iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 11 **Issue:** IV **Month of publication:** April 2023

**DOI:** https://doi.org/10.22214/ijraset.2023.50695

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 11 Issue IV Apr 2023- Available at www.ijraset.com*

# Big Data Stream Mining Using Integrated Framework with Classification and Clustering Methods

Mr. M. Nagasuresh[1], Ms. R. Roopa[2]
[1]*Assistant Professor, Department of Information Technology, Karpagam Institute of Technology, Coimbatore, Tamilnadu, India*
[2]*Assistant Professor, Department of AI & DS, Arjun College of Technology, Coimbatore, Tamilnadu, India*

***Abstract:*** ***The causes of numerous sorts of big data and data stream problems include the quick development of industry firms, the vast amount of data generated by these innovations, and the exponential growth of industrial company websites. There are numerous stream data mining algorithms for classification and grouping, each with its own unique set of attributes and important features. Ensemble classifiers aid in enhancing the greatest prediction performance results from these cutting-edge techniques. Ensemble approaches teach multiple types of classifiers and clusters rather than a single classifier. Their machine learning prediction findings are merged to form a voting schedule. This research offered a framework for stream data mining based on miss categorization stream data, utilizing the advantages of assembly technology. Real-world data streams are used in experiments. The experimental results are compared to modern popular ensemble techniques such as Boosting and Bagging. The test results show an increase in accuracy rate and decrease in classification time.***
***Keywords: Big Data, Bagging, Boosting, Data Stream Mining, Ensemble Classifiers, Misclassification Stream Data***

## I. INTRODUCTION

Every online application now has a substantial need to process massive amounts of data from many sources in the rapidly expanding big data era. This growth is accelerating swiftly and having a positive impact on all business and technological environments for the benefit of both organisations and individuals. Additionally, big data analysis aims to instantly extract statistical data using data mining algorithms that help with probability calculations, information discovery, categorization of recent events, and decision-making. However, the increase in classification speed comes at a price, with inaccurate relative class assignment regardless of the machine learning techniques utilised and a discrepancy in estimation from the original [1-3].

This research focuses on finding a solution to this issue by accelerating the mining of streaming data streams with high accuracy rate based on miss categorization data streams. The nature of big data and how it relates to applications in the real world are covered in Section 2. Modern data stream mining techniques are briefly discussed in Section 3's discussion of data stream mining. The prior research on data stream classification is presented in Section 4. Section 5 suggests a classification scheme. Results and experimental setups are shown in Section 6. Section 7 concludes by providing conclusions that compile results.

## II. BIG DATA

One definition of "big data" is the vast volume of data that exceeds a given threshold. However, there are numerous more ways to describe this term. Others are characterized as data that standard analytical tools like Microsoft Excel cannot handle. More widely used methods defined large data as having the characteristics of variety, velocity, and volume.

Big data analytics is a cutting-edge method that combines numerous methods and processes to glean priceless insights from unstructured data that, for whatever reason, is not appropriate for the conventional database system.

Applications for big data can be found in a variety of industries, including the financial, technological, electronic governmental, commercial, and healthcare sectors. Energy control also uses big data, anomaly detection, crime prediction, and risk management in other particular situations. Big data are a powerful control for all business types.

Information data can be characterized as a novel investment vehicle, an alternative kind of money, and an original source of priceless items. The ability of big data, which has effective corporate growth techniques, has been made public. However, it cannot be stated that all big data strategies cannot be applied to all business models. Regardless of the volume of data, it is a universal truth that a data information strategy is still beneficial. This vast amount of data in use creates brand-new, difficult detection challenges and encourages data stream mining [4].

*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 11 Issue IV Apr 2023- Available at www.ijraset.com*

## III. DATA STREAM MINING

Data mining and data streams are two disciplines related to data stream mining in computer science. It turns out to be fundamental to many computer science applications, including robotics, e-commerce, spam filtering, industrial engineering processes, credit card transaction flows, sensor networks, etc.

Although the data stream mining task differs significantly from traditional data mining tasks in terms of processing or execution, the goals are the same. Because of the following reasons, standard data mining algorithms cannot be applied directly to data streams:

*1)* Data streams can contain a significant amount of data and an essentially infinite number of components.
*2)* Data Streams are capable of arriving at the system quickly.
*3)* Data streams may be modified in a variety of ways during processing distribution times.

As a result, earlier information must be stored in compressed format structured by data stream techniques. The following groups comprise the most popular techniques for classifying data streams:

- Instance-Based Learning Methods
- Bayesian Learning Methods
- Artificial Neural Networks Learning Methods
- Decision Trees Learning Methods
- Ensemble Learning Methods

### *A. Instance-Based Learning Methods*

K-nearest neighbours learners are another name for instance-based learning classifiers. The prior data pieces must be stored in the memory because these instance classifiers may handle incremental learning methods. As a result, many typical learning techniques cannot be applied directly to data streams [5]. Instance-based classification techniques from every series were reported in [7].

### *B. Bayesian Learning Methods*

The common Bayesian theorem serves as the foundation for Bayesian learning classification techniques. Utilising the current training dataset, Bayesian learning seeks to assess the critical likelihoods. Then, new data is categorised using a learning algorithm; the group that maximises the following probability is assigned to an unlabeled or uncategorized element. The Naive Bayes method of learning is used incrementally. They must, however, have a set amount of memory. These Naive Bayes learning characteristics may be useful in the data stream mining process [5].

### *C. Artificial Neural Networks Learning Methods*

Learning algorithms for artificial neural networks are probably inspired by animal nervous systems. The most popular technique for learning classification is the multi-layer perceptron. When there are numerous data streams used for training, neural network learning can be changed to a single-pass incremental method. The amount of input synapses and neurons must remain constant throughout the learning process to maintain the same level of memory demand. Data streams may benefit from the neural network's aforementioned characteristics [5, 8].

### *D. Decision Trees Learning Methods*

Modern technology Data stream classification can be done using decision tree techniques. The Hoeffding trees approach provides the foundation for the algorithms of this type. Hoeffding tree selects an attribute that is suitable to split tree nodes for the static data. All of the data pieces of the node cannot be stored in memory due to the unlimited size of the data streams. As a result, data streams are handled by evolutionary learning algorithms. The VFDT algorithm is the most prominent technique for this type of learning [10]

### *E. Clustering Methods*

For unsigned instances with homogeneous groupings relating to their commonalities, clustering can be employed. Clustering can be done using streaming methods on two different levels: online and to get rid of similar data that is offline. The stream is effectively used to compute and update a set of very small clusters at the online level; offline, the micro clusters are processed using a traditional batch clustering method, such as k-means. Online level clustering simply uses one pass step for the input data, but offline level clustering involves multiple processing phases.

Considering that the offline processing can be divided into a number of small clusters and called when the stream is finished. Also, they can refresh the collection of discrete clusters on a regular basis in accordance with their required stream flows.

Due to its simplicity, the k-means clustering method is one of the most popular clustering techniques. The value of k is first picked at random to begin the clustering process, but most widely used established methods start with 1, and some start with 5 or 10. Each instance is then assigned to the closest centroid based on that centroid value. The cluster centroids are once more calculated using the assigned instance's centre of mass.

Once the desired criterion is met or the assignments cannot be modified, this computation is repeated. Because data streams require multiple passes to cluster, this routine cannot be used for data stream mining [16].

Nonetheless, the Bayesian Classifier, Ensemble Classifier, Decision Tree, and Cluster are the main topics of this work.

## IV. LITERATURE REVIEW

On the typical dataset of data stream mining, research for big data categorization has been the subject of hundreds of scholarly papers. These completed works are characterised in accordance with the classification of the aforementioned cutting-edge classification categories.

In paper [6], the author discusses the key capabilities of the various streaming data processing systems that have been established. They clearly outline the future directions for research on high-speed, large-scale mining techniques for streaming data from various angles, including procedures, implementation style, and performance evaluation analysis. They made it very apparent that although Instance-Based Classifiers take much longer to complete, they are more accurate than other classifiers. Thus, the authors in [17] introduced the nearest neighbour incremental classifier, taking advantage of distributed computing's benefits to execute faster updates.

An operational pattern-based Bayesian learning classifier for handling data streams was proposed in [18]. In order to learn more, the researchers in [19] applied the well-known and very effective "Naive Bayes Algorithm" on a vast amount of complex data. They suggested a reduction strategy to exclude comparable data, which sequentially decreases computing time, lowers memory space requirements, and improves the functionality of the Nave Bayes algorithm. Their research points to extremely effective Naive Bayes algorithms or massive data streams.

Numerous writers suggested using Neural Network Deep Learning techniques in various ways for data stream processing. Deep learning neural network topologies have the capacity for complex tasks and occasionally surpass humans in specific application domains.

Despite the obvious impressive progress in this field, training deep architectures presents an unsolved optimization challenge with a sizable number of hyper-parameters.

Due to this, a version of the Neural Network Integrated Framework has been introduced in [20] to enable highly scalable online calculation capabilities for mining data stream.

Many surveys and studies on data stream categorization and regression problems have been conducted by [13]. They reviewed a wide variety of ensemble data stream processing techniques and introduced novel learning strategies for processing imbalanced data streams, including complicated data representation, semi-supervised learning, structured outputs, and detection methods. Iterative Boosting Streaming Ensemble (IBS), which the authors of [16] suggest, is a novel ensemble learning approach that can categorise streaming data. In order to prevent the risks associated with vote-based separated ensembles, the authors in [21] offer a new distributed training strategy for ensemble classifiers. The name of this model is "LADEL."

## V. PROPOSED FRAMEWORK

Let DS = S1, S2,..., St, be a data stream input as a series of batches, with St = S1..., SN being an unlabeled batch. Consider the equivalent labelled set, defined as St = (S1, L1)..., (SN, LN), that can be employed during the training stages. Suppose the genuine class label Li of instance Si, for I = 1... N.

For incoming unlabeled real-world data, the class labels must be manually forecasted. The common stream dataset does not require this state an automated system for mining data streams that consistently performs well in terms of categorization accuracy, computation speed, and memory utilization.

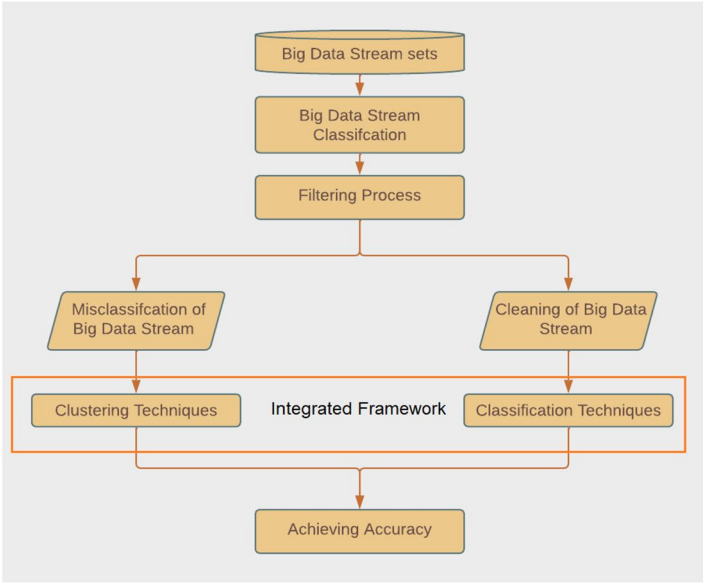



Fig. 1. Integrated Framework for Big Data Streams

The basic classifier classifies St's instances after labelling batches of data. There is a presumption that the labels' names will be known as soon as the classification procedure is complete, allowing for the prediction of this batch's miss classification mistake and the clear identification of the proper instance stream. Not all techniques are appropriate for all data streams, and the performance of the classifier can be decreased by the uncertain, ambiguous, incomplete, and subjective data. As a result, after the classification procedure, the misclassification data streams are filtered out. Following the creation of the filtering process, an assembly approach is developed utilising a basic cluster to enhance the model and concentrate on data streams that are challenging to categorise. Separated from correctly labelled streams, clean data streams CDS= "S1, S2..., St" from DS, are the miss categorised data streams, MDS = "S1, S2..., St." The left incorrect data streams are then labelled using an ensemble clustering method, or MDS. Following the clustering of MDS batches, the accuracy for each cluster is determined, and the overall accuracy of the DS can also be determined as shown in Fig. 1.

## VI. EXPERIMENT RESULTS

The ensemble effect is examined using the well-known data set of streams, real-world Electricity [22] data. The electrical market in New South Wales, an Australian State, served as the source of the data on electricity. Pricing in this market are erratic and dependent on supply and demand. It includes actual data that was gathered for two years and seven months at intervals of 30 minutes. There are 45,312 occurrences in this dataset, each with five different values for the time, day, period, and price. The change in price that corresponds to the moving average (MA) of the previous 24 hours is defined by the class label.

The classification of data streams is the topic of the experiments. The system has to know the label of the class in order to filter out the data that has been incorrectly categorised, hence the experiment design is built using the evaluate prequential approach. First, the results of the classification using the well-known classifiers Naive Bayes and VFDT are displayed in Table I.

TABLE I. PERFORMANCE MEASUREMENT COMPARISON FOR SINGLE STANDARD DATA STREAM CLASSIFIERS

| Name of Data Stream Algorithm | Classification Accuracy (%) | Kappa Statistic | Kappa Temporal Statistic | Elapsed Time (s) |
|---|---|---|---|---|
| Naive Bayes | 73.07 | 40.89 | -83.57 | 0.67 |
| VFDT | 72.23 | 43.59 | -89.30 | 0.52 |

The data streams are tested with the standard ensemble methods of Leveraging and Boosting. The results from the experiments are summarized in Table II.

TABLE II. PERFORMANCE MEASUREMENT COMPARISON FOR ENSEMBLE DATA STREAM CLASSIFIERS

| Name of Data Stream Algorithms | Classification Accuracy (%) | Kappa Statistic | Kappa Temporal Statistic | Elapsed Time (s) |
|---|---|---|---|---|
| LeveragingNB | 52.82 | 12.95 | -221.62 | 1.58 |
| LeveragingVFDT | 75.497 | 48.38 | -67.04 | 4.25 |
| OZOBoostNB | 74.322 | 44.38 | -75.04 | 1.02 |
| OZOBOOSTVFDT | 69.352 | 39.70 | -108.92 | 1.72 |

Next experiments are conducted utilising the straightforward KMeans clustering approach for the proposed framework's portion. Table III presents the data and outcomes. These added activities can ensure categorization accuracy but must be mindful of passing time. Table IV provides an illustration of the suggested ensemble method's overall performance.

TABLE III. PERFORMANCE MEASUREMENT COMPARISON FOR CLUSTERING FOR MISS DATA

| Data Stream | Classification Accuracy (%) | Data Count | Elapsed Time (s) |
|---|---|---|---|
| Miss data for NB | 45.57 | 12202 | 0.03 |
| Miss data for VFDT | 59.29 | 12583 | 0.08 |

TABLE IV. PERFORMANCE MEASUREMENT COMPARISON FOR PROPOSED ENSEMBLE DATA STREAM CLASSIFIERS

| Name of Data Stream Algorithms | Classification Accuracy (%) | Elapsed Time (s) | Improved Accuracy (%) |
|---|---|---|---|
| NB + KMeans | 88.04 | 0.7 | 14.97 |
| VFDT + KMeans | 88.69 | 0.6 | 16.46 |

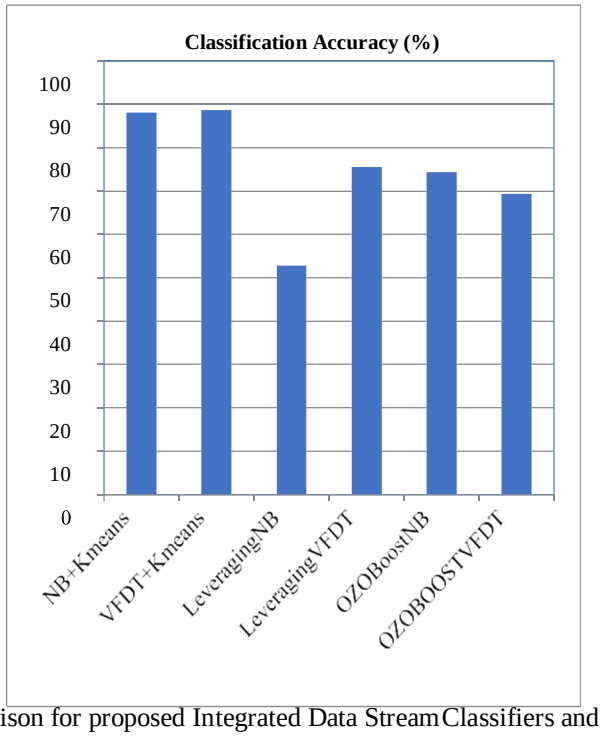



Fig. 2. Measurement Comparison for proposed Integrated Data Stream Classifiers and Standard Integrated Classifier

Then the comparison is carried out the proposed ensemblemethod and the state-of-the-art ensemble methods and resultsare shown in Fig. 2. From these results, it can be seen clearly that the proposed framework not only can increase the classification accuracy but also less than inelapsed time.

## VII. CONCLUSION

In this paper, the ensemble framework constructed from the data streams classifiers and simple K-Means clustering isproposed for mining data streams. The proposed framework of the ensemble learning classifiers, the combination of NaïveBayes and K-Means, and VFDT and K-Means, has been evaluated. Furthermore, the comparison of the proposed framework against state-of-the-art ensembles, Leveraging andBoosting using standard data stream set. The results clearly show that the proposed framework not only can improve the classification accuracy based on mis-classification data, but also can reduce the time taken than the above standard ensemble techniques. Future research will concentrate on learning the influence of the size of stream data and more effective ensemble mechanisms on accuracy of the ensembleclassifier.

## REFERENCES

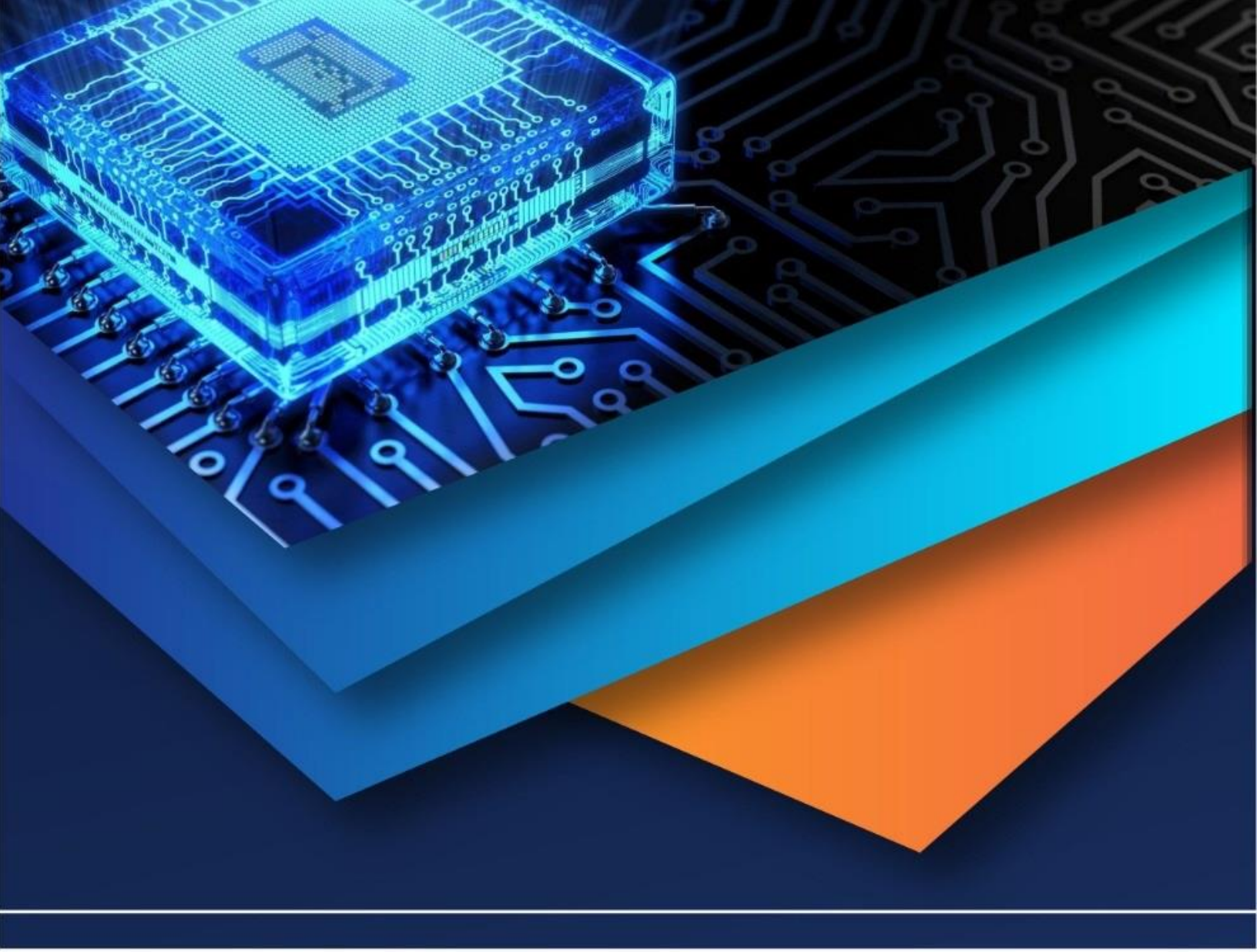
[1] N. Sun, B. Sun, J. Lin and M. Yu-Chi Wu, “Lossless Pruned Naive Bayes for Big Data Classifications,” Big Data Research, vol.14, pp. 27-36, December 2018. https://doi.org/10.1016/j.bdr.2018.05.007.
[2] C. Tsai, C. Lai, H. Chao and A. V. Vasilakos, “Big Data Analytics: A Survey,” Journal of Big Data, vol.2, A21, pp. 1-32, December 2015. https://doi.org/10.1186/s40537-015-0030-3.
[3] M. Marjani, F. Nasaruddin, A. Gani, A. Karim, A., I. Abaker Targio Hashem, A. Siddiqa, and I. Yaqoob, “Big IoT Data Analytics: Architecture, Opportunities, and Open Research Challenges,” IEEE Access, vol.5, pp. 5247-5261, May 2017. https://doi.org/10.1109/ACCESS.2017.2689040
[4] F. Corea, An Introduction to Data Everything You Need to Know About AI, Big Data and Data Science. ISBN 978-3-030-04467-1, Springer Nature Switzerland AG, 2019. https://doi.org/10.1007/978-3-030-04468-8.
[5] L. Rutkowski, M. Jaworski and P. Duda, Stream Data Mining: Algorithms and Their Probabilistic Properties, Studies in Big Data, Volume 56, ISSN 2197-6503, Springer Nature Switzerland AG: Springer International Publishing, 2020.
[6] B. Rohit Prasad and S. Agarwal, “Stream Data Mining: Platforms, Algorithms, Performance Evaluators and Research Trends,” International Journal of Database Theory and Application, vol. 9, No. 9, pp. 201-218, 2016. http://dx.doi.org/10.14257/ijdta.2016.9.9.19.
[7] D.W. Aha, D. Kibler and M.K. Albert, “Instance-Based Learning Algorithms,” Machine Learning, vol. 6, No. 1, pp. 37-66, 1991. https://doi.org/10.1007/BF00153759.
[8] J. Gama, P. Pereira Rodrigues, “Stream-Based Electricity Load Forecast,” Knowledge Discovery in Databases: PKDD 2007, Lecture Notes in Computer Science, vol. 4702, pp. 446-453, 2007, Springer, Berlin. https://doi.org/10.1007/978-3-540-74976-9_45.
[9] D. Jankowski, K. Jackowski and B. Cyganek, “Learning Decision Trees from Data Streams with Concept Drift,” International Conference on Computational Science 2016, ICCS 2016, San Diego, California, USA, 6–8 June 2016. Procedia Computer Science vol. 80, pp. 1682-1691, 2016. https://doi.org/10.1016/j.procs.2016.05.508.
[10] P. Domingos and G. Hulten, “Mining High-Speed Data Streams,” In Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston, Massachusetts, USA, pp. 71-80, 2000. https://doi.org/10.1145/347090.347107.
[11] J. N. van Rijn, G. Holmes, B. Pfahringer and J. Vanschoren, “The Online Performance Estimation Framework: Heterogeneous Ensemble Learning for Data Streams,” Machine Learning, vol. 107, No. 1, pp. 149-176, 2018. https://doi.org/10.1007/s10994-017-5686-9.
[12] L. I. Kuncheva, “Classifier Ensembles for Detecting Concept Change in Streaming Data: Overview and Perspectives,” In Proceedings of the 2nd Workshop SUEMA, ECAI, Patras, Greece, pp. 5-9, July 2008.
[13] B. Krawczyk, L. L. Minku, J. Gama, J. Stefanowski and M. Wo´zniak, “Ensemble Learning for Data Stream Analysis: A Survey,” Information Fusion, vol. 37, pp. 132-156, 2017.
[14] https://doi.org/10.1016/j.inffus.2017.02.004.
[15] N. C. Oza and R. Russell, “Online Bagging and Boosting,” In Eighth International Workshop on Artificial Intelligence and Statistics, pp. 105-112, January 2001, Morgan Kaufmann, Key West, Florida, USA.
[16] J. Roberto Bertini Junior and M. Carmo Nicoletti, “An Iterative Boosting-Based Ensemble for Streaming Data Classification,” Information Fusion, vol. 45, pp. 66-78, 2018.
[17] https://doi.org/10.1016/j.inffus.2018.01.003.
[18] A. Bifet, R. Gavaldà, G. Holmes and B. Pfahringer, Machine Learning for Data Streams: with Practical Examples in MOA. ISBN: 9780262037792. The MIT Press, 2018.
[19] S. Ram´ırez-Gallego, B. Krawczyk, S. Garc´ıa, M. Wo´zniak, J. Manuel Ben´ıtez and F. Herrera, “Nearest Neighbor Classification for High-Speed Big Data Streams Using Spark,” IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 47, issue. 10, pp. 2727- 2739, 2017. https://doi.org/10.1109/TSMC.2017.2700889.
[20] J. Yuan, Z. Wang, Y. Sun, W. Zhang, and J. Jiang, “An Effective Pattern-based Bayesian Classifier for Evolving Data Stream,” Neurocomputing, pp. 1-12, 2018.
[21] https://doi.org/10.1016/j.neucom.2018.01.016.
[22] S. David, K. Ranjithkumar, S. Rao, S. Baradwaj, and D. Sudhakar, “Classification of Massive Data Streams Using Naïve Bayes,” IAETSD Journal for Advanced Research in Applied Sciences, vol. 5, issue 4, pp. 208-215, 2018.
[23] M. Pratama, P. Angelov, J. Lu, E. Lughofer, M. Seera and C. P. Lim, “A Randomized Neural Network for Data Streams,” International Joint Conference on

Neural Networks (IJCNN), Anchorage, AK, USA, May 2017, pp. 14-19. https://doi.org/10.1109/IJCNN.2017.7966286.

[24] Sundaresan K, Nallakumar R, "DL-Based Human Face Recognition" International Journal of Research and Analytical Reviews" E-ISSN:2348-1269, P-ISSN:2349-5138,Volume 10, Issue 1 March, 2023.

[25] S. Khalifa, P. Martin, and R. Young, "Label-Aware Distributed Ensemble Learning: A Simplified Distributed Classifier Training Model for Big Data," Big Data Research, vol. 15, pp. 1-11, 2019. https://doi.org/10.1016/j.bdr.2018.11.001.

[26] M. Harries, "Splice-2 Comparative Evaluation: Electricity Pricing," Technical Report 9905, School of Computer Science and Engineering, University of New South Wales, Australia, 1999.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY