**International Journal of Environmental Sciences**
ISSN: 2229-7359
Vol. 11 No. 1s, 2025
https://www.theaspd.com/ijes.php

# Machine Learning Based Big Data Analytics Framework For Discovering Of Patterns From Multiple Data Sources

**Mohammad Islam[1], Dr. Ashish Sharma[2,3]*, Zahid Ahmed[4], Nausheen Khilji [5]**

[1] Research Scholar, Department of Computer Science, Maulana Azad University, Jodhpur-342802 Rajasthan, India, Email id - islamjodhpur@gmail.com, https://orcid.org/0009-0001-2548-4262

[2*] Associate Professor, HOD, Department of Computer Science, Maulana Azad University, Jodhpur-342802 Rajasthan, India, Email id - aashishid@gmail.com

[3] Professor, Department of Technology, JIET, Jodhpur 342802, Rajasthan, IndiaEmail id - aashishid@gmail.com https://orcid.org/0009-0007-6644-6362

[4]Assistant Professor, Department of Computer Science and Application, Vivekananda Global University, Jaipur - 303012 Rajasthan, India Email Id- zahidahmed59@gmail.com , https://orcid.org/0009-0002-4020-1002

[5] Assistant Professor, Department of Technology, JIET UNIVERSE, Jodhpur 342802, Rajasthan, India Email Id- naushy90@gmail.com, https://orcid.org/0000-0003-1750-2675

**Abstract**

*Organization that scatters over different regions that perform multi-state business transaction are always interested I identifying novel patterns of interest given augmenting nosiness volumes. On a daily basis, we encounter an unparalleled increase in data from many sources, which adds to the concept of big data in terms of its size, speed, and diversity. These datasets provide significant obstacles to analytics frameworks and processing resources, so making the total study arduous for extracting useful information promptly. Therefore, in order to overcome these types of difficulties, it is crucial to create a very effective framework for analyzing large amounts of data. Therefore, in order to tackle these difficulties by harnessing non-linear connections from extensive and complex information, analytics frameworks are using machine (ML) and (DL) methods. Apache Spark is widely recognized as the most efficient tool for processing large amounts of data. It is particularly useful for solving complex machine learning problems that need several iterations. Spark MLlib, a distributed machine learning library, is used for this purpose. When dealing with research problems in the real world, architectures like Long Short-Term Memory in deep learning are a useful method for addressing practical challenges like as decreased accuracy, long-term sequence dependence, and the problem of disappearing and expanding gradients in traditional deep architectures. This study proposes a very efficient analytics system that combines a progressive machine learning method with Spark-based linear models, Multilayer Perceptron , and Long Short-Term Memory. The accuracy of the system's predictions is improved by the use of a two-stage cascade structure. The architecture that we provide makes it possible to organize statistical analysis of large amounts of data in a way that is not only scalable but also effective. We applied the cascading structure on two different real-world datasets in order to illustrate the effectiveness of our framework. Furthermore, the results of the experiments show that our analytical framework is superior to the approaches that are considered to be state-of-the-art in terms of classification accuracy.*

***Keywords:*** *Big data, multiple data source, patterns, machine learning.*

## 1. INTRODUCTION

Big data refers to very large and complex collections of data (including real-time, unstructured, and streaming data) that pose challenges for traditional techniques of data management, analysis, storage, and retrieval in order to extract valuable insights [1][2]. In recent times, the field of BD and its associated tools and methods, such as Big Data Analytics (BDA), have revolutionized the operations of organizations and companies. This has brought up new and important opportunities for corporations, professionals, and academia [3]. In addition to corporations and research institutions, governmental and non-government organizations increasingly often produce vast amounts of data that are distinct in their scale and complexity [4][5]. The rising amount of data is the primary characteristic of "big data," a term that has become common parlance in the academic communities, organizations, and on the Internet. "Big data" is a jargon that has become a household word. The insights that may be gleaned from this data can be used in the construction of intelligent applications across a wide range of fields. Applications for real-time engineering are built on tools and strategies for handling the data and having the power to extract meaningful information or insights in a manner that is both intelligent and timely. Hence, it is crucial for organizations globally to extract significant information and substantial benefits from the abundant big data at their disposal. Nevertheless, the research indicates that effectively and proficiently extracting valuable insights from big data in a fast and effortless manner is a difficult task [6]. BDA has become crucially necessary in order to fully use the value of Big Data and enhance company performance and market share for most organizations. While many (AI) and (ML) techniques and platforms for doing Big Data Analytics are available for free, they need a unique skill set that is not often possessed by most professionals in this sector or IT departments of organizations [7][8]. Therefore, effectively integrating these tools and platforms into an organization's internal and external data on a shared platform is a difficulty. Over the last several years, there has been a significant accumulation of data across many domains such as medical, government, industry, chemistry, biology, and other academic sectors.

Web-based applications such as social computing, online texts and documents, and online search indexing often handle substantial amounts of data. Proficiency in diverse computing challenges, information security protocols, and computational methodologies might help alleviate obstacles encountered in the analysis of massive datasets [9]. For example, several statistical methods that are effective with data sets of intermediate size have difficulties when dealing with really large datasets. The same is true for the analysis of huge data, when many computing approaches that work well with little data encounter substantial difficulties. Numerous researchers [10] have examined the sector's responses to a wide range of difficulties. In order to make a decision or take some kind of action in a short period of time, real-time applications need immediate input and rapid analysis.

## 2. REVIEW OF LITERATURE

**Maroufkhani et al., (2023)[11]** analyzed the effect of Technological, Organizational, and Environmental (TOE) variables on the adoption of BDA among Small and Medium Enterprises (SMEs). This study challenges the assumption of independence among TOE elements. It was suggested that support from top management may operate as a mediator between the adoption of BDA and the technical and organizational variables involved. In addition, this research investigated and evaluated how environmental variables influence the connection between positional advantage, reliability, productivity, organizational preparedness, and adoption of BDA. The method of partial least squares was used to conduct the analysis on the data that was obtained from 171 small and medium-sized manufacturing companies. The data provided evidence that the TOE components are interconnected with one another. It's also possible for environmental factors to lessen the benefits of alignment and organizational preparation on support from top management. The results make the TOE model stronger by questioning the idea of independence that all TOE parts are based on.

**VentkateswaraRao et al., (2023)[12]** studied that the banking sector has seen substantial changes in terms of operational efficiency and service delivery in recent decades. The rising global population poses a significant challenge to the financial services infrastructure. By catering to a substantial portion of clientele, it enhances the customer base, online transaction volume, and generates substantial quantities of data. Currently, financial institutions in the United States and several other countries use the utilization of Big Data Analytics (BDA) on a regular basis in order to effectively manage this particular situation. The main goal of this system is to find different trends and patterns in organizational files so that businesses can make more money. This new technology called "big data" makes managing large amounts of data much easier and more useful. It can be used in risk management situations that need complex data analysis and a lot of data. This study talks about the structure of a banking investigation into credit and integrated risk administration system, with a focus on the big data parts that make it up. The comparisons and analyses conducted clearly demonstrate that the proposed system has improvement in both efficiency and security and exhibits superior performance as compared to other method.

**Qi et al., (2023)[13]** stated that the convergence of the expansion of big data with the proliferation of the IoT is clearly shown by the significant increase in the utilization of IoT-connected devices and the exponential development in data consumption. The deployment of a substantial quantity of sensors and devices within the industry sector has led to the generation of a significant volume of data in the context of the IIoT. The existing body of literature encompasses several studies that have been undertaken on the subject of BDA and Industrial Internet of Things (IIoT). However, there is a dearth of study pertaining to the primary obstacles that impede the advancement of intelligent IIoT systems. And introduced a novel approach that combines multi-objective optimization with ratio analysis and the complete multiplicative form (MULTIMOORA) technique. In addition, the architecture of q-rung orthopair fuzzy sets (q-ROFSs) contains criterion interaction through inter-criterion correlation (CRITIC). The proposed approach makes use of the CRITIC technique to calculate attribute weights and the MULTIMOORA algorithm to estimate where each option falls on the q-ROFSs. Within the framework of Industry 4.0, perform a case study to investigate the challenges presented by Big Data Analytics (BDA) in the creation of smart Industrial Internet of Things (IIoT) systems. The usefulness of the developed framework in setting priorities for intelligent IIoT systems is further shown by a comparison and sensitivity analysis of the proposed approach.

**Kanan et al., (2023)[14]** examined one of the most prevalent problems affecting businesses today is satisfaction detection. Therefore, the purpose of this research is to propose utilizing well-known machine learning and deep learning (DL) approaches to identify the Satisfaction tone that leads to customer satisfaction in Big Data originating from e-commerce platforms such as Facebook and Twitter. There is a dearth of relevant data sets for this area of study. As a direct result of this, we pruned the data collection that came from the internet of things. Companies can now use Big Data analytics on social media more easily thanks to the team's work on three well-known NLP technologies: stemming from, standardized procedures, and stop word removal. We found the F1-measure, Recall, and Precision to figure out how well the algorithms worked. Based on the results, the Random Forest classifier is the most suitable option among the various alternatives. The F1-measure achieved a classification accuracy of 99.1% using this strategy. The Support Vector Machine achieved the highest performance utilizing an F1-measure of 93.4%, without any pre-processing steps. Conversely, we use DL methods, namely Word Embedding and Bag of Words as part of our feature extraction strategy, on the dataset. According to the findings, the Deep Neural Networks (DNN) method performed better.

**Li et al., (2022) [15]** examined the primary purpose of this research is to guide the smart city toward better governance and more secure data processing by using BDA to the massive volumes of data generated by the IoT in the smart city. The paper talks about how to use CNN's distributed parallelism technique to use BDA to apply the deep learning (DL) approach on the multi-source data collected in the smart city. Digital twins (DTs) and multi-hop transfer technologies make it possible to build a DL-based IoT-BDA system to support a smart city. This makes it easier to test and simulate how well it works. According to research on low energy absorption, sending model data more efficiently gets better as it gets bigger. Still, a better power diversion factor is very important for making the transmission of data in the IoT-BDA system more energy efficient. The evaluation of the model's predictive accuracy shows that the system developed achieves an accuracy of 97.80%, surpassing the DL technique used by other researchers by at least 2.24 percentage points. When the probability of successful transmission is 100% and the distribution of exponential parameters are set to values between 0.01 and 0.05, the constructed system achieves the highest level of accuracy and maintains the lowest level of data delay, which is in the millisecond range. In conclusion, the DL method for enhancing the IoT-BDA system in a smart city may lessen the time it takes to send and receive data, increase the precision with which data predictions can be made, and provide real-world benefits, all of which serve as useful experimental benchmarks for the smart city's digital future.

**Awan et al., (2022)[16]** examined the number of people using various types of online social media is now at an all-time high. This led to a growth in the trend of false profiles, which is having a negative impact on both social and commercial organizations since fraudsters are using photographs of real individuals to create new phony accounts. The majority of these recommended approaches, on the other hand, are antiquated and do not provide sufficient precision, with an average accuracy of 83%. A project based on Spark ML that is capable of predicting false profiles with a greater level of accuracy than other existing techniques of profile recognition is the answer that we have suggested as a solution to this issue. Our work includes the creation of Spark ML libraries, which include the Random Forest Classifier and several other visualization tools. For your convenience, we have included a diagrammatic description of our suggested model, and we have made an effort to display our findings using graphical representations such as confusion matrices, learning curves, and ROC plots. The outcomes of the research conducted for this project demonstrate that the suggested method has an accuracy of 93% when it comes to locating phony accounts across social media networks. Although there is a 7% false positive rate, which means that sometimes our algorithm is unable to recognize a bogus profile effectively.

**Zhang et al., (2021) [17]** analyzed the use of big data, driven by information technology, presents promising opportunities for urban planning and construction. New technologies, like the Internet of Things (IoT) for collecting data and Artificial Intelligence (AI) for processing large amounts of data, make it easier to share and integrate data and make smart cities work better. In the last several years, the idea underlying the IoT has been a central focus of study in the advancement of smart cities, classrooms, factories, and businesses. Smart cities rely heavily on the services and applications of the IoT in order to provide a sustainable urban lifestyle. ICT in the internet of things makes smart city participants more knowledgeable, productive, and engaged. Since more smart city applications are being built on the Internet of Things, more data is being generated and processed. The city's stakeholders and governments also take preventative measures to ensure a sustainable city by analysing data from IoT devices and anticipating their potential effects. When it comes to boosting the effectiveness of fire danger detection in smart cities, artificial intelligence has emerged as a significant research approach that has been analyzed and shown to be the best by a number of researchers. This study proposes using a DBN in conjunction with a Recurrent LSTM Neural Network (R-LSTM-NN) to make predictions using large datasets amassed by IoT-based smart cities. The suggested methodology also focuses on estimating the fire risk data collected from IoT sensors in smart cities. In regards to accuracy, precision, memory, and F-1 score, the simulation results show that the suggested method does better than methods that are thought to be best in their field. The model that was suggested can pinpoint the start of a fire with an amazing 98.4% accuracy and a mere 0.14 % mistake rate. The suggested method could also be used to solve other predicting problems that smart cities face.

**Behera et al., (2021) [18]** analyzed of customer feedback that has been submitted on social media is vital for a variety of commercial applications. The number of reviews left by customers on social media platforms, as well as the significance of those evaluations, are growing at an exponential pace, which ultimately results in the accumulation of large amounts of data. When it comes to the sequential analysis of a long text, recurrent networks (LSTM) generally produce gratifying results, whereas deep convolutional networks have shown to be effective when it comes to identifying pertinent local features. The proposed Co-LSTM model for sentiment analysis has two primary objectives. To begin, it is very adaptable for evaluating massive quantities of social data while maintaining scalability. Secondly, in contrast to the traditional methods to machine learning, it is not restricted to any one specific topic. In order to train a model that is capable of handling the many types of dependencies that often crop up in a post, the experiment was carried out on four review datasets that came from a variety of different disciplines. The findings of the experiments demonstrate that the suggested ensemble model performs superiorly to alternative techniques to machine learning in terms of accuracy as well as other characteristics.

**Lopez-Martin et al., (2020) [19]** detailed the novel use of many deep reinforcement learning (DRL) algorithms on a labeled dataset for intrusion detection. Here, we describe a DRL-based methodology for supervised learning. Since there is no automatic mechanism to identify intrusions, it is very challenging for Intrusion Detection Systems (IDS) to develop a reward function that is linked with the detection of intrusions. Most often, this is a manual process, with the results documented in databases of network parameters related to intrusion occurrences. For the purpose of categorizing intrusion events, supervised machine learning algorithms are trained on these datasets. Application of DRL is shown using the NSL-KDD and AWID datasets. We've taken a novel approach by revising the standard DRL paradigm conceptually. An example of this change is the introduction of a sampling function of previously recorded training incursions in place of the original surroundings. The training dataset is sampled in this innovative pseudo-environment, and incentives are generated based on the errors in detection that are discovered during training. And demonstrated that by using our model and making some minor tweaks to the parameters, DRL, in contrast to other machine learning approaches, may provide superior results for intrusion detection.

**Jiang et al., (2019) [20]** examined a reliable and rapid disease detector in apples is still missing from the literature, which would be crucial for the industry's continued success. In order to identify illnesses in apple leaves in real time, the authors of this research suggest a deep learning method based on enhanced CNNs. Using data enhancement and image annotation tools, the first step of this project is to create the Apple Leaf Disease Dataset (ALDD). Within the dataset there are both controlled laboratory photos and detailed photographs taken in natural settings. Researchers found that the INAR-SSD model can quickly find objects on ALDD (23.13 frames per second), and it does this with 78.80% mean average accuracy. The findings show that the innovative INAR-SSD model offers an efficient approach for early identification of apple leaf diseases, outperforming prior techniques in both accuracy and speed of detection in real-time.

**Thamilarasu et al., (2019) [21]** suggested as more and more devices, apps, and communication networks are being linked to one another and integrated, the number of malicious cyberattacks targeting the IoT is expanding at an alarming pace. When assaults on IoT networks go unnoticed for extended periods of time, it has a negative impact on the availability of vital services for end users, it raises the amount of data breaches and identity thefts, it causes expenses to rise, and it has an adverse effect on income. In order to offer effective security and protection, it is very necessary to identify assaults on IoT systems in close to real time. We create a smart intrusion-detection system that works with the Internet of Things (IoT) in this study. To be more specific, we use a deep-learning system to analyze IoT network data in order to identify malicious activity. The detection solution enables compatibility across multiple network communication protocols that are used in the Internet of Things and delivers security as a service. We test our proposed detection system by employing simulation in order to provide evidence of its scalability as well as real-world network traces in order to provide a proof of concept for the framework. The results of our experiments demonstrate that the suggested method for detecting intrusions is capable of efficiently detecting intrusions in the actual world.

## 3. PROBLEM FORMULATION

In the era of massive data proliferation, organizations face the daunting challenge of extracting meaningful insights from diverse and voluminous datasets sourced from various channels. The advent of Big Data has accentuated the need for sophisticated analytics frameworks to unravel hidden patterns and valuable knowledge within this deluge of information. However, existing methodologies often struggle to cope with the complexity arising from the integration of multiple data sources. The problem lies in the absence of a unified and efficient framework that seamlessly integrates Machine Learning algorithms with Big Data analytics to discern patterns across disparate datasets. This deficiency impedes the potential for comprehensive analysis and hinders organizations from harnessing the full power of their data reservoirs. Addressing this problem is imperative for unlocking the true potential of Big Data, enabling organizations to make informed decisions, uncover correlations, and derive actionable insights from a multitude of interconnected data sources. Therefore, there is a critical need for the development of a robust and scalable Machine Learning-based Big Data analytics framework tailored to the discovery of patterns across diverse datasets, fostering a more nuanced understanding of complex relationships and trends within the data landscape.

## 4. RESEARCH METHODOLOGY

A large data analysis framework (Spark or Hadoop) and artificial intelligence (ml or dl) are used in the suggested architecture in order to determine real-life high-dimensional big data challenges. The limitations of both time and space make it difficult to find solutions to challenges of this kind. Big data consists of very huge quantities that are always expanding, yet it must still be processed by powerful computer hardware in order to enable learning architectures that can successfully handle the data while making effective use of particular resources.“Cascading the Spark MLlib data analytics framework with MLP and LSTM was one of the suggested framework's approaches toward overcoming these issues. Cascading between Spark MLlib and DL while making use of MLP and LSTM networks is the central idea behind the experiment, which is based on the core architectural framework that is provided here and which forms the basis of the study. Figure 1 depict the proposed framework structure. It contains two learning stage, outlined as follows.

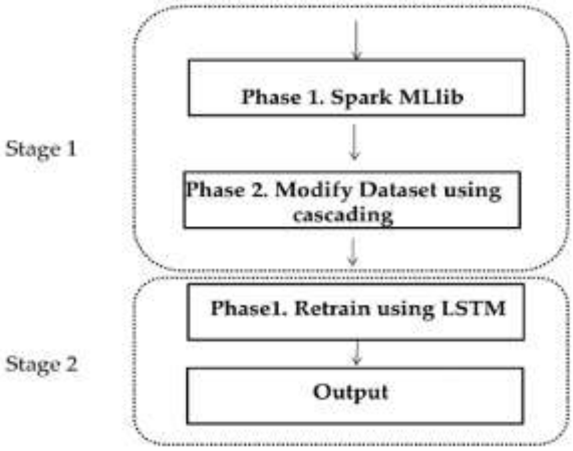

Stage 1
Phase 1. Spark MLlib
Phase 2. Modify Dataset using cascading
Stage 2
Phase1. Retrain using LSTM
Output

**Figure 1. Macro view of the stages in our proposed big data analytics framework.**

Stage 1
i. Phase 1: Big data analysis using spark MLib
ii. Phase 2: Cascading
Stage 2
i. Retrain the model using MLP and LSTM
ii. Output"
Using MLP and LSTM on high-dimensional data, the suggested method combines the benefits offered by the powerful Spark MLlib with those offered by deep learning. The following subsections provide a more in-depth explanation of each stage.

**A. Stage 1**
**Phase 1: Big data processing using spark MLib**
In light of the enormous amounts of data, many machine learning strategies were evaluated. We decided to create a SVM, DT, RF, and (LR) classifiers using the powerful Spark MLlib library. After being preprocessed, the data were run through these classifiers in order to build regression models. These models present the possibility that every data point is related to the binary class. The first step towards learning binary is this one. Train the model by feeding it the training data and utilizing the Spark MLlib to do so. Testing all of the data at once is the only way to generate a binary prediction.

**Phase 2: Cascading**
In the proposed architecture, cascading plays a crucial role in the process of linking the prediction data to the original dataset. When training the MLP and LSTM, we made use of the modified dataset that was generated from stage 2 (knowledge = prediction + original dataset).

**B. Stage 2**
**Retrain the model using Deep learning**
Before output, we have reached the second and last step of the framework learning process. Training the MLP and LSTM models requires utilizing a modified version of the dataset obtained through cascade. MLP and LSTM may be generated by either recycling steps 2–8 from stage 1 and exchanging ML for MLP by making use of the Spark library, or by originally forming MLP from the ANN. Both of these methods can provide the same results. In order to reduce the number of errors produced by the prediction process, MLP may be trained using a high-quality training back-propagation method. This can be done throughout the training process.
Several components of this structure, such as its classification structure and its recommendation tools, have been used in the past in a variety of big data and machine learning domains.

### Framework underlying logic

The previously mentioned structure design improves accuracy and system speed to tackle large data challenges, while also resolving traditional ML concerns and taking into account all ML situations. The next section elaborates on the primary reasoning behind the framework.

### Computation time

The approach that was suggested is quite effective in cutting down on computation time. It would be time-consuming and laborious to run the two-layer technique utilizing two DL models, and it would be fairly challenging to guarantee the quality of the results. Instead of using two different deep learning models simultaneously, we used Hadoop and Spark, which are both fast and efficient large data analysis tools. This allowed us to solve the computation time issue. Spark contains MLlib, which is especially useful for iterative machine learning tasks and greatly reduces the amount of time required for computation in comparison to more traditional machine learning techniques.

### Feature set

The accuracy of the framework is improved when following stages are fed with input comprised of the cascade-modified dataset. Consequently, using the DL model on the updated data set results in an improved feature set.

### Continuous learning improvement

The second step of the proposed architecture for DL combines MLP and LSTM, with back-propagation serving as the primary mechanism of model training. As a result, the model is constantly picking up new characteristics without being directed by an instructor. This makes it possible for the framework to rapidly grow into a model that is more accurate and trustworthy. The train prototype is kept throughout the back propagation learning process, which helps to improve the accuracy of predictions

### 4.1Technique

In this study, the following machine learning techniques are used which has been described given below.

- **Logistic regression**

Logistic regression is used to calculate the probability of predicting a binary dependent variable while considering the independent variable in a dataset. Logistic regression, like linear regression, constructs a curve by fitting a straight line, while linear regression also yields a curve. Logistic regression is a statistical method that examines the association between independent variables and a binary result. It does this by constructing logistic curves that depict the chance of the outcome being 1, ranging from 0 to 1 [22]. Figure 2 depict the diagram of logistic regression as shown below.

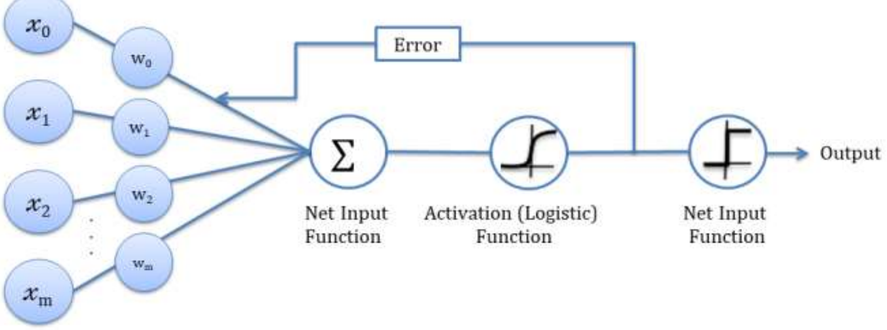


**Figure 2. Diagram of Logistic regression**

- **Decision tree**

Decision trees (DT) are a kind of decision assistance tool that use a hierarchical structure, like a tree. Each node in the tree corresponds to a test conducted on a certain attribute, while each branch indicates a potential outcome. Predictive models are used in data mining and machine learning tasks due to their objective of categorizing observations into distinct and non-overlapping groupings. Decision trees are used in predictive decision support to establish a correlation between facts and likely outcomes. The decision tree approach has several advantages. Upon first observation, the notable level of adaptability offered by the non-parametric technique, which lacks any notion of data distribution, becomes evident. Figure 3 illustrates a decision tree purpose-built diagram, including probability and uncertainty [23].

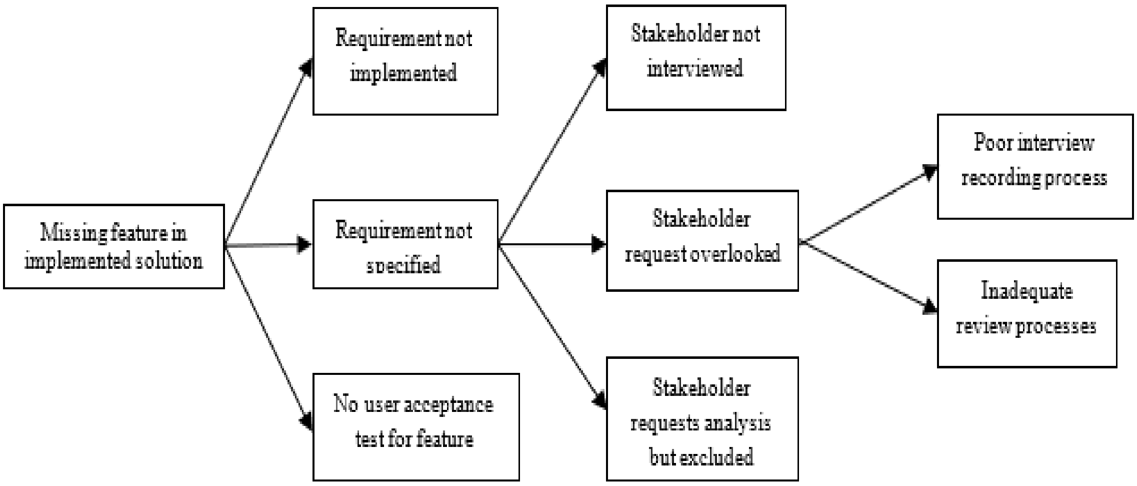

Missing feature in implemented solution
Requirement not implemented
Requirement not specified
No user acceptance test for feature
Stakeholder not interviewed
Stakeholder request overlooked
Stakeholder requests analysis but excluded
Poor interview recording process
Inadequate review processes

**Figure 3. Schematic diagram of Decision tree**

- **Random forest**

This is an ensemble methodology exclusively used to enhance both the efficacy and accuracy of machine learning algorithms in artificial intelligence. Utilizing a random forest approach may assist in identifying the most suitable independent variables that the system can use for optimal operation. Furthermore, several studies have previously shown the importance of considering multiple options for each shrub. Empirical research has also shown that this approach leads to optimum prediction accuracy [24]. The reason for the name of this forest is its abundant population of high-quality trees. The data collected from these trees is then merged to provide the most precise projections attainable. Figure 4 displays a decision tree that includes both a classification model and a regression model to predict the dependent variables [25].

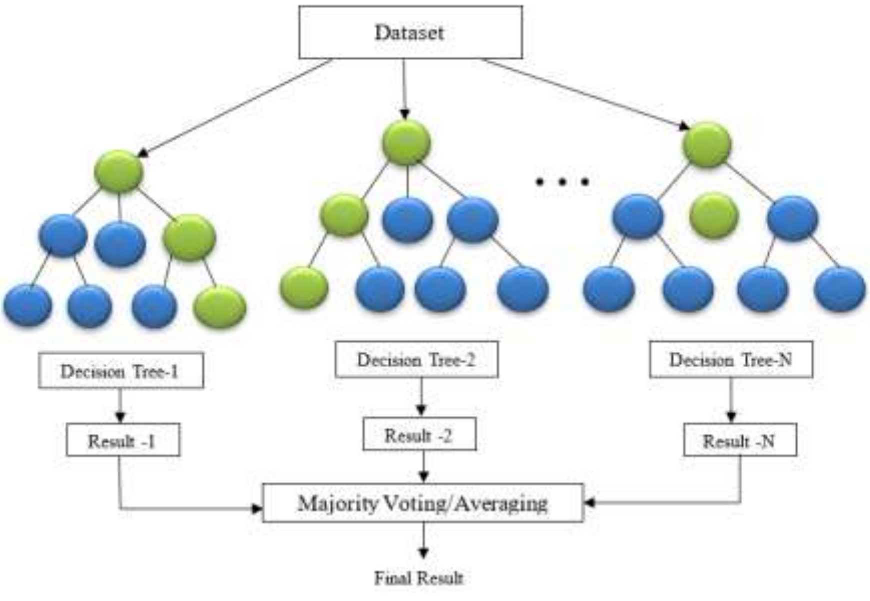
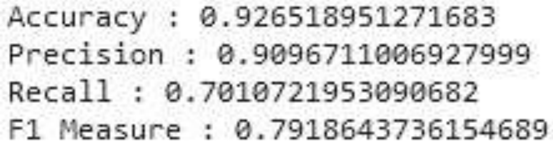


**Figure 4. Schematic diagram of Random Forest**

## 5. RESULT AND DISCUSSION

This section examined findings of this study which is described in performance metrics such as accuracy, precision, recall, and f1-score as shown below.

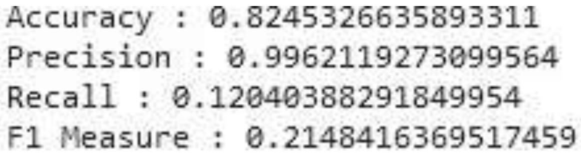
Figure 5 depict the performance metrices of decision tree method. From Fig.5, the decision tree obtained the accuracy is 0.9265, precision is 0.909, recall is 0.7010, and F1 measure is 0.7918 as shown below.



```
Accuracy : 0.926518951271683
Precision : 0.9096711006927999
Recall : 0.7010721953090682
F1 Measure : 0.7918643736154689
```

**Figure 5. Performance metrices of Decision tree**

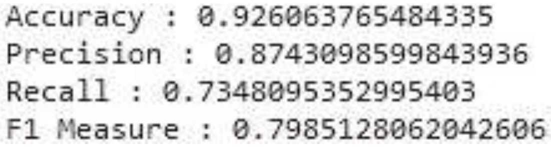
Figure 6 depict the performance metrices of Logistic Regression method. From Fig.6, the logistic regression obtained the accuracy is 0.8245, precision is 0.996, recall is 0.1204, and F1 measure is 0.2148 as shown below.



```
Accuracy : 0.8245326635893311
Precision : 0.9962119273099564
Recall : 0.12040388291849954
F1 Measure : 0.2148416369517459
```

**Figure 6. Performance metrics of logistic regression**

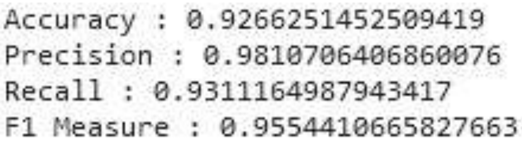
Figure 7 illustrate the performance metrices of Random Forest method. From Fig.7, the Random Forest attained the accuracy is 0.9260, precision is 0.8743, recall is 0.7348, and F1 measure is 0.79851 as shown below.



```
Accuracy : 0.926063765484335
Precision : 0.8743098599843936
Recall : 0.7348095352995403
F1 Measure : 0.7985128062042606
```

**Figure 7. Performance metrics of Random Forest**

Figure 8 demonstrate the performance metrics of proposed (MLP+LSTM) method. From Fig. 8, the proposed method attained the maximum accuracy 0.9266251, precision is 0.98107, recall is 0.93111 and f1 score is 0.95544 which is high all compared to other methods as shown below.



```
Accuracy : 0.9266251452509419
Precision : 0.9810706406860076
Recall : 0.9311164987943417
F1 Measure : 0.9554410665827663
```

**Figure 8. Performance metrics of Proposed (MLP+LSTM) method.**

Table 1 demonstrate the comparison table of proposed method with other methods as well as shown the comparison graph of proposed method in Figure 9. From Table 1, Decision tree method obtained the accuracy is 0.9265, precision is 0.909, recall is 0.7010, and F1 measure is 0.7918. Random Forest attained the accuracy is 0.9260, precision is 0.8743, recall is 0.7348, and F1 measure is 0.79851. Logistic Regression obtained the accuracy is 0.8245, precision is 0.996, recall is 0.1204, and F1 measure is 0.2148.

The proposed method (MLP+LSTM) achieved the accuracy 0.92662, precision is 0.9810, recall is 0.9311, and f1 score is 0.9554 which is higher than to other methods as well as shown in Fig 9. Therefore, it is clear that the proposed method gives superior performance in all terms as compared to other all methods as shown in table 1 and Fig.9.

**Table 1. Comparison table of proposed method**

| Models | Accuracy | Precision | Recall | F1 score |
|---|---|---|---|---|
| Logistic regression | 0.824533 | 0.996212 | 0.120404 | 0.214842 |
| Decision tree | 0.926519 | 0.791864 | 0.909671 | 0.701072 |
| Random Forest | 0.926064 | 0.798513 | 0.874310 | 0.734810 |
| Proposed method (MLP+LSTM) | 0.92665 | 0.955441 | 0.981071 | 0.931116 |

Figure 9 depict the comparison graph of proposed method with other method. The proposed method attained high accuracy, precision, recall and f1 score than to all other methods which give outperformance as shown in Fig.9.

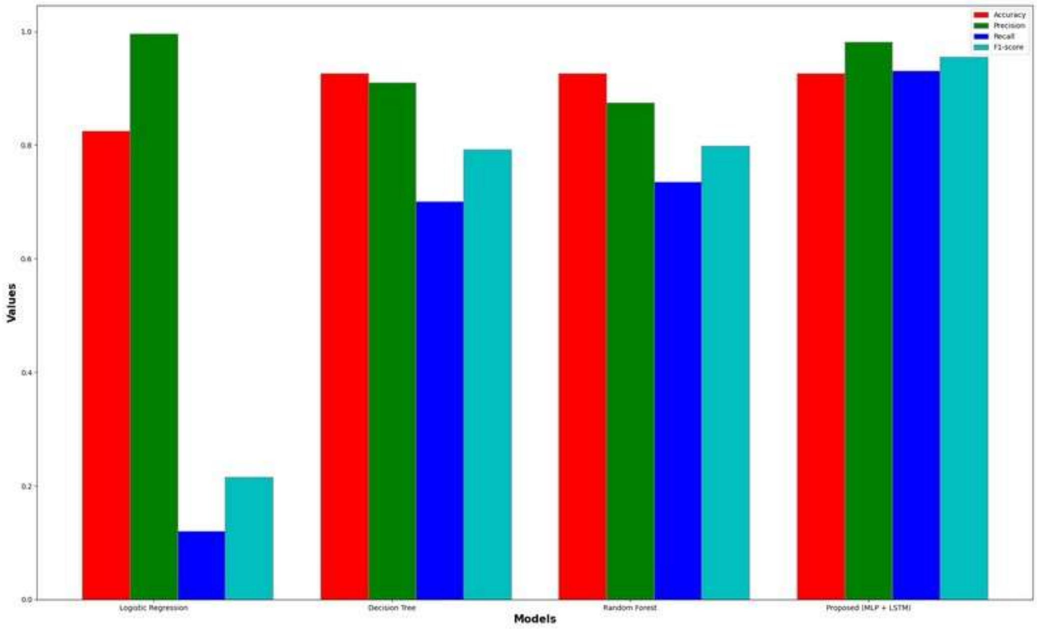


**Figure 9. Comparison graph of proposed method with other method.**

## 6. CONCLUSION AND FUTURE WORK

This essay provides a solid template for performing accurate and scalable big data analytics. We suggested combining Apache Spark and deep learning (DL) into the machine learning (ML) system as part of our plan. In comparison to current state-of-the-art methods, this mix produced much better outcomes. Based on a two-stage cascade design, our frameworks were able to solve multiclass and binary classification problems correctly. Coupled with limited computer power, the two-step framework for big data analysis suggested could handle a huge amount of data chores in a short amount of time. Specifically, it was able to do this by reducing the amount of computational complexity involved and considerably improving the accuracy of the datasets for cardiac arrhythmia and URL Reputation. Our future plans include expanding the framework by including other elements and parameters to address various real-world research challenges using strong deep architectures.

## REFERENCES

[1]. Nair, Lekha R., Sujala D. Shetty, and Siddhanth D. Shetty. "Applying spark-based machine learning model on streaming big data for health status prediction." Computers & Electrical Engineering 65 (2018): 393-399.

[2]. Hbibi, Lamyae, and Hafid Barka. "Big data: Framework and issues." In 2016 International Conference on Electrical and Information Technologies (ICEIT), pp. 485-490. IEEE, 2016.

[3]. Elgendy, Nada, and Ahmed Elragal. "Big data analytics: a literature review paper." In Advances in Data Mining. Applications and Theoretical Aspects: 14th Industrial Conference, ICDM 2014, St. Petersburg, Russia, July 16-20, 2014. Proceedings 14, pp. 214-227. Springer International Publishing, 2014.

[4]. Sun, Zhaohao, Lizhe Sun, and Kenneth Strang. "Big data analytics services for enhancing business intelligence." Journal of Computer Information Systems 58, no. 2 (2018): 162-169

[5]. Debortoli, Stefan, Oliver Müller, and Jan vom Brocke. "Comparing business intelligence and big data skills: A text mining study using job advertisements." Wirtschaftsinformatik 56 (2014): 315-328

[6]. Acharjya, Debi Prasanna, and Kauser Ahmed. "A survey on big data analytics: challenges, open research issues and tools." International Journal of Advanced Computer Science and Applications 7, no. 2 (2016): 511-518.

[7]. Divya, K. Sree, Peyakunta Bhargavi, and S. Jyothi. "Machine learning algorithms in big data analytics." Int. J. Comput. Sci. Eng 6, no. 1 (2018): 63-70.

[8]. Mittal, Shweta, and Om Prakash Sangwan. "Big data analytics using machine learning techniques." In 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), pp. 203-207. IEEE, 2019.

[9]. Kashyap, Ramgopal. "Big Data Analytics challenges and solutions." In Big Data Analytics for Intelligent Healthcare Management, pp. 19-41. Academic Press, 2019.

[10]. Nambiar, Raghunath, Ruchie Bhardwaj, Adhiraaj Sethi, and Rajesh Vargheese. "A look at challenges and opportunities of big data analytics in healthcare." In 2013 IEEE international conference on Big Data, pp. 17-22. IEEE, 2013

[11]. Maroufkhani, Parisa, Mohammad Iranmanesh, and Morteza Ghobakhloo. "Determinants of big ata analytics adoption in small and medium-sized enterprises (SMEs)." Industrial Management & Data Systems 123, no. 1 (2023): 278-301

[12]. VenkateswaraRao, M., SaiSrinivas Vellela, Venkateswara Reddy, Nagagopiraju Vullam, Khader Basha Sk, and D. Roja. "Credit Investigation and Comprehensive Risk Management System based Big Data Analytics in Commercial Banking." In 2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS), vol. 1, pp. 2387-2391. IEEE, 2023

[13]. Qi, Quansong, Zhiyong Xu, and Pratibha Rani. "Big data analytics challenges to implementing the intelligent Industrial Internet of Things (IIoT) systems in sustainable manufacturing operations." Technological Forecasting and Social Change 190 (2023): 122401

[14]. Kanan, Tarek, Ala Mughaid, Riyad Al-Shalabi, Mahmoud Al-Ayyoub, Mohammed Elbes, andOdai Sadaqa. "Business intelligence using deep learning techniques for social media contents." Cluster Computing 26, no. 2 (2023): 1285-1296

[15]. Li, Xiaoming, Hao Liu, Weixi Wang, Ye Zheng, Haibin Lv, and Zhihan Lv. "Big data analysis of the internet of things in the digital twins of smart city based on deep learning." Future Generation Computer Systems 128 (2022): 167-177

[16]. Awan, Mazhar Javed, Muhammad Asad Khan, Zain Khalid Ansari, Awais Yasin, and Hafiz Muhammad Faisal Shehzad. "Fake profile recognition using big data analytics in social media platforms." International Journal of Computer Applications in Technology 68, no. 3 (2022): 215-222

[17]. Zhang, Yongchang, Panpan Geng, C. B. Sivaparthipan, and Bala Anand Muthu. "Big data and artificial intelligence based early risk warning system of fire hazard for smart cities." Sustainable Energy Technologies and Assessments 45 (2021): 100986

[18]. Behera, Ranjan Kumar, Monalisa Jena, Santanu Kumar Rath, and Sanjay Misra. "Co-LSTM: Convolutional LSTM model for sentiment analysis in social big data." Information Processing & Management 58, no. 1 (2021): 102435

[19]. Lopez-Martin, Manuel, Belen Carro, and Antonio Sanchez-Esguevillas. "Application of deep reinforcement learning to intrusion detection for supervised problems." Expert Systems with Applications 141 (2020): 112963

[20]. Jiang, Peng, Yuehan Chen, Bin Liu, Dongjian He, and Chunquan Liang. "Real-time detection of apple leaf diseases using deep learning approach based on improved convolutional neural networks." IEEE Access 7 (2019): 59069-59080

[21]. Thamilarasu, Geethapriya, and Shiven Chawla. "Towards deep-learning-driven intrusion detection for the internet of things." Sensors 19, no. 9 (2019): 1977

[22]. Hosmer Jr, David W., Stanley Lemeshow, and Rodney X. Sturdivant. Applied logistic regression. Vol. 398. John Wiley & Sons, 2013.

[23]. Kirkos, Efstathios, Charalambos Spathis, and Yannis Manolopoulos. "Data mining techniques for the detection of fraudulent financial statements." Expert systems with applications 32, no. 4 (2007): 995-1003

[24]. Patil, Suraj, Varsha Nemade, and Piyush Kumar Soni. "Predictive modelling for credit card fraud detection using data analytics." Procedia computer science 132 (2018): 385-395.

[25]. Yeh, I-Cheng, and Che-hui Lien. "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients." Expert Systems with Applications 36, no. 2 (2009): 2473-2480

## Biography of Authors

**Mr. Mohammad Islam**, born in 1983 in Jodhpur, Rajasthan India, is a research scholar at Maulana Azad University, Jodhpur, Rajasthan, India. He holds a master's degree in Master in Computer Applications, completed in 2012. With over 05 years' experience in the IT industry and 11 years of experience in the Teaching, he currently serves as an Associate Professor in Department of Computer Science, Maulana Azad University, Jodhpur. His expertise includes C, C++, JAVA, Artificial intelligence, Machine learning, Big Data, Data Science, Python, PHP as well as automation and IoT. He can be contacted at email: islamjodhpur@gmail.com

**Dr. Ashish Sharma** btained his Ph.D. Computer Science & Engineering in the year 2016 and M.Tech. (Gold Medalist) Computer Science & Engineering in the year 2012. His area of interest includes programming languages such as C, C++, JAVA, and Design & Analysis of algorithms, cryptography & cloud computing. Apart from this he has published various papers in national and international journals. He has also attended various national and international conferences. He has more than 18 years' experience of teaching and research. Many research scholars have completed their research work under his supervision. He can be contacted at email: aashishid@gmail.com

**Mr.Zahid Ahmed** is an accomplished Assistant Professor in Computer Science and Application with a strong academic background and deep expertise in Artificial Intelligence (AI). With a rich professional experience in both teaching and research, Zahid has developed a comprehensive skill set in AI technologies, applications, and theoretical aspects. He can be contacted at email: zahidahmed59@gmail.com

**Ms. Nausheen Khilji** received her B.Tech and M.Tech degrees in Computer Science and Engineering with Honors. She has over seven years of experience as an Assistant Professor in the domain of Computer Science and Engineering. Her research interests include Big Data, Cloud Computing, and Image Processing. She has published research work in reputable journals and has participated in various national and international conferences. In addition to her academic contributions, she has also mentored foreign batch students, demonstrating her commitment to inclusive and global education. She can be contacted at email: naushy90@gmail.com