as depicted in Eq. 12 in Figure 3, is utilized. Its output ranges from 0 to 1, where 0 indicates blocking and 1 signifies allowing complete passage.

$$f(t) = \sigma(wf.[ht-1, xt\,] + bf) \text{ Eq. 6}$$

$$I(t) = \sigma(wf.[ht-1, xt\,] + bi) \text{ Eq. 7}$$

$$\tilde{c}(t) = tanh(wc.[ht-1, xt\,] + bc) \text{ Eq. 8}$$

$$c(t) = fi * ct - 1 + It * \tilde{c}t \text{ Eq. 9}$$

$$o(t) = \sigma(wo.[ht-1, xt\,] + bo) \text{ Eq. 10}$$

$$h(t) = ot * \tanh(c1) \text{ Eq. 11}$$

$$sigmoid\ (x) = \frac{1}{1 + e^{-x}} \text{ Eq. 12}$$

$$tanh\ (x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \text{ Eq. 13}$$

To solve the gradient vanishing problem, the hyperbolic tangent function in Eq. 13 is utilized [37, 38]. The output and input of the LSTM network structure shown in Figure 3 are defined by Eqs. 6-13.

4. Methodology

Hadoop can be defined as an open-source software system designed to store, manage, and analyze large volumes of data. It addresses the scalability and cost-effectiveness challenges of traditional relational database management systems (RDBMS) in handling massive amounts of data. Unlike RDBMS, Hadoop excels at managing unstructured and semi-structured data. Its flexibility allows seamless integration with other big data technologies, making it an ideal choice for data administration and analysis [39]. Deep learning efficiently manages large-scale data within time and cost constraints. By leveraging advanced algorithms and abundant data, it enhances decision-making speed and accuracy, revolutionizing various sectors [40]. CNN excels in feature extraction, while LSTM processes spatially encoded features sequentially [41, 42]. Combining them into a hybrid CNN-LSTM model optimizes their advantages. The structure of the hybrid DL model is shown in Figure 5.

For testing the big data set on the suggested system, it was obtained from a free online source. The local system might include the dataset. Preprocessing the dataset in the first stage involves addressing missing values using an ordinal encoder as well as a normalization scaler. The Hadoop stage begins with loading the dataset from the local system into the Hadoop cluster. Next, the system divides the input data set's total size by the block size that was previously defined in the Hadoop configuration files, resulting in data splits that are sequentially transferred from the local machine into HDFS. Based on the replication factor parameter dfs, each split of data will be replicated inside the cluster, depending on the parameter of replication factor ***dfs.replication***. Replication was enabled in the ***hdfs-site.xml*** systems configuration file to prevent data loss in the event of a node failure; Figure 4 shows how data loads into HDFS.

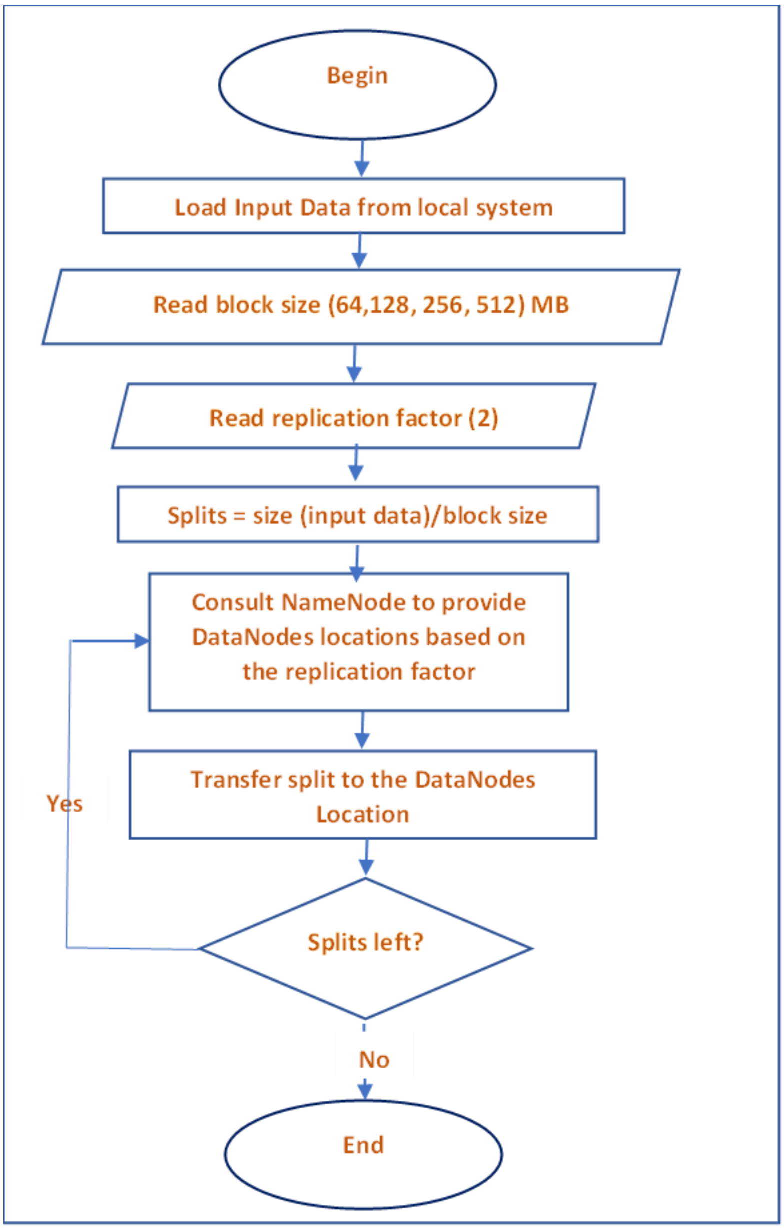

Begin
Load Input Data from local system
Read block size (64,128, 256, 512) MB
Read replication factor (2)
Splits = size (input data)/block size
Consult NameNode to provide DataNodes locations based on the replication factor
Transfer split to the DataNodes Location
Yes
Splits left?
No
End

Figure 4: Loading data set into the *HDFS*

The next step in the proposed system is to build the data warehouse, which will serve as the primary source layer for the DL stage. After loading the data set into HDFS and replicating it on the cluster, the hybrid DL model will run on the data warehouse to extract the desired business insights. Deep learning is a potent solution for efficiently managing vast data volumes under tight time and cost constraints. Deep learning efficiently analyzes complex information, improving decision-making speed and accuracy by leveraging advanced algorithms and massive amounts of data. Deep learning has the potential to revolutionize various sectors, driving significant improvements in production, profitability, and efficiency [43]. Common deep learning algorithms include CNN and LSTM. One-dimensional convolutional neural network (1D CNN) excels at feature extraction [44], whereas LSTM sequentially processes spatially encoded 1D CNN features. Combining these models yields an effective hybrid CNN-LSTM structure, as shown in Figure 5.

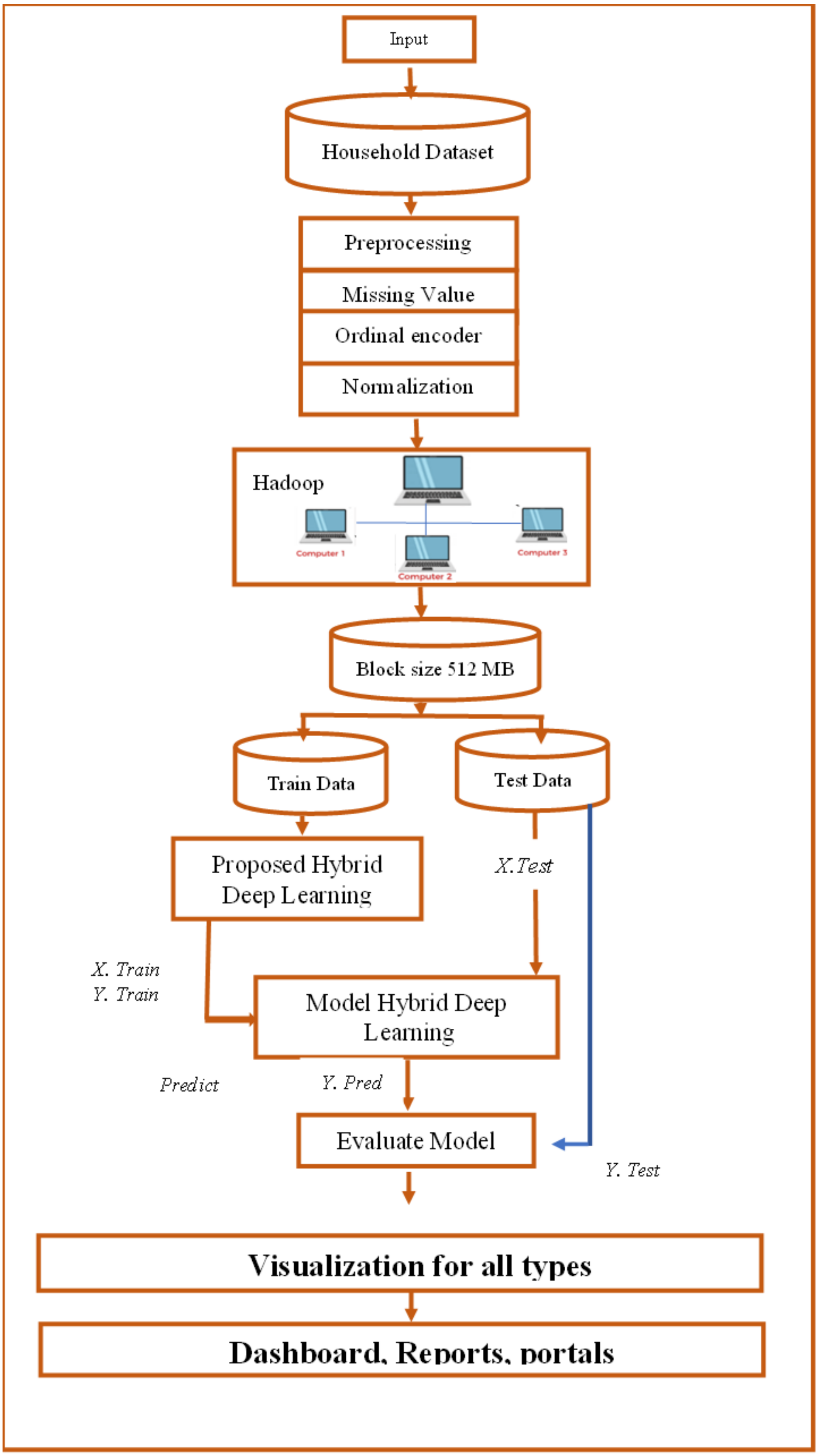


Figure 5: Hybrid Deep Learning Model Structure

4.1 Proposed Hybrid Deep Learning Model

While the LSTM captures temporal features and uses them for prediction, the CNN captures spatial features. The hybrid DL model uses a combination of temporal and spatial data to improve prediction accuracy. The four stages of the suggested system are data collection and preprocessing, Hadoop framework and data warehouse, DL, and BI, which is the final level and allows for the visualization of the insights using any visualization tool, such as a dashboard. The system was developed to address the difficulty of organizing and interpreting vast volumes of data. To do this, it makes use of the DL and Hadoop platforms. First, it loads the desired block size from the HDFS-stored data warehouse after splitting the data block into a 30% test dataset and a 70% training dataset. The training dataset is used to train the model, while the test dataset is used to evaluate it. The CNN+LSTM hybrid DL model was assessed using MSE and MAE. When a BI system incorporates a DL regression model, it becomes a predictive system that aims to make precise predictions. The method provides insightful information and enables well-informed decision-making depending on projected outcomes by utilizing such predictions. Twenty layers make up the hybrid DL model, as seen in Figure 6. The four layers make up the CNN component: 16 convolution filters of size 3 × 1, strides of 1, and valid padding in the first layer; 32 convolution filters of size 3 × 1, strides of 1, and valid padding in the second layer; 64 convolution filters of size 3 × 1, strides of 1, and valid padding in the third layer; and 32 convolution filters of size 3 × 1, strides of 1, and valid padding in the fourth layer. Three layers make up the LSTM component: 16 units and true return sequences in the first layer, 64 units and true return sequences in the second layer, and 32 units and false return sequences in the third layer. Six max-pooling layers are also included, with pool size and strides set to 1, using the Leaky Rectified Linear Unit (Leaky ReLU) activation function and Adam optimization. With a single class and a linear activation function, the final layer is dense. Please refer to **Table 1** for more information on the parameter settings.

Table 1: Parameter Setting of The Hybrid Deep Learning Model

Parameters	Value
Convolutional Layer Filters (conv-1, conv-2, conv-3, conv-4)	(16, 32, 64, 32)
Convolutional Layer kernal_size	3
Convolutional Layer Stride	1
Convolutional Layer Padding	valid
Pooling Layer pool_size	1
Pooling layer activation functions	Leaky Relu
LSTM Layer (output, return sequences)	((16, True), (64, True), (32, False))
LSTM activation function	Linear
Learning Rate	0.001
Optimizer	Adam
Epochs	100
Batch_size	64
Loss function	(MAE, MSE, RMAE)

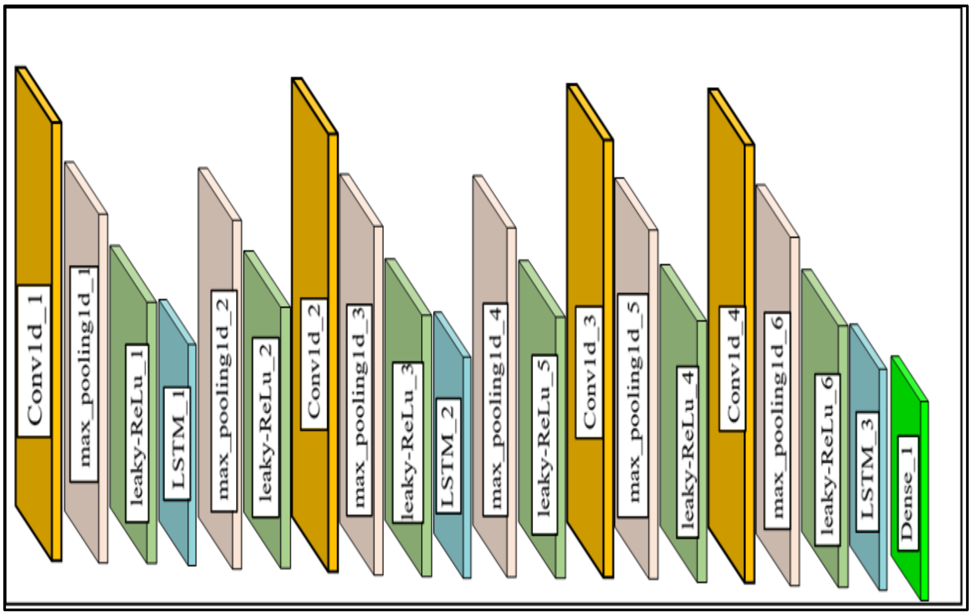


Figure 6: The Proposed Model Structure of Hybrid Deep Learning

5. Results and Discussions

5.1Evaluation of the models

Once the model's structure is defined, it is trained using the training set until convergence is achieved [45]. Error metrics, also referred to as performance metrics or evaluation metrics, serve as quantitative measures to assess the accuracy and quality of predictions made by a model. Error metrics play a significant role in quantitatively evaluating the accuracy and performance of models. They enable meaningful comparisons between different models and facilitate the assessment of model enhancements. By utilizing these metrics, researchers can make informed decisions regarding model selection, fine-tuning, and overall improvement. To assess the model's effectiveness, three error metric formulas were employed in this work: MAE, MSE, and RMAE. Let Xi, i = 1,..., n denote the n validation data prediction scores derived for prediction methods (e.g., deep neural networks or machine learning methods) built using the training dataset. Further, let Y i, i = 1,…, n denote the n actual values of the target variable contained in the validation dataset. Last, let Zi = Yi - Xi, i = 1,..., n denotes the n scoring (prediction) errors (over the validation dataset) associated with the chosen prediction method [46]. The validation dataset is commonly evaluated using the following scoring measures to assess prediction methods:

$$\mathrm{MAE} = \frac{\sum_{\mathrm{i=1}}^{\mathrm{n}} |\mathrm{Z}_i|}{n} \quad \text{Eq. 14}$$

$$\mathrm{MSE} = \frac{\sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{Z}_i^2}{n} \quad \text{Eq. 15}$$

$$\mathrm{RMAE} = \sqrt{\mathrm{MAE}} \quad \text{Eq. 16}$$

5.2 Performance Comparison of Hybrid Deep Neural with Machine Learning Models

To assess the efficiency of the suggested hybrid CNN-LSTM deep learning model, it was compared against two different machine learning models. The models were implemented using the same data block size of 512 MB of training and test datasets. Table 2 provides error metrics used for the hybrid deep learning model CNN+LSTM for data block sizes 64 and 512 MB to achieve optimal performance, while Table 3 provides error metrics used for each machine learning model for block size 512 MB to achieve optimal performance. The results are presented in Table 4 and Table 5. Figure 7 shows the error metric in the hybrid deep neural learning model.

Table 2: Comparison of Hybrid Deep Learning Model on block size 512 MB within attributes of transaction file

	1D-CNN+LSTM MODEL	BLOCK SIZE 512			BLOCK SIZE 64		
No.	TRANSACTION FILES	MAE	MSE	R M A E	MAE	MSE	R M A E
1	RETAIL_DISC	0.000590111	6.65959E-06	0.02429219	0.00632282	0.000149882	0.079516164
2	TRANS_TIME	4.361611377	3.020255649	2.088447121	4.360771573	3.020257343	2.088246052
3	QUANTITY	0.013351193	0.014453723	0.115547363	0.011944454	0.008232062	0.109290687
4	DAY	0.702847513	0.676221558	0.838360014	0.702961958	0.676249268	0.838428267
5	WEEK_NO	0.000258706	1.09563E-07	0.016084328	0.000144298	2.74764E-08	0.012012397
6	BASKET_ID	2.509011634	7.632322688	1.583985996	2.509013653	7.632322711	1.583986633
7	PRODUCT_ID	2.259307867	6.700851097	1.50309942	2.25938953	6.700860581	1.503126585
8	STORE_ID	0.677580037	1.348844612	0.823152499	0.00700992	0.000505149	0.083725263
9	SALES_VALUE	0.004379282	0.000250901	0.066176141	0.002896977	0.00019385	0.053823577
10	HOUSEHOLD-KEY	0.63335951	5.306799835	0.795838872	0.633360978	5.306799359	0.795839794
11	COUPON_DISC	2.04709E-05	1.29E-07	0.004524483	1.69775E-05	3.45539E-07	0.004120379
12	COUPON_MATCH_DISC	2.6716E-06	6.61799E-10	0.001634504	5.92061E-05	1.34805E-07	0.007694549

Table 3: ML Models on data block size 512 MB within attributes of transaction file

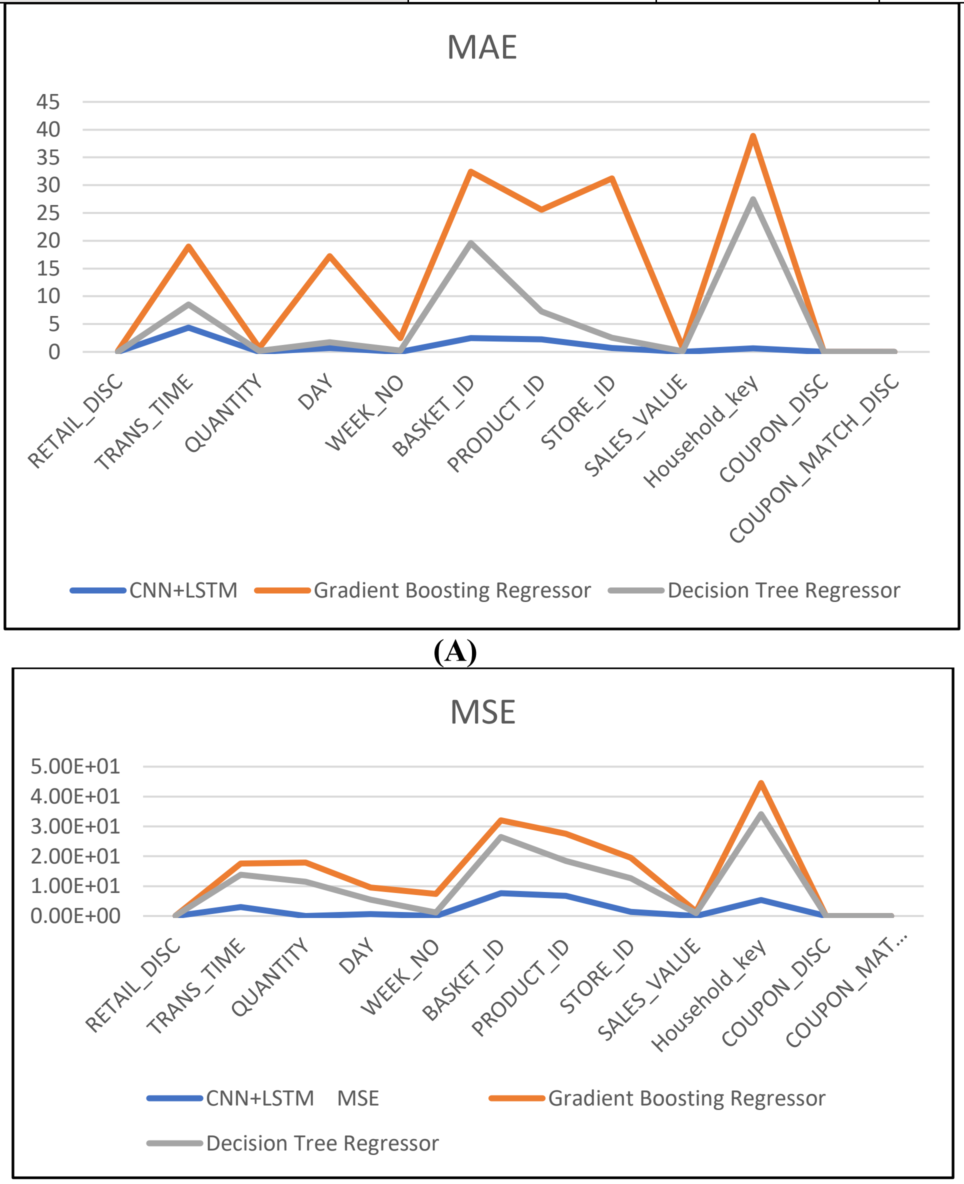
No.	TRANSACTION FILES	Decision Tree Regressor			Gradient Boosting Regressor		
		MAE	MSE	RMAE	MAE	MSE	RMAE
1	RETAIL_DISC	0.026115044	0.05487987	0.161601497	0.183801246	0.137621408	0.428720476
2	TRANS_TIME	8.545125654	13.86720645	2.923204689	18.96464525	17.52599899	4.354841587
3	QUANTITY	0.187550568	11.48618866	0.433071089	0.69350636	17.89754949	0.832770293
4	DAY	1.73408212	5.515774367	1.316845519	17.22265916	9.510863288	4.150019175
5	WEEK_NO	0.216256126	1.13504933	0.465033467	2.464597943	7.442102527	1.569903801
6	BASKET_ID	19.61018769	26.43753307	4.428339157	32.43437794	32.01680414	5.695118782
7	PRODUCT_ID	7.239965668	18.47223366	2.69071843	25.60235168	27.58682195	5.059876647
8	STORE_ID	2.563953005	12.66814886	1.601234838	31.24105221	19.47591899	5.589369572
9	SALES_VALUE	0.097067209	0.941483774	0.311556109	0.760306908	1.520290069	0.871955795
10	HOUSEHOLD-KEY	27.47605665	34.09864827	5.241760835	38.76128725	44.52990844	6.225856346
11	COUPON_DISC	0.001149976	0.005247464	0.033911294	0.001799456	0.006385639	0.042419991
12	COUPON_MATCH _DISC	0.000143174	4.92E-05	0.011965534	0.000311238	3.83E-05	0.017641946

Table 4: Comparison of Hybrid DL Model with ML Models on data block size 512MB within attributes of transaction file for Mean Absolute Error (MSE)

No.	TRANSACTION FILES	1D CNN+LSTM	Gradient Boosting Regressor	Decision Tree Regressor
1	RETAIL_DISC	0.000590111	0.183801246	0.026115044
2	TRANS_TIME	4.361611377	18.96464525	8.545125654
3	QUANTITY	0.013351193	0.69350636	0.187550568
4	DAY	0.702847513	17.22265916	1.73408212
5	WEEK_NO	0.000258706	2.464597943	0.216256126
6	BASKET_ID	2.509011634	32.43437794	19.61018769
7	PRODUCT_ID	2.259307867	25.60235168	7.239965668
8	STORE_ID	0.677580037	31.24105221	2.563953005
9	SALES_VALUE	0.004379282	0.760306908	0.097067209
10	HOUSEHOLD	0.63335951	38.87612872	27.47605665
11	COUPON_DISC	2.04709E-05	0.001799456	0.001149976
12	COUPON_MATCH_DISC	2.6716E-06	0.000311238	0.000143174

Table 5: Comparison of Hybrid DL Model with ML Models on data block size 512MB within attributes of transaction file for Mean Squared Error (MSE)

No.	TRANSACTION FILES	1D CNN+LSTM	Gradient Boosting Regressor	Decision Tree Regressor
1	RETAIL_DISC	6.65959E-06	0.137621408	0.05487987
2	TRANS_TIME	3.020255649	17.52599899	13.86720645
3	QUANTITY	0.014453723	17.89754949	11.48618866
4	DAY	0.676221558	9.510863288	5.515774367
5	WEEK_NO	1.09563E-07	7.442102527	1.13504933
6	BASKET_ID	7.632322688	32.01680414	26.43753307
7	PRODUCT_ID	6.700851097	27.58682195	18.47223366
8	STORE_ID	1.348844612	19.47591899	12.66814886
9	SALES_VALUE	0.000250901	1.520290069	0.941483774
10	HOUSEHOLD-KEY	5.306799835	44.52990844	34.09864827
11	COUPON_DISC	1.29E-07	0.006385639	0.005247464
12	COUPON_MATCH_DISC	6.61799E-10	3.83E-05	4.92E-05



(A)

(B)

Figure 7: Represents the performance Evaluation Results of Predicting Models:

(A) : MAE

(B) : MSE

6. Conclusion

This study proposes a hybrid DL regression model (1D CNN-LSTM) for prediction. After testing the model using 512-data block size transaction files of the household data set and with two separate ML models (DT Regressor as well as Gradient Boosting Regressor), the accuracy of each model was determined, and the hybrid 1D CNN+LSTM model yielded the best results through, firstly, overcoming limitations in data storage capacity often encountered in traditional BI systems, thereby enhancing their analytical capabilities. Secondly, this combination of deep learning techniques with the BI framework improves decision-making processes by enhancing predictive capabilities, reducing error rates, and ultimately leading to more robust analytical outcomes. Lastly, the proposed system comprehensively integrates BI with big data, Hadoop, and deep learning techniques. This combination emphasizes the critical role of flexible data visualization. Such flexibility allows for clear and precise representation of results and predictions, along with explicit quantification of accuracy and error rates. Additionally, by creating a model that incorporates RNN, long short-term memory, and convolutional neural networks CNN-LSTM-RNN, we hope to increase the prediction accuracy even further.

Based on our analysis, we recommend the following actions and potential paths for further research and development looking into the future:

1. To ensure that the model remains correct and relevant as the dataset changes or new information becomes available, perform continuous model monitoring and updates. Periodically retraining the model can help maintain its accuracy and effectiveness over time.
2. Use data augmentation strategies to improve overall robustness by increasing the model's generalization on large-scale data sets.
3. To address unbalanced data in a dataset, investigate techniques such as oversampling, undersampling, or alternate assessment metrics to ensure that the final results are more accurate and reliable.
4. To improve the performance of the hybrid 1D CNN-LSTM model, it would be worth exploring hyperparameter fine-tuning while considering various structures, learning rates, and regularization strategies. This could potentially lead to significant improvements in the model's accuracy and overall performance.

REFERENCES

[1] B. Kapoor and J. Sherif, "Human resources in an enriched environment of business intelligence," *The International Journal of Systems & Cybernetics,* vol. 41, no. 10, pp. 1625-1637, Oct. 2012. doi: 10.1108/03684921211276792.

[2] A. H. Ali, S. A. Dheyab, A. H. Alamoodi, A. A. Magableh, Y. Gu, " Leveraging AI and Big Data in Low-Resource Healthcare Settings," *Mesopotamian Journal of Big*, vol. 2024, pp. 11–22, Feb 2024, doi:10.58496/MJBD/2024/002.

[3] S. Gholami, E. Zarafshan, R. Sheikh, and S.S. Sana, "Using deep learning to enhance business intelligence in organizational management," *Data Science in Finance and Economics*, vol. 3*,* no. 4, pp. 337-353, Oct. 2023. doi: 10.3934/DSFE.2023020.

[4] V. Sontakke and R. B. Dayanand, "Optimization of Hadoop MapReduce Model in Cloud Computing Environment," *International Conference on Smart Systems and Inventive Technology (ICSSIT)*, Tirunelveli, pp. 510-515, Nov. 2019. doi: 10.1109/ICSSIT46314.2019.8987823.

[5] K. Rattanaopas and S. Kaewkeeree, "Improving Hadoop MapReduce performance with data compression: A study using wordcount job," *International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, Phuket, pp. 564-567, Jun. 2017. doi: 10.1109/ECTICon.2017.8096300.

[6] C. Vorapongkitipun and N. Nupairoj, "Improving performance of small-file accessing in Hadoop," *11th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, Chon Buri, pp. 200-205, May. 2014. doi: 10.1109/JCSSE.2014.6841867.

[7] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A Fadhel, M. Al-Amidie, and L. Farhan "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions." *Journal of Big Data*, vol. 8, no. 1, March. 2021. doi: 10.1186/s40537-021-00444-8

[8] M. Ahsan and K. E. Nygard. "Convolutional Neural Networks with LSTM for Intrusion Detection," *In Proceedings of 35th International Conference on Computers and Their Applications* (*CATA)*, vol. 69, pp. 69-79. 2020. doi:10.13140/RG.2.2.24796.82567.

[9] G. Huang, Q. Shen, G. Zhang, P. Wang, Z. YuHuang, and Guohua, "LSTM CNN succ: a bidirectional LSTM and CNN-based deep learning method for predicting lysine succinylation sites," *BioMed Research International,* vol. 2021, no. 3, pp. 1-10, May. 2021. doi: 10.1155/2021/9923112.

[10] K. Ullah and M. Qasim, "Google Stock Prices Prediction Using Deep Learning," *10th International Conference on System Engineering and Technology (ICSET)*, *Shah Alam,* pp. 108-113, 2020. doi: 10.1109/ICSET51301.2020.9265146.

[11] A. Begum, V. D. Kumar, J. Asghar, D. Hemalatha, and G. Arulkumaran, "A Combined Deep CNN: LSTM with a Random Forest Approach for Breast Cancer Diagnosis", *Hindawi: Complexity*, vol. 2022, Sep. 2022. doi: 10.1155/2022/9299621

[12] H. F. Alaskar, and T. Saba, "Application of Business Intelligence Solution Development and Implementation in a Small-Sized Enterprise," *First International Conference of Smart Systems and Emerging Technologies (SMARTTECH)*, Riyadh, pp. 183-190, 2020 doi: 10.1109/SMART-TECH49988.2020.00051.

[13] Z. Desai, K. Anklesaria and H. Balasubramaniam, "Business Intelligence Visualization Using Deep Learning Based Sentiment Analysis on Amazon Review Data," *12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pp. 1-7, 2021. doi: 10.1109/ICCCNT51525.2021.9579786.

[14] Z. Huang, K.S. Savita, and J. Zhong-Jie, "The Business Intelligence impact on the financial performance of start-ups," *Information Processing & Management*, vol. 59, no. 1, Jan. 2022. doi: 10.1016/j.ipm.2021.102761.

[15] T. Kanan, and A. Mughaid," Business intelligence using deep learning techniques for social media contents" *Cluster Computing*, vol. 26, pp. 1285–1296, 2023. doi: 10.1007/s10586-022-03626-y.

[16] V. Sontakke and R. B. Dayanand, "Optimization of Hadoop MapReduce Model in Cloud Computing Environment," *2019 International Conference on Smart Systems and Inventive Technology (ICSSIT)*, Tirunelveli, pp. 510-515, 2019. doi: 10.1109/ICSSIT46314.2019.8987823.

[17] T. L. S. R. Krishna, T. Ragunathan and S. K. Battula, "Performance Evaluation of Read and Write Operations in Hadoop Distributed File System," *Sixth International Symposium on Parallel Architectures, Algorithms and Programming, Beijing,* pp. 110-113, 2014, doi: 10.1109/PAAP.2014.49

[18] O. Haddad, F. Fkih, and M. N. Omri, "Towards a Prediction Approach based on Deep Learning in Big Data Analytics," *Neural Computing and Applications*, vol. 35, pp. 6043–6063, Nov.2022. doi: 10.1007/s00521-022-07986-9.

[19] I. S. Thaseen, V. Mohanraj, S. Ramachandran, Ki. Sanapala and S. Yeo," A Hadoop-Based Framework Integrating Machine Learning Classifiers for Anomaly Detection in the Internet of Things," *Electronics*, vol. 10, no. 16, Aug. 2021. doi: 10.3390/electronics10161955.

[20] M. T. Islam, S. Karunasekera and R. Buyya, "Performance and Cost-Efficient Spark Job Scheduling Based on Deep Reinforcement Learning in Cloud Computing Environments," in *IEEE Transactions on Parallel and Distributed Systems*, vol. 33, no. 7, pp. 1695-1710, July 2021. doi: 10.1109/TPDS.2021.3124670.

[21] Dunnhumby the Complete Journey, "transaction_data.csv," [Online]. Available: https://www.kaggle.com/datasets/frtgnn/dunnhumby-the-complete-journey?resource=download.

[22] S. A. Razoqi, and G. A.A. Al-Talib, " A Survey Study on Proposed Solutions for Imbalanced Big Data," *Iraqi Journal of Science*, vol. 65, no. 3, pp: 1648-1662, 2024. doi: 10.24996/ijs.2024.65.3.37.

[23] J. W. Graham, "Missing data analysis: Making it work in the real world," *Annual review of psychology*, vol. 60, pp. 549-576, Jan. 2009. doi: 10.1146/annurev.psych.58.110405.085530.

[24] R. J. Little, R. D'Agostino, M. L. Cohen, K. Dickersin, S. S. Emerson, J.T. Farrar, C. Frangakis, J. W. Hogan, G. Molenberghs, S. A. Murphy, J. D. Neaton, A. Rotnitzky, D. Scharfstein, W. J.

Shih, J.P. Siegel, and H. Stern, "The prevention and treatment of missing data in clinical trials," *New England Journal of Medicine*, vol. 367, no. 6, pp. 1355-1360, Oct. 2012. doi: 10.17226/12955.

[25] X. Meng, C. Fan, Y. Ming, and H. Yu, "CORNet: Context-Based Ordinal Regression Network for Monocular Depth Estimation," *in IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 7, pp. 4841- 4853, July 2022. doi: 10.1109/TCSVT.2021.3128505.

[26] K. Bringmann, M. Künnemann, and K. Węgrzycki, "Approximating APSP without scaling: equivalence of approximate min-plus and exact min-max," *Computer Science - Data Structures and Algorithms (CS-DS)*, pp. 943-954, June. 2019. doi: 10.48550/arXiv.1907.11078.

[27] X. Wu, "A MapReduce Optimization Method on Hadoop Cluster," *International Conference on Industrial Informatics - Computing Technology, Intelligent Technology, Industrial Information Integration*, Wuhan, pp. 18-21, 2015. doi: 10.1109/ICIICII.2015.92.

[28] T. Lăpușan," Hadoop MapReduce deep diving and tuning," *Today Software Magazine (TSM*), no. 33, 2020.

[29] S. Mehtab and J. Sen, "Stock Price Prediction Using CNN and LSTM-Based Deep Learning Models," *International Conference on Decision Aid Sciences and Application (DASA), Sakheer,* vol. 1, pp. 447-453, Nov. 2020. doi: 10.1109/DASA51403.2020.9317207.

[30] H. A. Ahmed and E. A. Mohammed, "Detection and Classification of The Osteoarthritis in Knee Joint Using Transfer Learning with Convolutional Neural Networks (CNNs)," *Iraqi Journal of Science*, vol. 63, no. 11, pp. 5058–5071, Nov. 2022, doi: 10.24996/ijs.2022.63.11.40..

[31] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M.A. Fadhel, M. Al-Amidie, and L. Farhan**,"** Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *Journal of Big Data*, vol. 8, no. 53, Mar. 2021. doi: 10.1186/s40537-021-00444-8.

[32] Z., Qi, M. Qian, Z. Zheng, D. Tong-Yi, W. Zhiguo, C. Xiaoyu, L. Yuanning, and F. Xiaoya, "Prediction of plant-derived xenomiRs from plant miRNA sequences using random forest and one-dimensional convolutional neural network models," *BMC Genomics*, vol. 19, no. 839, Nov. 2018. doi: 10.1186/s12864-018-5227-3.

[33] I. N. Yulita, D. F. I. Manurung, and I. Suryana, "Machine Learning Approach for New COVID-19 Cases Using Recurrent Neural Networks and Long-Short Term Memory," *Iraqi Journal of Science*, vol. 64, no. 11, pp: 5887- 5895, 2023. doi: 10.24996/ijs.2023.64.11.34.

[34] H. S. Abdullah, N. H. Ali, and N. A.Z. Abdullah, "Evaluating the Performance and Behavior of CNN, LSTM, and GRU for Classification and Prediction Tasks," *Iraqi Journal of Science*, vol. 65, no. 3, pp: 1741-1751, 2024. doi: 10.24996/ijs.2024.65.3.43.

[35] M. A. Istiake Sunny, M. M. S. Maswood and A. G. Alharbi, " Deep Learning-Based Stock Price Prediction Using LSTM and Bi-Directional LSTM Model," *2nd Novel Intelligent and Leading Emerging Sciences Conference (NILES)*, Giza, pp. 87-92, 2020. doi: 10.1109/NILES50944.2020.9257950.

[36] C. Jiang, Y. Chen, S. Chen, Y. Bo, W. Li, T. Wenxin and J. Guo," A Mixed Deep Recurrent Neural Network for MEMS Gyroscope Noise Suppressing," *Electronics*, vol. 8, no. 2, Feb. 2019. doi: 10.3390/electronics8020181.

[37] S. Selvin, R. Vinayakumar, E. A. Gopalakrishnan, V. K. Menon, and K. P. Soman, "Stock price prediction using LSTM, RNN and CNN-sliding window model," *International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Udupi, pp. 1643-1647, 2017. doi: 10.1109/ICACCI.2017.8126078.

[38] K. Magulova, and A. P. James, "A survey on LSTM memristive neural network architectures and applications," *The European Physical Journal Special Topics*, vol. 228, pp. 2313-2324, October 2019. doi: 10.1140/epjst/e2019-900046-x.

[39] P. M. Bante and K. Rajeswari, "Big Data Analytics Using Hadoop Map Reduce Framework and Data Migration Process," *International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, Pune, pp. 1-5, 2017. doi: 10.1109/ICCUBEA.2017.8463824.

[40] D. Goularas and S. Kamis, "Evaluation of Deep Learning Techniques in Sentiment Analysis from Twitter Data," *International Conference on Deep Learning and Machine Learning in Emerging Applications (Deep-ML)*, pp. 12-17, 2019. doi: 10.1109/Deep-ML.2019.00011.

[41] A. M. Dhayea, N. K. El Abbadi, and Z. G. Abdul Hasan "Human Skin Detection and Segmentation Based on Convolutional Neural Networks," *Iraqi Journal of Science*, vol. 65, no. 2, pp. 1102-1116, 2024. doi: 10.24996/ijs.2024.65.2.40.

[42] I. N. Mahmood, and H. S. Abdullah “Telecom Churn Prediction based on Deep Learning Approach,” *Iraqi Journal of Science*, vol. 63, no. 6, pp. 2667-2675, 2022. doi: 10.24996/ijs.2022.63.6.32.

[43] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *in Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278-2324, Nov. 1998. doi: 10.1109/5.726791.

[44] T. Kim, and S. Cho, "Predicting residential energy consumption using CNN-LSTM neural networks. Energy," *Elsevier: Science Direct*, vol. 182, no. 10 pp. 72-81, Sep. 2019. DOI:10.1016/j.energy.2019.05.230.

[45] S. Jayawardena, J. Epps, and E. Ambikairajah, **"**Evaluation Measures for Depression Prediction and Affective Computing**,**" *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton,* pp. 6610-6614, 2019. doi: 10.1109/ICASSP.2019.8682956.

[46] L. Skylar and M. R. Smith, "Evaluation of several Efron bootstrap methods to estimate error measures for software metrics," *Canadian Conference on Electrical and Computer Engineering (CCECE2002),* vol. 2, no. 6, pp. 703-708, 2002. doi: 10.1109/CCECE.2002.1013027.