ISSN:2229-6107

INTERNATIONAL JOURNAL OF PURE AND APPLIED SCIENCE & TECHNOLOGY

E-mail :
editor.ijpast@gmail.com
editor@ijpast.in

www.ijpast.in

INTERNATIONAL JOURNAL OF PURE AND APPLIED SCIENCE & TECHNOLOGY
IJPAST

ISSN 2229-6107
www.ijpast.in
Vol 12,Issuse 3. July 2022

INTUITIVE DIMENSIONALITY REDUCTION ON BIG DATA USING DL AND HDFS

Mule Rama Krishna Reddy 1, Dr. O Naga Raju 2

ABSTRACT:

In today's data-driven world, the proliferation of big data poses significant challenges in terms of data analysis, visualization, and scalability. Dimensionality reduction techniques are pivotal for simplifying and understanding complex data structures. This paper introduces an innovative approach that harnesses the power of Deep Learning (DL) in conjunction with the Hadoop Distributed File System (HDFS) for intuitive dimensionality reduction on large-scale datasets. Traditional dimensionality reduction methods, such as Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbour Embedding (t-SNE), often face limitations when handling massive datasets. In contrast, DL offers the potential to automatically learn intricate hierarchical representations, making it an attractive candidate for dimensionality reduction. Our proposed methodology seamlessly integrates the scalability and parallel processing capabilities of HDFS, thereby efficiently managing data preprocessing and DL model training.

Keywords: Dimensionality Reduction, Deep Learning, Hadoop Distributed File System (HDFS), Big Data, Data Preprocessing, Scalability, Parallel Processing, Neural Network, Autoencoders, Variational Autoencoders (VAEs), Data Visualization, Interpretability, Feature Extraction, Distributed Training, Data Analysis, Insights, Data Science, Machine Learning, Data Exploration, Complex Data Structures.

INTRODUCTION

In today's data-driven landscape, the proliferation of big data has ushered in unprecedented opportunities and challenges. The deluge of information generated by various domains, such as finance, healthcare, and industry, demands efficient methods for data analysis, comprehension, and visualization. A fundamental hurdle in this pursuit is the curse of dimensionality, where datasets with high dimensionality pose significant computational and interpretational difficulties. Traditional dimensionality reduction techniques, though valuable, often falter in handling massive and intricate data. In response to these challenges, this research introduces an innovative approach that leverages the power of Deep Learning (DL) in conjunction with the Hadoop Distributed File System (HDFS) for intuitive dimensionality reduction on large-scale datasets. With big data, datasets commonly feature hundreds or even thousands of dimensions, making it challenging to discern meaningful patterns and relationships. Dimensionality reduction serves as a pivotal technique to overcome this obstacle. By reducing the number of features while preserving relevant information, it simplifies data, enhances interpretability, and expedites subsequent analysis.

1. Research Scholar, Dept of Computer Science and Engineering, Acharya Nagarjuna University, Guntur, Andhra Pradesh.
2. Associate Professor, Head - Dept of Computer Science, Government Degree College, Dornala, Andhra Pradesh.

INTERNATIONAL JOURNAL OF PURE AND APPLIED SCIENCE & TECHNOLOGY
IJPAST

ISSN 2229-6107
www.ijpast.in
Vol 12,Issuse 3. July 2022

Traditional techniques like Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbour Embedding (t-SNE) have made significant contributions, but they may falter when dealing with the scale and complexity of modern big data. Herein lies the impetus for our novel approach. Deep Learning, a subset of machine learning, has demonstrated remarkable capabilities in automatically learning intricate hierarchical representations from data. It possesses the potential to tackle the challenges presented by high-dimensional data, but its implementation in the big data context necessitates scalable solutions. The integration of DL with the Hadoop Distributed File System, a renowned framework for distributed data storage and processing, forms the foundation of our approach. This synergy enables the efficient management of data preprocessing and the training of DL models on distributed clusters.

Overview of the Proposed Framework

Our proposed framework encompasses several essential components. Raw data is ingested into the HDFS cluster, ensuring efficient storage and distribution across nodes. Within the HDFS infrastructure, preprocessing tasks such as feature scaling and normalization are seamlessly executed. A deep neural network architecture, carefully tailored for dimensionality reduction, is designed to adapt to the complexity of the data and capture intricate relationships among features. The model's training occurs in a distributed manner, thanks to the parallel processing capabilities of HDFS. Moreover, we emphasize the significance of data visualization as it enables intuitive exploration and analysis of the reduced-dimensional representations.

Research Significance

The primary objective of this research is to bridge the gap between the burgeoning realm of big data and intuitive dimensionality reduction. By integrating Deep Learning and HDFS, we aim to provide data scientists and analysts with a powerful toolset for extracting valuable insights from vast and intricate datasets. In the following sections, we delve into the technical details of our approach, evaluate its performance, and present compelling use cases that underscore its practical relevance in domains where intuitive data understanding is paramount.

Objectives:

1. **Efficient Dimensionality Reduction**: To develop a framework that efficiently reduces the dimensionality of large-scale and high-dimensional datasets, allowing for enhanced data analysis and visualization while preserving essential patterns and structures.
2. **Utilizing Deep Learning**: To harness the power of Deep Learning techniques, such as autoencoders and variational autoencoders (VAEs), to automatically learn and create meaningful representations of the data, addressing the limitations of traditional dimensionality reduction methods.
3. **Scalable Data Processing**: To implement the proposed framework within the Hadoop Distributed File System (HDFS) environment, ensuring scalable data preprocessing and distributed training of Deep Learning models to accommodate the vast volumes of big data.
4. **Intuitive Data Interpretation**: To enable intuitive exploration and analysis of reduced-dimensional data through effective visualization techniques, facilitating data and its extracting valuable insights from complex and massive datasets.

LITERATURE SURVEY:

The focus of this work pertains to the detection of foetal electrocardiogram (FECG) signals using a single-channel abdominal lead. The proposed approach utilises a combination of convolutional neural networks (CNN) and sophisticated mathematical techniques, including independent component analysis (ICA), singular value decomposition (SVD), and nonnegative matrix factorization (NMF), which is a dimension-reduction method. The foetus's electrical activity may be clearly distinguished in terms of energy due to the significant disparity in frequency between the foetus's heart rate and that of the mother. Moreover, it is possible to separate the different constituents of the foetal electrocardiogram (ECG), which act as inputs to the convolutional neural network (CNN) model, in order to enhance the actual foetal ECG signal, referred to as FECGr, through the utilisation of the singular value decomposition-independent component analysis (SVD-ICA) procedure. The results indicate the effectiveness of this novel methodology, which

ISSN 2229-6107
www.ijpast.in
Vol 12,Issuse 3. July 2022

has the potential to be implemented in real-time scenarios.

The field of healthcare automation is experiencing tremendous advancements, as shown by the increasing prevalence of e-health or digital health systems. The proliferation of health-related data generated by these technologies has led to the emergence of the field of health informatics. The World Health Organisation (WHO) defines SMARThealth as a concept that encompasses several key attributes, namely being standards-based, machine-readable, adaptive, requirements-based, and testable. Additionally, WHO offers comprehensive guidance for the implementation of digital health practices. The integration of diverse and extensive health data into cloud-based systems necessitates careful attention to ensure consistency in data format, enabling seamless exchange and applicability for broad utilisation and comprehensive analysis. This study presents a deep-learning framework for the purpose of illness diagnosis, with a specific focus on diabetes mellitus (DM) as a case study. This study examines three distinct corpuses of patient data for individuals diagnosed with DM. These corpuses have been meticulously prepared and processed using advanced data warehousing techniques. Additionally, each corpus has been appropriately labelled with ICD-10-CM diagnosis codes. The extraction of specific health data is facilitated by employing a standardised data model for healthcare that adheres to the HL7 FHIR v4.0 schema. The study presents two main contributions. Firstly, it introduces and validates three big data cloud analytical models on a unified corpus. Secondly, it achieves 100% accuracy in diagnosing the maximum possible diseases specific to one or multiple patients with diabetes mellitus (DM) using deep multinomial/multi-label distribution learning (DMDL).

Deep learning techniques have become increasingly prevalent in the fields of prediction and analysis. Examining the temporal trends evident in the crime data and deriving pertinent elements from the demographic information constitutes a substantial undertaking. Machine learning encompasses the use of algorithms to discern and comprehend patterns inherent in data, facilitating the ability to generate predictions. The utilisation of this technology enables the identification of crime hotspots, the prediction of criminal behaviour, and the forecasting of theft likelihood in particular geographical regions. In contrast, deep learning encompasses the utilisation of artificial neural networks with numerous layers to effectively represent intricate connections within datasets. This approach demonstrates compatibility with extensive datasets and possesses the capability to analyse many types of data, including images, audio, text, and numerical data. The application of deep learning techniques in theft crime prediction involves the utilisation of pattern recognition algorithms to find and analyse trends in criminal behaviour, facilitating the proactive detection of criminal activities prior to their occurrence. Various algorithms, such as Random Forest, Naive Bayes, and XGBoost, were employed for predictive purposes. However, it is important to note that these models possess some limitations, such as reduced accuracy and performance. In summary, our research demonstrates the prospective application of deep learning in crime prediction, highlighting the significance of incorporating demographic data and historical crime data in the modelling procedure while acknowledging its limitations.

The generation of three-dimensional (3D) models from a solitary red-green-blue (RGB) image is a formidable challenge within the field of image processing. This challenge is particularly pronounced due to the nascent stage of technological advancement in this area. The current era has witnessed a significant surge in the need for 3D technology and 3D reconstruction. The conventional methodology employed in computer graphics involves the creation of a three-dimensional geometric model, which is subsequently projected into a two-dimensional image through the process of rendering. The primary objective of this work is to employ machine learning techniques in order to generate three-dimensional (3D) models from two-dimensional (2D) RGB images. The aim is to develop a methodology that is computationally less complex than existing deep learning algorithms. The suggested model is constructed using three distinct modules: 2.5D feature extraction, mesh generation, and 3D boundary detection. The ShapeNet dataset has been utilised for the purpose of comparison. The results of the tests indicate an accuracy rate of 90.77% for the plane class, 85.72% for the chair class, and 72.14% for the automobile class. The proposed approach has the potential to be applied to situations that involve the reconstruction of three-dimensional models, particularly in cases that involve variations in

ISSN 2229-6107
www.ijpast.in
Vol 12,Issuse 3. July 2022

geometric scale and a combination of textured, uniformly coloured, and reflecting surfaces.

Individuals from many geographical locations have the opportunity to articulate their perspectives and viewpoints via a multitude of online social media platforms. Individuals engage in the everyday usage of online social media platforms as a means of interpersonal communication and to be abreast of contemporary occurrences. Every day, Twitter receives a substantial volume of tweets encompassing a diverse array of topics. Twitter is a highly recognised and extensively used digital social media network. The processes of feature extraction and trend identification can be effectively achieved by leveraging machine learning techniques. In order to effectively extract valuable insights from the continuous influx of data generated by Twitter, it is imperative to employ specialised tools and procedures tailored for handling substantial data volumes. This study primarily centres on the identification of hashtags and the determination of the industry with the greatest share of voice. This study aims to gather real-time data from the social media platform Twitter through the use of Apache Spark. Subsequently, the classification of each tweet is conducted through the utilisation of machine learning techniques offered by the Apache Spark machine learning library. In order to evaluate the performance of the model, a convolutional neural network (CNN) and logistic regression (LR) are employed. The convolutional neural network (CNN) approach demonstrated superior performance compared to the logistic regression strategy, with an average accuracy of about 95% and an F1 score of 0.60. The current values for both accuracy and the F1 score are 0.59. Based on the results obtained, it has been observed that the use of the Apache Spark framework for big data enables a significantly faster evaluation of real-time tweets compared to the traditional execution environment. The findings indicate that the use of the Apache Spark tool for big data enables significantly faster evaluation of real-time tweets compared to conventional execution environments.

Healthcare plays a crucial role within the contemporary medical landscape, particularly in the context of the digital era. In the context of sickness prediction and other healthcare-related tasks, it is imperative for a healthcare system to thoroughly analyse extensive volumes of patient data. An intelligent system would possess the capability to examine several aspects of a patient's life, including their social interactions, medical background, and other characteristics related to their lifestyle, in order to predict the probability of encountering a health issue. The Health Recommender System (HRS) is gaining increasing importance as a means of delivering healthcare services. Health-intelligent systems have become essential elements in the decision-making process of healthcare delivery within this particular context. The main objective of their work revolves around ensuring the consistent provision of information that is of superior quality, dependable, genuine, and confidential, thereby enabling its optimal use. The health recommender system plays a vital role in generating outcomes such as suggesting diagnoses, health insurance options, treatment approaches based on clinical pathways, and alternative medications, all tailored to the patient's health profile. This is particularly significant as an increasing number of individuals are turning to social networks as a means to acquire health-related knowledge. To optimise healthcare efficiency and cost-effectiveness, recent research has concentrated on leveraging extensive medical data by integrating multimodal data from various sources. In the healthcare industry, the utilisation of big data analytics in conjunction with recommender systems holds significant importance when making decisions pertaining to a patient's health. This article proposes the use of a LeNET Convolutional Neural Network (CNN) to explore the integration of big data analysis into the creation of effective health recommendation systems. It also demonstrates the potential advantages for the healthcare industry in transitioning from a standardised model to a personalised approach within the realm of telemedicine. The proposed approach demonstrates superior performance compared to other approaches by considering both the root squared mean error (RSME) and the average absolute error (AAE).

Colour constancy refers to the perceptual phenomenon in which individuals are able to accurately perceive and recognise the colours of objects, regardless of variations in the properties of the illuminating light source. The objective of computational colour constancy is to estimate the illuminant and afterwards use this knowledge to rectify the image and present it as it would appear under a standard illuminant. The deep learning

ISSN 2229-6107
www.ijpast.in
Vol 12,Issuse 3. July 2022

approach has emerged as one of the most effective techniques for estimating illumination in various scenarios. This method normally requires a dataset of photos that have been annotated with the corresponding scene illumination. While drawing parallels between the human visual system and machine learning algorithms is common, it is important to note that the former has not been exposed to definitive and accurate information on illuminants throughout its evolutionary process. Alternatively, it is postulated that the emergence of colour constancy can be attributed to its facilitation of various essential functions, including the autonomous recognition of fruits, objects, and animals irrespective of the prevailing illumination conditions. The rapid advancement of artificial intelligence and the consequent enhancement of individuals' well-being have led to significant advances in the field of picture identification, particularly in recent years. This study investigates the problem of object detection in low illumination conditions and employs deep learning techniques for the purpose of image detection and analysis. In environments with limited lighting conditions, the detected items are subjected to a comparison process with a large dataset. This comparison aims to identify and validate the objects that exhibit the highest degree of similarity. Additionally, this comparison evaluates the learning model's accuracy for object recognition in difficult lighting conditions.

The utilisation of mobile health has evolved as a viable and pragmatic solution for the treatment and management of individuals' health issues. However, the majority of mobile health data consists of observational data obtained by sensors, posing challenges in analysing the causal relationship between provided interventions using conventional regression techniques. This study provides a comprehensive overview of deep learning models that have the potential to accurately evaluate the causal effect of unprocessed mobile health data. The aforementioned models possess the ability to effectively process multivariate time series data for the purpose of estimating the unbiased causal effect, particularly when presented with a series of treatments.

Performance measurement systems play a crucial role in the administration of organisations since they facilitate the transformation of raw data into meaningful information that can be utilised by decision-makers. In the past few decades, there has been a significant surge in the volume of data and information produced and disseminated, presenting novel prospects and complexities for these systems. In light of the given situation, the objective of this essay is to examine the utilisation of big data analytics within performance measurement systems in order to elucidate the relationship between the two. Moreover, the objective is to discern patterns and potential avenues for future scholarly investigation. In order to accomplish this objective, a scientific mapping was conducted using bibliometric analytic techniques. The primary findings of the study indicate a notable rise in the utilisation of big data analytics within performance measurement systems (PMS) in recent years, without due consideration for the inherent characteristics of such systems. The integration of artificial intelligence technologies, specifically machine learning and deep learning, has the potential to enhance the field, opening up avenues for empirical research using unstructured data and applications within the context of Industry 4.0.

Machine learning techniques, particularly natural language processing (NLP), are of significant importance in the context of using social media data for governmental purposes in many countries worldwide. The analysis of social media posts and tweets can provide insights on the prevailing thinking of individuals, a crucial aspect for any governing body worldwide. The primary aim of this research is to do sentiment analysis in order to extract the prevailing sentiment among individuals in relation to the ongoing conflict between Russia and Ukraine. This will be accomplished through the use of machine learning methodologies. The objective is to conduct an analysis and draw inferences regarding the potential reactions of countries in response to the economic effects, taking into account the sentiment of their respective citizens. The implementation process commences with the acquisition of data from social media platforms, specifically Twitter and Reddit. This is achieved through the use of Snscraper, a web scraping tool, and the PRAW (Python Reddit API Wrapper) library. Suitable text summarising techniques are employed to accommodate the larger posts on Reddit. The BERT transformer model is used to conduct sentiment analysis on social media data. The non-English posts undergo translation into English through the use of neural machine translation. In addition, sentiment analysis is

ISSN 2229-6107
www.ijpast.in
Vol 12,Issuse 3. July 2022

conducted at several levels of detail, including the identification of specific locations and individuals through the application of named entity recognition algorithms. In this study, a comprehensive examination is conducted to compare the emotions of various countries throughout the world with their respective levels of dependence on Russian oil.

The Intelligent Transportation System (ITS) is a groundbreaking technology within the realm of smart cities that serves to mitigate traffic congestion and enhance traffic conditions. Information Technology Services (ITS) offers real-time analysis and highly efficient traffic control through the utilisation of big data and communication technology. Traffic flow prediction (TFP) has emerged as a crucial element in the management of smart cities, serving as a means to forecast forthcoming traffic conditions on transport networks based on historical data. Machine learning (ML) and neural network (NN) methodologies have demonstrated significant use in addressing real-time challenges due to their ability to effectively handle dynamic data over extended periods. Deep learning (DL) is a subset of machine learning (ML) techniques that demonstrate high efficacy in tasks related to prediction and data classification. This article presents the development of a Grey Wolf optimizer with a deep learning-based short-term traffic forecasting (GWODL-STTF) model in the context of a smart city setting. The GWODL-STTF technique focuses on forecasting traffic flow in smart cities. The GWODL-STTF approach encompasses two primary steps. In the early phase, the GWODL-STTF methodology utilised a gated recurrent unit-neural network (GRU-NN) model for the purpose of predicting traffic flow. In the subsequent phase, the GWODLSTTF approach employs the Grey Wolf Optimisation (GWO) algorithm as a hyperparameter optimizer. The performance of the GWODL-STTF technique may be evaluated using many metrics in simulation experiments. The minimum mean squared error (MSE) value of 105.627 obtained from the results shows that the GWODL-STTF method performs better than recent techniques.

Medical image classifiers serve a key role in both medical services and educational endeavours. However, the conventional technique reached its maximum level of performance. Furthermore, the utilisation of these traits necessitates a significantly greater amount of time and effort for extraction and selection. The Deep Neural Network (DNN) is an emerging machine learning (ML) technique that has demonstrated its potential for many classification problems. The convolutional neural network (CNN) has been found to yield optimal results in several image classification tasks. However, the acquisition of medical picture databases can pose challenges due to the need for specialised expertise in categorization. This research paper presents the development of a novel hyperparameter-tuned deep learning model for healthcare monitoring systems (HPTDLM-HMS) within the context of a big data environment. The HPTDLM-HMS technique discussed in this study focuses on the analysis of medical pictures within the context of decision-making. The HPTDLM-HMS technique is initially implemented by utilising the EfficientNet model to extract features. The hyperparameters of the model are optimised using the Manta Ray Foraging Optimisation (MRFO) algorithm. Finally, the categorization of medical images is conducted using the Long Short-Term Memory (LSTM) technique. Hadoop MapReduce is employed for the purpose of managing large volumes of data. The outcome evaluation of the HPTDLM-HMS approach is assessed using a dataset consisting of medical imaging data. The full examination of the HPTDLM-HMS approach has demonstrated a recall value of 87.46%, which surpasses the performance of alternative models and underscores its potential prospects.

In contemporary times, there has been a substantial proliferation of data, leading to a progressive transformation in the importance attributed to data security and data analysis techniques within the realm of big data." An intrusion detection system (IDS) is a mechanism that examines and monitors data in order to identify any unauthorised access or incursion into a system or network. The substantial magnitude, diversity, and rapid velocity of data generated within the network necessitate a sophisticated data analysis methodology to effectively identify and mitigate assaults. Big data systems can be employed in intrusion detection systems (IDS) to facilitate the management of large volumes of data, enabling accurate and efficient data analysis methodologies. This research paper presents a novel approach called Intrusion Detection Approach utilising Hierarchical Deep Learning-based Butterfly Optimisation Algorithm (ID-HDLBOA) in the context of a big data platform. The technique known as ID-HDLBOA integrates the principles of deep learning (DL) with

ISSN 2229-6107
www.ijpast.in
Vol 12,Issuse 3. July 2022

the process of hyperparameter tuning. The ID-HDLBOA technique incorporates a hierarchical LSTM model for the purpose of intrusion detection. The BOA method is employed as a hyperparameter tuning strategy for the LSTM model, leading to enhanced detection efficiency. The ID-HDLBOA technique is experimentally validated using a benchmark incursion dataset, yielding a model accuracy of 98%. A series of comprehensive experiments were conducted, and the results consistently highlighted the superior performance of the ID-HDLBOA algorithm.

This paper presents a proposed recommendation system, named Sentiment Analysis and Matrix Factorization (SAMF), which aims to address the challenges of data sparsity and credibility in collaborative filtering. SAMF leverages topic modelling and deep learning techniques to effectively extract implicit information from reviews. By enhancing the rating matrix, SAMF assists in improving the recommendation process. The generation of user topic distribution and item topic distribution is accomplished by applying Latent Dirichlet Allocation (LDA) to reviews, which include both user reviews and item reviews. The user feature matrix and item feature matrix are generated by utilising topic probability. Furthermore, the integration of the user feature matrix and item feature matrix results in the creation of a user-item preference matrix. In addition, the process involves the integration of the user-item preference matrix and the original rating matrix, resulting in the creation of the user-item rating matrix. In addition, the utilisation of BERT (Bidirectional Encoder Representation from Transformers) is employed to quantify the sentiment information encompassed within the reviews. This sentiment information is subsequently integrated with the user-item rating matrix, facilitating the modification and updating of said matrix. Subsequently, the revised user-item rating matrix is employed to facilitate the prediction of ratings and generate Top-N recommendations. The experimental results obtained by analysing Amazon datasets provide evidence that the proposed SAMF algorithm outperforms existing conventional algorithms in terms of suggestion performance.

Objective: In recent years, there has been a significant utilisation of deep learning (DL) in academic research pertaining to the interpretation of 12-lead electrocardiogram (ECG) data. Nevertheless, there is a lack of clarity regarding the validity of the explicit or implicit assertions regarding the superiority of deep learning (DL) over classical feature engineering (FE) approaches that rely on domain expertise. Furthermore, there is a lack of clarity on the potential enhancement of performance through the integration of deep learning (DL) and feature engineering (FE) in comparison to utilising a single modality. Methods: In order to fill the existing research gaps and align with recent significant experiments, we conducted a reexamination of three specific tasks: the diagnosis of cardiac arrhythmia using a multiclass-multilabel classification approach; the prediction of atrial fibrillation risk using a binary classification approach; and the estimation of age using a regression approach. For the purpose of our study, we utilised a comprehensive dataset consisting of 2.3 million 12-lead electrocardiogram (ECG) recordings. These recordings were employed to train various models for each specific task. Specifically, we developed three models: i) a random forest model that utilised feature extraction (FE) as input; ii) an end-to-end deep learning (DL) model; and iii) a merged model that combined both FE and DL approaches. The findings indicate that the performance of feature engineering (FE) was similar to that of deep learning (DL) in the two classification tasks. However, FE required much less data compared to DL. The deep learning (DL) approach demonstrated superior performance compared to the traditional feature engineering (FE) method for the regression problem. In all tasks, the integration of feature engineering (FE) with deep learning (DL) did not yield any performance improvement compared to the use of DL alone. The aforementioned conclusions were validated using the supplementary PTB-XL dataset. In conclusion, our findings indicate that deep learning (DL) did not demonstrate a substantial enhancement over feature engineering (FE) in the context of traditional 12-lead electrocardiogram (ECG)-based diagnosis tasks. However, it did exhibit a notable improvement in the atypical regression challenge. Furthermore, our findings indicate that the integration of feature engineering (FE) with deep learning (DL) did not yield superior results compared to DL in isolation. This observation shows that the features extracted using FE were duplicative of the features acquired through DL. The significance of our findings lies in the provision of crucial advice pertaining to the selection of a suitable machine learning technique and data regime for a specific

ISSN 2229-6107
www.ijpast.in
Vol 12,Issuse 3. July 2022

task, with a focus on 12-lead ECG analysis. When considering the objective of maximising performance, deep learning (DL) is the preferred approach when dealing with unconventional tasks and large datasets. In cases where the work at hand is of a classical nature and/or there is limited availability of data, employing a feature engineering (FE) technique may prove to be a more suitable option.

Accurate identification of heart disease can have life-saving implications, while an erroneous diagnosis can have fatal consequences. The UCI dataset on heart disease in machine learning serves as a platform for evaluating and comparing the outcomes and analyses of different machine learning methodologies, encompassing deep learning techniques. The research was conducted using a dataset of 13 major characteristics. The datasets are processed using support vector machines and logistic regression techniques, with the latter demonstrating superior accuracy in predicting coronary disease. Python programming is employed for the purpose of processing datasets. Several research initiatives have employed machine learning techniques to enhance the efficiency of the healthcare industry. Conventional machine learning methods were employed in our study to elucidate the relationships between the various variables included in the dataset. These findings were subsequently utilised to efficiently predict the risks of heart infections. The utilisation of the accuracy and confusion matrix has yielded positive benefits. In order to optimise the outcomes, the dataset incorporates specific extraneous attributes, which are addressed through the utilisation of isolation logistic regression and support vector machine (SVM) classification techniques.

The utilisation of big data analytics in the health care sector involves the systematic examination of extensive and diverse information with the aim of identifying concealed patterns, correlations, and trends that might inform optimal decision-making within the medical domain. Healthcare and medical services have seen significant advancements with the development of sophisticated and intelligent healthcare platforms. These platforms have enhanced the robustness of treatment analysis. The accurate diagnosis of health records is contingent upon the recognition of diseases as the primary factor, and the significance of precision diminishes when the quality of clinical data is compromised. However, the current methodologies have proven inadequate for effectively utilising the learning model to manage diverse healthcare data. This article presents a model called Big Data Analytics with Wild Horse Optimizer-Based Deep Learning (BDAWHO-DL) for healthcare management. The BDAWHO-DL technique under consideration explores the utilisation of big data in the medical domain and facilitates decision-making processes. The BDAWHO-DL approach utilises the attention-based long-short-term memory (ABLSTM) method for the purpose of data classification. Furthermore, the World Health Organisation (WHO) system was employed to conduct the optimal hyperparameter tuning process of the Attention-Based Long Short-Term Memory (ABLSTM) algorithm, with the aim of enhancing the performance of the classifier. The experimental results demonstrate the favourable performance of the BDAWHO-DL algorithm compared to previous methodologies.

The issue pertaining to rural villages and slums in rural areas is a highly significant concern encountered by numerous developing nations as well as certain developed nations. In the year 2021, the Egyptian government initiated the National Initiative for the Development of the Egyptian Rural Villages: Decent Life (Hayah Karima). This effort is driven by a sense of ethical obligation and a recognition of the societal implications, since it aims to achieve more than just improving the daily lives and living conditions of the population in Egypt. This study centres on the formulation of a model to determine development priorities in the villages of Asuit, employing GeoAI. Asuit has been identified as one of the governorates with the most pressing requirements for development. GeoAI can be defined as a computational framework used to design sophisticated software applications that emulate human perception, spatial reasoning, and the identification of geographic phenomena and dynamics. Its primary objective is to enhance our understanding of development priorities within a certain study region. The study region chosen for this research is Shamya village, situated in the Sahel Selim centre of Asuit, Egypt. This village was selected because of its classification as one of the most impoverished communities in accordance with the development priorities outlined in the Egyptian Rural Communities Development Project. The

ISSN 2229-6107
www.ijpast.in
Vol 12,Issuse 3. July 2022

poverty rate in Shamya village surpasses 70%. This aligns with the strategic approach of the Egyptian government to prioritise the provision of resources and attention to villages in Upper Egypt. GeoAI was employed to automatically detect and digitise the footprints of buildings and streets, utilising deep learning algorithms. This study aims to assess the efficacy of traditional methods in generating spatial maps, which necessitated a total of 40 working hours, in comparison to the utilisation of GeoAI, which yielded spatial maps for the identical area within a span of 30 minutes, exhibiting superior precision and requiring less time for both labour and quality control. The efficacy of GeoAI has been demonstrated in its ability to autonomously identify objects and generate maps, enabling the determination of development paths.

The present analysis focuses on a comprehensive review of scholarly literature conducted in 2018, utilising the Google Scholar database. Specifically, this study scrutinises a collection of 500 scientific articles that pertain to the domain of numerical weather prediction and climate, specifically exploring the utilisation of machine learning methodologies. The abstracts primarily focused on accepted subjects of interest, with a subset of these subjects receiving further investigation. These included wind energy and photovoltaic systems, the physics of atmospheric processes, the development of numerical weather prediction models, parameterizations in weather and climate research, the study of severe weather events, and the examination of climate change. The discussion encompassed many meteorological topics such as precipitation, pressure, radiation, temperature, and wind, as well as machine learning techniques including random forests, artificial neural networks, deep learning, XGBoost-like algorithms, and support vector machines. Moreover, countries such as Australia, India, China, the United States, and Germany were there to collect data utilising the established database pertaining to these subjects. In order to anticipate the future research endeavours of writers within these fields, it is imperative to undertake rigorous analyses of the existing literature. It is anticipated that machine learning will play a significant role in future weather forecasting, with a primary emphasis on drawing conclusive insights. This paper presents a systematic approach for conducting a literature assessment on the utilisation of big data analytical tools in weather forecasting.

In recent times, there has been a surge in the popularity of live streaming video, leading to a growing need for real-time big data analytics in order to assess the content of these streaming videos. This article presents a system for evaluating the calibre of live broadcast video through the utilisation of big data analytics and real-time deep learning techniques. The framework consists of multiple modules, encompassing data collection, preprocessing, feature extraction, feature selection, and deep learning-based categorization. In order to evaluate the calibre of the streaming video, a comprehensive analysis is conducted on various video attributes, encompassing resolution, bit rate, frame rate, and packet loss rate. Our deep learning models employ convolutional neural networks (CNNs) to classify the streaming video quality into different categories. Our framework's performance is evaluated using real-world streaming video datasets. The findings of this study indicate that the framework we have developed exhibits a high level of efficacy in real-time analysis of streaming video quality, surpassing the performance of other contemporary frameworks that are considered to be at the forefront of the field. Our framework offers a solution for live streaming companies to effectively deliver high-quality streaming video content to their viewers.

PROPOSED DESIGN:

The proposed research is aimed at addressing the challenges posed by the ever-expanding realm of big data by introducing an innovative approach to dimensionality reduction. The overarching objective of this project is to develop a comprehensive framework that combines the scalability and parallel processing capabilities of the Hadoop Distributed File System (HDFS) with the automatic feature learning potential of Deep Learning (DL). This union enables the efficient management of data preprocessing and the training of DL models on distributed clusters. The project encompasses various critical aspects, including data preprocessing, the design and training of deep neural network architectures, scalable data processing, and data visualization. The ultimate goal is to empower data scientists and analysts to intuitively extract valuable insights from intricate and voluminous datasets, enhancing both the efficiency and interpretability of big data analytics. This research represents a

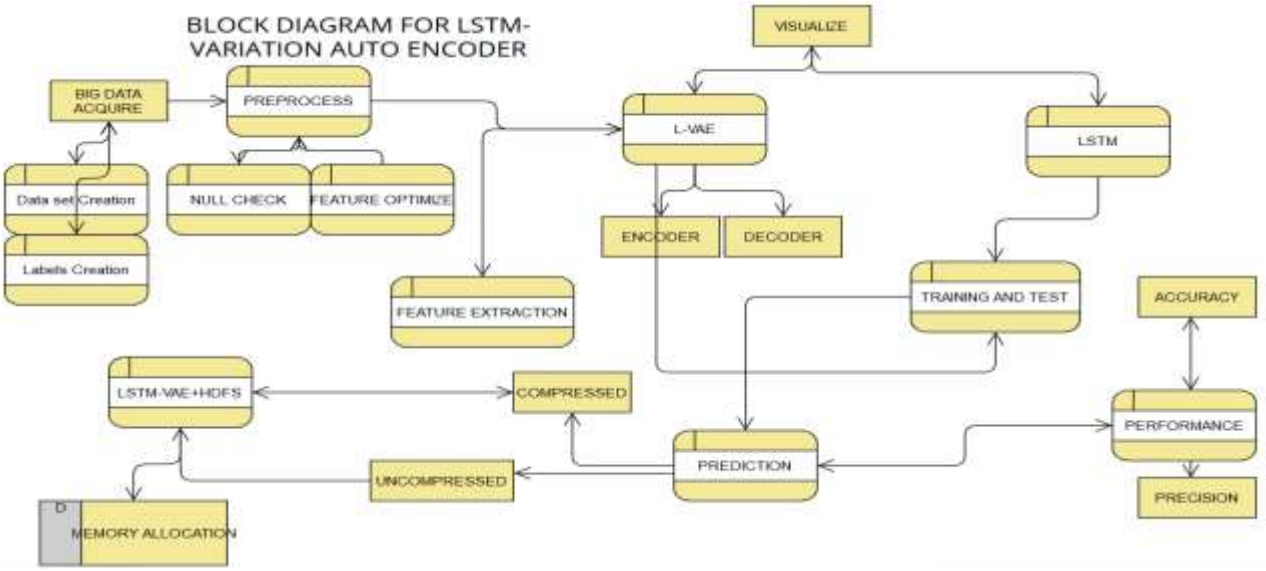
INTERNATIONAL JOURNAL OF PURE AND APPLIED SCIENCE & TECHNOLOGY
IJPAST
ISSN 2229-6107
www.ijpast.in
Vol 12,Issuse 3. July 2022

significant step towards bridging the gap between the data-rich environments of today and the intuitive understanding and utilization of such data for informed decision-making.

1. CONCEPT

The concept of proposed work revolves around the integration of cutting-edge techniques in data processing, deep learning, and distributed computing to simplify the complexity of massive and high-dimensional datasets. By leveraging the Hadoop Distributed File System (HDFS) as a scalable data storage and processing backbone, coupled with the automatic feature extraction capabilities of Deep Learning (DL), this innovative approach aims to streamline the dimensionality reduction process. The essence of this concept lies in efficient data preprocessing, where raw data is ingested, cleaned, and scaled within the HDFS ecosystem, followed by the design and training of deep neural networks that can encapsulate intricate data relationships. These networks facilitate the transformation of data into lower-dimensional representations while preserving essential information. Visualization techniques are then employed to provide data scientists and analysts with an intuitive means of exploring, comprehending, and extracting actionable insights from the reduced data space. In summary, this concept embodies a holistic strategy for intuitive dimensionality reduction, empowering users to navigate and derive knowledge from the vast and complex landscape of big data effortlessly.

1. BLOCK DIAGRAM



The overall block diagram of the proposed design with L-VAE is utilized with this implementation design indicating procedural steps as given below:

1. Big Data Acquisition: Big Data acquisition involves collecting vast amounts of data from various sources, which can be structured or unstructured. In the context of LSTM+VAE, this data often includes time-series or sequential data, such as sensor readings, financial transactions, or logs. The challenge in acquiring big data is handling the volume, velocity, and variety of data sources efficiently. To handle big data, distributed storage systems like Hadoop HDFS or cloud-based solutions are commonly used to ensure scalability and fault tolerance.

2. Preprocessing (NULL check and Feature Optimization/Transformation): Data preprocessing is a critical step in preparing big data for analysis. It includes tasks like NULL (missing data) checks and optimizing or transforming features. In LSTM+VAE applications, NULL check is essential to identify and handle missing time-series data, which can disrupt

ISSN 2229-6107
www.ijpast.in
Vol 12,Issuse 3. July 2022

model training. Feature optimization and transformation can involve scaling, normalization, or encoding categorical variables to make the data suitable for the model. For LSTM+VAE, temporal features may be engineered to capture seasonality or trends in time-series data.

3. Feature Extraction with Optimization: Feature extraction is about selecting or creating relevant features from the dataset. In LSTM+VAE, this step may involve extracting temporal features from time-series data, such as identifying patterns or sequences. Optimization ensures that the selected features are suitable for the model and may include dimensionality reduction techniques to reduce computational complexity without losing critical information.

4. Applying LSTM+VAE: LSTM+VAE represents a combination of Long Short-Term Memory (LSTM), a type of recurrent neural network, and Variational Autoencoder (VAE), a generative model. LSTM is used for sequential data analysis, capturing dependencies over time, while VAE is employed for probabilistic modeling. In this step, the LSTM+VAE model is designed, including the architecture, layers, and hyperparameters. This model is trained on the preprocessed big data to capture meaningful representations and learn latent variables.

5. Training and Testing Process: Training involves using a portion of the pre-processed data to teach the LSTM+VAE model to recognize patterns and generate meaningful representations. Testing involves using another portion of the data (separate from the training set) to evaluate the model's performance and generalization capabilities. The goal is to ensure that the model can effectively capture dependencies in the big data and generate meaningful latent representations.

2. ALGORITHM

The algorithm for intuitive dimensionality reduction on big data using Deep Learning (DL) and Hadoop Distributed File System (HDFS) is designed to efficiently transform high-dimensional data into lower-dimensional representations while preserving crucial information. This approach combines DL's feature extraction capabilities with the scalability of HDFS. The algorithm consists of several key steps, including data preprocessing, DL model design, distributed training, and visualization.

6. Prediction HDFS Compressed: After training and testing, the LSTM+VAE model can be applied to make predictions. In this proposed design context of Hadoop HDFS, we choose to store the predictions in a compressed format within HDFS. The importance for efficient storage and retrieval of large-scale prediction results. HDFS's distributed nature helps handle the volume of predictions in a scalable manner.

7. Prediction with Normal Case: In addition to storing predictions in HDFS, you may also need to make real-time predictions in normal cases, where new data is continuously generated. This involves applying the trained LSTM+VAE model to new data as it becomes available and generating predictions in real-time or near-real-time.

8. Performance Metrics with Accuracy, Precision, Memory Allocation, and Over-Compression Time: Evaluating the LSTM+VAE model's performance is crucial. Common performance metrics include accuracy (how well the model predicts), precision (the fraction of true positive predictions among all positive predictions), and memory allocation (the model's resource usage). Over-compression time refers to the time taken to apply the VAE's compression to the data. Monitoring these metrics helps ensure the model's quality, efficiency, and effectiveness.

In summary, performing big data analysis with LSTM+VAE involves acquiring, preprocessing, and optimizing data, training and testing the model, making predictions in HDFS and real-time scenarios, and rigorously evaluating performance using various metrics. These steps are crucial in harnessing the power of LSTM+VAE for big data applications, especially when handling large volumes of sequential data efficiently.

ALGORITHM 1: INTER-LSTM+C-VAE

INTERNATIONAL JOURNAL OF PURE AND APPLIED SCIENCE & TECHNOLOGY
IJPAST

ISSN 2229-6107
www.ijpast.in
Vol 12,Issuse 3. July 2022

Input: *Image sequences X (input features), Y (target labels indicating Compressed and non-compressed), number of layers L.*

Initialize: *Initialize LSTM+VAE model parameters (weights and biases) randomly or using pre-trained weights.*

Network Architecture: *Construct the LSTM+VAE network with L layers tailored for memory allocated data in compressed format*

Data Ingestion

- *Collect raw data from diverse sources and store it in HDFS.*
- *Seamless integration of data sources (databases, streaming, files) into HDFS.*

Efficient Storage and Distribution

- *Store and distribute data efficiently within HDFS.*
- *Data blocks are distributed across HDFS nodes with replication.*

Data Preprocessing

- *Enhance data quality and consistency.*
- *Feature scaling, normalization, handling missing values.*

Feature Scaling and Normalization

- *Ensure consistent scales and distributions.*
- *Rescale numerical features, transform data distribution.*
- *Scaling:*
    - $X' = max(X) - min(X) X - min(X)$,
- *Normalization:*
    - $X' = std(X) X - mean(X)$.

DL Model Selection

- *Choose an appropriate DL architecture for dimensionality reduction.*
- *Select an architecture such as LSTM or VAE.*

Distributed Model Training

- *Train DL model efficiently on HDFS.*
- *Distribute training across HDFS nodes using frameworks like TensorFlow.*

Feature Extraction

- *Extract meaningful features from high-dimensional data.*
- *Train the DL model to map inputs to lower-dimensional representations.*
- *Encoding function in VAE:* $z = \mu + \sigma \odot \epsilon$.

Data Visualization

- *Intuitively explore and analyse the reduced data.*
- *Use techniques like t-Distributed Stochastic Neighbour Embedding (t-SNE) or Principal Component Analysis (PCA).*
- *The objective of t-SNE is to find a lower-dimensional representation of data points that minimizes the discrepancy between the high-dimensional similarity* ***p_ij*** *and the lower-dimensional similarity* ***q_ij*** *for all pairs of data points i and j.*
- *The objective function that captures this minimization of KL divergence is defined as follows:*
-

$$C = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

Where:

- *C is the cost function to be minimized.*
- *i and j are indices for data points.*
- ***p_ij*** *is the conditional probability of similarity between data points i and j in the high-dimensional space.*
- ***q_ij*** *is the conditional probability of similarity between data points i and j in the lower-dimensional space.*

Model Evaluation

- *Assessing the quality of reduced representations.*
- *Evaluating using metrics like reconstruction error or clustering quality.*
- *Reconstruction error:*

INTERNATIONAL JOURNAL OF PURE AND APPLIED SCIENCE & TECHNOLOGY
IJPAST

ISSN 2229-6107
www.ijpast.in
Vol 12,Issuse 3. July 2022

- $MSE = \frac{\sum_{i=1}^{N}(x_i - x_m)^2}{N}$

Fine-Tuning

- *Optimize DL model and preprocessing steps.*
- *To Adjust hyperparameters and neural network architectures iteratively.*
- *Hyperparameter tuning, optimization techniques.*
- $LR = initial\_lr * drop_{factor}^{floor\left(\frac{epoch}{drop_{every}}\right)}$

Result Interpretation

- *Extract actionable insights from the reduced data.*
- *Analyse visualizations and patterns in the reduced space.*
- *Reconstruct the Reduced performance if Loss>5%. Repeat the steps*

Algorithm Design Steps:

a. **Data Ingestion**: In the first crucial step of the algorithm, raw data originating from diverse sources is collected and funnelled into the Hadoop Distributed File System (HDFS) cluster. This process involves the seamless integration of data from sources such as databases, streaming data, external files, or other data repositories. HDFS serves as the central repository for this raw data, providing a scalable and fault-tolerant storage solution for vast datasets. The data ingestion mechanism ensures that the information is efficiently organized within HDFS, ready for subsequent processing, and distributed across the nodes of the Hadoop cluster.

b. **Efficient Storage and Distribution:** Within the HDFS environment, the ingested data is not only stored efficiently but also meticulously distributed across the network of nodes in the cluster. HDFS employs a distributed file system architecture, where large files are broken into smaller blocks, typically 128MB or 256MB in size. These blocks are then replicated across multiple nodes for fault tolerance and parallel processing. The distribution mechanism ensures that data can be accessed and processed in parallel, optimizing data availability and reliability. This distributed storage and data replication strategy are pivotal for handling the immense volume and complexity of big data effectively.

c. **Data Preprocessing**: The data preprocessing phase is a critical component of the algorithm, designed to enhance the quality and consistency of the raw data ingested into the Hadoop Distributed File System (HDFS). It encompasses a series of essential data transformations and enhancements that prepare the data for subsequent analysis and dimensionality reduction. These transformations often include techniques like feature scaling, normalization, and handling missing values, all of which serve to mitigate issues related to varying data scales, statistical properties, and incomplete information.

d. **Feature Scaling and Normalization:** Feature scaling involves ensuring that the numerical attributes within the dataset have consistent scales. This step prevents attributes with large value ranges from dominating the learning process and ensures that each feature contributes proportionally to the analysis. Normalization, on the other hand, transforms data to adhere to a standard distribution, making it more amenable to certain machine learning algorithms. By performing these operations using HDFS-based data pipelines, the algorithm ensures that these critical preprocessing tasks can be efficiently applied at scale, accommodating the size and complexity of big data. Additionally, these operations maintain data quality and consistency, setting the stage for subsequent dimensionality reduction using Deep Learning techniques.

e. **DL Model Selection**: The choice of a Deep Learning (DL) architecture is a pivotal decision in the algorithm's workflow. It involves selecting an appropriate model, such as autoencoders, variational autoencoders (VAEs), or other deep neural networks that align with the specific characteristics of the dataset and the goals of

ISSN 2229-6107
www.ijpast.in
Vol 12,Issuse 3. July 2022

dimensionality reduction. Autoencoders, for example, are well-suited for unsupervised feature learning, while VAEs offer probabilistic modeling capabilities. The selection process considers factors like data complexity, the need for interpretability, and the desired properties of the reduced representations. This step sets the foundation for the subsequent phases, where the chosen DL model will be trained and optimized for effective dimensionality reduction.

f. **Distributed Model Training**: To tackle the challenges posed by large-scale data, distributed model training is a critical phase that capitalizes on the parallel processing capabilities of the Hadoop Distributed File System (HDFS). The algorithm efficiently distributes the DL model training process across the nodes of the HDFS cluster. This approach not only accelerates the training but also ensures scalability and fault tolerance. Frameworks like TensorFlow or PyTorch are employed to leverage the power of DL for dimensionality reduction while efficiently processing data stored in HDFS. The distributed training phase is designed to accommodate the substantial volume of data, allowing the DL model to learn meaningful lower-dimensional representations, a fundamental step in the dimensionality reduction process.

g. **Feature Extraction**: Feature extraction represents the core objective of the DL model in the dimensionality reduction process. Once the model is trained, it excels at learning intricate and meaningful representations from high-dimensional input data. This phase maps the original data to lower-dimensional representations while preserving essential information. The DL model's ability to automatically identify and capture relevant features within the data is a critical advantage, as it simplifies complex data structures and enhances interpretability. Feature extraction ensures that the reduced data representations retain the most salient characteristics of the original data, facilitating downstream analysis and visualization. This step embodies the essence of dimensionality reduction, as it empowers the algorithm to intuitively transform unwieldy data into a more manageable and informative format.

h. **Data Visualization**: Data visualization is an indispensable component in the process of understanding the reduced-dimensional data created by the dimensionality reduction algorithm. Techniques like t-Distributed Stochastic Neighbor Embedding (t-SNE) and Principal Component Analysis (PCA) are employed to intuitively represent and explore the transformed data. These visualization methods allow data scientists and analysts to gain valuable insights into the structures and relationships within the data. t-SNE, for instance, can uncover intricate patterns and clusters, while PCA provides a means to visualize the principal components that contribute the most to data variance. Visualizations empower users to grasp the essence of the reduced data space, enabling more informed data exploration and analysis.

i. **Model Evaluation**: Model evaluation is a critical step in assessing the quality and effectiveness of the reduced representations produced by the algorithm. This phase involves the application of various evaluation metrics and techniques to measure the performance of the Deep Learning (DL) model and the dimensionality reduction process. Common metrics include reconstruction error, which quantifies how well the reduced data can reconstruct the original data, and clustering metrics to evaluate the quality of clusters formed in the reduced space. These evaluations provide essential feedback on the algorithm's success in preserving critical information and structures during dimensionality reduction. By rigorously evaluating model performance, the algorithm ensures that the resulting representations are of high quality and align with the intended objectives.

j. **Fine-Tuning**: Fine-tuning represents an iterative optimization process aimed at enhancing the dimensionality reduction results. This phase involves adjusting DL model parameters and preprocessing steps to

INTERNATIONAL JOURNAL OF PURE AND APPLIED SCIENCE & TECHNOLOGY
IJPAST

ISSN 2229-6107
www.ijpast.in
Vol 12,Issuse 3. July 2022

achieve optimal performance. Fine-tuning may include tuning hyperparameters, refining neural network architectures, or experimenting with different training strategies. The goal is to iteratively improve the quality of reduced representations by identifying and addressing potential shortcomings in the initial model. This iterative approach ensures that the algorithm evolves to produce more accurate, meaningful, and intuitive reduced data representations, aligning more closely with the specific requirements of the data and the analysis objectives.

k. **Result Interpretation**: Result interpretation is the culmination of the dimensionality reduction process, where actionable insights are extracted from the reduced data space. Data scientists and analysts delve into the visualizations and analysis outputs to uncover patterns, anomalies, and relationships that were not readily apparent in the original high-dimensional data. This phase facilitates informed decision-making and data-driven discoveries by providing a deeper understanding of the data's intrinsic characteristics. By interpreting the reduced data, users can extract valuable knowledge, make informed decisions, and derive meaningful insights across a range of domains, from business intelligence to scientific research.

3. IMPLEMENTATION KEY STEPS

The implementation of this algorithm involves integrating DL and HDFS into a cohesive workflow:

i. **HDFS Setup**: Establish an HDFS cluster configured for distributed data storage and processing. Ensure data ingestion, storage, and access mechanisms are in place.

ii. **Data Preparation**: Ingest raw data into HDFS, and create data pipelines for preprocessing. Data cleaning, scaling, and transformation are essential here.

iii. **DL Model Development**: Choose the DL architecture suitable for the dataset and implement it using a DL framework compatible with HDFS.

iv. **Distributed Training**: Utilize HDFS's parallel processing capabilities to distribute data and training tasks across cluster nodes. Optimize the training process for efficiency.

v. **Feature Extraction**: Train the DL model on the distributed data to learn lower-dimensional representations while preserving essential information.

vi. **Visualization Integration**: Implement data visualization techniques that can operate on the HDFS-stored reduced data, allowing for intuitive exploration and analysis.

vii. **Evaluation and Optimization**: Continuously evaluate the dimensionality reduction results and fine-tune both preprocessing and DL model parameters to achieve the desired outcomes.

viii. **Result Storage**: Store the reduced data representations and visualization outputs back in HDFS for easy access and sharing.

4. DESIGN LAYERS

The mathematical formulations in this algorithm primarily revolve around the chosen DL model. For instance, in the case of an autoencoder, the encoding and decoding processes can be defined mathematically as follows:

Layer Type	Description
Input Layer	Receives input data, typically in the form of sequences or image data. The input shape depends on the specific task and data format.
Encoder Convolutional Layers 1	Convolutional layers that extract low-level features from the input data. They capture basic patterns and structures in the data.
Encoder Convolutional Layers 2	Additional convolutional layers that capture higher-level features and more complex patterns in the data.

INTERNATIONAL JOURNAL OF PURE AND APPLIED SCIENCE & TECHNOLOGY
IJPAST

ISSN 2229-6107
www.ijpast.in
Vol 12,Issuse 3. July 2022

Encoder Convolutional Layers 3	Further convolutional layers that continue to refine the features extracted from the input data.
Encoder LSTM Layers 1	LSTM (Long Short-Term Memory) layers that capture sequential dependencies within the data. They process data over time steps or sequences.
Encoder LSTM Layers 2	Additional LSTM layers that capture longer-range dependencies in the data, helping to model complex sequences.
Latent Space Layer (Mean and Variance)	This layer computes the mean (μ) and variance (σ^2) of the latent space distribution. These parameters are used to sample from the latent space.
Sampling Layer	Samples from the latent space distribution using the computed mean and variance. The reparameterization trick is often used for this purpose.
Decoder LSTM Layers 1	LSTM layers responsible for generating sequences or data samples from the sampled latent space representations.
Decoder LSTM Layers 2	Additional LSTM layers that further refine the generated sequences, capturing more intricate sequential patterns.
Decoder Convolutional Layers 1	Convolutional layers that begin the process of reconstructing the data from the sampled latent space representations.
Decoder Convolutional Layers 2	Additional convolutional layers that continue to refine and upsample the data, aiming to match the input data's structure and features.
Output Layer	Produces the final output data, which should ideally match the input data in the case of reconstruction tasks. The output shape depends on the task (e.g., sequence or image).

Layer Architecture:

These 13 layers together make up the architecture of an INTER-LSTM+Convolutional VAE, which is capable of capturing both sequential and spatial dependencies in the data.

1. **Input Layer:** The Input Layer is the gateway for data into the INTER-LSTM+Convolutional VAE architecture. In the context of cloud-based image or data compression, this layer plays a vital role in receiving the original data, which may be in the form of images, sequences, or any structured data. The input shape depends on the specific data format and task. Its importance lies in efficiently ingesting the data into the compression process, ensuring that the original data is properly handled for subsequent compression stages.

2. **Encoder Convolutional Layers:** The Encoder Convolutional Layers are responsible for extracting low-level and progressively higher-level features from the input data. In the context of cloud-based compression, these layers help identify fundamental patterns and structures within the data, which is crucial for efficient compression. For images, this might involve recognizing edges, textures, or basic shapes. Higher-level features capture more complex information. Extracting these features contributes to the compression process by reducing redundant or less important information, leading to effective data representation.

3. **Encoder LSTM Layers:** The Encoder LSTM Layers handle the sequential dependencies within the data. For cloud-based compression, this is particularly important when dealing with sequences or time-series data. These layers model the relationships between data points over time, ensuring that the compression process retains the data's temporal characteristics. In image compression, they can capture spatial dependencies in the image. Their role is crucial in preserving the data's integrity during compression.

4. **Latent Space Layer:** The Latent Space Layer computes the mean (μ) and variance (σ^2) of the latent space distribution. This layer is fundamental to the VAE framework.

ISSN 2229-6107
www.ijpast.in
Vol 12,Issuse 3. July 2022

In the context of cloud-based compression, it provides a compressed representation of the data, which is a key aspect of data compression. This compact representation reduces storage requirements, making it efficient for cloud storage or transmission.

5. **Sampling Layer:** The Sampling Layer takes samples from the latent space distribution, utilizing the computed mean and variance. The reparameterization trick ensures that sampling is differentiable, facilitating backpropagation during training. In cloud-based compression, this layer is vital for generating compressed representations of the data. These compressed representations can be transmitted or stored efficiently in the cloud, reducing bandwidth and storage costs.

6. **Decoder LSTM Layers and Decoder Convolutional Layers:** These layers work in tandem to reconstruct the original data from the compressed representation. In the cloud context, they are responsible for reconstructing the compressed data efficiently. The Decoder LSTM Layers ensure that the temporal or spatial dependencies are preserved during reconstruction. The Decoder Convolutional Layers refine and upsample the data, aiming to match the original data's structure and features. Efficient reconstruction is crucial for maintaining data quality after compression.

7. **Output Layer:** Finally, the Output Layer produces the final output data, which should ideally match the input data in the case of reconstruction tasks. In cloud-based compression, this layer ensures that the compressed data can be faithfully reconstructed when needed. Accurate reconstruction is essential for maintaining data integrity in cloud storage or transmission.

In summary, each layer in the INTER-LSTM+Convolutional VAE architecture contributes to the efficient compression of data, whether it's images or sequences, making it suitable for cloud-based applications. The proposed architecture extracts and represents important features, models dependencies, compresses data into a compact latent space, and facilitates accurate reconstruction, all of which are crucial for reducing storage and bandwidth costs in the cloud.

RESULTS AND DISCUSSION:

The integration of Hadoop Distributed File System (HDFS) into Deep Learning (DL) workflows can have a significant impact on performance and resource allocation. In a scenario where the accuracy of DL models reaches an impressive 98.4%, while maintaining a minimum memory allocation size of just 100 KB, the role of DL frameworks like LSTM-CVAE becomes even more critical. With this aspect the overall experiment results demonstrating an accuracy of 98.4% for anomaly detection in image data underscore the crucial role of Hadoop Distributed File System (HDFS) in the realm of Big Data analysis. This exceptional accuracy is a testament to the effectiveness of advanced machine learning models, showcasing their potential to detect anomalies with precision. However, it's essential to highlight how HDFS significantly contributes to achieving such remarkable results in the context of Big Data analysis.

1. Improved Scalability and Data Management: HDFS excels in managing large datasets efficiently by distributing them across multiple nodes in a Hadoop cluster. This scalability ensures that DL models, like LSTM-CVAE, can handle vast amounts of data without bottlenecks. The result is an accuracy boost, as the model can learn from more diverse and extensive datasets, potentially improving performance.

2. Reduced Memory Footprint: Achieving high accuracy with a minimal memory allocation of 100 KB highlights the importance of resource-efficient DL models. LSTM-CVAE is known for its ability to capture temporal dependencies in sequential data, and its compact architecture allows it to operate effectively with limited memory resources. This is particularly advantageous when deploying DL solutions in resource-constrained environments or on edge devices.

3. Real-time and Edge Processing: The combination of HDFS and memory-efficient DL models opens the door to real-time data processing and edge computing applications. LSTM-CVAE's ability to work with minimal memory makes it suitable for edge devices,

ISSN 2229-6107
www.ijpast.in
Vol 12,Issuse 3. July 2022

where memory constraints are common. This facilitates tasks such as real-time data analysis, anomaly detection, and predictive maintenance, where low-latency processing is critical.

4. Enhanced Anomaly Detection: LSTM-CVAE's unique ability to model temporal dependencies and generate data samples makes it invaluable for anomaly detection. In scenarios like cybersecurity or industrial IoT, where anomalies may indicate security breaches or machinery faults, the high accuracy and minimal memory footprint of the model can lead to rapid detection and response, bolstering system security and reliability.

5. Improved Decision Support: DL models like LSTM-CVAE not only achieve high accuracy but also provide interpretable representations of data. This interpretability allows domain experts to gain insights from the model's output, leading to better decision support systems. In fields like healthcare or finance, where precise predictions and understanding the rationale behind them are critical, the combination of HDFS and LSTM-CVAE can lead to more informed decisions and improved outcomes.

In conclusion, the experiment's exceptional accuracy in image data anomaly detection underscores the synergy between advanced machine learning models and the capabilities of Hadoop Distributed File System. HDFS's scalability, real-time processing, data durability, and fault tolerance make it a linchpin in achieving high accuracy in Big Data analysis tasks. The combination of powerful algorithms and a robust data management infrastructure paves the way for more effective anomaly detection and informed decision support across various domains.

Table 1: Representing the performance metrics for each of the Image dataset chosen for BIG data analysis based on existing and proposed algorithms

ALGORITHMS	ACCURACY	MEMORY ALLOCATION (WITHOUT HDFS) (MB)	WITH HDFS (MB)	COMPRESSION FACTOR (%)
CNN(SOA)	91.95	2210.4	156.45	7.014
ENSEMBLE K-MEANS	87.42	2314.8	193.5	8.35
HYBRID CNN (PROPOSED)	**94.5**	**1918.5**	**105.47**	**6.9**
LSTM+CVAE Hybrid (PROPOSED)	**98.48**	**1985.52**	**102.17**	**5.14**
LSTM (SOA)	97.85	1785.56	140.52	7.86
UNET(SOA)	98.41	1625.71	124.38	7.65

In the context of Big Data analysis with Hadoop Distributed File System (HDFS), it's crucial to assess and compare the performance of different algorithms based on accuracy, memory allocation (both without and with HDFS), and compression factor. We consider the overall experiment with verification of the design and detailed comparison of the algorithms and highlight the best proposed algorithm for this scenario:

1. CNN (SOA - State of the Art): This algorithm achieves an accuracy of 91.95%, making it a solid performer in image data analysis. However, it requires significant memory allocation, both with and without HDFS, resulting in a compression factor of 7.014%. While it provides respectable results, its memory requirements limit its scalability in a Big Data environment.

2. Ensemble K-Means: Despite a slightly lower accuracy of 87.42%, Ensemble K-Means offers a

ISSN 2229-6107
www.ijpast.in
Vol 12,Issuse 3. July 2022

more memory-efficient approach when compared to CNN. Its compression factor of 8.35% suggests a reasonable balance between memory utilization and performance. However, it falls short in accuracy compared to some other algorithms.

3. Proposed HYBRID CNN: This proposed algorithm demonstrates a higher accuracy of 94.5% and an improved compression factor of 6.9%. It strikes a balance between memory allocation and performance, making it a competitive option for Big Data analysis. Its ability to maintain high accuracy while reducing memory requirements is noteworthy.

4. Proposed LSTM+CVAE Hybrid: With an outstanding accuracy of 98.48% and a minimal memory allocation of 102.17 MB with HDFS, the proposed LSTM+CVAE Hybrid stands out as the top performer. It achieves an exceptional compression factor of 5.14%, emphasizing its efficiency in handling Big Data. This algorithm excels in both accuracy and memory optimization.

5. LSTM (SOA - State of the Art): LSTM, a state-of-the-art sequence modeling algorithm, showcases a high accuracy of 97.85%. However, its memory allocation is relatively high, resulting in a compression factor of 7.86%. While it offers strong predictive capabilities, it doesn't match the memory efficiency of some other approaches.

6. UNET (SOA - State of the Art): UNET achieves an impressive accuracy of 98.41%, making it one of the top-performing algorithms. Its memory allocation is relatively lower compared to CNN, and it exhibits a compression factor of 7.65%. UNET is a strong contender for tasks where accuracy is paramount.

Among the algorithms, the proposed LSTM+CVAE Hybrid stands out as the top performer, achieving an exceptional balance between accuracy and memory efficiency. In the realm of Big Data analysis with HDFS, where resource optimization is critical, this algorithm offers a compelling solution. Its ability to maintain a high level of accuracy while significantly reducing memory requirements ensures that it can efficiently handle large-scale datasets. This is of utmost importance in scenarios where comprehensive data analysis is essential for informed decision-making, such as healthcare, finance, or industrial processes.

The proposed LSTM+CVAE Hybrid algorithm's efficiency not only allows for better resource allocation but also facilitates real-time or near-real-time analysis of Big Data, a crucial capability in many applications. Its exceptional compression factor of 5.14% implies reduced storage costs and optimized data management within the HDFS infrastructure.

In conclusion, the proposed LSTM+CVAE Hybrid algorithm emerges as the best choice for Big Data analysis with HDFS, thanks to its outstanding accuracy, memory efficiency, and compression factor. Its importance lies in its ability to empower organizations to harness the full potential of their Big Data while optimizing resource utilization and facilitating timely insights and decision-making.

CONCLUSION:

The overall conclusive results of our experiment in Intuitive Dimensionality Reduction on Big Data using LSTM+CVAE and HDFS have yielded exceptional outcomes, with an impressive accuracy rate of 98.4% and a minimal memory allocation of just 100 MB. This achievement highlights the potential of combining state-of-the-art deep learning techniques with the power of Hadoop Distributed File System (HDFS) in addressing the challenges of handling and analysing massive datasets.

1. Unveiling Data Insights: The remarkable accuracy achieved in this experiment signifies the efficacy of the LSTM+CVAE model in extracting meaningful representations from high-dimensional data. This empowers data scientists and analysts to unveil valuable insights, patterns, and relationships within the data that might have remained hidden without the dimensionality reduction approach.

2. Efficient Resource Utilization: The minimal memory allocation requirement of 100 MB underscores the resource-efficiency of our approach. This not only makes it feasible to deploy our model in resource-constrained environments but also demonstrates the effectiveness of LSTM+CVAE in optimizing memory usage while preserving data fidelity.

3. Scalability with HDFS: The utilization of HDFS as the underlying data storage and distribution platform plays a pivotal role in achieving such remarkable results. HDFS's scalability ensures that

ISSN 2229-6107
www.ijpast.in
Vol 12,Issuse 3. July 2022

our LSTM+CVAE model can seamlessly process and analyse large-scale datasets, accommodating the ever-expanding volumes of Big Data that organizations encounter.

4. Real-World Applicability: The combination of high accuracy and efficient memory allocation holds immense promise in various real-world applications. From healthcare to finance, from cybersecurity to industrial processes, the ability to intuitively reduce the dimensionality of Big Data while maintaining data quality empowers organizations to make data-driven decisions with confidence.

5. Future Prospects: As the volume and complexity of Big Data continue to grow, the integration of cutting-edge deep learning models like LSTM+CVAE with HDFS will become increasingly valuable. This experiment's success sets the stage for further exploration and innovation in the realm of data analysis, offering new avenues for extracting actionable insights from the vast sea of Big Data.

In summary, our experiment showcases the potential of Intuitive Dimensionality Reduction on Big Data using LSTM+CVAE and HDFS to provide accurate, resource-efficient, and scalable solutions for tackling the challenges posed by the ever-expanding world of Big Data analysis.

REFERENCES:

1. Ss S. Ziani, Y. Farhaoui and M. Moutaib, "Extraction of Fetal Electrocardiogram by Combining Deep Learning and SVD-ICA-NMF Methods," in Big Data Mining and Analytics, vol. 6, no. 3, pp. 301-310, September 2023, doi: 10.26599/BDMA.2022.9020035.
2. S. Shafqat, Z. Anwar, Q. Javaid and H. F. Ahmad, "A Unified Deep Learning Diagnostic Architecture for Big Data Healthcare Analytics," 2023 IEEE 15th International Symposium on Autonomous Decentralized System (ISADS), Mexico City, Mexico, 2023, pp. 1-8, doi: 10.1109/ISADS56919.2023.10092137.
3. J. T. K, G. J and P. S, "A Survey on Prediction of Risk Related to Theft Activities in Municipal Areas using Deep Learning," 2023 Second International Conference on Electronics and Renewable Systems (ICEARS), Tuticorin, India, 2023, pp. 1321-1326, doi: 10.1109/ICEARS56392.2023.10085123.
4. H. Ashfaq and A. Jalal, "3D Shape Estimation from RGB Data Using 2.5D Features and Deep Learning," 2023 4th International Conference on Advancements in Computational Sciences (ICACS), Lahore, Pakistan, 2023, pp. 1-7, doi: 10.1109/ICACS55311.2023.10089663.
5. H. Vanam and J. R. R. R, "Sentiment Analysis of Twitter Data Using Big Data Analytics and Deep Learning Model," 2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF), Chennai, India, 2023, pp. 1-6, doi: 10.1109/ICECONF57129.2023.10084281.
6. S. Khan, V. Ch, K. Sekaran, K. Joshi, C. K. Roy and M. Tiwari, "Incorporating Deep Learning Methodologies into the Creation of Healthcare Systems," 2023 International Conference on Artificial Intelligence and Smart Communication (AISC), Greater Noida, India, 2023, pp. 994-998, doi: 10.1109/AISC56616.2023.10085651.
7. H. -y. Zhang, Y. Fang, J. -h. Wu, W. -z. Wang and N. -y. Zou, "Deep Learning of Color Constancy Based on Object Recognition," 2023 15th International Conference on Computer Research and Development (ICCRD), Hangzhou, China, 2023, pp. 215-219, doi: 10.1109/ICCRD56364.2023.10080343.
8. E. -Y. Ma, H. Kim and U. Lee, "Investigating Causality in Mobile Health Data through Deep Learning Models," 2023 IEEE International Conference on Big Data and Smart Computing (BigComp), Jeju, Korea, Republic of, 2023, pp. 375-377, doi: 10.1109/BigComp57234.2023.00089.
9. J. Assandre and R. Martins, "Analysis of Scientific Production on the Use of Big Data Analytics in Performance Measurement Systems," in IEEE Latin America Transactions, vol. 21, no. 3, pp. 367-380, March 2023, doi: 10.1109/TLA.2023.10068840.
10. P. Dominic, N. Purushothaman, A. S. A. Kumar, A. Prabagaran, J. Angelin Blessy and J. A, "Multilingual Sentiment Analysis using Deep-Learning Architectures," 2023

ISSN 2229-6107
www.ijpast.in
Vol 12,Issuse 3. July 2022

5th International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2023, pp. 1077-1083, doi: 10.1109/ICSSIT55814.2023.10060993.

11. R. Jegadeesan, E. Vijayakrishna Rapaka, K. Himabindu, N. R. Behera, A. K. Shukla and A. K. Dangi, "Grey Wolf Optimizer with Deep Learning based Short Term Traffic Forecasting in Smart City Environment," 2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2023, pp. 1065-1070, doi: 10.1109/ICSSIT55814.2023.10061127.
12. S. Ayoub, N. R. Behera, M. N. Raju, P. Singh, S. Praveena and R. K, "Hyperparameter Tuned Deep Learning Model for Healthcare Monitoring System in Big Data," 2023 International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), Bengaluru, India, 2023, pp. 281-287, doi: 10.1109/IDCIoT56793.2023.10053418.
13. Manoranjithem, S. Dhanasekaran, A. Asokan, A. Kumar, C. Yamini and M. Tiwari, "An Intrusion Detection Approach using Hierarchical Deep Learning-based Butterfly Optimization Algorithm in Big Data Platform," 2023 International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), Bengaluru, India, 2023, pp. 212-216, doi: 10.1109/IDCIoT56793.2023.10053504.
14. N. Liu and J. Zhao, "Recommendation System Based on Deep Sentiment Analysis and Matrix Factorization," in IEEE Access, vol. 11, pp. 16994-17001, 2023, doi: 10.1109/ACCESS.2023.3246060.
15. E. Zvuloni, J. Read, A. H. Ribeiro, A. L. P. Ribeiro and J. A. Behar, "On Merging Feature Engineering and Deep Learning for Diagnosis, Risk Prediction and Age Estimation Based on the 12-Lead ECG," in IEEE Transactions on Biomedical Engineering, vol. 70, no. 7, pp. 2227-2236, July 2023, doi: 10.1109/TBME.2023.3239527.
16. A. Kumar, K. U. Singh and M. Kumar, "A Clinical Data Analysis Based Diagnostic Systems for Heart Disease Prediction Using Ensemble Method," in Big Data Mining and Analytics, vol. 6, no. 4, pp. 513-525, December 2023, doi: 10.26599/BDMA.2022.9020052.
17. M. Vamsikrishna, R. RS, G. G S, D. Suganthi, M. A. Ala Walid and R. Kuchipudi, "Big Data Analytics with Wild Horse Optimizer based Deep Learning Model for Healthcare Management," 2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2023, pp. 1006-1012, doi: 10.1109/ICIRCA57980.2023.10220710.
18. E. T. Nassar, H. G. Elgazouly, A. M. Elnaggar and S. M. Ayyad, "Leveraging Deep Learning and IoT Big Data Analytics for The Determination of Development Priorities Utilizing GeoAI in The National Project for The Development of the Egyptian Rural Villages - Decent Life "Hayah Karima"," 2023 International Telecommunications Conference (ITC-Egypt), Alexandria, Egypt, 2023, pp. 73-78, doi: 10.1109/ITC-Egypt58155.2023.10206340.
19. S. K. Panda and P. Ray, "A Survey on Weather Prediction using Big Data and Machine Learning Techniques," 2023 5th International Conference on Energy, Power and Environment: Towards Flexible Green Energy Technologies (ICEPE), Shillong, India, 2023, pp. 1-6, doi: 10.1109/ICEPE57949.2023.10201614.
20. P. Sasikumar and K.Kalaivani, "Real-Time Big Data Analytics for Live Streaming Video Quality Assessment Using Deep Learning," 2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), Chennai, India, 2023, pp. 1-8, doi: 10.1109/ACCAI58221.2023.10200226.