EAI Endorsed Transactions
on Energy Web
Research Article EAI.EU

Unsupervised Approach for Email Spam Filtering using Data Mining

Mehdi Ebady Manaa[1*], Ahmed J. Obaid[2], and Mohammed Hussein Dosh[3]

[1]Department of Information Networks, College of Information Technology, University of Babylon, Iraq.
[2]Department of Computer Science, Faculty of Computer science and Mathematics, University of Kufa, Iraq.
[3]College of Education for Girls, University of Kufa, Iraq.
E-mail: [1]It.mehdi.ebady@itnet.uobabylon.edu.iq, [2]ahmedj.aljanaby@uokufa.edu.iq, [3]mohammedh.dosh@uokufa.edu.iq

Abstract

The computer networks overwhelm with unwanted emails, which are called spam emails. This email brings financial damage to companies and losses of user reputation. In this paper, the increasing volume of these emails has created the intense need to design and implement robust anti-spam filtering using the vector space model and Machine Learning (ML). ML algorithms have successfully used to detect and filter spam emails that jeopardize the network resources and consume the bandwidth. The main objective is to apply unsupervised learning M-DBSCAN to classify spam and ham emails. A robust method using the Modified Density-Based Spatial Clustering of Applications with Noise (M-DBSCAN) is implemented. The extracted N- representative points from each cluster are applied in the online test. These points represent the cluster objects to detect spherical and non-spherical clusters. These N-representative points are formed from the training step to detect spam email using distance measures. The data set used from the Kaggle website included many objects of ham and spam emails. The results show good performance accuracy with 97.848% in M-DBSCAN compared with 95.918% for standard DBSCAN accuracy and efficient values in false-negative rate, false-positive rate, f-score and online time detection.

Keywords: Spam Emails, Vector Space Model, Data Security, Machine Learning, M-DBSCAN

Received on 09 December 2020, accepted on 06 March 2021, published on 09 March 2021

doi: 10.4108/eai.9-3-2021.168962

*Corresponding Author. Email: it.mehdi.ebady@itnet.uobabylon.edu.iq

1. Introduction

The computer networks have recently jeopardized with unwanted commercial bulks emails. These emails are spam emails and spammer who's sending them. There are a set of reasons that harm the losing of the company reputation and network consumption. Some of these reasons are inadequate staff training, virus, worm, theft of data, unauthorized copying of data, and program alteration [1]. Kaspersky reported the volume of spam emails sent within two years (2016 - 2018), see figure (1).

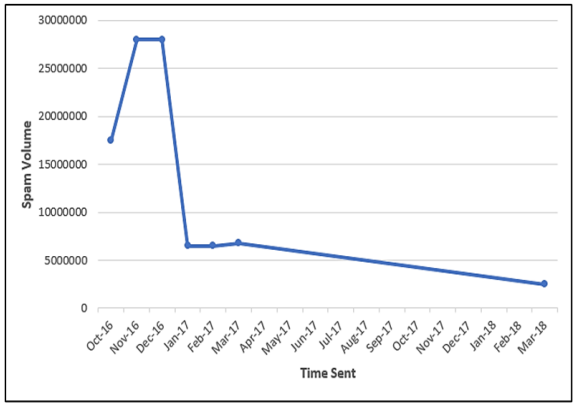

Figure 1. The volume of Spam Emails from 2016 to 2018 [2]

Furthermore, pernicious attachments of spam emails such as malware, Marcos script, and Javascript were illustrated in this report. The cybercriminals tried to send blackmail to the users in 2019. Theoretically, dealing with a case opened against the message recipient to the storage and distribution of pornographic images of minors to pay $10,000 [2][3]. There are four spam email filtering approaches content-based, sample-based, heuristic-based, and trained-based. E-commerce is the main motivation that needs to avoid a lack of trust in online marketing. A spammer sends a large number of spam emails, which attached harmful scripts. This could overwhelm the network and prevent legitimate users from accessing the resources securely [4][5].

Machine learning has launched to play an integral role to identify and prevent spam email. In general, data mining tasks can be classified into two main categories: descriptive data mining or predictive data mining. The first category summarizes the general purpose of the data succinctly. We can apply the methods of statistical analysis to describe the main purpose of the data. For instance, a histogram can create a picture of the data distribution as a graphical display. Moreover, it is possible to utilize frequency to extract the number of data iterations. In the second category, predictive data mining aims to predict the data model and then identify particular data's behaviour [6][7].

2. Data Mining Techniques

A data mining technique may use one or more of the following data technique association, classification and clustering methods: -

2.1. Association Methods

These methods work potentially to find the correlation between the tuples or sets of data records. It mainly depends on the rule expression form. An association rule consists of deriving the set of rules in the form of $X \rightarrow Y$. Where X and Y are sets of attributes values with $X \cap Y \neq 0$. It is commonly used in market data transaction [8].

2.2 Classification methods

It is a predictive model which has been involved in organizing data into classes. It comprises two steps to achieve classification, training and testing steps. In the first step, the classification model was trained using one classification technique, such as the neural network and decision tree. This would have occurred in the presence of the target (class). In the second step, training features could be harnessed to classify unknown data or points. For example, the decision tree features are used to classify the data point with the unknown class to one class obtained in the training step [9]. Many performance metrics are utilized to calculate the accuracy, false-positive rate, false-negative rate, and time. The confusion matrix is used to calculate the metrics above [10].

2.3 Clustering methods

It is unsupervised descriptive data mining methods to extract new knowledge and clustering groups of data into sub-clusters. Objects in each cluster similar to each other and different from other objects in the other clusters. One of the distance-measures was recruited for efficient classification. The distance measured (X, Y) takes two variables in the space and returns a numeric distance between these two arguments. Figure (2) illustrates the output clusters using one of the cluster analysis methods, and it is three groups (A, B and C) in coordinate space. Furthermore, some cluster analysis is applied to produce clusters with different size and densities [11].

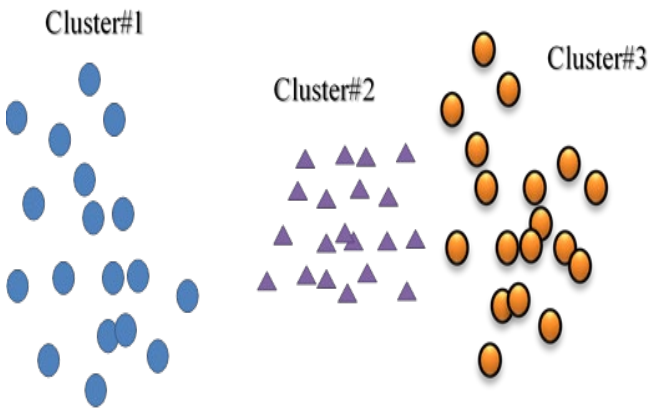

Figure 2. The landscape of three different clusters in coordinate space

In this study, the M-DBSCAN has been proposed to classify the spam, and hame emails using the Kaggle dataset, which consists of spam and ham emails with (4993) uniques values spam emails 29% and the ham emails with 71%.

3. Related works

Spam emails are among the most complicated email services problems because they jeopardize network resources and consume bandwidth. A lot of work has been done on spam filtering. Most dominant spam filters are based on machine learning classification techniques. Classification plays an important role in detecting the target type at the training and testing steps, including finding fraud, checking attacks, intruder and early diseases. These algorithms are supported by vector machine, neural networks, Naive Bayes and Decision tree.

Therefore, most researchers are interested in finding the best classifier for spam detection.

In a published study [12], the clustering algorithms were carried out, the digest algorithm represented emails, and then the emails were clustered using the DBSCAN. The results showed that accuracy was improved by 30% compared with other Standard DBSCAN. The Naïve Bayes, KNN, and reverse DBSCAN classifiers are implemented to classify the spam emails based on the two public datasets [13].

The results present good performance in terms of accuracy, recall, precision, and F-Measure. The comparative study was conducted between the classification algorithms in terms of using n-gram or without through the use of public data in 2007 TREC Public Spam Corpus [31].

The study introduced that the results were outperformed in terms of n-gram and combined datasets for the classifier naïve base, decision tree, artificial neural network, random forest, and linear regression. On the other hand, another study [14] shows spam emails can be distinguished across some features because they know if they used the same set of features for a long time. The anti-spam companies could develop tools as anti-spam filtering. Since the spam emails may be prone to so-called "concept-drift", the study proposed Ensemble-based Lifelong Classification using Adjustable Dataset Partitioning (ELCADP).

The results have shown that this model outperforms several data mining algorithms in terms of accuracy performance. Spam emails are clustered using K-means data mining cluster analysis in [32].

The proposed work consists of four steps, and the first step is tokenized the incoming message. Information gain is calculated in the second step to select the best features from the incoming email message, while the feature vector created in the third step [33]. Finally, the K-means is applied using to detect the spam cluster from the ham cluster. Naïve bases are used to detect spam emails using two public datasets, Spam Data and SPAMBASE datasets [14]. The datasets were evaluated using accuracy, recall, and precision performance metrics [34].

4. The Proposed System

Currently, there is an increasing concern in data mining approaches for spam emails detection. Specifically, we concern about lower time and accurate results for a large amount of spam emails dataset. These approaches have been designed and implemented to identify the knowledge discovery from the spam emails dataset. Data mining's main steps are data selection, preprocessing, transformation, data mining models, and evaluation [10]. Figure (3) shows the main step of the proposed system.

Figure 3 shows the main step of the proposed system.

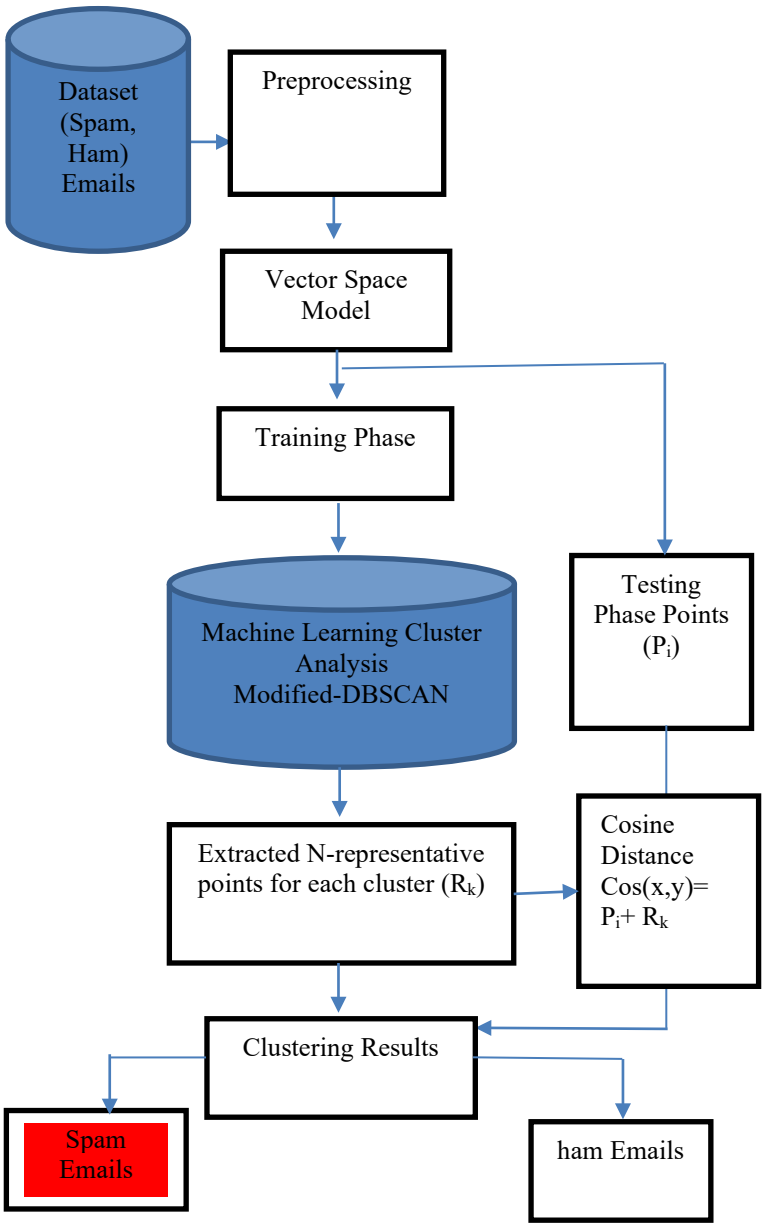

Figure 3. The proposed spam detection system

The dataset source has been from the website https://www.kaggle.com/venky73/spam-mails-dataset/data Enron-1 spam emails with spam emails and ham emails. The spam email dataset consists of two columns, the emails data and label_num, which either be (1) for spam and (0) for ham with the total unique values (4993), where the emails percents were (29% and 71%) for the spam and ham emails respectively [15];[20]. It is clearly shown in figure (3) that after getting the spam email dataset, the preprocessing step is used to remove the stop words and special characters and then we used the Vector Space Model (VSM), which Salton introduces to measure the similarity between texts [16];[21]. We replace each word in a vector to binary representation

[0,1] and then, the cosine distance between two vectors space is calculated from the formula (1).

$$\text{Similarity}(x, y) = \cos(\theta) = \frac{\sum_{i=1}^{n} x_i \times y_i}{\sqrt{\sum_{i=1}^{n} x_i} \times \sqrt{\sum_{i=1}^{n} y_i}} \quad (1)$$

Where x, y refers to a vector of n points for each. Where the cosine distance can be found by multiplying both vectors and dividing by multiplying each vector's sum, values range between -1 and 1, where -1 is perfectly dissimilar, and 1 is perfectly similar. The main idea of DBSCAN identifies a set of objects in the data space based on the cluster in data space is a high point density [22]. The intercluster objects have a high point density during the intracluster with low points density. Parameters (MinPts) refers to a minimum of points (threshold) for cluster objects to be dense, while the parameter (eps) is a distance measure to locate neighbouring points for any point (P) [19]. there are three points in DBSCAN core, border and noise. The core point refers to a Point P with at least m points with distance n from itself. Point P is called border if it has at least one core point at a distance n. Point p is called noise if it is neither a Core nor a Border [17][18]. Figure (4) shows the main concepts of core, noise, and border points [23].

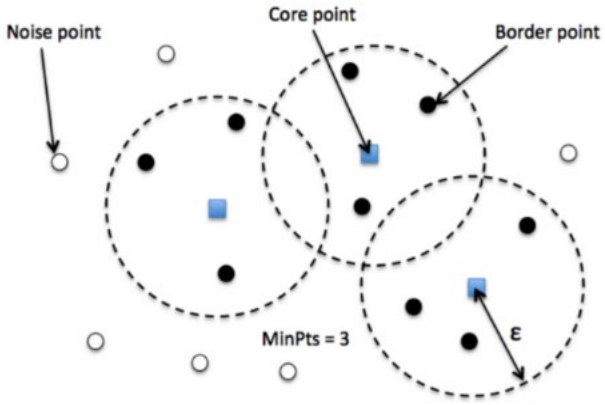

Figure 4. The Core, Border, and Noise Points in DBSCAN

This paper introduced and applied state of the art modified technique based on the DBSCAN concept. But we extracted N-representative points (Rk) [24]. These points are chosen as far as from cluster centroids to be used in the online test. Algorithm (1) shows the major steps of the modified-DBSCAN [25].

Algorithm 1: **Algorithm** 1: Modified DBSCAN (D, eps, MinPts, distFunc)
Input : training database (D) , neighborhood radius (eps) , density threshold (MinPt) **Output** : clustering the data with Spam, Ham, Representative Points (R_k)

Begin

Training Phase

Convert all emails into the binary vector (x_i , y_j)

The dataset D includes class label, which is not included in the training processing

Class label is not included in the training phase.

initialize cluster counter *c* =0

for each point p ∈ database (D) {

  if label (p) ≠ undefined, **then continue**

    Neigbours Nr=**RangeQuery** (D , distFunc, **P** , **c**, eps , **MinPts**)

      If $|Nr|$< MinPts then

        label(P) = Noise

          continue }

      c=c+1;

      label(P)=c;

    Seed Set ss=Nr/ {P}

  foreach point q ∈ seed sets (ss) {

    Extract noise point and border points

      If label (q) = Noise then label (q)= c

      If label(q) ≠ UNCLASSIFIED **then continue**

        label(q)=c

        Neigbours Nr= RangeQuery (D, disFunc, q, c, eps,

MinPts)

        If $|N| \geq$ MinPts then

          ss= ss ∪ N

  foreach cluster c

    R_k =Extract N-represented points.

end

return R_k

Testing Phase

Begin

foreach x ∈ D // testing data points

  foreach r ∈ R_k // N-representative points in each cluster

    Find cosine distance (x,c)

  end

label x with same class to N- representative

points have min distances with **x**

35 end

36 End

5. Results and Discussion

Dataset was labeled as ham and spam. N-representative points were extracted based on two clusters (ham and spam) by setting the N-representative points= 5 [26]. The testing points are assigned the class label having the minimum distance of N-representative points [27]. Quantitative statistical analysis was performed based on the confusion matrix, and four metrics were evaluated (accuracy, precision, recall, and f-score) as shown in the following formula (2-5) [28].

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP} \quad (2),$$

The formula (3) is illustrated the precision.

$$precision\ (Pr) = \frac{TP}{TP+TN} \quad (3)$$

The formula (3) is illustrated the recall

$$Recall\ (Re) = \frac{TP}{TP+FN} \quad (4)$$

The relation between precision and recall is calculated in the formula (5) by f-score

$$F1\ score = 2 * \frac{Pr * Re}{Pr+Re} \quad (5)$$

Table (1) shows the main results using the Modified DBCAN by setting the N-representative point=5 for each cluster. We find the N-representative point=5 satisfied the best results for performance metrics above using trial and error [29].

Table 1. The evaluation Results of the Modified-DBSCAN with N-representative point=5

Train Size	DBSCAN Results	Modified DBSCAN				
		Accuracy	Precision	Recall	f1-score	Time
10%	90.121	90.287	93.254	89.254	91.2102	2.7
20%	89.561	92.244	93.661	87.365	90.4035	8.6
30%	88.015	93.512	93.454	86.868	90.0407	16.8
40%	88.358	97.847	97.684	87.919	92.5446	27.6
50%	92.999	97.111	97.254	88.215	92.5142	38.5
60%	91.531	96.121	96.218	88.861	92.3933	48.6
70%	95.265	97.231	97.666	88.254	92.7218	58.8
80%	95.683	97.321	97.351	80.234	87.9676	72.8
90%	95.918	97.848	97.561	88.358	92.7317	90

It is clearly shown from the table (6) that the increasing training percent would not have an impact on the accuracy of the proposed system. However, time is increased for the Modified DBSCAN [30]. The results show that the proposed system displays superior performance in accuracy metrics compared with the k-means, which has an accuracy (93%) and time (88 secs). In contrast, the DBSCAN shows less accurate results than the modified DBSCAN. The reason that the k-means can't detect the cluster with non-spherical shapes compared with Modified-DBSCAN.

6. Conclusion

This study was intended to intensify intruders who send spam emails by using one of the unsupervised data mining algorithms. The main role of the extracted N-representative points from each cluster is their implementation in online testing. Each test point is classified in a proactive way for spherical and non-spherical cluster shapes. The modified M-DBSCAN algorithm's evaluation results involved relatively higher rates in terms of accuracy compared to the results with the original algorithm DBSCAN and K-means. A total number of (747) spam emails and (4825) ham emails have been used, and the results proved the efficiency of the proposed algorithm with accuracy and precision of more than 97% and a recall rate over 88%. It is thereby concluded that the modified M-DBSCAN deals with non-spherical shapes (clusters) with very high accuracy and detection rates.

References

[1] E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi, and O. E. Ajibuwa, "Machine learning for email spam filtering: review, approaches and open research problems," Heliyon, vol. 5, no. 6, 2019.

[2] S. Puri, D. Gosain, M. Ahuja, I. Kathuria, and N. Jatana, "Comparison and Analysis of Spam Detection Algorithms," vol. 2, no. 4, pp. 1–7, 2013.

[3] T. K. Maria Vergelis, Tatyana Shcherbakova, Tatyana Sidorina, "Spam and phishing in 2019," Secure List.

[4] Z. S. Torabi, M. H. Nadimi-Shahraki, and A. Nabiollahi, "Efficient Support Vector Machines for Spam Detection: A Survey," IJCSIS) Int. J. Comput. Sci. Inf. Secur., vol. 13, no. 1, pp. 10–28, 2015.

[5] T. Gangavarapu, C. D. Jaidhar, and B. Chanduka, Applicability of machine learning in spam and phishing email filtering: review and approaches, no. 0123456789. Springer Netherlands, 2020.

[6] W. Hämäläinen, "Descriptive and Predictive Modelling Techniques for Educational Technology, Thesis," Msc. Thesis, University of Joensuu, Finland., 2006.

[7] N. Jain and V. Srivastava, "Data Mining Techniques: a Survey Paper," Int. J. Res. Eng. Technol., vol. 2, no. 11, pp. 116–119, 2013.

[8] D. Barbará, J. Couto, S. Jajodia, L. Popyack, and N. Wu, "ADAM: Detecting Intrusions by Data Mining," in International Conference on information Assurance and Security, 2001, no. June, pp. 11–16, doi: 10.1145/604264.604268.

[9] A. Tiwari and S. Sharma, “A review of data mining techniques,” Int. J. Sci. Res. Technol., vol. 1, no. 1, pp. 28–33, 2015.
[10] M. J. Zaki and J. W. Meira, Data mining and analysis: Fundamental Concepts and Algorithms, Book. Cambridge University Press, 2014.
[11] A. H.Ahmed and M. Mikki, “Improved Spam Detection using DBSCAN and Advanced Digest Algorithm,” Int. J. Comput. Appl., vol. 69, no. 25, pp. 11–16, 2013, doi: 10.5120/12126-8300.
[12] A. Harisinghaney, A. Dixit, S. Gupta, and A. Arora, “Text and image based spam email classification using KNN, Naïve Bayes and Reverse DBSCAN algorithm,” in ICROIT 2014 - Proceedings of the 2014 International Conference on Reliability, Optimization and Information Technology, 2014, pp. 153–155,
[13] N. Sutta, Z. Liu, and X. Zhang, “A Study of Machine Learning Algorithms on Email Spam Classification,” in Proceedings of 35th International Confer- ence on Computers and Their Applications A, 2020, vol. 69, pp. 170–159, doi: 10.29007/qshd.
[14] G. Venkatesh, “SMS Spam Collection Data Set,” Spam Mails Datase, 2019. [Online]. Available: https://www.kaggle.com/venky73/spam-mails-dataset/activity.
[15] S. Al-Anazi, H. Almahmoud, and I. Al-Turaiki, “Finding Similar Documents Using Different Clustering Techniques,” Procedia Comput. Sci., vol. 82, no. March, pp. 28–34, 2016, doi: 10.1016/j.procs.2016.04.005.
[16] M. Daszykowski and B. Walczak, “Density-Based Clustering Methods,” in International Conference Knowledge Discovery and Data Mining, 1996, vol. 2, pp. 226–231.
[17] Leo Willyanto Santoso, Bhopendra Singh, S. Suman Rajest, R. Regin, Karrar Hameed Kadhim (2020), “A Genetic Programming Approach to Binary Classification Problem” EAI Endorsed Transactions on Energy, Vol.8, no. 31, pp. 1-8. DOI: 10.4108/eai.13-7-2018.165523
[18] Jalil, N. A., & Kian Yeik, K. (2019). Systems, design and technologies anxieties towards use of self-service checkout. In ACM International Conference Proceeding Series. https://doi.org/10.1145/3371647.3371664
[19] Jalil, N. A., Prapinit, P., Melan, M., & Mustaffa, A. Bin. (2019). Adoption of business intelligence - Technological, individual and supply chain efficiency. In Proceedings - 2019 International Conference on Machine Learning, Big Data and Business Intelligence, MLBDBI 2019. https://doi.org/10.1109/MLBDBI48998.2019.00021
[20] Ravi Kumar Gupta, “Employment Security and Occupational Satisfaction in India,” Journal of Advanced Research in Dynamical & Control System, Vol. 10, Issue 10, pp. 244-249, 2018.
[21] T. A. Al-asadi and A. J. Obaid, "Object Based Image Retrieval Using Enhanced SURF," Asian Journal of Information Technology, vol. 15, no. 16, pp. 2756-2762, 2016.
[22] C. Meshram, R. W. Ibrahim, A. J. Obaid, S. G. Meshram, A. Meshram and A. M. Abd El-Latif, "Fractional chaotic maps based short signature scheme under human-centered IoT environments," Journal of Advanced Research, 2020.
[23] A. J. Obaid, K. A. Alghurabi, S. A. K. Albermany and S. Sharma, "Improving Extreme Learning Machine Accuracy Utilizing Genetic Algorithm for Intrusion Detection Purposes," in Advances in Intelligent Systems and Computing , Springer, Singapore, 2021, pp. 171-177.
[24] A. J. Obaid, "Critical Research on the Novel Progressive, JOKER an Opportunistic Routing Protocol Technology for Enhancing the Network Performance for Multimedia Communications," in Research in Intelligent and Computing in Engineering. Advances in Intelligent Systems and Computing, vol 1254. Springer, Singapore., Springer, Singapore, 2021, pp. 369-378.
[25] S. Sharma and A. J. Obaid, "Mathematical modelling, analysis and design of fuzzy logic controller for the control of ventilation systems using MATLAB fuzzy logic toolbox," Journal of Interdisciplinary Mathematics, vol. 23, no. 4, pp. 843-849, 2020.
[26] S. Sharma and A. J. Obaid, "Contact-mechanics and dynamics analysis of three-different ellipsoidal raceway geometries for deep Groove ball bearing using Abaqus 6.13 version FEA simulation for high load-bearing as well as speed-rotating applications," International Research Journal of Multidisciplinary Science and Technology, vol. 3, no. 5, pp. 36-43, 2020.
[27] Ahmed, E. R., Islam, A., Zuqibeh, A., & Alabdullah, T. T. Y. (2014). Risks management in Islamic financial instruments. Advances in Environmental Biology, Vol. 8, no. 9, pp. 402-406.
[28] Singla M.K., Gupta J., Nijhawan P., Ganguli S., Rajest S.S. (2020) Development of an Efficient, Cheap, and Flexible IoT-Based Wind Turbine Emulator. In: Haldorai A., Ramu A., Khan S. (eds) Business Intelligence for Enterprise Internet of Things. EAI/Springer Innovations in Communication and Computing. Springer, Cham
[29] Rajasekaran R., Rasool F., Srivastava S., Masih J., Rajest S.S. (2020) Heat Maps for Human Group Activity in Academic Blocks. In: Haldorai A., Ramu A., Khan S. (eds) Business Intelligence for Enterprise Internet of Things. EAI/Springer Innovations in Communication and Computing. Springer, Cham
[30] Bhopendra Singh, S. S Rajest, K. Praghash, Uppalapati Srilakshmi and R. Regin (2020) Nuclear structure of some even and odd nuclei using shell model calculations. Proceedings of the 2020 2nd International Conference on Sustainable Manufacturing, Materials and Technologies. AIP Conference Proceedings, 2020, https://aip.scitation.org/doi/abs/10.1063/5.0030932
[31] S. S Rajest, D.K. Sharma, R. Regin and Bhopendra Singh, “Extracting Related Images from E-commerce Utilizing Supervised Learning”, Innovations in Information and Communication Technology Series, pp. 033-045, 28 February, 2021.
[32] Souvik Ganguli, Abhimanyu Kumar, Gagandeep Kaur, Prasanta Sarkar and S. Suman Rajest, “A global optimization technique for modeling and control of permanent magnet synchronous motor drive”, Innovations in Information and Communication Technology Series, pp. 074-081, 28 February, 2021.
[33] Jappreet Kaur, Tejpal Singh Kochhar, Souvik Ganguli and S. S Rajest, “Evolution of Management System Certification: An overview”, Innovations in Information and Communication Technology Series, pp. 082-092, 28 February, 2021.
[34] R. Regin, S. S Rajest and Bhopendra Singh, “Spatial Data Mining Methods Databases and Statistics Point of Views”, Innovations in Information and Communication Technology Series, pp. 103-109, 28 February, 2021.