A Modified DBSCAN Algorithm for Anomaly Detection in Time-series Data with Seasonality

Praphula Jain, Mani Shankar Bajpai, and Rajendra Pamula
Department of Computer Science and Engineering, Indian Institute of Technology (Indian School of Mines), India

Abstract: *Anomaly detection concerns identifying anomalous observations or patterns that are a deviation from the dataset's expected behaviour. The detection of anomalies has significant and practical applications in several industrial domains such as public health, finance, Information Technology (IT), security, medical, energy, and climate studies. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) Algorithm is a density-based clustering algorithm with the capability of identifying anomalous data. In this paper, a modified DBSCAN algorithm is proposed for anomaly detection in time-series data with seasonality. For experimental evaluation, a monthly temperature dataset was employed and the analysis set forth the advantages of the modified DBSCAN over the standard DBSCAN algorithm for the seasonal datasets. From the result analysis, we may conclude that DBSCAN is used for finding the anomalies in a dataset but fails to find local anomalies in seasonal data. The proposed Modified DBSCAN approach helps to find both the global and local anomalies from the seasonal data. Using normal DBSCAN, we are able to get 19 (2.16%) anomaly points. While using the modified approach for DBSCAN, we are able to get 42 (4.79%) anomaly points. In comparison, we can say that we are able to get 2.21% more anomalies using the modified DBSCAN approach. Hence, the proposed Modified DBSCAN algorithm outperforms in comparison with the DBSCAN algorithm to find local anomalies.*

Keywords: *Anomaly detection, data mining, DBSCAN, modified DBSCAN, seasonal data, time series.*

Received October 18, 2019; accepted January 19, 2021
https://doi.org/10.34028/iajit/19/1/3

1. Introduction

Anomaly detection is the task of finding patterns that deviate from the expected behaviour of the data [4]. It can be applied to various problems such as intrusion detection systems, military surveillance, finding instances of fraud for credit cards, health care, insurance, and fault detection in safety-critical systems before any major harm occurs [4]. Their ability to translate the detected anomaly into actionable information makes it important for data analysis. For instance, sensitive data being sent out from a hacked computer to an unauthorized destination could lead to abnormal traffic patterns in the computer network, its detection can prevent further damage [27]. The abnormality in an MRI image might be symptomatic of malignant tumors [24]. The aberrant readings from sensors of a spacecraft could indicate some flaws in its components. The outliers could identify a credit card or identity theft in a credit card transaction [16].

In recent years, there has been an exponential increase in the availability of time-series data [15]. Real-time data sources such as sensors connected to the enormous number of applications and the rise of the Internet of Things (IoT) have helped produce data that vary with time. Analyzing of these data can effectively provide insights important for any application or use case [19].

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is one of the most common clustering algorithms and is most cited in the scientific literature [3]. The DBSCAN algorithm mostly related clusters as a dense area of events in the data samples isolated by low-density regions. Due to the advantages of the DBSCAN algorithm, many researchers considered it for their research purpose. A unique algorithm based on DBSCAN for anomaly detection is presented in [9]. Blockchain technology is a recent research trend, and a researcher used the DBSCAN algorithm for anomaly detection in the Bitcoin price [8].

In this paper, a modified DBSCAN is proposed for identifying anomalies in a seasonal time-series dataset. For instance, in the monthly temperature data, DBSCAN would be able only to detect global anomalies, perceiving the data as a whole, but would fail to identify local anomalies, i.e., for individual months. Since the temperature data would have a strong seasonal component, one viable approach could be deseasoning the data in the pre-processing step to eliminate the seasonal part, as explained in [26]. The modified approach adds additional attributes to each instance of the dataset. The features are added based on a circular coordinate system that repeats after a given interval. Further, with the optimized value of the parameters, the results for our modified approach is

obtained. The experimental result helped to find out the global as well as local anomalies. Modified DBSCAN outperforms compared with the standard DBSCAN method for anomaly detection in seasonal data.

We have compared our algorithm output with the DBSCAN algorithm and mentioned it as a comparison. As such, no related work has been implemented on the same datasets, so we did not include a comparison with the existing paper. We first implemented DBSCAN, then implemented our proposed algorithm, and comparisons of both algorithms are highlighted properly.

The rest of the paper is organized as follows: Section 2 presents the different approaches researchers use in anomaly detection with a brief literature review. Section 3 presents the DBSCAN algorithm, and section 4 describes the proposed modified DBSCAN approach. Section 5 shows the obtained experimental results analysis. Finally, section 6 finalizes this paper with a conclusion, limitations, and future research potentials.

2. Literature Review

In the literature, for anomaly detection, there are different techniques employed by researchers [4]. Anomaly detection uses both supervised, for example, support vector machines or decision trees [4], or unsupervised (e.g., clustering) data mining techniques depending on the various domain and use cases. Anomaly detection in time-series data has been a heavily studied research topic in the machine learning and data science domain [12].

Conventionally, concerning an expected behaviour model, the events with low probability are considered anomalies [21], whereas this task is performed by a maximum posterior estimation in [20]. A probabilistic, nonparametric approach using a squared-loss objective function for anomaly detection is proposed in [23]. Deep learning is a neural network-based computational model composed of multiple processing layers that learn data representation with various abstraction levels [18]. A convolutional autoencoder for automated video surveillance by applying the reconstruction error of each frame as an anomaly score and proposing a method for aggregating high-level spatial and temporal features input frames is presented in [25]. Marchi *et al*. [22], described denoising autoencoder that uses Bidirectional Long Short-Term Memory (Bi-LSTM) to process auditory spectral features to LSTM can maintain long-term memory, and stacking recurrent hidden layers in such networks enables the learning of higher-level temporal features for faster learning with sparser representations and can be used for anomaly detection in time-series [11] and internet traffic [6, 28].

In the industry, Netflix uses its Robust Principal Component Analysis (RPCA) method [1]. For scalability, Yahoo Extensible Generic Anomaly Detection System (EGADS) [17] deploys an anomaly filtering layer on a series of anomaly detection and forecasting models for identifying anomalies on time-series data. Both require analyzing the complete dataset, whereas, prior to anomaly detection, Symbolic Aggregate Approximation (SAX) [7] generates symbols by decomposing the full-time series. Akouemo and Povinelli [2], employed a probabilistic approach with anomalies being discovered using a linear regression model with weather inputs, and a Bayesian classifier tested the anomalies for false positives. Ahmad *et al*. [1] and Pimentel *et al*. [23], authors propose an unsupervised algorithm for anomaly detection in real-time online streaming data to tackle real-time anomaly detection challenges. Few more studies related to anomaly detection are found in [5, 10, 13, 14].

2.1. Comparisons

The dataset contains monthly average data of 73 years, and thus we have ~876 data points. Using normal DBSCAN, we are able to get 19 (2.16%) anomaly points. While using the modified approach for DBSCAN, we are able to get 42 (4.79%) anomaly points. In comparison, we can say that we are able to get 2.11% more anomalies using the modified DBSCAN approach.

3. DBSCAN Algorithm

Density-based spatial clustering of application with noise, DBSCAN [13] is a data clustering algorithm that forms clusters with a maximal set of density-connected points. Clusters in the data space are typically high-density regions separated by lower object density regions.

DBSCAN defines the density in terms of the following:

1. ε-Neighborhood: Objects within a radius of ε (eps) from an object and can be represented by the relation,

$$N_{\epsilon}(p): \{q \mid d(p, q) \leq \epsilon \quad (1)$$

Where *p, q* are data points in the space and *d*(*p, q*) represents the separation between the data points.

2. High density: ε-Neighborhood of an object containing at least minpts of data points.

The algorithm requires two parameters: the neighbourhood distance ε (eps) and the minimum number of the points needed to form a high-density region minpts. The parameters categorize the data points as core points, border points, and outlier points. A core point has more than minpts number of points within the ε (eps) distance and lies at the cluster's interior. A border point is in the neighbourhood of a core point but has fewer than minpts number of points within eps. Outlier points are the anomalous points

that are neither a core point nor a border point and do not fit any cluster.

The DBSCAN algorithm works as follows. An arbitrary point that has not been visited yet is selected, and its ε-neighborhood is retrieved. If the number of neighborhood points is greater than the minpts, a cluster is started; else, the point is marked as noise. If the point being noise is later found to lie in the ε-neighborhood of some other point with apt size, it would be made part of that cluster. If a point lies in a cluster's high-density zone, then its ε-neighbourhood is also a part of that cluster. All points found within the ε-neighbourhood are added to the cluster, as is their own ε-neighborhood if they are dense until it is found that the density-connected cluster is complete. Again, an unvisited point is retrieved and processed as stated above, leading to the determination of a further cluster or noise.

4. Proposed Modified DBSCAN Algorithm

In a time-series dataset with a seasonal component, DBSCAN works well to detect global anomalies but fails to identify the local anomalies. The dataset under consideration is a monthly average temperature data having a strong seasonal component. The global anomalies, in this case, would refer to the anomalies of the data as a whole, and the local anomalies depict the outliers for a particular month. To find local anomalies as well, the DBSCAN is modified to add some features to the dataset. The average monthly temperature data would be having a period of 12. Using this, each data point can be assigned additional coordinates projecting the 1-dimensional data to 3- dimensional domain. It is better represented in Figure 1.

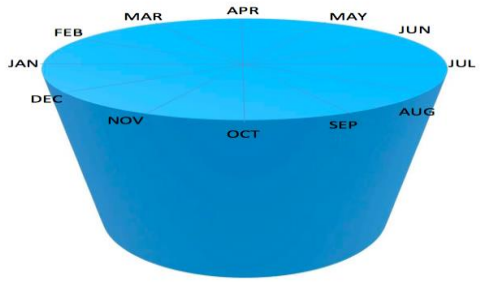


Figure 1. Data points conversion.

Applying DBSCAN to this modified data helps us to find local anomalies of the dataset. With the modification, neighbour months are equidistant, and DBSCAN can be applied individually to identify anomalies of each month. The pseudo-code for the modified DBSCAN algorithm is as below.

5. Analysis of Experimental Results

This section presents the evaluation of the results obtained with the modified DBSCAN on the monthly average temperature dataset and its comparison with the normal DBSCAN algorithm. For the results, seasonality is taken into consideration. With the help of additional coordinates, as stated in the above section, an attempt is made to detect the data's local month-wise anomalies.

5.1. Data Set Description

The dataset used in this study consists of monthly average temperature data collected at Will Rogers World Airport (U.K.). The data were converted to monthly averages by averaging daily values for each month. The data were available for 73 years (from 1938 to 2011).

5.2. Anomaly Detection Using DBSCAN Algorithm

While applying normal anomaly detection, minpts are taken as 4, and eps are taken as 0.3. Experimental results obtained by applying DBSCAN Algorithm are illustrated in Figure 2. The no. of anomalies varies with different minpts and eps values. The same values for these parameters are taken for the normal DBSCAN algorithm to solve this issue. The modified algorithm as DBSCAN is a distance-based clustering algorithm that also helps determine anomalies in the data.

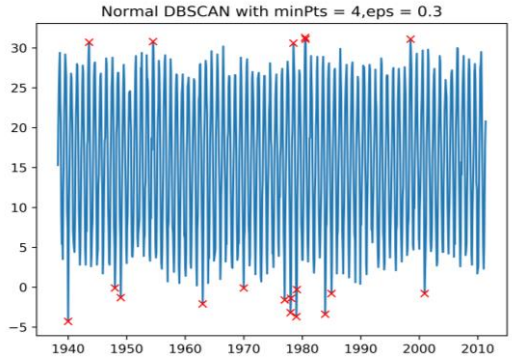
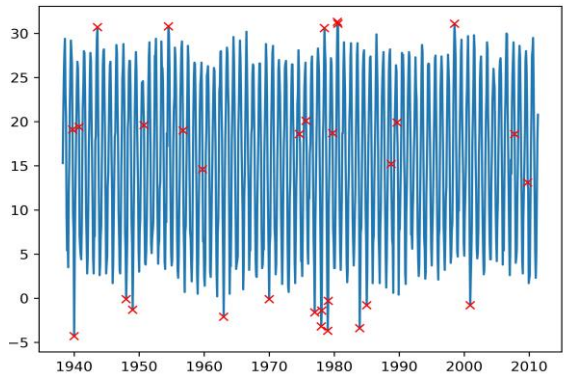


Figure 2. Experimental results over DBSCAN Algorithm.

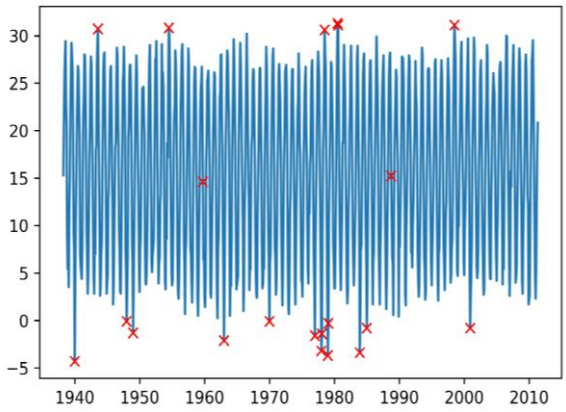
5.3. Anomaly Detection Using Proposed Modified DBSCAN Algorithm

The main aim of this modified DBSCAN algorithm is to determine local anomalies in the data. As can be seen, the normal DBSCAN algorithm only gives anomalies that have extreme values. As only the extreme anomalies can be identified, the anomalies of months with values lying in midrange temperature cannot be detected, which is a reasonable possibility. For example, in September, temperature lays in the range of 21-26 ^{0}C. But when the temperature is below or above that range, there may be a chance of anomaly. This modified approach would help find those points that lie in the midrange but significantly deviate from their respective month values.

The result is shown for different radius values of cylindrical coordinates. When radius value is low, adjacent months are highly dependent on each other, and they easily form clusters. But for a higher value of the radius, months and their neighbours are separated at equal distances. The result is shown in Figures 3-a) and 3-b) for different values of radius as follows. This helped to find anomalies of that particular month and also consider the influence of the neighbour months.



a) MinPts=4, eps=0.3,Radius=0.2.

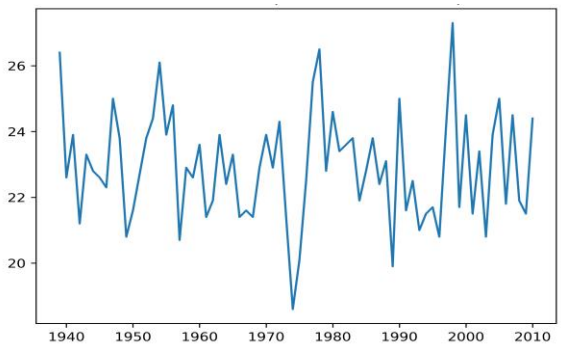


b) MinPts=4, eps=0.3,Radius=0.15.

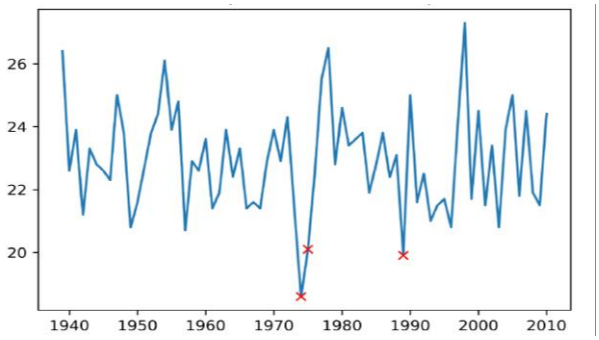
Figure 3. Result analysis using proposed DBSCAN algorithm.

The change in radius also changes the result, which can be seen in Figures 3-a) and 3-b). With decrement in the radius's value, the number of anomalies decreases Figure 3-b), and with increment in radius, the number of anomalies increases refer Figure 3-a). For the midrange value of the radius, the desired results are obtained. The result obtained can be better visualized when the results for individual months are analyzed.

The results of individual months present the various aspects of the modified DBSCAN algorithm as shown in Figures 4-a) and 4-b). The main consequence is shown for September using DBSCAN and modified DBSCAN algorithm presented in Figures 4-a) and 4-b). A comparison between Figures 4-a) and 4-b) shows that modified DBSCAN outperforms. But still, they were considered anomalies because their values were not in the range as per their individual month values and neighbours.



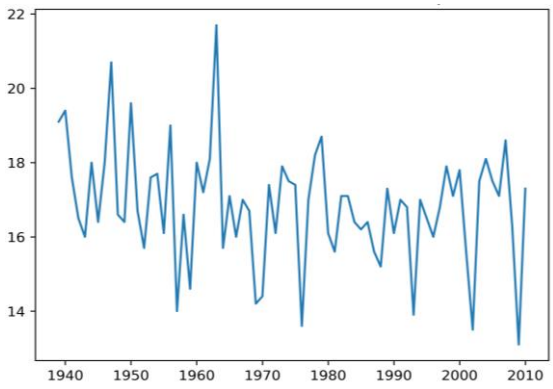
a) Using DBSCAN algorithm with minPts=4, eps=0.3.



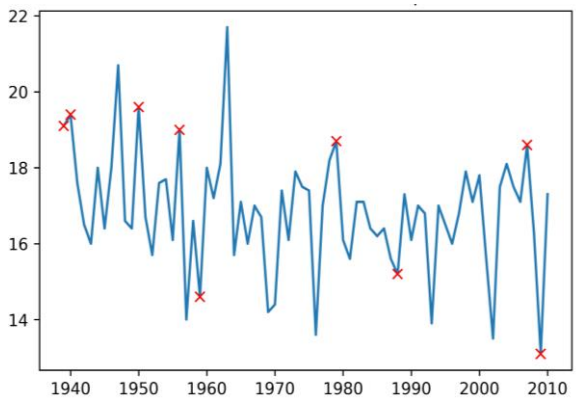
b) Using the proposed DBSCAN algorithm with minPts=4, eps=0.3.

Figure 4. Results analysis for September month.

The proposed algorithm gives better results for all months except October because the month had a high range of temperature values presented in Figures 5-a) and 5-b) This can be considered a limitation of this approach.



a) Using the DBSCAN algorithm with minPts=4, eps=0.3.



b) Using Modified DBSCAN algorithm minPts=4, eps=0.3.

Figure 5. Results analysis of October month.

6. Conclusions

Anomaly detection in time-series data has been an essential task using case cutting across various industries. In this paper, a modified approach for using DBSCAN for seasonal time-series datasets are presented, enabling the algorithm to detect local anomalies and global anomalies. Conventionally, DBSCAN fails to see the abnormalities in data with seasonality, and it requires the data to be free of such trends. Thus, the modified algorithm reduces the tedious task of pre-processing the data to remove the seasonal trends.

This study considered only seasonal data collected from a single source. The proposed approach can be applied to data collected from different sources and various applications in future work. It will increase the applicability of the proposed work to more datasets. Furthermore, this work can be extended for the data which follows linear or nonlinear trends.

References

[1] Ahmad S., Lavin A., Purdy S., and Agha Z., "Unsupervised Real-Time Anomaly Detection for Streaming Data," *Neurocomputing*, vol. 262, pp. 134-147, 2017.

[2] Akouemo H. and Povinelli R., "Probabilistic Anomaly Detection in Natural Gas Time Series Data," *International Journal of Forecasting*, vol. 32, no. 3, pp. 948-956, 2016.

[3] Birant D. and Kut A., "St-dbscan: An algorithm for Clustering Spatial-Temporal Data," *Data and Knowledge Engineering*, vol. 60, no. 1, pp. 208-221, 2007.

[4] Chandola V., Banerjee A., and Kumar V., "Anomaly Detection: A Survey," *ACM Computing Surveys*, vol. 41, no. 3, pp. 1-58, 2009.

[5] Chandola V., Mithal V., and Kumar V., "Comparative Evaluation of Anomaly Detection Techniques for Sequence Data," *in Proceedings of 8th IEEE International Conference on Data Mining*, Pisa, pp. 743-748. 2008.

[6] Cheng M., Xu Q., Jianming L., Liu W., Li Q., and Wang J., "Ms-Lstm: A Multi-Scale Lstm Model for Bgp Anomaly Detection," *in Proceedings of 24th International Conference on Network Protocols*, Singapore, pp. 1-6, 2016.

[7] Devarajan R. and Rao P., "An Efficient Intrusion Detection System by Using Behaviour Profiling and Statistical Approach Model," *The International Arab Journal of Information Technology*, vol. 18, no. 1, pp. 114-124, 2021.

[8] Dokuz A., Celik M., and Ecemi A., "Anomaly Detection in Bitcoin Prices Using Dbscan Algorithm," *European Journal of Science and Technology*, pp. 436-443, 2020.

[9] Emadi H. and Mazinani S., "A Novel Anomaly Detection Algorithm Using Dbscan and Svm in Wireless Sensor Networks," *Wireless Personal Communications*, vol. 98, no. 2, pp. 2025-2035, 2018.

[10] Ester M., Kriegel H., Sander J., and Xu X., "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," *in Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Portland, pp. 226-231, 1996.

[11] Feng C., Li T., and Chana D., "Multi-Level Anomaly Detection in Industrial Control Systems via Package Signatures and Lstm Networks," *in Proceedings of 47th Annual IEEE/IFIP International Conference on Dependable Systems and Networks*, Denver, pp. 261-272, 2017.

[12] Fox A., "Outliers in Time Series," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 3, pp. 350-363, 1972.

[13] Gama J., *Knowledge Discovery from Data Streams*, CRC Press, 2010.

[14] Jain P. and Pamula R., "Two-Step Anomaly Detection Approach Using Clustering Algorithm," *in Proceedings of International Conference on Advanced Computing Networking and Informatics*, Springer, pp. 513-520, 2019.

[15] Jain P., Quamer W., and Pamula R., "Electricity Consumption Forecasting Using Time Series Analysis," *in Proceedings of International Conference on Advances in Computing and Data Sciences*, Dehradun, pp. 327-335, 2018.

[16] Kalid S., Ng K., Tong G., and Khor K., "A Multiple Classifiers System for Anomaly Detection in Credit Card Data with Unbalanced and Overlapped Classes," *IEEE Access*, vol. 8, pp. 28210-28221, 2020.

[17] Laptev N., Amizadeh S., and Flint I., "Generic and Scalable Framework for Automated Time-Series Anomaly Detection," *in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, pp. 1939-1947, 2015.

[18] LeCun Y., Bengio Y., Hinton G., "Deep learning," *Nature*, vol. 521, pp. 436-444, 2015.

[19] Lee I. and Lee K., "The Internet of Things (Iot): Applications, Investments, and Challenges for Enterprises," *Business Horizons*, vol. 58, no. 4, pp. 431-440, 2015.

[20] Li S., Liu C., and Yang Y., "Anomaly Detection Based on Maximum A Posteriori," *Pattern Recognition Letters*, vol. 107, pp. 91-97, 2018.

[21] Liu L., Fan J., Qiao S., Song J., and Guo R., "Efficiently Mining Outliers from Trajectories of Unrestraint Movement," *in Proceedings of 3rd International Conference on Advanced*

Computer Theory and Engineering, Chengdu, 2010.

[22] Marchi E., Vesperini F., Eyben F., Squartini S., and Schuller B., "A Novel Approach for Automatic Acoustic Novelty Detection Using A Denoising Autoencoder with Bidirectional Lstm Neural Networks," *in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, South Brisbane, pp.1996-2000, 2015.

[23] Pimentel M., Clifton D., Clifton L., and Tarassenko L., "A Review of Novelty Detection," *Signal Processing*, vol. 99, 215-249, 2014.

[24] Quinn J. and Sugiyama M., "A Least-Squares Approach to Anomaly Detection in Static and Sequential Data," *Pattern Recognition Letters*, vol. 40, pp. 36-40, 2014.

[25] Rajiah P., Fulton N., and Bolen M., "Magnetic Resonance Imaging of the Papillary Muscles of the Left Ventricle: Normal Anatomy, Variants, and Abnormalities," *Insights into Imaging*, vol. 10, no. 1, pp. 1-17, 2019.

[26] Ribeiro M., Lazzaretti A., and Lopes H., "A Study of Deep Convolutional Autoencoders for Anomaly Detection in Videos," *Pattern Recognition Letters*, vol. 105, pp. 13-22, 2018.

[27] Tan P., Steinbach M., Kumar V., Potter C., Klooster S., and Torregrosa A., "Finding Spatio-Temporal Patterns in Earth Science Data," *in KDD 2001 Workshop on Temporal Data Mining*, pp. 1-12, 2001.

[28] Yang C., "Anomaly Network Traffic Detection Algorithm Based on Information Entropy Measurement Under the Cloud Computing Environment," *Cluster Computing*, vol. 22, no. 4, pp. 8309-8317, 2019.

Praphula Jain is a Ph.D. researcher at the Department of Computer Science and Engineering, Indian Institute of Technology(ISM), Dhanbad, JH, INDIA. He obtained a B.E. degree in Computer Science and Engineering at the Faculty of Computer Science and Engineering, RGPV, Bhopal, MP, India, and an M.Tech degree in Computer Science and Engineering at the Indian Institute of Technology(ISM), Dhanbad, JH, INDIA. His Ph.D. research focuses on Machine Learning methods for various applications.

Mani Shankar Bajpai is an undergraduate student at the Department of Computer Science and Engineering, Indian Institute of Technology(ISM), Dhanbad, JH, INDIA. His research area includes computer vision, deep learning, and machine learning.

Rajendra Pamula is a Professor at the Department of Computer Science and Engineering, Indian Institute of Technology (ISM), Dhanbad, JH, INDIA. He obtained Ph.D. from the Department of Computer Science and Engineering, Indian Institute of Technology, Guwahati, Asam, INDIA. He has ten years of teaching experience and has published many technical articles in international scholarly journals.