International Journal of Advanced Science and Technology
Vol. 29, No. 7s, (2020), pp. 3038-3046

Hybrid Approach for Fraud Detection in Financial Sector

[1] A.Varija Naga Soujanya, [2] Dr. C. Karthikeyan, [3]Ch. Manasa, [4]Y. Narendra Surya

[1234] *Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Guntur, A.P India.*

Abstract

Credit card cheat exchanges are turning out to be progressive step by step and it is getting increasingly hard for people to break down fake exchanges by examining subsequently and it has gotten vital for people to build up a canny framework to decide fake exchanges. A few keen calculations can be utilized right now peculiar identification. In this paper we executed Neural Networks, Decision Tree and Random Forest to figure out which calculation is best fit as far as time taken, recall, accuracy, precision and exactness. We had the option to define consequences of 284,407 exchanges over a time of two days in September 2013. We had the option to recognize that three models are practically equivalent with regards to exactness yet arbitrary backwoods is increasingly exact. Our goal is to look at models that foresee which exchanges could be fake with high exactness and propose a hybrid technique that can be more precise than existing ones and help in easy implementation.

Keywords: *Credit Card frauds, fraud or outlier detection, finance or banking sectors, hybrid approach.*

1. Introduction

The significance of Data mining is the route toward making sense of huge enlightening files to see plans and set up relationship to deal with issues through data assessment. Data mining instruments license relationship to foresee future examples. The upsides of Data Mining are perceiving structures in the data and exploring the lead of the customer. Data mining techniques are used in many research locales, including science, human administrations, and inherited characteristics, banking and advancing.

Nowadays, Banks have comprehended that customer associations are a critical factor for their success. [10]Customer relationship the officials (CRM) is a procedure that can help them with building solid relationship with their customers and augmentation their wages and advantages. CRM in the budgetary portion is of increasingly conspicuous noteworthiness. The CRM focus is moving from customer acquiring to customer upkeep and ensuring the fitting proportions of time, money and authoritative resources are focused on both of these key tasks.[2] Standard systems for data assessment have for a long while been used to recognize deception. They require perplexing and repetitive assessments that oversee different spaces of data like cash related, budgetary viewpoints, key methodologies and law. The machine learning and deep learning algorithms have been widely used and applied on many problems so that the solution can be obtained easily and accurately, facial emotion reorganization and facial detection is one of the mostly studied topics in this field [3]. Image fusion is another technique which is now widely used in image processing to fuse two images to extract more information from the images [11]. With the help of IoT and image processing a new devices are develop in medical field to assist the specialist for the decision making in operating the patient had made easy [12][13][14].

Conventional techniques for information investigation have for some time been utilized to distinguish misrepresentation [15]. They require complex and tedious examinations that manage various spaces of information like budgetary, financial aspects, strategic policies and law. In creating nations like India, Bankers face more issues with the fraudsters [19]. Utilizing information mining procedure, it is easy to construct a fruitful prescient model and imagine the report into significant data to the client.

2. Literature Survey

2.1. Need for Fraud Detection

Banks are experiencing challenges in guaranteeing the on the web/web banking channel. The test is in keeping customer's record secure while avoiding multifaceted design in the login procedure. Anyway the store of passwords, hardware token devices and other out-of-bound specific mechanical assemblies introduced by specific banks has inconceivably incapacitated a couple of customers.[1]Well unmistakably security focuses less on customer comfort, yet there can regardless be an improvement.

2.2. Information Mining Techniques

Financial frauds could be creating stress with route showing up at results inside the organization, association affiliations, and record trade. In Today's world high dependence on web development has gotten a kick out of intensified financial trades. Regardless, banking portion coercion had conjointly stimulated as on the web and separated trade. As trades become no matter how you look at it strategy for portion, focus has been given to progressing system methodology to manage the distortion downside. There are various distortion acknowledgments courses of action and programming structure which prevents fakes in associations like MasterCard, retail, online business, assurance and adventures [16]. Data mining methodology is one famous and typical strategies used in affirmation banking part distortion recognizable proof disadvantage. It's unreasonable to be sheer sure concerning reality desire related right behind an application or trade. In fact, to pursue out possible affirmations of distortion from the open data using logical counts is the best suitable possibility.

Ideas of administered and unaided taking in are gotten from the study of AI, which has been known as a sub-region of man-made reasoning. Man-made reasoning methods the execution and investigation of frameworks that show self-sufficient insight or conduct of their own. AI manages systems that empower gadgets to gain from their own exhibition and adjust their own working. Information mining applies AI ideas to information.[5]
Supervised learning is otherwise called coordinated learning. The learning procedure is coordinated by a recently known ward characteristic or target. Guided information mining endeavors to clarify the conduct of the objective as a component of a lot of autonomous traits or indicators [17]. Managed adapting commonly brings about prescient models. [6]This is rather than solo realizing where the objective is design recognition.

Unsupervised learning is non-coordinated. There is no differentiation among needy and free qualities. There is no beforehand realized outcome to direct the calculation in building the model [18].

2.3. Methods:

2.3.1. Classification

Classification is the most normally applied information mining strategy, which utilizes a lot of pre-grouped information to build up a model that can order the number of inhabitants in records on the loose. Fraud recognition and credit hazard applications are especially appropriate to this kind of investigation. The information grouping process includes learning and order.

2.3.2 Clustering

Clustering can be said to be the ID of comparable classes of items. Right now, with comparative conduct are gathered into one gathering. Grouping can be utilized as a preprocessing approach for property subset determination and characterization. k-Means is a separation based grouping calculation that parcels the information into a foreordained number of bunches[20]. Each group has a centroid (focus of gravity). Cases (people inside the population) that are in a bunch (cluster) are near the centroid

2.3.3. Association Rule

The main errand of this is to discover sets of parallel factors that co-happen together every now and again in an exchange database , while the objective of highlight determination issue is to distinguish clusters that are firmly associated with one another with a particular objective variable. Affiliation rule has the few calculations like: CDA, APRIORI and DDA. These rules are on the off chance that/at that point articulations that help reveal connections between apparently inconsequential information in a social database or other data storehouse.

2.3.4. Prediction

The expectation as its name infers is one of the information mining strategies that find connection between free factors and connection between subordinate factors. Relapse examination can be utilized to display the connection between at least one free factor and ward factors. In information mining, free factors are characteristics definitely known and reaction factors are what we need to foresee. Lamentably, some certifiable issues are not just unsurprising.

2.3.5. Sequential Patterns

Sequential patterns analysis is one of data mining techniques that seek to discover similar patterns in data transaction over a business period. The uncovered patterns are used for further business analysis to recognize relationships among data.

2.3.6. Artificial Neural Network

It consolidates the considering power human mind with computational intensity of machine. It utilizes neurons as the choosing destinations and the edges between neurons to compute the commitment of every neuron in the past layer in the choice and result at the present neuron. It depends on design acknowledgment. Earlier year's information is taken care of into the system and afterward dependent on that information it perceives another approaching exchange to be extortion or real one. Its preparation can either be administered for example the result is now known for a given exchange and the normal yield is contrasted with real with train the framework [21-28].

2.3.7. Decision Tree

It is a computational apparatus for grouping and forecast. A tree includes interior hubs which indicate a test on a property, each branch signifies a result of that test and each leaf hub (terminal hub) holds a class mark. It recursively segments a dataset utilizing either profundity first covetous methodology or broadness first insatiable methodology and stops when all the components have been appointed a specific class. For the parcel rule to be proficient it must separate the information into bunches where a solitary class prevails in each gathering.

2.3.8. Fuzzy Logic

It is utilized in the situations when we don't have discrete truth esteems i.e., they are persistent. It is a multivalve logic. There is sure arrangement of rules dependent on which an exchange is named a real or extortion one. There are three significant segments in fluffy logics that should be executed in the expressed request:

2.3.9. Support Vector Machines

It is a directed learning calculation where given a dataset it isolates them into various classes utilizing a hyperplane. The objective of SVM is to discover this hyperplane. There could be numerous hyperplanes however we are resolved to locate an ideal hyperplane. The focuses nearest to the hyperplane in the various classes are known as help vectors and these help vectors are utilized to anticipate the class of new information focuses. Another approaching point is put on the condition of the hyperplane and afterward is named to which class it has a place based on which side of hyperplane it falls on the vector space.[9] To prepare our machine we feed directed information for example information with results definitely known. It learns the conduct of misrepresentation and real exchanges and afterward it can order new exchange about which class it has a place.

2.3.10. Bayesian Network

It depends on the Bayes Theorem of restrictive likelihood; thus it is a probabilistic model that is utilized for computerized discovery of different occasions. It comprises of hubs and edges, wherein the hubs speak to the irregular factors and the edges between the hubs speak to the connections between these arbitrary factors and their probabilistic appropriation. We compute predefined least and most extreme estimation of probabilities of an exchange being extortion or legitimate. At that point for another approaching exchange we see that whether it's likelihood of being legitimate is not exactly the base characterized an incentive for lawful exchange and is more noteworthy than the greatest characterized an incentive for an extortion exchange. In the event that genuine, at that point the exchange is named a fraud exchange.

2.3.11. K-Nearest Neighbor

It is one of the most utilized calculations for both characterization and relapse prescient issues. Its presentation relies upon three factors: the separation measurements, the separation rule and the estimation of K.[8] Separation rule encourages us to arrange the new information point into a class by contrasting its highlights and that of information focuses in its neighborhood. Also, the estimation of K chooses the quantity of neighbors with whom to look at. The significant inquiry is how would we pick the factor K? So as to acquire the ideal estimation of K, the preparation and approval is isolated from the underlying dataset. Presently a diagram dependent on the approval mistake bend is plotted to accomplish the estimation of K. This estimation of K ought to be utilized for all forecasts.

2.3.12. Hidden Markov Model

There is a difference in state with time thus the name Markov. The states are covered up thus can't be watched legitimately. In any case, something corresponded to them can be watched and dependent on that grouping of perceptions we foresee the request for state changes. We first train our model dependent on given arrangement of parameters like way of managing money of cardholder.[7] Beginning arrangement of probabilities are picked dependent on this profile. At that point any new approaching exchange is dissected by our model and delegated fake on the off chance that it fluctuates from the general profile and conduct of a cardholder by in excess of edge esteem and subsequently it can't be acknowledged by the states in concealed Markov model.

2.3.13. Logistic Regression

To battle the peculiarities of straight relapse where it gave values more prominent than 1 and under 0, calculated relapse becomes an integral factor. In spite of the name being relapse, LR is utilized for characterization issues for foreseeing binomial and multinomial results, having the objective of assessing the estimations of parameter's coefficients utilizing the sigmoid capacity. Strategic relapse is utilized for grouping and when an exchange is progressing it looks at the estimations of its properties and tells whether the exchange ought to continue or not.

2.3.14. Anomaly Detection Techniques

Anomaly location procedures can be isolated into three mode bases on the accessibility to the names:

Supervised Anomaly Detection: This sort of peculiarity discovery systems have the supposition that the preparation informational collection with precise and delegate names for ordinary case and abnormality is accessible.[4]

Semi-Supervised Anomaly Detection: This sort of system accepts that the train information has named occurrences for simply the typical class. Since they don't request names for the irregularity, they are generally appropriate than directed techniques. For model, Reza use semi-managed calculation to anomaly in online informal organization.

Unsupervised Anomaly Detection: These procedures needn't bother with preparing informational index and in this way are most generally utilized. Solo irregularity recognition techniques can

"pretend" that the whole informational collection contains the ordinary class and build up a model of the typical information and respect deviations from that point typical model as inconsistency.

3. Proposed Method

We propose a hybrid method that uses and combines the advantages of k means clustering and SVM and the solo strategies for the discovery of creditcard fraud exchanges. Clusters are formed using the k means clustering process repetitively. Distance based technique is used to find outliers or frauds from each cluster respectively. Threshold value is different for each cluster as we take different set of transaction data each in each cluster that has been divided.

Description:

1. Dataset: as referenced previously.
2. Preprocessing: Information in genuine world is grimy, boisterous, conflicting, and fragmented. So we need to do preprocessing. We have done information cleaning and information decrease in week's apparatus.
3. Apply proposed calculation as expressed previously.
4. Apply each progression as expressed above to distinguish exceptions. In the wake of applying proposed calculation, we get 13000 and 492 exceptions. As notice in above three cases Threshold esteem are distinctive according to client input. Number of Outliers is changing if the estimation of limit is change. Number of anomalies is relying upon estimation of edge.

Steps in Hybrid Method

1. Obtain exchange information
2. Preprocess information to change over straight out credits to numerical traits
3. Normalize the numerical information utilizing Min-Max Normalization.
4. Create preparing and testing information records utilizing the SVM group
5. Select a property a1
6. Apply K-Means Clustering as for the trait a1
7. Select the subsequent level property a2
8. For each bunch c got a. Apply K-Means Clustering regarding the quality a2
9. End for
10. Supply the preparation record to SVM
11. Set the qualities for C and V.
12. Obtain results utilizing the present C and V pair
13. Perform stage 6 and 7 till palatable outcomes are gotten from the preparation set
14. Test the precision utilizing the test document
15. For each pernicious exchange got, a. Discover bunches that contains the present exchange b. Utilizing aggregate creature conduct, check for a comparative example in the groups c. In the event that the example comparability surpasses the limit t think about the exchange as ordinary d. Think about the exchange as vindictive
16. End.

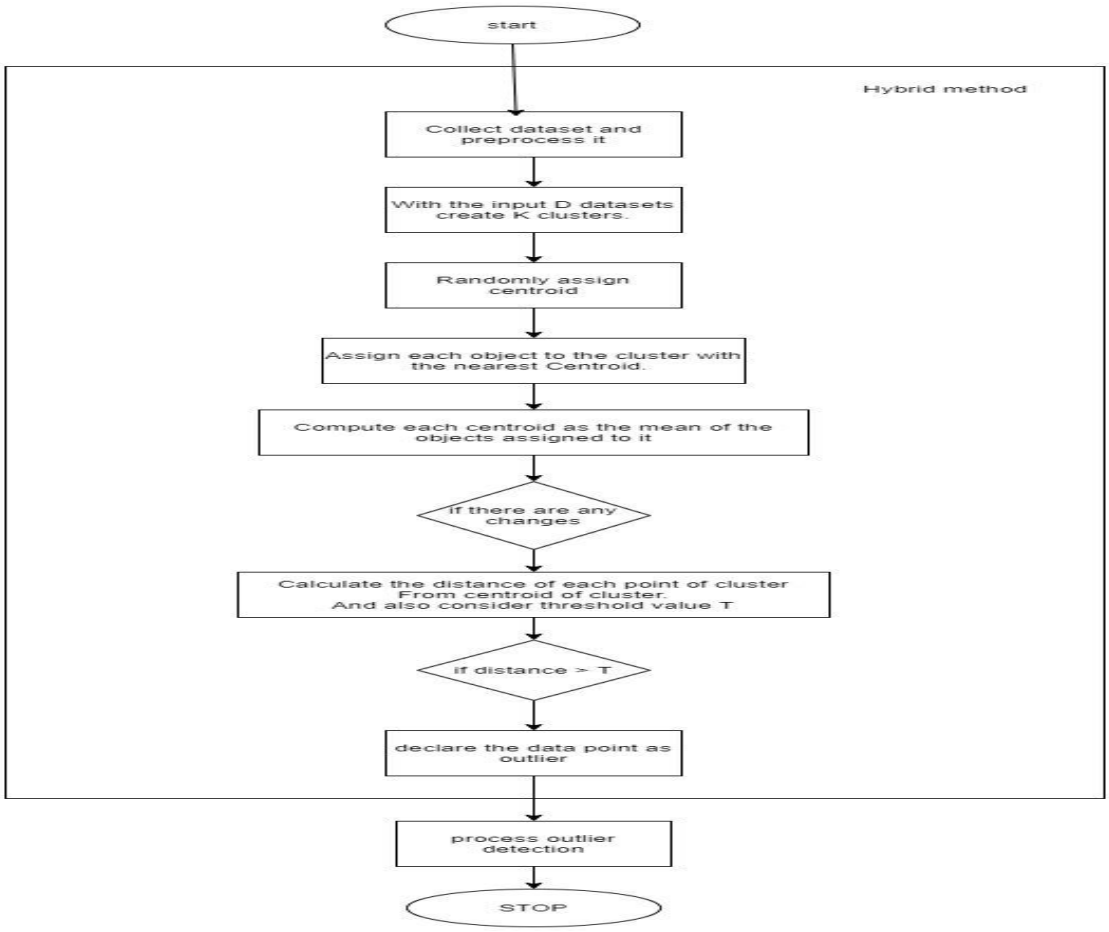

Fig.1. Flow Chart for the Hybrid Method

4. Results and Discussion

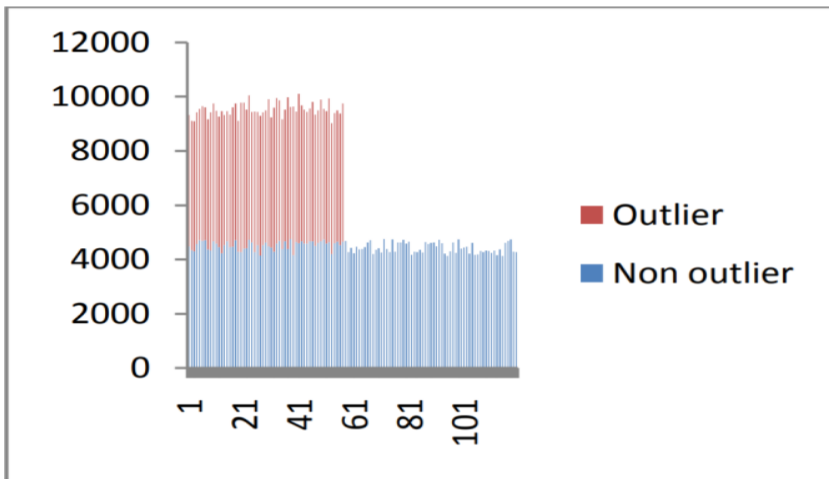

Fig.2. Chart Analysis for Outlier and Non Outlier

Here x axis is num of instances and y represents value of each instance and outputs ; number of outliers or frauds in around 5000 instances that belong to a single cluster are approximately 60 and non frauds are 120.

Table.1. Result Analysis

Parameters	Prediction on Test Set			
	Decision Tree	Neural Network	Random forest	Proposed Method
Accuracy	0.9993	0.9995	0.9995	0.9995
Precision	0.8409	0.8496	0.9417	0.9871
Recall	0.7551	0.8163	0.7687	0.8256
Total score	0.7957	0.8421	0.8464	0.8957

Table 1. Shows the parameters such as accuracy, precision, recall and total score values of proposed method are compared with the existing algorithms Decision Tree, Neural Network and Random forest. Within the existing algorithms random forest has the highest precision and accuracy when compared to all the others with a precision of 94%,accuracy of 99.5%.

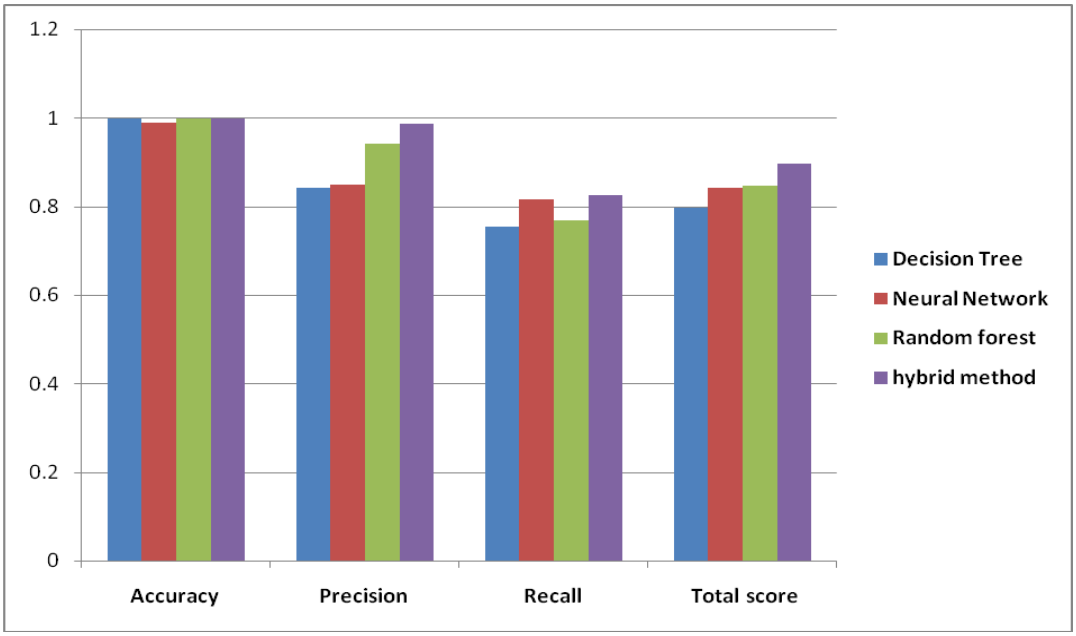

Fig.3. Chart representation for Result Analysis

5. Conclusion with Future Works

In budgetary administrations, identifying extortion in charge card is intense issue. Right now have talked about different methods of information mining through which we can distinguish the extortion. Different strategies like Neural Network Hidden Markov Model, Decision Tree, Fusion approach due utilizing dumpster, K-means clustering, SVM, K-nearest neighbor Bayesian Network, and Logistic Regression are utilized. Irregular Forest has the most elevated precision yet needs speed when there are a ton of occasions in the dataset which are alluded as exchanges. Our proposed strategy can't distinguish fraud 100% but can give same outcomes inside less time and most likely better examination is done. To recognize the misrepresentation precisely and productively, it is essential that the genuine information is accessible. The future work will be to utilization of various calculations to build exactness. For this the distinctive dataset ought to be accessible to improve credit card frauds.

References

[1] Performance of machine learning techniques in the detection of financial frauds I.SADGALIa, N.SAELa, F.BENABBOUa .

[2] Expert Systems in Finance: Smart Financial Applications in Big Data Environments edited by Noura Metawa, Mohamed Elhoseny, Aboul Ella Hassanien, M. Kabir Hassan.

[3] Jahnavi, P., Vamsidhar, E., Karthikeyan, C., "Facial expression detection of all emotions and face recognition system" , International Journal of Emerging Trends in Engineering Research, December 2019, DOI: 10.30534/ijeter/2019/087122019.

[4] A. O. Adewumi, A. A. Akinyelu, "A survey of machine-learning and nature-inspired based credit card fraud detection techniques", Int. J. Syst. Assurance Eng. Manage., vol. 8, no. 2, pp. 937-953, 2017.

[5] Card Fraud Detection Using AdaBoost and Majority Voting by Kuldeep Randhawa ; Chu Kiong Loo ; Manjeevan Seera ; Chee Peng Lim.

[6]Text Mining for Big Data Analysis in Financial Sector: A Literature Review by Mirjana Pejić Bach, Živko Krstić, Sanja Seljan and Lejla Turulja.

[7] The fraud diamond: element in detecting, financial statement of fraud by RR. Maria Yulia Dwi Rengganis, Maria Mediatrix Ratna , I.G.A.N Budiasih, I Gde Ary Wirajaya, Herkulanus Bambang Suprasto.

[8] Adetunmbe Awoyeme, Oluwadari, "Credit card fraud detection using machine learning techniques: A comparative analysis", 2017 International Conference on Computing Networking and Informatics (ICCNI).

[9] S. Velliangiri, P. Karthikeyan & V. Vinoth Kumar (2020) Detection of distributed denial of service attack in cloud computing using the optimization-based deep networks, Journal of Experimental & Theoretical Artificial Intelligence, DOI: 10.1080/0952813X.2020.1744196.

[10] Praveen Sundar, P.V., Ranjith, D., Vinoth Kumar, V. et al. Low power area efficient adaptive FIR filter for hearing aids using distributed arithmetic architecture. Int J Speech Technol (2020). https://doi.org/10.1007/s10772-020-09686-y.

[11] Vinoth Kumar V, Karthikeyan T, Praveen Sundar P V, Magesh G, Balajee J.M. (2020). A Quantum Approach in LiFi Security using Quantum Key Distribution. International Journal of Advanced Science and Technology, 29(6s), 2345-2354.

[12] Subramanian R, Karthikeyan C, Siva Nageswara Rao G, Mariappan R, "Design and Implementation device for monitoring Fetal ECG", in Advances in Intelligent Systems and Computing, 2020 https://doi.org/10.1007/978-3-030-37218-7.

[13] K Ravindranath "Security key provided for Group Data Sharing in Cloud Computing "in the International Journal of Advanced Sciences and Technology (IJAST), ISSN: 2005-4238, November-2019.

[14] K Ravindranath, et.al., "An Advanced Secured Privacy Preserving Techniques for Cloud Using Numerical SQL Query's " in the International Journal of Advanced Sciences and Technology (IJAST), ISSN:2005-4238 , November-2019.

[15] P S RajaKumar, et.al., "Optimized and Efficient Computation of Big Data in Heterogeneous Internet of Things " in the International Journal of Engineering and Advanced Technology (IJEAT), ISSN:2249:8958,Volume -9, Issue-1, PP:6005-6010, October-2019.

[16] Karthikeyan, T., Sekaran, K., Ranjith, D., Vinoth kumar, V., Balajee, J.M. (2019) "Personalized Content Extraction and Text Classification Using Effective Web Scraping Techniques", International Journal of Web Portals (IJWP), 11(2), pp.41-52 .

[17] G Sreeram, et.al., "Improving Cloud Data Storage Performance Based on Calculating Score using Data Transfer Rate Between the Internetwork Drives" in the International Journal of Engineering and Advanced Technology (IJEAT),ISSN: 2249 -8958, Volume-8 Issue-4,Page No: 1830-1835, April 2019.

[18] L.Hemanth Reddy, et.al., "Deployment of a Secured Web Application using Cryptanalysis in Cloud Environment" in the International Journal of Engineering and Advanced Technology (IJEAT),ISSN: 2249 -8958, Volume-8 Issue-4, Page No: 1841-1844, April 2019

[19] K Sreenivasa Rao, et.al., "Detecting Fake Account On Social Media Using Machine Learning Algorithms" in the International journal of Control and Automation "2005-4297, Vol. 13, No. 1s, (2020), pp. 95-100, April 2020.

[20] G.Sreeram, et.al., "Efficiency and Stationing in Edge Computing "in the International Journal of Advanced Sciences and Technology (IJAST), ISSN: 2005-4238, Vol.29, 9s, (2020) pp.112-119.

[21] Umamaheswaran, S., Lakshmanan, R., Vinothkumar, V. et al. New and robust composite micro structure descriptor (CMSD) for CBIR. International Journal of Speech Technology (2019), doi:10.1007/s10772-019-09663-0

[22] Vinoth Kumar, V., Arvind, K.S., Umamaheswaran, S., Suganya, K.S (2019), "Hierarchal Trust Certificate Distribution using Distributed CA in MANET", International Journal of Innovative Technology and Exploring Engineering, 8(10), pp. 2521-2524

[23] Maithili, K , Vinothkumar, V, Latha, P (2018). "Analyzing the security mechanisms to prevent unauthorized access in cloud and network security" Journal of Computational and Theoretical Nanoscience, Vol.15, pp.2059-2063

[24] V.Vinoth Kumar, Ramamoorthy S (2017), "A Novel method of gateway selection to improve throughput performance in MANET", Journal of Advanced Research in Dynamical and Control Systems,9(Special Issue 16), pp. 420-432

[25] Dhilip Kumar V, Vinoth Kumar V, Kandar D (2018), "Data Transmission Between Dedicated Short-Range Communication and WiMAX for Efficient Vehicular Communication" Journal of Computational and Theoretical Nanoscience, Vol.15, No.8, pp.2649-2654

[26] Kouser, R.R., Manikandan, T., Kumar, V.V (2018), "Heart disease prediction system using artificial neural network, radial basis function and case based reasoning" Journal of Computational and Theoretical Nanoscience, 15, pp. 2810-2817

[27] Shalini A, Jayasuruthi L, Vinoth Kumar V, "Voice Recognition Robot Control using Android Device" Journal of Computational and Theoretical Nanoscience, 15(6-7), pp. 2197-2201

[28] Jayasuruthi L,Shalini A,Vinoth Kumar V.,(2018) " Application of rough set theory in data mining market analysis using rough sets data explorer" Journal of Computational and Theoretical Nanoscience, 15(6-7), pp. 2126-2130