

# A study on high dimensional big data using predictive data analytics model

Nivethitha Krishnadoss, Lokesh Kumar Ramasamy

School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India

## Article Info

### Article history:

Received Jul 6, 2022

Revised Nov 8, 2022

Accepted Nov 18, 2022

### Keywords:

Dimension reduction

Hyperparameter optimization

Machine learning

Predictive big data analytics

Splitting random forest

## ABSTRACT

A massive bulk of data is being created due to digitalisation in various industries, including medical, manufacturing, sales, internet of things (IoT) devices, the web, and businesses. To find data patterns for data attributes machine learning (ML) algorithms are used. In this fast-growing world, we can see that data is generated in abundance by people, machines, and corporations. With the increase in computer science market, researchers are integrating heterogeneous and diverse data into accurate patterns by applying machine learning algorithms and complex strategies on data sets. The overabundance of high-dimensional big data has made it more difficult for scientists to extract important information from these data efficiently. Conventional data mining approaches are ineffective when dealing with large amounts of data. As big data increase exponentially, predictive analytics has become widely known. To evaluate a large number of data patterns, data driven technology predictive big data analytics (PBA) can be used and ML algorithms to investigate the present and future data based on the records of data patterns. In this research paper, predictive analysis on big data has been proposed using the splitting random forest (SRF) methodology with help of hyperparameter optimization and dimension reduction technique.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



## Corresponding Author:

Lokesh Kumar Ramasamy

School of Computer Science and Engineering, Vellore Institute of Technology

Vellore, Tamilnadu, India

Email: lokeshkumar.r@vit.ac.in

## 1. INTRODUCTION

The exponential rise of internet community, digital assistants, sensing devices, and the internet of things (IoT) has accompanied a major increase in data. As these are increasing tremendously which gives rise to big data. According to a Gartner research from 2017, the volume of data had increased dramatically from 17.3 billion in 2016 to 21.9 billion in 2021. Compared to conventional data, big data refers to excessive data growth in heterogeneous formats. Because it is so large and complex that standard data processing tools cannot cope with applications of big data [1], [2].

The task becomes challenging when data volume, variety, processing, and utilizing grows, to deal with such a challenging environment, many techniques came into existence. Big data refers to significant data expansion in a variety of formats. Big data analytics studies show massive and diverse data sets from roots to identify information such as covered patterns and unknown relationships, so better choices can be made. For storing data and processing of data a scalable architecture is needed. Big data refers to significant data expansion comprising various formats and is extremely large. As we know, the data size is so large that it cannot be processed in simple computational manual methods. Data processing is to be done based on the volume, velocity, variety [3]-[5].

The analysis of big data is established on: i) to speculate the outcomes and building models the statistical algorithm is used, ii) to figure out data patterns and their correlation data mining is used and iii) machine learning (ML) to solve the complexity of new models and new data.

Textual data and speech recognition are used to examine the free-form text and spoken language. Figure 1 illustrates how the four types of big data analytics tools are grouped to give detailed information related to the field. The types are as:

- Descriptive analytics is used to answer "What Happened?" from historical data. It helps find data trends and is utilized in business intelligence. Visualization is done through pie charts, tables, and graphs.
- Diagnostic analytics addresses "Why?" using drill down, data discovery, data mining, and correlations. Data mining is used to extract information from unstructured data. Attempts to find systematic relationships in data help us determine the goal.
- Predictive analytics tells an organization "What is likely to happen?" Once the corporation knows the above two analytics models, predictive analytics helps gather data to check what has transpired. It uses regression analysis, and pattern matching. To complete the assignments, you must know statistics and programming.
- Prescriptive analytics teaches us "what to do" and is an advanced level. Techniques are used to analyze graphs, simulation results, complicated events, neural networks, and machine learning. Data architecture and implementation are needed for good data quality.

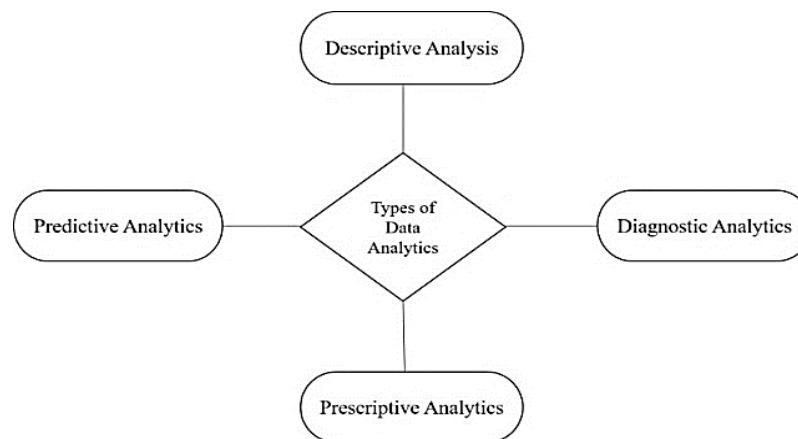


Figure 1. Types of data analytics [6]

Figure 2 shows the primary domains where data analysis is needed and where big data analytics can be employed, including healthcare, transport, E-commerce, banking and finance, and manufacturing. All fields have advanced greatly due to such analytics. Predictive big data analytics (PBA) helps solve large-scale data with hidden patterns, uncover opportunities and predictable outcomes, and act as data-driven technology. PBA uses machine learning techniques to forecast future occurrences by analyzing current and historical data. According to predictive analytics, machine learning is an intelligent tool for business that helps extract useful insights from enormous datasets for pioneering attempts. Traditional methodologies confront significant hurdles and become computationally impractical when it comes to huge data. However, when the data size expands, the algorithm's speed becomes hard to change the data across the system's processing units. Some efficient statistically machine learning methods are necessary to cope with large data while requiring minimal resources such as memory [7]-[10].

There are many issues related to the PBA system as it helps in extracting a large amount of knowledge with huge sample sizes and the problems arise when high dimensionality combined with its high computational cost and algorithmic instability and considered as the drawbacks of the PBA system [4]. One of its solutions is by increasing the size of big data that help in offering performance efficient outcome. Many efforts are being made to overcome these shortcomings, identify the accurate solution to such problems, and make a way out of the dimensions of datasets with the same virtue of data. One of the operations done on big data is to decrease the number of characteristics in data sets by maintaining main variables using a dimension reduction technique. When dimension reduction techniques are used, first data is fed to the machine learning prototype. In the PBA system, overcoming these challenges might be a significant exercise that must be undertaken. Effective predictive analytics necessitates a fast model design and the development of reliable prediction models. For a better understanding of high dimensional big data, researchers have the main goal

for developing a better PBA system. Machine learning regression algorithm is the fundamental strategy that helps make correct decision-making processes. In the PBA system, splitting random forest (SRF) regression is known for the machine learning algorithm.

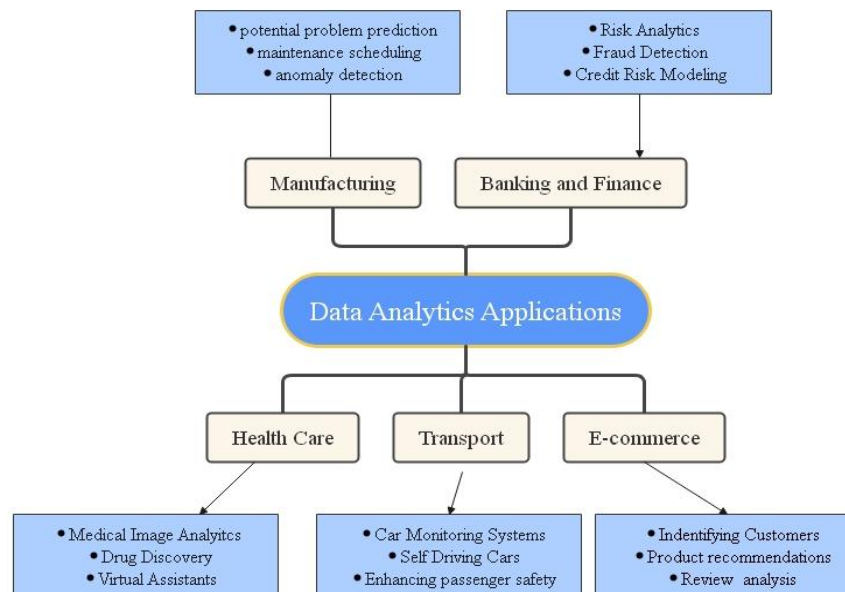


Figure 2. Applications of data analytics [11]

This study makes efforts for the PBA system to work onto the two major issues. First, with SRF, parameterization is substantially connected with prediction scores but is narrowly engrossed on ideal parameters. Accordingly, selecting an ideal collection of hyperparameters necessitates using an efficient model selection procedure. To fulfill behavior and trends from high dimensional large data SRF [4], [12], [13] prediction models are built. SRF optimizes hyperparameters intending to obtain adequate generalization performance. The suggested method will be able to track all hyperparameter combinations in SRF as their associated prediction scores and make a trade-off between predictive power and computing time [5], [14], [15].

One more is that emergence data, it comprises a huge number of complicated and high-dimensional characteristics. There is a greater impact on predictive power with some characteristics of data. Traits of data that have no bearing on prediction accuracy must be removed. Overhead processing time can be decreased with the dimension reduction [16] approach in developing random forest (RF) models. To assess the system's performance, two different methodologies, principal component analysis (PCA), and information gain (IG), can be studied, which are different dimension reduction approaches. The system data nature is then recognized, and a more advantageous approach is chosen.

The main goal of the big data approach is to handle data in a way that is valuable to the enterprise. Otherwise, the expense of storing and maintaining data outweighs the value of processing it. The most difficult aspects of big data analysis are successfully processing the data to obtain useful information and using the processed data for decision-making. Many technologies are available on the market for analyzing and managing big data. We must select both practical and effective instruments for the research endeavor. The new 5V model is used to define big data. This model is built on the fundamental volume, velocity, and variety (3V) paradigm. The quality of big data is connected to value and veracity. Data storage, data processing, data quality are relevant, the main matter lies in privacy, security, and scalability of data.

## 2. LITERATURE REVIEW

PBA systems are concerned with data complexities, variability, privacy, portability, and voluminous data. Furthermore, high-dimensional data management and computational optimization issues have recently risen in popularity. This research paper identifies one of the PBA systems with high-dimensional data and prediction. Researchers have described various developments in their previous work for boosting PBA system execution, presenting the system's design and prediction performance. Many traditional analytics

approaches, have difficulty learning on large-scale, complicated, and varied data. Predictive analytics system for data utilized in the metalwork sector to overcome the large-scale data analytics challenge [4], [17], [18]. To anticipate the long-term performance of power consumption, they employed the backpropagation neural networks technique. Their proposed approach provided a useful pattern of data in an unknown data correlation, and the learning parameters must be maintained so that the prediction score does not suffer. To recognize extreme occurrences, Shenoy and Gorinevsky [19] used Bayesian formulation which is nonparametric for the PBA system, Martino *et al.* [20] and Zhang *et al.* [21] used the PCA methodology for ensemble learning using Apache Spark this research has compared the performance of various big data machine learning models but there were large data sets with a variety of characteristics within the data so the analysis was not executed properly.

A spark-based parallel random forest (PRF) technique was introduced by Chen *et al.* [22]. They used a hybrid strategy to improve the suggested PRF to overcome the challenge of high-dimensional data. However, depending on the type of datasets, PCA and IG techniques significantly impact the PBA system's prediction performance while doing dimension reduction [7], [22], [23]. It is necessary to investigate the performance efficiencies of various dimensionality reduction approaches for the PBA system.

Along with commercial well-being, the digital age brings challenges and concerns. Massive data as in Table 1 comparison dimensions influence the PBA system's machine learning efficiency. High dimensional data are difficult to deal with existing well-known learning algorithms. Based on SRF, a systematic PBA system is proposed [8], [24]. To handle complicated problems high dimensional data is needed to get in presentable manner, the proposed PBA system enhance the tree-based approach. Enhancing the tree-based technique with help of hyperparameters optimization and dimension reduction technique. In this proposed model there is one source from where the data is input by various means like social media, mobile apps, sensing devices, IoT, and many others. Then this raw data is passed through the data pre-processing unit where cleaning generation and selection is done based on the type of data then data is sent to the model generation and then the data undergoes the process of hyperparameter optimization and dimension reduction techniques and then at the final stage the predictions are made.

Table 1. Proposed predictive big data analytics model

Author [Year]	Technique and methodology used for Research	Challenges faced in Research
Hernandez and Zhang [1]	Conventional Analysis technique.	Needs to perform for a large scale.
Shenoy and Gorinevsky [19]	Bayesian formulation for PBA system.	Solves the problem of excellent performance but needs to tackle PBA system with high dimensional data.
Zhang and Yang [21]	PCA based dimension reduction technique to reduce the dimension of big data.	Focused on performance results with accuracy and processing time.
Shin <i>et al.</i> [25]	Predictive analysis used for metal cutting for back propagation neural networks.	Parameters were not managed because of unknown data correlation.
Ntaliakouraset <i>al.</i> [26]	Decision tree algorithm with pre spark.	Accuracy was not maintained for forecasting of tourism demand.
LakshmiPadmaja <i>et al.</i> [27]	Random subset feature selection (RSFS).	Needs advancement in RSFS for handling high dimensional data.

Data storage receives massive volumes of numeric data and links with computer server nodes for quick processing. Hadoop distributed file system (HDFS) is used to offer fault-tolerant manageable storage. When HDFS receives large amounts of data, it divides it into discrete chunks into multi-user computer nodes in a cluster. Data storage is intended to be cost-effective and scalable. Furthermore, it has been deliberately designed to be very fault-tolerant. At the serves, replicate data is sent and distributed to them. As an outcome, crush data on nodes can be discovered on other nodes in a cluster. While the data is being retrieved, the processing part continues in this process [9]. The data processing unit is the most important component of the suggested system for obtaining superior computing infrastructure and therefore mining and analyzing enormous amounts of data in a timely and effective way.

Data analytics is a critical component of the suggested approach for achieving high predicted accuracy. It is divided into two stages: data pre-processing and system prediction model construction. Because big data can have inaccurate and redundant data, a data cleaning phase is conducted to eliminate or minimize noise using smoothing techniques. It eliminates outliers and corrects irregularities with missing value treatments. The standard normalizing approach is used for data processing and reduction. It contributes to a more intelligible pattern for prediction. SRF can be employed in the prediction model building phase to enable correct decision-making for the developed framework. The SRF predictor is a well-known indicator for the PBA system [10].

### 3. HYPER-PARAMETERS OPTIMIZATION

Hyperparameters are tunable parameters that must be fine-tuned to attain exceptional performance from a model. The reliability of the model depends on hyper-parameter optimization. The process of discovering the most optimal hyperparameter is called hyperparameter optimization. Every machine learning system contains hyperparameters, and the most fundamental goal in automated machine learning (AutoML) is to adjust these hyperparameters automatically to improve performance. Recent deep neural networks, in particular, rely heavily on a wide variety of hyperparameter options for the neural network's construction, regularization, and optimization. To analyze massive data efficiently, SRF needs the ideal value of hyperparameters [14], [28].

For some predictions, the default values can be used as they do not fulfil the requirement of data sets. By optimizing the parameters SRF can be improved as these parameters setting may be tweaked before training for maximum performance. If we take SRF as a predictive tool for optimizing hyperparameters then the amount of work done will be reduced. Researchers have found many conceivable hyperparameter combinations, and an individual SRF prediction model has been created for each pair. SRF creates a forest of trees by optimizing two hyperparameters: the tree's size and the tree's maximum depth. The number of trees (NumTrees) governs the computation cost and prediction model presentation [29], [30].

The maximum depth of the tree controls the depth of each tree with an exponential increase in time. Fundamentally, the model building procedures are built utilizing the greedy method to discover the optimal combination of hyperparameters and discard models. All possible hyperparameter combination sets can be produced and run to pick the ideal parameters for developing the best model for each dataset. To develop the decision model, we can employ the RF model creation from Spark MLlib and the relevant data pipelines and use a relevant model.

#### 3.1. The number of trees in the SRF

The ideal number of trees governs the cost of computations and execution of the prediction model. Using both more and fewer trees may cause issues so choosing the appropriate number of trees is tricky one. Splitting in SRF is referred to as random number of features for each tree. This study has a variable number of trees at exponential rates in base two i.e.,  $L=2^j$ ,  $j=1,2,\dots, 11$ , SRF is built and tested.

#### 3.2. SRF maximum depth

The maximum depth of the tree determines the depth of each tree in the forest, and the running time grows exponentially with tree depth. It decreases the complexity of learning models and the risk of overfitting. Overfitting occurs when there is too much depth. RF, on the other hand, overcomes this problem, and proper tree depth can give good performance for error reduction.

#### 3.3. Dimension reduction

The dataset's dimension point to the number of characteristics shown in the datasets. To eliminate processing time overhead during the model construction phase, some characteristics that do not affect the model are noticed and subsequently decreased utilizing dimension reduction techniques. The amount of characteristics given in a data collection is referred to as its dimensionality. Reducing data dimensionality has become a significant problem undertaking effective analysis in a distributed context. The combination of hyperparameter tuning and dimension reduction techniques can greatly improve the model's prediction performance [13], [31], [32]. Dimension reduction has emerged as a critical problem for achieving efficient analytics in a distributed context. There are already a variety of machine learning approaches that take feature significance into account. This study compares two common feature reduction approaches, principal component analysis and information gain, to validate the efficiency of dimension reduction strategies. This research provides a high-dimensional big data predictive analysis method based on enhancing scalable random forest (ESRF), which is utilized to analyze high-dimensional vast data, to further increase classification accuracy and stability. The combination of parameter optimization and dimensionality reduction dramatically increases the system's prediction performance. To avoid the processing time overhead of the model creation stage, the system needs to adopt a greedy technique to determine the optimal hyperparameter combination of SRF, which helps predict trends and behavioral patterns from high-dimensional big data and reduce data sets using PCA and IG technologies [33], [34]. The experimental findings suggest that the PBA system described in this study may show high predictive ability and an effective performance with the shortest processing time is the complete experimental data set.

##### 3.3.1. Dimension reduction with PCA

This approach works by putting vast data into a subspace where changes may be detected, reducing the data's huge dimensionality. To find k-dimensions and represent them into a new set of variables PCA is

used, known as principle components [35]. The first component gives the largest possible variance of all data points and is important.

### 3.3.2. Dimension reduction using IG

IG is an approach that uses feature selection that helps minimise datasets' dimensions by determining the original data relevance. The primary concept behind this method is to sort the characteristics of each feature variable by computing the gain ratio value. The remaining ones are chosen from additional variables whereas the principal variable is the topmost feature variable. The greatest amount may be simply selected using the IG. To avoid overfitting, the feature variable with the highest value is replaced with the gain ratio value [33].

## 4. RESULTS AND ANALYSIS

In the previous work, it was seen that PBA systems face many challenges in the area of data complexity, heterogeneity, privacy, maintaining a large volume of data, with this, there are problems of arranging and managing high dimensional data, and also there are issues of data computational. This proposed system of PBA can deal with both high dimensional and can also help in predictive analysis. In this proposed system, we can examine with big data analytics platform with one master node and three other nodes and can use any processor with 8 GB memory for an individual node. Five real-world datasets can be taken for examination like a credit card (U.C.I. means data repository of machine learning databases), and progressive web apps (PWA) like high-performance computing center north.

It is proposed that for prediction accuracy mean absolute error (MAE) and root mean square error (RMSE) can be used as evaluating metrics in the Table 2. Maybe after performing hyperparameter optimization on datasets, MAE. results could be better than the default parameter. Suppose if MAE with default parameter for credit card data is nearly around 0.1019 then after hyperparameter optimization, it can be near around 0.0017.

Table 2. Mean absolute error (MAE) comparison table for ESRF with dimension reduction

Datasets	ESRF	ESRF with DR_PCA	ESRF with DR_IG
Database autonomy service (DAS)	2.1300	1.0899	1.9999
Susy	0.3400	0.3399	0.3199
High performance computing (HPC)	1.0890	1.0299	0.8199
Knowledge discovery in databases (KDD)	0.7876	0.7855	0.6699
Credit-Card	0.1019	0.1030	0.0022

The effectiveness of the PBA system affects by excessive data dimension so it is proposed that reduction techniques can be applied to reduce computational time and get the lowest MAE values for the datasets in Table 3. The lowest MAE and RMSE values of effect splitting random forest (ESRF) with DR\_PCA for DAS datasets are 1.0907 and 2.0950 shown in Figure 3. The maximum values are 2.256 and 3.0465. If the MAE value for DSA dataset is 2.1300 then after applying the dimension reduction technique we can get around 1.0899 we can see how innovative researchers are trying to improve the accuracy of the PBA system by using different types of techniques and methods.

Table 3. RMSE comparison of ESRF with dimension reduction

DATASETS	ESRF	ESRF WITH DR_PCA	ESRF WITH DR_IG
DAS	3.7995	2.0950	3.0879
SUSY	0.3940	0.3989	0.3199
HPC	1.5467	1.0650	0.8939
KDD	0.7779	0.7850	0.6699
CREDIT-CARD	0.1019	0.0350	0.0249

In this paper, the high dimensional big data and dimension reduction are proposed for an effective PBA system by using the SRF model for decision tree. The Figure 4 illustrates the tradeoff between the datasets and processing time. We need to maintain accuracy and efficiency for its design and implementation. In the DAS dataset, RF spends 129 seconds training and ESRF spends 71 seconds predicting the number of processors for upcoming workload traces.

In the HPC2N dataset, ESRF forecasts that the required processors would distribute resources efficiently in 64 seconds, whereas RF requires 121 seconds. In the Susy dataset, RF takes 139 seconds to

anticipate the signal process, but ESRF takes just 76 seconds. In a credit-card dataset, ESRF can identify whether a card is counterfeit or real in 100 seconds, whereas RF takes 154 seconds. In the KDD dataset, ESRF outperforms RF in terms of processing speed. The suggested PBA system employs ESRF to transform a vast amount of disparate data into timely insights for speedier decision making.

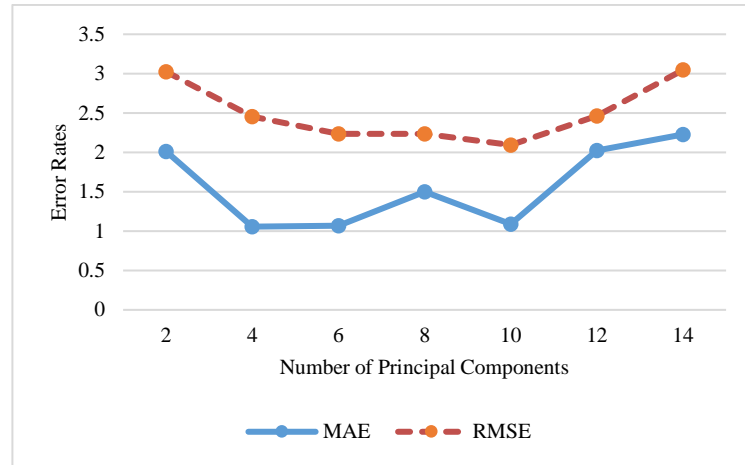


Figure 3. Datasets error rate chart

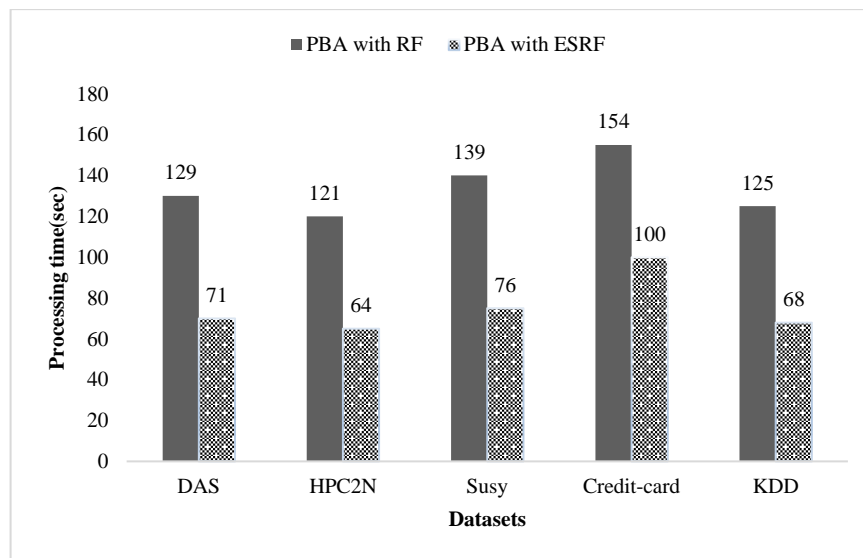


Figure 4. Comparison of RF and SRF

## 5. CONCLUSION

Data scientists' primary focus is on the growth of big data which are high dimensional from numerous data sources. Predictive big data analytics is crucial for extracting business knowledge from this data and forecasting results. To develop an effective and scalable PBA system capable of handling high-dimensional large data this is critical. This study suggests a PBA system based on ESRF to deal with large amounts of high-dimensional data. It can also help in eliminating the time of processing in model construction. Using dimension reduction techniques to decrease irrelevant feature variables in datasets, ESRF with DR PCA and DR IG approaches can be utilized. In the DAS dataset, the ESRF prediction models with DR PCA achieve strong prediction scores (MAE 1.091 and RMSE 2.095) and reduce execution time from 129 to 69 seconds. In the credit-card dataset, the suggested PBA system may deliver good prediction performance while minimizing processing time. In summary, an appropriate technique for determining optimum hyperparameters can be established. The two most commonly used dimension reduction approaches

for the PBA system can be used. As a result of the predictive analytics data nature, the advantage of information gain theory can be used for dimension reduction of the suggested system. The key conclusion of this work is that optimizing hyperparameters in SRF in conjunction with reduction techniques may considerably improve the system's prediction performance. To get accurate outcomes, the proposed PBA system can outperform the findings.

## REFERENCES





- [1] I. Hernandez and Y. Zhang, "Using predictive analytics and big data to optimize pharmaceutical outcomes," *American Journal of Health-System Pharmacy*, vol. 74, no. 18, pp. 1494–1500, Sep. 2017, doi: 10.2146/ajhp161011.
- [2] E. J. De Fortuny, D. Martens, and F. Provost, "Predictive modeling with big data: Is bigger really better?," *Big Data*, vol. 1, no. 4, pp. 215–226, Dec. 2013, doi: 10.1089/big.2013.0037.
- [3] J. Z. Huang, W. Huang, and J. Ni, "Predicting bitcoin returns using high-dimensional technical indicators," *Journal of Finance and Data Science*, vol. 5, no. 3, pp. 140–155, Sep. 2019, doi: 10.1016/j.jfds.2018.10.001.
- [4] Y. Zhang *et al.*, "A predictive data feature exploration-based air quality prediction approach," *IEEE Access*, vol. 7, pp. 30732–30743, 2019, doi: 10.1109/ACCESS.2019.2897754.
- [5] H. Bastani, "Predicting with proxies: Transfer learning in high dimension," *Management Science*, vol. 67, no. 5, pp. 2964–2984, May 2021, doi: 10.1287/mnsc.2020.3729.
- [6] P. S. Deshpande, S. C. Sharma, and S. K. Peddoju, "Predictive and prescriptive analytics in big-data era," in *Studies in Big Data*, vol. 52, 2019, pp. 71–81.
- [7] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *International Journal of Information Management*, vol. 35, no. 2, pp. 137–144, Apr. 2015, doi: 10.1016/j.ijinfomgt.2014.10.007.
- [8] E. T. Bradlow, M. Gangwar, P. Kopalle, and S. Voleti, "The role of big data and predictive analytics in retailing," *Journal of Retailing*, vol. 93, no. 1, pp. 79–95, Mar. 2017, doi: 10.1016/j.jretai.2016.12.004.
- [9] M. Z. H. Jesmeen *et al.*, "A survey on cleaning dirty data using machine learning paradigm for big data analytics," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 10, no. 3, pp. 1234–1243, Jun. 2018, doi: 10.11591/ijeecs.v10.i3.pp1234-1243.
- [10] M. Seyedan and F. Mafakheri, "Predictive big data analytics for supply chain demand forecasting: methods, applications, and research opportunities," *Journal of Big Data*, vol. 7, no. 1, p. 53, Dec. 2020, doi: 10.1186/s40537-020-00329-2.
- [11] K. Vassakis, E. Petrakis, and I. Kopanakis, "Big data analytics: applications, prospects and challenges," in *Lecture Notes on Data Engineering and Communications Technologies*, vol. 10, 2018, pp. 3–20, doi: 10.1007/978-3-319-67925-9\_1.
- [12] J. Chen *et al.*, "A Parallel random forest algorithm for big data in a spark cloud computing environment," *IEEE Transactions on Parallel and Distributed Systems*, vol. 28, no. 4, pp. 919–933, Apr. 2017, doi: 10.1109/TPDS.2016.2603511.
- [13] Y. Perwej, E. Bhuvanawari, S. Kumar, V. Arulkumar, and P. Nancy, "Unsupervised feature learning for text pattern analysis with emotional data collection: a novel system for big data analytics," in *2022 International Conference on Advanced Computing Technologies and Applications (ICACTA)*, Mar. 2022, pp. 1–6, doi: 10.1109/ICACTA54488.2022.9753501.
- [14] M. J. Sousa, A. M. Pesqueira, C. Lemos, M. Sousa, and Á. Rocha, "Decision-making based on big data analytics for people management in healthcare organizations," *Journal of Medical Systems*, vol. 43, no. 9, p. 290, Sep. 2019, doi: 10.1007/s10916-019-1419-x.
- [15] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, 2005, vol. 1, pp. 886–893, doi: 10.1109/CVPR.2005.177.
- [16] A. C. Wilkerson, H. Chintakunta, and H. Krim, "Computing persistent features in big data: A distributed dimension reduction approach," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 11–15, doi: 10.1109/ICASSP.2014.6853548.
- [17] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013, doi: 10.1109/TPAMI.2013.50.
- [18] V. Arulkumar, S. Sridhar, G. Kalpana, and K. S. Guruprakash, "Real-time big data analytics for improving sales in the retail industry via the use of internet of things beacons," in *Lecture Notes in Networks and Systems*, vol. 444, 2022, pp. 111–126.
- [19] S. Shenoy and D. Gorinevsky, "Predictive analytics for extreme events in big data," *IEEE First International Conference on Big Data Computing Service and Applications*, pp. 184–193, 2015, doi: 10.1109/BigDataService.2015.66.
- [20] B. Di Martino, R. Aversa, G. Cretella, A. Esposito, and J. Kołodziej, "Big data (lost) in the cloud," *International Journal of Big Data Intelligence*, vol. 1, no. 1/2, p. 3, 2014, doi: 10.1504/ijbdi.2014.063840.
- [21] T. Zhang and B. Yang, "Big data dimension reduction using PCA," *IEEE International Conference on Smart Cloud (SmartCloud)*, pp. 152–157, 2016, doi: 10.1109/SmartCloud.2016.33.
- [22] J. Chen *et al.*, "Big data challenge: a data management perspective," *Frontiers of Computer Science*, vol. 7, no. 2, pp. 157–164, Apr. 2013, doi: 10.1007/s11704-013-3903-7.
- [23] B. H. Brinkmann, M. R. Bower, K. A. Stengel, G. A. Worrell, and M. Stead, "Large-scale electrophysiology: acquisition, compression, encryption, and storage of big data," *Journal of Neuroscience Methods*, vol. 180, no. 1, pp. 185–192, May 2009, doi: 10.1016/j.jneumeth.2009.03.022.
- [24] H. Zou, Y. Yu, W. Tang, and H. M. Chen, "Improving I/O Performance with adaptive data compression for big data applications," in *2014 IEEE International Parallel & Distributed Processing Symposium Workshops*, May 2014, pp. 1228–1237, doi: 10.1109/IPDPSW.2014.138.
- [25] Seung-Jun Shin, Jungyub Woo, Sudarsan Rachuri, "Predictive analytics model for power consumption in manufacturing," *Procedia CIRP*, vol. 15, pp. 153–158, 2014, doi: 10.1016/j.procir.2014.06.036.
- [26] N. Ntaliakouras, G. Vonitsanos, A. Kanavos and E. Dritsas, "An apache spark methodology for forecasting tourism demand in greece," *10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, 2019, pp. 1-5, doi: 10.1109/IISA.2019.8900739.
- [27] D. Lakshmi padmaja and B. Vishnuvardhan, "Classification performance improvement using random subset feature selection algorithm for data mining," *Big Data Research*, vol. 12, pp. 1-12, 2018, doi: 10.1016/j.bdr.2018.02.007.
- [28] S. Lakshminarasimhan *et al.*, "Compressing the incompressible with ISABELA: In-situ reduction of spatio-temporal data," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6852 LNCS, no. PART 1, 2011, pp. 366–379, doi: 10.1007/978-3-642-23400-2\_34.







- [29] J. P. Ahrens, J. Woodring, D. E. DeMarle, J. Patchett, and M. Maltrud, "Interactive remote large-scale data visualization via prioritized multi-resolution streaming," in *Proceedings of the 2009 Workshop on Ultrascale Visualization, UltraVis '09*, 2009, pp. 1–10, doi: 10.1145/1838544.1838545.
- [30] C. Bi, K. Ono, K. L. Ma, H. Wu, and T. Imamura, "Proper orthogonal decomposition based parallel compression for visualizing big data on the K computer," in *IEEE Symposium on Large Data Analysis and Visualization 2013, LDAV 2013 - Proceedings*, Oct. 2013, pp. 121–122, doi: 10.1109/LDAV.2013.6675169.
- [31] H. Zou, Y. Yu, W. Tang, and H. W. M. Chen, "FlexAnalytics: a flexible data analytics framework for big data applications with I/O performance improvement," *Big Data Research*, vol. 1, pp. 4–13, Aug. 2014, doi: 10.1016/j.bdr.2014.07.001.
- [32] K. Ackermann and S. D. Angus, "A resource efficient big data analysis method for the social sciences: the case of global IP activity," *Procedia Computer Science*, vol. 29, pp. 2360–2369, 2014, doi: 10.1016/j.procs.2014.05.220.
- [33] C. Yang *et al.*, "A spatiotemporal compression based approach for efficient big data processing on Cloud," *Journal of Computer and System Sciences*, vol. 80, no. 8, pp. 1563–1583, Dec. 2014, doi: 10.1016/j.jcss.2014.04.022.
- [34] A. Monreale *et al.*, "Privacy-preserving distributed movement data aggregation," in *Lecture Notes in Geoinformation and Cartography*, vol. 2013-January, 2013, pp. 225–245.
- [35] O. Al Shorman, B. Al Shorman, M. Al-Khassawneh, and F. Alkahtani, "A review of internet of medical things (IoMT) - Based remote health monitoring through wearable sensors: A case study for diabetic patients," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 20, no. 1, pp. 414–422, Oct. 2020, doi: 10.11591/ijeecs.v20.i1.pp414-422.

## BIOGRAPHIES OF AUTHORS



**Nivethitha Krishnadoss**     received her B.Tech degree in Information technology in 2004 and her M.E degree in Computer science and engineering from S.K.P engineering college affiliated to Anna University, Chennai in 2007. She has been an assistant professor at various engineering colleges since May 2007. She is having 10 years of experience in teaching. She is presently pursuing Ph.D. degree in Vellore institute of technology, Vellore. Her area of interest includes big data analytics and machine learning. She can be contacted at email: nivethitha.k2020@vitstudent.ac.in.



**Dr. Lokesh Kumar Ramasamy**     has 12 years of academic and one year of Industry experience. He completed undergraduate in computer science degree from Anna University Chennai and post-graduation in Information Technology at the Anna University of Technology. He is proficient in Web Data Mining, Big data Analytics. I am currently working in the field of Data Science. I have published and presented around 55 papers in International Conferences and Reputed Journals. He has received a Young Scientist Award from Tamil Nadu State Council for Science and Technology, India, for a fellowship in the data analytics field. He was elevated and Privileged to become IEEE senior member in the Year 2019 and also a lifetime member of ISTE ACM, SCRS, and IAENG. He can be contacted at email: lokeshkumar.r@vit.ac.in.