Latest Tools for Data Mining and Machine Learning

Kanupriya Verma, Sahil Bhardwaj, Resham Arya, Mir Salim Ul Islam, Megha Bhushan, Ashok Kumar, Piyush Samant

Abstract*: *Nowadays, Data Mining is used everywhere for extracting information from the data and in turn, acquires knowledge for decision making. Data Mining analyzes patterns which are used to extract information and knowledge for making decisions. Many open source and licensed tools like Weka, RapidMiner, KNIME, and Orange are available for Data Mining and predictive analysis. This paper discusses about different tools available for Data Mining and Machine Learning, followed by the description, pros and cons of these tools. The article provides details of all the algorithms like classification, regression, characterization, discretization, clustering, visualization and feature selection for Data Mining and Machine Learning tools. It will help people for efficient decision making and suggests which tool is suitable according to their requirement.

Keywords*: *Data mining, Open source tools, Licensed tools, Machine learning

I. INTRODUCTION

Data Mining (DM) is the procedure for programmed revelation of abnormal state learning by acquiring data from the genuine world, extensive as well as complex informational indexes. It is advancement towards more extensive process, called Knowledge Discovery Databases (KDD) [1]. It is a process of finding naturally occurring information from databases which presents an exceptionally alluring and testing assignment, both for the scholarly world and industry. DM and ML tools are used to find the best suitable model through mechanized procedures (called machine realizing) which seek through the dataset to distinguish designs.

DM and Machine learning (ML) deals with various techniques like regression, classification, visualization and feature selection. Regression and classification techniques are categorized as supervised as well as unsupervised learning. Classification is used to predict the class labels, so it can be used to categorize various datasets. It is based on the model of applying mapping function on the dependent variable that can be used to predict the independent variable [2]. On the other hand, regression can be applied to continuous data instead of discrete data as in classification. Further, it can be classified as linear regression based on a single independent variable whereas polynomial regression is based on multiple independent variables. Therefore, DM and ML goes hands in hands to achieve a system to produce the program based on the input, statistical analysis and the predicted outcomes [3].

This paper an overview of the existing tools and technologies used for DM and ML. A description of open-source and licensed tools is provided based on the types of data that can be mined along with the respective application domain where they can be used.

The organization of the paper is as follows: Section II focuses on the description of DM and ML tools. Section III includes related work based on tools. Section IV incorporates a descriptive study of tools. Finally, conclusion is outlined in Section V.

II. TOOLS DESCRIPTION

This section focuses on description of DM and ML Tools.

D2K (Data to Knowledge) [2] toolbox gives a visual programming condition and a lot of layouts expected to associate it with other standard bundles. It gives bundles to perform picture and content mining and furthermore, offers an outside arrangement of transformative strategies for building up some essential hereditary calculations.

KNIME (Konstanz Information Miner) [4] is easy to use, secluded and provides an open-source information coordination, preparation, examination and investigation stage. KNIME contains devices for information pre-handling, changing, grouping, affiliation leads etc. The benefit of the tool is that WEKA can be coordinated and broaden for conceivable outcomes with KNIME by different administrators.

WEKA (Waikato Environment for Knowledge Analysis) [5] is a ML tool. It consists of all ML algorithms which are used to solve the real-life application problems.

RapidMiner (RM - some time ago YALE) [6,7] is a free, adaptable and open-source tool executed in Java. It is a tool for ML, DM, image processing and business analytics.

ORANGE [8] uses the Python language which helps in the visualization of data in DM. It helps in predictive modelling, analysis, selection of subset and empirical analysis. This tool performs tasks like data manipulation and data transformations.

KEEL [2] is an open source software. It stands for knowledge extraction based

Revised Manuscript Received on June 04, 2019.

Kanupriya Verma, Deepak Kumar, **Resham Arya**, **Mir Salim Ul Islam**, **and Megha Bhushan*, Ashok Kumar**, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India.
Piyush Samant, Chandigarh University, Punjab India.

on evolutionary learning. It is a JAVA tool which is used for tasks like data discovery and knowledge extraction. Further, dataset repository is available for performing classification and regression technique.

SciKit Learn [9] is supported by python language, grouped with NumPy and SciPy. It is used for plotting outlines and performing DM calculations.

III. LITERATUTE REVIEW

This section includes articles related to the DM and ML tools.

DM tool is presented as open source software with three features [2]. The key points of this tool include the dataset of KEEL which is a repository that includes the partitions of information sets in KEEL format. In this dataset, results of some algorithms are shown. This tool provides the guidelines for using new algorithm. As KEEL is not dependent on any operating system, so it can be used by anyone. This study concludes that Hider method is the best method with respect to the other methods used for analysis.

RapidMiner is a tool that presents few characteristics of extraction operators and individual operations applied on the extension [10]. It is formerly known as YALE, available for data analysis as a stand-alone application. It can also be integrated with other applications as DM engine. It can run on major platform or operating system, since it is a flexible, free and Open Source platform which is implemented in Java.

Mikut and Markus have discussed various historical developments and presented wide range of current DM and related tools for supporting decision making process [2]. Nine various types of tools are presented in their work. These tools are BIS, DMS, INT, mats, RES, EXT, libs, sols and specs. These vary in characteristics, for example, intended user groups, data structure, implemented tasks and methods, interaction styles such as export and import abilities, license policies and platforms are adjustable. Large dataset with single feature, unstructured data-like texts and time series can be managed using current tools as well as in absence of comprehensive and powerful DM tools for datasets which are multidimensional such as videos and images.

Various algorithms of clustering have been discussed using by DM using WEKA tool [5]. The main key points include explaining the comparison of various algorithms for clustering of WEKA and concluded the best algorithm for users. This tool was chosen by the user because it can be used without having a detailed knowledge of DM techniques. They have worked only on the clustering algorithm using WEKA tool.

Hirudkar and Shereka have given an evaluation of database systems and comparative analysis of DM techniques and tools [4]. An overview is provided with steps included in mining data and methods. A comparative study of freely available tools such as WEKA tools, RapidMiner Tool and NetTool Spider for web mining has been provided. Further, predicted the behavior's and future trends that help organizations to make heedful knowledge driven decisions. Software becomes highly robust for various users using WEKA tool. RapidMiner tool is used by the users having programming skills. KNIME tool provides aid to the users without the knowledge of programming skills. For web mining purposes, NetTool Spider mining tool is used.

MATLAB and TANGARA have been used for a comparative study of classification techniques [11]. The performance of different classification techniques is analyzed for set of data. Medical diagnosis is an important factor for obtaining the important parameters of the disease as without diagnosis it is difficult to identify parameter of disease. Tests are conducted which include clusters and classifications techniques. However, many tests could complicate the process of diagnosis and it would be difficult to obtain results. Thus, to overcome ML tools are used. The extraction of information from large datasets and the correlation of an element in data set will help to analyze the results. Fuzzy proposition establishes some sort of relation between input and output fuzzy set using fuzzy logic. The decision rules are implemented for control output value and input parameters to find the result of a diabetic person. The result could be negative or positive.

A theory has been proposed on Unified DM with analysis of DM tools, as there is a huge amount of data which has been stored in repository, cloud or databases [12]. There is a need to evaluate an efficient data pattern for decision making. So, for predicting the best patterns out of many datasets, the tools are required for different data types. It also provides knowledge on unification theory. For development of unification process some sort of measure was suggested which could be used on set of database and domains. This process performed all tasks of mining, classification, clustering and visualization in group or in unified method instead of performing each task individually. The four algorithms were also used i.e., zero rule, one rule, decision tree and KNN (K-nearest neighbor). The tools offer the Functionalities like API support and Graphical Presentation. The algorithm applied over the dataset and percentage accuracy was served for measuring performance. WEKA is better as it provides zero's and one's implementation.

A comparison of various open source DM tools has been presented in [13]. Their work described the technical specifications, features, and specialization for each selected tool along with its applications. By employing this study, the choice and selection of tools can be made easy.

The details of the open source tools have been provided for supporting the more advanced and specialized research topics like big data, data streams, text mining, etc. [14].

A comparison of DM algorithms and techniques like clustering, visualization has been given [8]. A comparison has been made for tools with respect to community support. In DM tools, advancements were made and it gives the features and quality of WEKA and KNIME. Some of the tools such as RapidMiner and KNIME are graphically integrated which help to enable connection, dragging and component placement. Structured view of all supported functionalities is offered by Orange Tool which was grouped into different categories i.e., unsupervised learning, prototype implementation, visualization using Qt, data operation and classification.

The efficiency of tools can be improved.

A methodology has been proposed to monitor the plan and execution of Crime location and identified the criminals in Indian urban areas utilizing DM strategies [15]. Their methodology is separated into six modules: Data Extraction (DE), Grouping, Data Pre-processing (DP), Google Delineate, Classification and WEKA execution. DE extricates the unstructured and undefined criminal dataset from different crime Web Sources. DP cleans, incorporates and lessens separated criminal info to organize number of criminal occurrences. They have resolved these cases utilizing 35 predefined criminal characteristics. Rest four modules were helpful to identify the Crime location, identification of the criminal and expectation, and verification of the crime, separately. Criminal identification and expectation were cracked by utilizing KNN classification. Crime verification is done by the results generated by WEKA. The proposed scenario improved the public lifestyle by helping the authorities in crime discovery as well as identification of criminals and hence, diminishing the crime rates.

A Binary Classifier tool was used for the diagnosis of patients suffering from brain disorders [9]. This tool provided a nonexclusive classifier to help and analyze patients experiencing cerebrum issue. They have manufactured a tool which utilizes ML taking in calculations from WEKA, Caret and SciKit Learn from Java, R and Python separately and joins the three bundles into one R bundle which helps in arranging the patients experiencing cerebrum issue. This tool can be utilized as an independent application for arrangement of any paired class information.

The compounds have been evaluated for their pharmacological and toxicological properties which are of extraordinary significance for industry and administrative offices [16]. In this investigation, a methodology utilizing open source programming and open access databases to assemble screening devices for receptor-interceded impacts is introduced. The retinoic corrosive receptor (RAR), as a pharmacologically and toxicologically applicable target, was chosen for this examination. RAR agonists were utilized in the treatment of various dermal conditions and explicit kinds of malignancy, for example, intense promyelocytic leukemia.

The Source ext Rewriting has been used to improve the quality of Machine Translation (MT) [17]. It has been characterized, the undertaking of transformation of substance starting with one dialect then onto the next. In Indian society, interpretation started with the interpretation of Holy Scriptures into Pali, Prakrit, Devanagari and other local dialects. It helped in transmission of good qualities, ethos, custom, convictions and culture over the globe. Indeed, even at present, when web has united individuals near one another, the job of interpretation has turned out to be much huger than past years.

A review has been presented on DM tools which are used to mine educational dossier [18]. Their work focused on EDM (educational DM tool) to perform EDM analysis rather than more traditional or modern statistical analysis. Emerging methods can be reviewed not only at theoretical level but at practical level also. Analyzing data sets from beginning to end, different tools are uniquely defined for different tasks, for example, SQL is used to select data for only particular month or year, EXCEL is used to refine data set and to calculate total employee time before fitting a predictive model to RapidMiner. NodeXL analyzed relationship between posts and all-over textual quality of posts replied by that employee with the help of Coh Metrix and finally with Gephi tool. Researchers can visualize the most interesting clusters of employees found within the social network. Each tool has its own strength and weakness. Efficient discoveries can be made using a combination of these tools.

The components of microstructure of compact graphite iron based on alloying elements have been identified to find the thickness and effect [19]. Linear regression models, Segmented regression models with MAR Splines algorithm, Artificial Neural Network (ANN), Classification and Regression Tree (CART) were used for conducting this study.

Siddique and Ahmad have stated that everything is going to be computerized in the present era of software [20]. For software organizations, it is challenging task to develop standard software within estimated cost on time. DM plays a highly important role in mining software repositories using tools like Apfel, Chianti, Dynamine, Hipikat, Kenyon and Softchange. The dimensions of these tools are to be intended, informative, infrastructure, effective, interactive, materialistic and language dependent.

The articles from 2007-2017 years under Fundamental Concepts of DM, KR (Knowledge Representation), CI (Computational Intelligence), Classification and Predication have been reviewed in [21].

Kodati and Vivekanandam have presented a paper on Orange and WEKA tools of DM for analyzing heart disease [7].

IV. VARIOUS DM AND ML TOOLS

This section provides a description of open source and licensed tools of DM and ML.

Fig. 1 shows various Open Source and Licensed Tools. Tables 1 and 2 describe Open Source tools and Licensed Tools, respectively. WEKA is the most efficient tool for the educational purpose and frees to use. Yellowfin tool is the best tool according to this descriptive study due to its quick response, simple to use and highly capable for Big Data integration with excellent capacity.

V. CONCLUSION

DM is one of the most popular techniques used for Information retrieval and for better decision making. Till date, various open source and licensed tools like WEKA, Rapid Miner, KNIME, Orange and many more have been developed for generating predictions. This review paper is focused on various available tools for DM and ML. A study

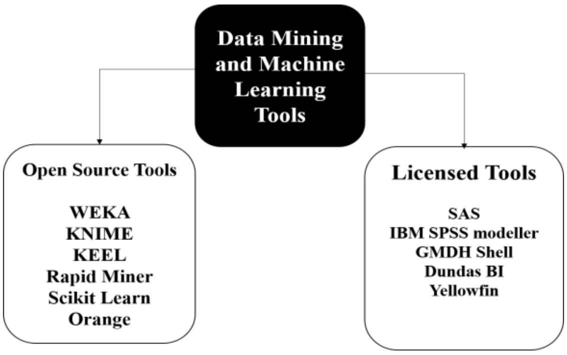

Fig. 1 Data Mining and Machine Learning Tools

Table 1. Description of various open source tools for DM and ML

Tool Name	**Year**	**Latest Version**	**Language used**	**Method Purpose**	**Application/ Area**	**Advantages**	**Disadvantages**
WEKA [5]	1997	3.6.11	Java	-Easy analytics of data and predictive modelling	Educational Purposes	-Free Extensible -ARFF, CSV, C4.5, binary are formats used to load files	-Weak in statistical analysis -For parameter optimization of machine learning (No automatic facility)
KNIME [2]	2004	2.9	Java	-Enables user to visually create data flows easily -Interactive data models	Pharmaceutical Research	-Easily visualization of molecular data	-No methods for data wrapper -Not automatic facility for parameter Optimization
KEEL [20]	2004	2.0	Java	-Evolutionary Algorithms for DM big problems	Used in Scientific research	-Contains big data preened libraries for analysis, processing prediction	-Less efficient due to large numbers of algorithms.
Rapid Miner [7]	2006	6.0	Java	-Text Mining, results in visualization, Model validation and optimization	Business, Training, Education, Data	-Full Faculty Model Evaluation Offers more procedures Over 1500 methods for data integration, analysis, visualization Compatible for large users.	-Only capable of SQL statements -Working with only database files.
SciKit Learn [2]	2007	0.14.1	Python	-Add on machine learning package	Machine Learning	-Include some libraries that are suitable for audio/video files	-Time Consumption -Less durable
Orange [20]	2009	2.7	C++ and python	-Data Pre-processing, filtering, and modelling Techniques	Data Visualization	-Debugging is better -Categorization problems like scripting DM are simple.	-Weak in statistical analysis -Limited capabilities of visual representations of data mode

Table 2. Description of various licensed tools for DM and ML

Tool Name	**Language**	**Main purpose**	**Application Area**	**Year**	**Latest version**	**Advantages**	**Disadvantages**

SAS [22]	HTML	Extraction, Formatting and cleansing to data analysis, building sophisticated models	Clinical Research And forecasting	1976	9.4	-Drag and drop interface is great -Very accurate -With good graphic design & Speed of processing.	-Although software is perfect but requires some work to create a model with incoming, read -time data
IBM SPSS modeller [22]	Python/ Java	Interactive and Statistical analysis	Forecasting, Healthcare, Risk Management	2010	25	-Not necessary to use complex Knowledge to encode data when we use qualitative data.	-To have better quality graphics or mainly we want to make. -Presentation it's not suitable
GMDH Shell [22]	C/C++	Knowledge discovery, prediction, complex system modelling, optimization	Forecasting and Business purpose	2009	3.8.3	-Takes raw or messy data set in CSV and provide a predictive mode more Quickly at reasonable cost	-With value very close to absolute max, the model does not converge to max but to local max.
Dundas BI [23]	C, C#, Java, C++	creating and viewing interactive dashboards, reports, scorecards	Small Businesses, Mid-size Business, Enterprise	1992	6.0.0 Revision 3	- On any device it allows users to integrate and connect with data source in real time.	-DO not have an interactive user community -tool doesn't support the 3D Charts
Yellowfin [23]	JAVA	Creating scorecards and dashboards, online analytical processing analyses, predictive analytics	Counting, advertising, agriculture, banking, insurance, manufacturing, media, marketing	2003	7.4.7	-Very fast and very simple product to create reports and panels -Integration capacity with Big Data is excellent.	-There are issues related to pulling data and displaying data set from a system.

on these tools has been done to study their respective pros and cons along with the application areas where they can be more beneficial. This review provides details of all the algorithms like classification, regression, characterization, discretization, clustering, visualization and feature selection for DM and ML tools. It will benefit people in efficient decision making and suggest which tool is more appropriate as per their requirement. There are always different approaches used to solve different problems, each with their own particular strengths and weaknesses. Therefore, by using a combination of aforementioned tools and algorithms, complex analyses can be done by future researchers which will result in useful discoveries on data.

REFERENCES

1. U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," in *AI Magazine*, 1996, vol. 17, no. 3, pp. 37–54.
2. J. Alcala-Fdez et al. "KEEL: a software tool to assess evolutionary algorithms for data mining problems," in *Soft Computing*, 2009, vol. 13, pp. 307-318.
3. R. Mikut, and R. Markus, "Data mining tools," in *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2011, vol. 1, pp. 431-443.
4. A. M. Hirudkar, and SS. Sherekar, "Comparative analysis of data mining tools and techniques for evaluating performance of database system," in *International Journal of Computer Science Appllications*, 2013, vol. 6, pp. 232-237.
5. N. Sharma, A. Bajpai, and R. Litoriya, "Comparison the various clustering algorithms of weka tools," in *facilities*, 2012, vol.4, pp. 78-80.
6. https://rapidminer.com/glossary/data-mining-tools/
7. Q. M. Yas, A.A. Zaidan, B.B. Zaidan, B. Rahmatullah and H.A. Karim,"Comprehensive insights into evaluation and benchmarking of real-time skin detectors: Review, open issues & challenges, and recommended solutions," in *Measurement*,2018, vol. 114, pp. 243-260
8. A. Jovic, B. Karla, and B. Nikola, "An overview of free software tools for general data mining," in IEEE *37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2014, pp. 1112-1117.
9. N. Borude, C. Maher, V. Sarda, and A. Santra, "Generic binary classifier tool for diagnosis of patients suffering from brain disorders in R," in *International Conference on Computing, Analytics and Security Trends (CAST)*, 2016, IEEE pp. 173-178.
10. B. Radim, K. Jan, S. Zdeněk, U. Václav, and D. Otto, "Rapidminer image processing extension: A platform for collaborative research," in *33rd International Conference on Telecommunication and Signal Processing*, TSP, 2010, pp. 114-118.
11. R.M. Rahman, and F. Afroz, "Comparison of various classification techniques using different data mining tools for diabetes diagnosis," in *Journal of Software Engineering and Applications*, 2013, vol. 6, p. 85.
12. H. Solanki, "Comparative study of data mining tools and analysis with unified data mining theory," in *International Journal of Computer Applications*,2013, vol. 75, pp. 23-28.
13. K. Rangra, and K. L. Bansal, "Comparative study of data mining tools," in *International journal of advanced research in computer science and software engineering*, 2014, vol. 4.
14. X. Wu, X. Zhu, G.Q. Wu, and W. Ding, "Data mining with big data," in *IEEE Transactions on Knowledge and Data Engineering*, 2013, vol. 26, pp. 97-107.
15. DK. Tayal, A. Jain, S. Arora, S. Agarwal, T. Gupta, and N. Tyagi, "Crime detection and criminal identification in India using data mining techniques," in *AI & society*, 2015, vol. 30, pp. 117-127.

16. FP. Steinmetz, CL. Mellor, T. Meinl, and MT. Cronin, “Screening Chemicals for Receptor-Mediated Toxicological and Pharmacological Endpoints: Using Public Data to Build Screening Tools within a KNIME Workflow,” in *Molecular informatics*, 2015, vol. 34, pp. 171-178.
17. D. Chopra , N. Joshi , and I. Mathur, “Improving Quality of Machine Translation Using Text Rewriting,” in *Computational Intelligence & Communication Technology (CICT),* Second International Conference on IEEE, 2016, pp. 22-27.
18. S. Slater, S. Joksimovic, V. Kovanovic, RS. Baker, D. Gasevic, “Tools for educational data mining: A review,” in *Journal of Educational and Behavioral Statistics*, 2017, vol. 42 ,pp. 85-106.
19. D. Wilk-Kolodziejczyk, K. Regulski, G. Gumienny, B. Kacprzyk, S. Kluska-Nawarecka, K. Jaskowiec, “Data mining tools in identifying the components of the microstructure of compacted graphite iron based on the content of alloying elements,” in *The International Journal of Advanced Manufacturing Technology*, 2018, vol. 95, pp. 3127-3139.
20. T. Siddiqui, and A. Ausaf, “Data mining tools and techniques for mining software repositories: A systematic review,” in *Big Data Analytics. Springer, Singapore*, 2018, pp. 717-726.
21. R. Alcalá, MJ. Gacto, J. Alcalá-Fdez “Evolutionary data mining and applications: A revision on the most cited papers from the last 10 years (2007–2017),” in *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2018 ,vol. 8 , e1239.
22. S. Yefimenko, “Advances in GMDH-based Predictive Analytics Tools for Business Intelligence Systems,” in *International Conference Proceedings, ACIT, 2018*, pp. 254-257.
23. https://www.softwareadvice.com/bi/dundas-bi-profile/

AUTHORS PROFILE

Kanupriya Verma is presently pursuing Masters of Engineering from Chitkara University, Punjab Campus and also working as an Assistant Lecturer. She has done her Bachelor of Technology from Punjabi University, Punjab, India. She has also done diploma from Government Polytechnic College for Girls, Punjab, India. She has published one paper in National Conference.

Sahil Bhardwaj is presently pursuing Masters in Computer Science and Engineering from Chitkara University Punjab and also working as an Assistant Lecturer. He has done Bachelors of Engineering in CSE from Chitkara University. He has industrial experience in Android app Development.

Resham Arya is presently a Ph. D Scholar in Computer Science and Engineering department from Chitkara University, Punjab. She has done Masters of Engineering in CSE from Chitkara University and Bachelor of Technology from Chitkara Institute of Engineering and Technology. She has 3 years of work experience as an Assistant Lecturer and 1.5 years of Assistant Professor in Teaching. Her research interests include affective computing, machine learning and working with physiological signals.

Mir Salim Ul Islam is presently pursuing Ph. D in Computer Science and Engineering from Chitkara University Punjab and also working as an Assistant Professor - Research in Chitkara Research Innovation Network. He has done Masters of Engineering in CSE from Kurukshetra University Haryana and Bachelor of Technology in CSE from Kashmir University. His research areas include physiological signals, machine learning and deep learning. He has industrial experience of 4 years in software and solution development and has hands on experience in Microsoft Dot Net technologies, SQL server and open source technologies.

Dr. Megha is an Assistant Professor in the Department of Computer Science & Engineering, Chitkara University, Punjab, India. She is a recipient of Grace Hopper Celebration India (GHCI). She was awarded with fellowship by University Grants Commission (UGC), Government of India, in 2014. In 2017, she was a recipient of Grace Hopper Celebration India (GHCI), fellowship. She has worked as Junior Research Fellow under UGC, New Delhi, Government of India from 2014 to 2016. She has also worked as Senior Research Fellow under UGC, New Delhi, Government of India from 2016 to 2018. Dr. Megha has published many research articles on the area of software reuse in international journals and conferences of repute. Her research interest includes Software quality, Software reuse, Ontologies, Artificial Intelligence, and Expert systems. After obtaining Bachelors of Technology degree in Information Technology from Himachal Pradesh University, India in 2010, she obtained her Masters in Engineering degree with specialization in Software Engineering from Thapar University, India in 2012. She continued her research in software product line combined with expert systems and ontologies obtaining her Ph.D. in 2018 from Thapar University, Punjab, India. She is also the reviewer and editorial board member of many international journals.

Ashok Kumar is currently an Assistant Professor in Chitkara University Research and Innovation Network (CURIN) Department, Punjab. He is PhD in Computer Science and Engineering from Thapar University, Punjab, India. He has 15+ years of teaching and research experience. He has number of publications in International Journals and Conferences of repute. His current areas of research interest include Cloud Computing, Internet of Things, and Mist Computing. His teaching interest includes Python, Haskell, Java, C/C++, Advanced Data structures and Data Mining.

Retrieval Number: I10030789S19/19©BEIESP
DOI: 10.35940/ijitee.I1003.0789S19