Survey on IoT Data Preprocessing

V.A. Jane[a] and Dr. L. Arockiam[b]

[a] Department of computer Science, St. Joseph's College, Trichy, Tamilnadu, India 62001.
[b] Department of computer Science, St. Joseph's College, Trichy, Tamilnadu, India 62001.

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 20 April 2021

Abstract: Internet of Things (IoT) is a growing technology in all fields of science and engineering. The amount of data emitted by the sensors used in the various fields is high. Therefore, efficient knowledge from such large datasets is a clear requirement of many users. This large data is far from perfect; it has many defects (such as noise, missing values, outliers etc.) and is not suitable for analysis because it can lead to incorrect conclusions. So, data preprocessing is a required technique for such data. Data preprocessing is an important and essential step, the main goal of which is to dedicate techniques to clean, refine, repair and improve that raw data. This paper proposes a survey on IoT data preprocessing and its techniques. This paper discusses exiting research on data preprocessing in the IoT context and, introduces the background of IoT data preprocessing and present literature reviews of the advanced research on data preprocessing techniques. The classification of various preprocessing approaches with techniques is clearly depicted in the figure. Various approaches of preprocessing cleaning, transformation, reduction and integration are described. In addition, methods for such approaches in IoT data preprocessing are also discussed. IoT Data preprocessing techniques on various applications are tabulated. Finally, issues and challenges, most useful in future work, are discussed.

Keywords: IoT, Preprocessing, Data Cleaning, Noise handling

1. Introduction

Internet of Things basically refers to a network of objects that are connected to the Internet. It is an excellent computerization and analysis system across various industries such as agriculture, medical, transport, city, etc., [1]. Being connected to the Internet, one can collect data and send it over the internet, receive information from the internet, or do both. In the Internet of Things (IoT), the connected devices / sensors generate data enormously. These data are transferred to the cloud database for analyzing and to create smart applications. Data analytics is a very important technique to find insights from these data [2]. Before analyzing the data, data preprocessing plays a vital task owing to such kind of data with many defects like missing, noise, and inconsistent data. It is a kind of key stages in knowledge discovery process [3]. Low-quality data can undermine the effectiveness of successive learning algorithms. Therefore, avoiding the impact in quality, improves reliability of successive automated innovations and enhances decisions by taking appropriate preprocessing methods. There are various techniques involved in it, namely, data transformation, data reduction, data normalization, data cleaning, and data integration [5]. These techniques simplify the data by selecting or eradicating unnecessary features and dividing difficult constant feature spaces. During this process, the original input construction needs to be maintained and processing time need to be considered. Some benefits of data preprocessing are rapid training of learning methods, advanced generalization skills, as well as better understanding and easy interpretation of results [6]. This paper aims to survey on data preprocessing, its techniques and existing contributions of data preprocessing. This survey constructed as follows: In part II, The related works on data preprocessing in IoT environments and its techniques are discussed. Part III summarizes the techniques in various IoT based application, and part V concludes this work.

2. Related work

Hui et al., [7] reviewed the physical sensor errors that occur during the data-collection process. This paper described types of physical sensor errors, various error-detection mechanisms, error-correction techniques and also explained the differences between the techniques. Among error-detection and correction mechanisms, Principal Component Analysis (PCA) and Artificial Neural Network (ANN) provided better results.

Mathew et al., [8] compared various preprocessing techniques, namely, Kalman filter, z-scoring and moving Average filter. Firstly, preprocessing techniques were applied to the chemical sensor data to clean it. After that, the dataset is cleaned and evaluated using different classifiers such as Linear Discriminant Analysis (LDA), K Nearest Neighbor (KNN), and Support Vector Classifier (SVC). Finally, the performances of the various preprocessing techniques were calculated. Among these, it was observed that the Kalman filter technique provided better result than others.

Zena et al.,[9] reviewed methods of selecting and extracting features for high-dimensional Microarray cancer dataset. The writer discussed about the problems of irrelevant and redundant features in the micro array dataset. Also, the importance of dimensionality reduction, its advantages and drawbacks were discussed.

Chao et al.,[10] explained the process of data transmission in IoT environment. A preprocessing technique was adopted for reducing transmission time and increasing processing speedbut this work, only focused on reducing data transmission time.

Evgeniy [11] proposed architecture for preprocessing sensor data. Various preprocessing techniques suitable for proposed architecture were found. Streaming sensor data, the Univariate time series dataset, was utilized in this architecture.

Natarajasivan et al.,[12] proposed filter-based monitoring system for IoT Context. Sensors utilized in this work for sensing acceleration, position, vision, audio, temperature and direction. Kalman filter was utilized to process the collected data from those sensors and to evaluate the results using SVM. The proposed system consumed more time.

Cleber et al.,[13] surveyed all IoT application papers published since 2015. The author numbered the IoT application based on the usage. Smart home applications are used widely used by the researchers when compared with others. The sensor used in the smart environments is also discussed.

Rajalakshmi et al.,[14] discussed the function of IoT in smart appliances and summarized the problems such as data aggregation, scalability, data fusion, de-noising, heterogeneity, data outlier detection, real-time processing and missing data imputation. The author explained the usage of cloud, fog and edge computing in IoT to improve the analytics process and described the IoT data analytics process using a drone for traffic-monitoring system.

David et al.,[15] reviewed the data management problems in IoT environment, namely data collection, cleaning, integration, migration and processing. The author discussed the advanced data-processing technologies such as AI, machine learning, deep learning, and data mining.

Karinaer al.,[16] presented a survey on preprocessing techniques with relevant issues related to data mining. The fundamental concepts of data mining, preprocessing techniques and its issues were explained in detail. Moreover, it offered various solutions and discussed future directions.

García et al.,[17] proposed data preprocessing methods for big data era. The key areas of data preprocessing and current open challenges were explained. Moreover, the different data preprocessing techniques, namely, normalization, discretization, subset selection and extraction, feature indexers and encoders. In addition, other techniques for text mining were reviewed. Also, major issues of big data preprocessing were highlighted.

Jayaram et al.,[18] presented a study on data preprocessing methods. The main aim was to provide solutions for various problems of data preprocessing. The author focused data cleaning methods that includes filter, imputation, hybrid, wrapper and ensemble methods. The process and uses of each methods were described with examples. In particular, noise, data handling were considered and explanations regarding how to detect and treat it were given. Finally, the challenges while dealing with data cleaning at different fields were illustrated.

Huma Jamshed et al.,[19] discussed various big data Preprocessing techniques to clean data for further mining and analysis tasks. Initially, the important steps involved during the data preprocessing were explained. Then, a framework for web data preprocessing was proposed and each step was explained one by one. Finally, the simple text data was applied on the framework and preprocessing steps, like noisy removal, tokenization, normalization, were done.

Categories of preprocessing techniques

Data preprocessing is the process to make real world data more suitable for data mining process [20]. Real-world data is more noisy, contains missing values and a lot of ambiguous information, and these data are large in size. These factors cause the deterioration of the quality of the data during the result that obtaining after the mining or modeling. Therefore, before mining or modeling the data, it must be passed through improvement techniques known as data preprocessing. There are different techniques to perform such kind of process to make the data suitable for analyzing purposes. The categories of data pre-
processing techniques are shown in fig 1.

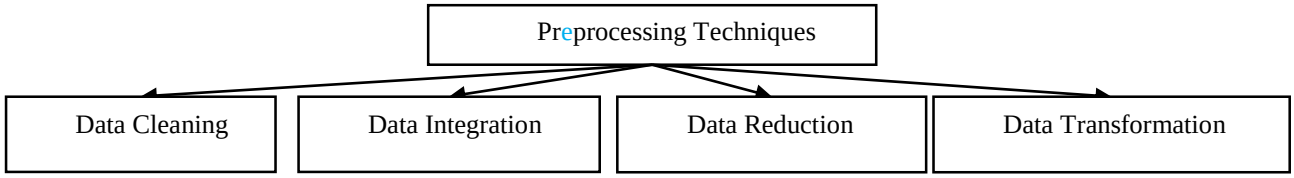

Fig: 1 Categories of data preprocessing Techniques

Data cleaning can be defined as the process of eliminating the erroneous and missing part in the data. The process of handling these noisy and missing values can be achieved by various ways, that shown in fig 2.

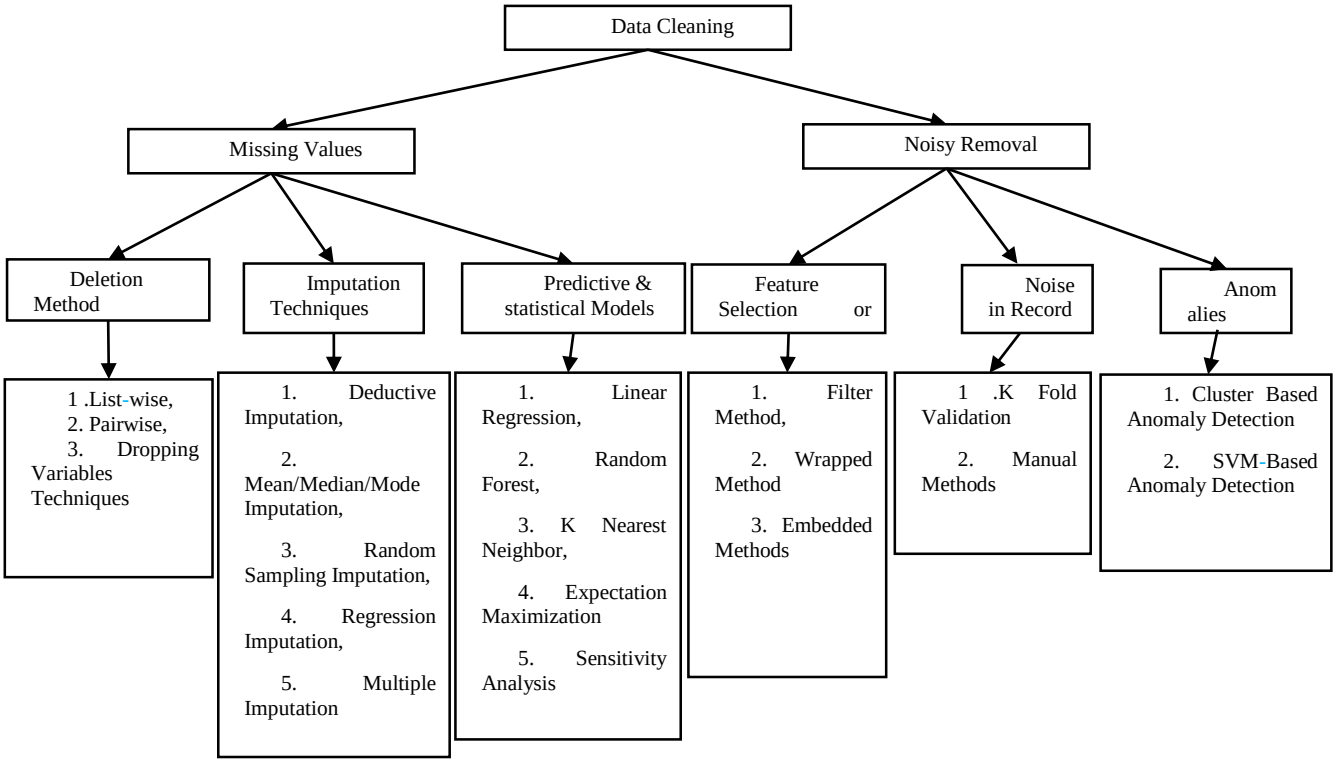

Fig: 2 Data Cleaning Techniques

Data integration is one important technique in preprocessing which combines data from different source and giving users an integrated view of this data. Mainly, Data integration is done through two main approaches, that are explains in the following fig 3.

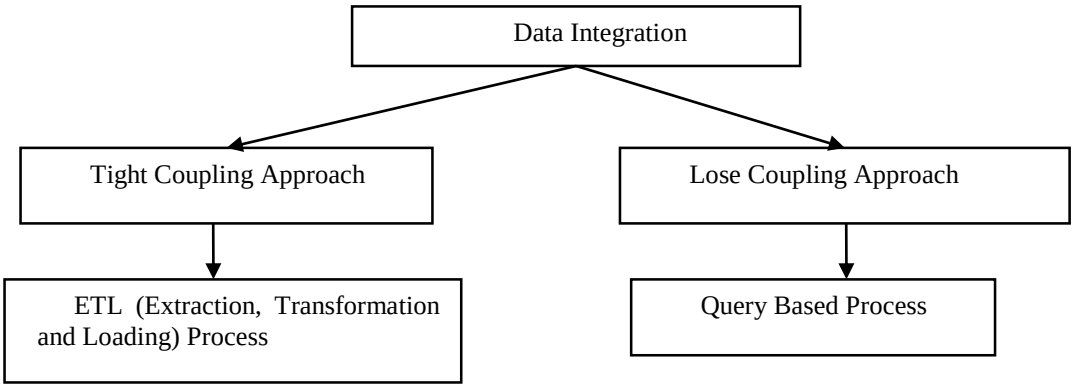

Fig: 3 Data Integration Techniques

Data reduction techniques can be used to obtain a data set, which are very small in size but yield, the same analytical results. Data reduction approaches utilized to diminish the unnecessary data as well as improve analytical process. Traditional, data reduction approaches are depicted in fig 4.

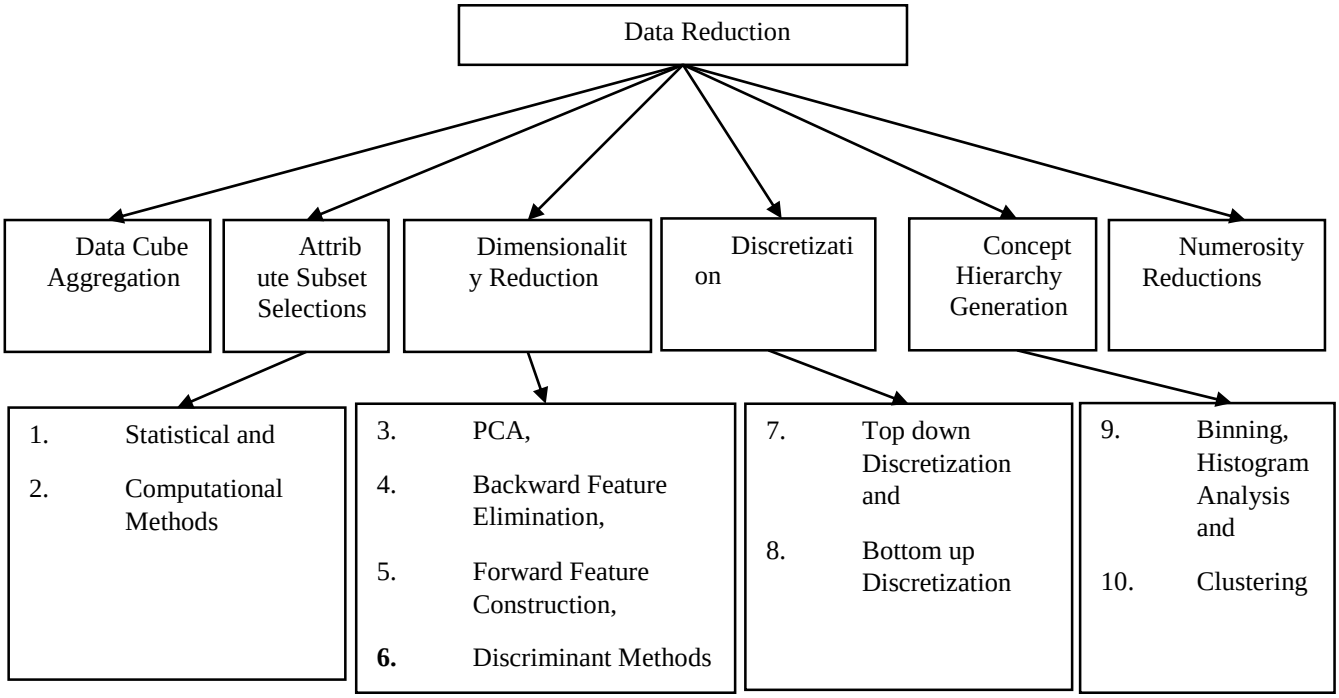

Figure: 4 Data Reduction Techniques

Data transformation is the process of converts' data from one format to another format. Data transformation includes various functions to achieve the perfect format that shown in fig 5.

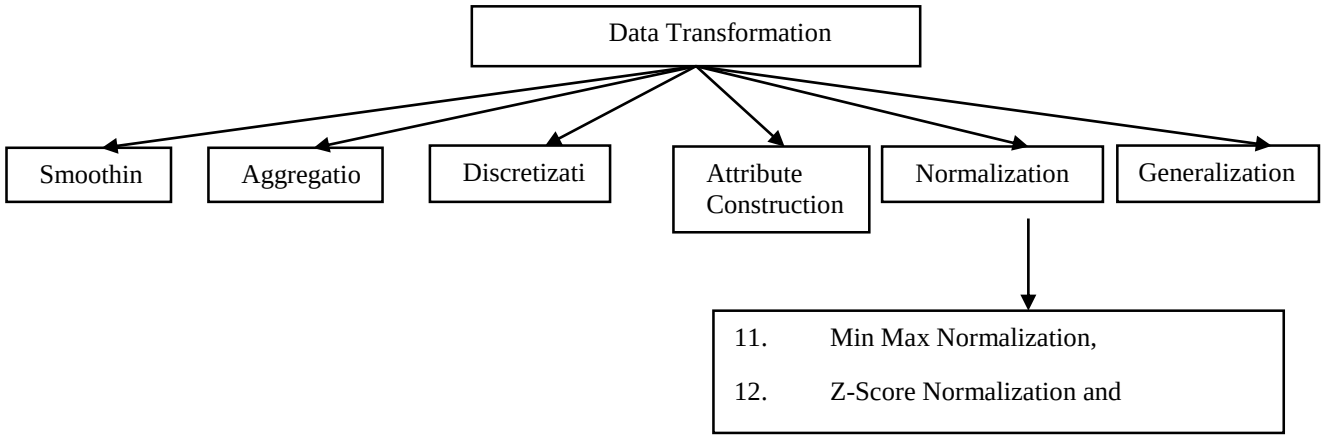

Figure: 5 Data Transformation Techniques

The above discussed techniques are mostly used for reducing the defects in dataset. By applying these techniques, the process of analytical models can be improved.

Moreover, the related work on preprocessing in various IoT-based applications are surveyed and listed in the following table 1.

Table 1: Uses of preprocessing Techniques in IoT-based Applications

Author Name & Year	Objective	Technique /Algorithm/Tool	Application Domain / Dataset
DiviyaPrabha 2016 [21]	Discuss various Technology that used in IoT for Data Collection and Data processing	Eclipse, KinomaJS, M2MLabs Mainspring, Node-RED, Raspberry Pi, RFID, QT (Quick Response), NFC (Near Field Communication), BLE (Bluetooth Low Energy), ZigBee	General IoT Environment

Peter 2017[22]	Overview Data Mining (DM) the Internet of Things (IoT), Preprocessing, Predictive Analytics	Machine Learning, Deep Learning, Natural Language Processing(NLP)	IoT Data
Brink2017 [23]	Provide solutions to Preprocessing problems	Modified Traditional Kalman Filter, intermittent Schmidt–Kalman filter (ISKF), the fixed-weight partial-update Schmidt–Kalman filter (FPSKF), and the partial-update Schmidt–Kalman filter (PSKF)	IMU camera Data
Bhavana 2017[24]	Survey & Discussion	IoT, Traditional Database management, Cloud, Sensor Data.	IoT Data
Shobanadevi 2017 [25]	Explain Role of Data Mining and Big Data in IoT	MapReduce, Appache Hadoop, KMeans, KNN(K nearest Neighbor), SVM(Support Vector Machine), Random Forest, Apriori	Health Care, Home Automation, Smart City
Akshat 2018[26]	Review	Data mining, IoT, Knowledge Discovery in Databases (KDD), Machine Learning	IoT Data
Pavithra 2019 [27]	Elaborate Role of Big Data in IoT to job and Market	Streaming Analytics, Spatial Analytics, Time Series Analytics, Prescriptive Analysis	General IoT Environment
Sandip 2019[28]	Survey	IoT, Radio Frequency Identification (RFID),Cloud, Machine to Machine Communication, Sensors and Actuators, Network Connectivity, Data Mining Preprocessing.	Smart application Data
Alcalde 2019 [29]	Library	Data Stream Library for Big Data Preprocessing DPASF	Streaming Big Data
Shivani 2019 [30]	Comparative Study	Reviewed all papers related with Noisy Data Between January 1993 to July 2018	Real world Data

Conclusion

Big data is now rapidly expanding across all domains such as education, agriculture, healthcare, institutions, web mining etc., Learning knowledge from this massive data is an interesting task as well as challenging one. Knowledge gaining from large sets of data brings significant opportunities and transformational potential to different sectors. But, the massive data comes with imperfection like noisy, missing values etc., this can lead to decrease in the efficiency and accuracy of decision making. So, refinement of data is required. This work offers the systematic flow of survey on data preprocessing techniques in the area of IoT and big environments. In which, the fundamentals of data preprocessing was covered, and literature reviews that related to the data preprocessing techniques were described. The classification of various pre-processing approaches with techniques was clearly depicted by the figure. Various approaches of preprocessing cleaning, transformation, reduction and integration with methods or techniques were illustrated. Data preprocessing techniques on various application were tabulated. Finally, issues and challenges, which need to be taken attention of in the future, were presented.

References

1. Bramer, Max. “Data for data mining”, In Principles of data mining*”*, pp. 9-19. *Springer,London*, 2016.
2. Alasadi, Suad A., and Wesam S. Bhaya.”Review of data preprocessing techniques indata mining, *Journal of Engineering and Applied Sciences* 12, no. 16 (2017): 4102-4107, 2017.
3. Cordón, Ignacio, Julián Luengo, Salvador García, Francisco Herrera, andFrancisco Charte. "Smartdata: Data preprocessing to achieve smart data in r." Neurocomputing 360, 1-13, 2019.
4. Hu, Hanqing, and Mehmed Kantardzic. "Smart preprocessing improves data streammining." In 2016 49th Hawaii International Conference on System Sciences (HICSS), pp.1749-1757. IEEE, 2016.
5. Shi, F.; Li, Q.; Zhu, T.; Ning, H.,“A survey of data semantization in internet of things”,*Sensors*, 18, 313,2018.
6. Shah, S. H., & Yaqoob, I, “A survey: Internet of Things (IOT) technologies, applications and challenges“,*IEEE Smart Energy Grid Engineering SEGE)*. doi:10.1109/sege.2016.7589556, 2016.
7. Teh, HuiYie, Kempa-Liehr, Andreas W, Wang, Kevin I-Kai, "Sensor data quality: a systematic review", *Journal of Big Data*, 7(1), 11–60,2020, doi:10.1186/s40537-020-0285-1
8. Weiss, Matthew, Wiederoder, Michael S, Paffenroth, Randy C, Nallon, Eric C, Bright, Collin J, Schnee, Vincent P, McGraw, Shannon; Polcha, Michael, Uzarski, Joshua R, "Applications of the Kalman Filter to Chemical Sensors for Downstream Machine Learning", *IEEE Sensors Journal*, (), 1–1, 2018, doi:10.1109/JSEN.2018.2836183
9. Hira, Z. M., &Gillies, D. F, "A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data", *Advances in Bioinformatics*, 1–13, 2015,doi:10.1155/2015/198363
10. Xu, C., Yang, H. H., Wang, X., &Quek, T. Q. S, "On Peak Age of Information in Data Preprocessing enabled IoT Networks", *IEEE Wireless Communications and Networking Conference (WCNC)*, 2019, doi:10.1109/wcnc.2019.8885690
11. EvgeniyLatyshev, "Sensor Data Preprocessing, Feature Engineering andEquipment Remaining Lifetime Forecasting for PredictiveMaintenance", *Data Analytics and Management in DataIntensive Domains (DAMDID/RCDL’2018)*,226-231, 2018.
12. D. Natarajasivan and M. Govindarajan, "Filter Based Sensor Fusion for Activity Recognition using Smartphone", *International Journal of Computer Science and Telecommunications* Volume 7, Issue 5, 2016.
13. Morais, C. M. de, Sadok, D., &Kelner, J, "An IoT sensor and scenario survey for data researchers", *Journal of the Brazilian Computer Society*, 25(1), doi:10.1186/s13173-019-0085-7, 2019.
14. RajalakshmiKrishnamurthi, Adarsh Kumar, DhanalekshmiGopinathan, AnandNayyar, and Basit Qureshi, "An Overview of IoT Sensor Data Processing, Fusion, and Analysis Techniques", *Sensors*, 20, 6076; doi:10.3390/s20216076.
15. Gil, D., Johnsson, M., Mora, H., & Szymanski, J, "Advances in Architectures, Big Data, and Machine Learning Techniques for Complex Internet of Things Systems", *Complexity*, 1–3, doi:10.1155/2019/4184708,2019.
16. Gibert, Karina, Miquel Sànchez–Marrè, and Joaquín Izquierdo. "A survey on pre-processing techniques: Relevant issues in the context of environmental data mining." *AI Communications* 29, no. 6 (2016): 627-663.
17. García, Salvador, Sergio Ramírez-Gallego, Julián Luengo, José Manuel Benítez, and Francisco Herrera. "Big data preprocessing: methods and prospects." *Big Data Analytics* 1, no. 1 (2016): 9.
18. Hariharakrishnan, Jayaram, S. Mohanavalli, and KB Sundhara Kumar. "Survey of pre-processing techniques for mining big data." In *2017 International Conference on Computer, Communication and Signal Processing (ICCCSP)*, pp. 1-5. IEEE, 2017.
19. Jamshed, Huma & Khan, M. & Khurram, Muhammad & Inayatullah, Syed & Athar, Sameen. (2019). Data Preprocessing: A preliminary step for web data mining. 206-221. 10.17993/3ctecno.2019.specialissue2.206-221.
20. Shobanadevi, A., & Maragatham, G. Data mining techniques for IoT and big data — A survey*, ‘‘International Conference on Intelligent Sustainable Systems (ICISS)’’,* ISBN:978-1-5386-1959-9,doi:10.1109/iss1.2017.8389260,2017.

21. V. Diviya Prabha R. Rathipriya, IoT Data and its Application-A Preliminary Study,"International Journal of Computational Intelligence and Informatics", Vol. 6: No. 1, 2016.
22. Peter Wlodarczak, Mustafa Ally, Jeffrey Soar,"Data Mining in IoT", *In Proceedings of 2nd Int. Workshop on Knowledge Management of Web Social Media, Leipzig, Germany, August 2017 (KMWSM '17)*, ISBN 978-1-4503-4951, https://doi.org/10.1145/3106426.3115866, 2017.
23. Brink, K. M, " Partial-Update Schmidt–Kalman Filter", *Journal of Guidance, Control, and Dynamics*, 40(9), 2214–2228, doi:10.2514/1.g002808,2017.
24. Bhavana Bachhav, Parikshit N. Mahalle, "Data Management for Internet of Things: A Survey and Discussion", International Research Journal of Engineering and Technology (IRJET) ISSN: 2395-0056, Volume: 04, Issue: 11, 2017
25. Shobanadevi, A., & Maragatham, G. Data mining techniques for IoT and big data — A survey*, "International Conference on Intelligent Sustainable Systems (ICISS)",* ISBN:978-1-5386-1959-9,doi:10.1109/iss1.2017.8389260,2017.
26. Akshat Savaliya, Aakash Bhatia, Jitendra Bhatia, "Application of Data Mining Techniques in IoT: A Short Review", Volume 4, Issue 2, ISSN: 2395-1990,2018.
27. A.Pavithra, C.Anandhakumar, V.Nithin Meenashisundharam, Internet of Things with BIG DATA Analytics – A Survey, "International Journal of Scientific Research in Computer Science Applications and Management Studies IJSRCSAMS",Volume 8, Issue 1,ISSN 2319 – 1953, 2019.
28. Sandip Sonawane, "Survey on Technologies, uses and Challenges of IoT",International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 8 Issue 12, 2019.
29. Alcalde-Barros, A., García-Gil, D., García, S., & Herrera, F.,"DPASF: a flink library for streaming data preprocessing,"*Big Data Analytics*, 4(1). doi:10.1186/s41044-019-0041-8, 2019.
30. Gupta, S., & Gupta, A. Dealing with Noise Problem in Machine Learning Data-sets: A Systematic Review." *Procedia Computer Science*, 161, 466–474. doi:10.1016/j.procs.2019.11.146, 2019.