International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-1, November 2019

Identification of Information Leakage and Guilt Agent by using MAC-IP Binding and Recursive Partitioning Algorithm to Modulate the Uncertainty in the Organization's Network

B. Raja Koti, G.V.S. Raj Kumar, K. Naveen Kumar

Abstract*: In this modern era, all organizations depend on internet and data so, maintaining of all data is done by the third party in large organizations. But in this present on-developing world, one have to share the data inside or outside the organization which incorporates the sensitive data of the venture moreover. Data of the organization have sensitive data which should not share with any others but unfortunately, that data was there in the third party hands so; we need to protect the data and also have to identify the guilt agent. For this, we propose a model that would evaluate and correctly identifies guilt agents, for which a recursive partitioning has been created which is a decision tree that spills data in to the sub partitions and does the easiest way to get alert and at least one specialist or it can autonomously accumulate by some different means. The main intention of the model is to secure sensitive information by recognizing the leakage and distinguish the guilt agent.

Keywords: Information protection, Sensitive Data, Recursive partitioning, Data-Leak Detection, Guilt Agent.

I. INTRODUCTION

Information protection identifies with how a piece of data or information should be distributed with a view of its relative significance [1]. For example, you likely wouldn't see any problems with imparting your name to an outsider during the time spent presenting yourself, however there's other data you wouldn't share, in any event not until the point that you turn out to be more familiar with that individual. Open another financial balance, however, and you'll likely be solicited to share a huge sum from individual data, well past your name.

In the advanced era, we ordinarily apply the idea of information protection to basic individual data, otherwise called by Personally Identifiable Information (PII) and Personal Health Information (PHI) [2]. This can incorporate Social Security numbers, wellbeing and medicinal records, money related information, including ledger and charge card numbers, and even fundamental, yet at the same time sensitive, data, for example, full names, addresses and birthdates [3]. Intellectual Property which has Product design documents, Source code, Process documentation. In case of an Enterprise, Information is like to be financial documents, Employee details, Future plans. Not only has this in Banking Customer Information Credit/Debit card numbered, Individual details, Bank statements [4]. All this sensitive information should be in discloser only at the point when information that needs to be kept private gets in the wrong hands, terrible things can happen [5]. An information break at an administration office can, for instance, put top secret data in the hands. A disagreement at an organization can put exclusive information in the hands of a candidate who is not related to that organization and it was done by the member of its organization. So we did know who did it so we need to identify that and protect the information.

II. PRELIMINARIES

In this proposed strategy, we utilize the method of MAC-IP restricting jointly with keeping up a log-file at the server side to distinguish the guilt agent and secure the information that is being exchanged with no confirmation [6]. Despite the fact that, it is secure, there is a possibility of danger of information. To dispense with this hazard, our model proposes a methodology [7]. We have to ensure copyrighted data against security dangers that are finished by approved worker opportunity of development and new communication channels. As we have the log-file it's very difficult to identify the guilt agent, so for that we applied the Recursive partitioning [8] which is a method that includes developing a choice tree by partitioning an informational collection into subsets as indicated by descriptors, or rules, which separate between various classes of information. This makes it very easy to identify in the sub partition set data when compared to the completed data set.

Revised Manuscript Received on November 30, 2019.
***** Correspondence Author
B. Raja Koti*, Research Scholar, Department of Computer Science and Engineering, GITAM Institute of Technology, GITAM, Visakhapatnam, Andhra Pradesh, India. (raja.badugu@gitam.edu)
Dr. G. V. S. Raj Kumar, Professor, Department of Computer Science and Engineering, GITAM Institute of Technology, GITAM, Visakhapatnam, Andhra Pradesh, India. (gvsrajkumar.ganapavarapu@gitam.edu)
Dr. K. Naveen Kumar, Asst. Professor, Department of Computer Science and Engineering, GITAM Institute of Technology, GITAM, Visakhapatnam, Andhra Pradesh, India. (naveenkumar.kuppili@gitam.edu)

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license http://creativecommons.org/licenses/by-nc-nd/4.0/

A. Proposed Algorithm

In our proposed methodology, we utilize the system of MAC-IP binding together with keeping up a log document at the server side to recognize the guilt agent and ensure the information that is being exchanged without confirmation.

Despite the fact that, it is secure, there is a shot of danger for information. To dispose of this hazard, our model proposes a methodology. We have to ensure copyrighted data against security dangers that are finished by approved worker opportunity of development and new correspondence channels.

```
S ← Server Starting
sc ← socket created
L ← Listen Connection

∀nc: <c, nc> ∈ SNW do
    MAC ←Get MAC Address
    IP ← Get IP Address
    B ← MAC+IP hash
    f ← B hash values store in File

C← Client
    S ← rq client request to the server for connection
    If  ∀rq ← < nc, f > ∈ f  then
            Output : rq Accept
            Input  : commands , functions
            exit ();
        end

    else ∀rq!= f  then
            Output: rq Accept
            Input:  get file
            Enc ← rf    encrypt the requested file
            D ← Enc download encrypted file to client
            exit ();
        end
end
exit ();
```

Step 1: Get IP, MAC and Bind and store all the values in server
Step 2: Any request comes from client then Server check bind and response
Step 3: If bind match server will respond to client
Step 4: If bind is not matched, server is alerted and response to client
Step 5: For that client, Server gets IP, MAC from it
Step 6: If client sent any file to another client
Step 7: Server records all the move of its users and identifies that miss match request
Step 8: . And checks the all log records of it
Step 9: Finalize the data leaked or not.

At first, the IP address and the MAC address of the system are combined or bind and put in the host server as a check document. The idea of MAC-IP binding decreases the modification of IP or MAC addresses meaning a framework's MAC with an IP address interfaces with the system just with the limited IP. The information in an association is transmitted all through utilizing organization. The file records every single move of the information in the convention and the clients that recover this information gives a most extreme affirmation that this information remains inside the verified organization.

The limited MAC-IP addresses that have been kept as log file in the host server are currently placed with each record of information development in to file. For example, in the event that User1 gets to the information, his IP and MAC address are connected with this log recorded file and activity was record in to recorded file in server. Presently, when the MAC-IP address is connected with the log file in the server, when there is an unapproved record of move(s) in the convention, that record is recognized. Utilizing the MAC-IP address connected to the specific log file and the timestamps assigned by these records, the guilt agent can be distinguished when he makes a move without the authorization of the super-client. Protecting the leaked data indeed, even subsequent to finding the guilt agent; the exchange of information is fruitful to the outside individual/organization. To shield this information from being leaked to by the unapproved clients, our technique proposes the accompanying: The MAC-IP addresses that are connected to the information in the log file dependably checks for the right MAC and IP address. In the event that it is spilled and exchanged to another organization or individual that are endeavoring to get to this information, the file of information distinguishes a befuddle in its own particular MAC and IP addresses and acknowledges it as out of its unapproved client or approved client. Presently, this unapproved client that gets files containing in that information is either defiled or encoded utilizing encryption algorithms.

In the Fig.1, we should expect that two distinct kinds of associations are there which was associated by web organize. Two sorts of ways drawn, those are the correspondence in the middle of them green line was right correspondence on that the predicament worth was coordinated with the goal that information will share among them.

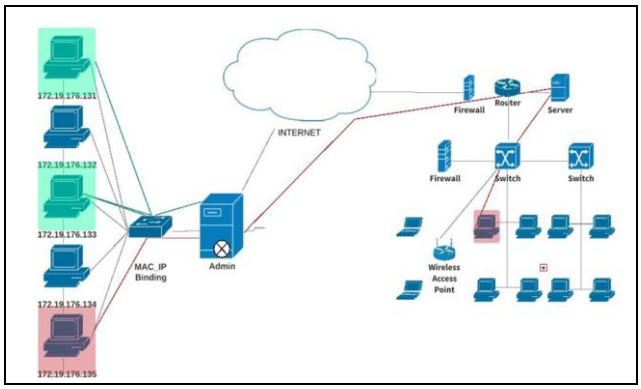

Fig. 1.The proposed methodology diagram

The red line shows confound of mismatch esteem so Admin will record those two client IP and MAC address so as to check that what they were sharing among them and furthermore distinguish if client is guilt agent or not. So every time Admin will get this notification alarms with the goal that it tends to be checked and distinguish information that leaked and furthermore who did it.

B. The Recursive Partitioning Algorithm

The essential thought is that every node is connected with a particular model. In a particular case of a part belonging to the node, it is necessary to consider a fluctuation test for finding out the parameter instability. During instability of separating variables, Zj divide the node into B in the neighborhood optimal segments and replication of the process.

If no significant instabilities are found, the recursion stops and returns a tree where each terminal node (or leaf) is associated with a model of type M (Y, θ) was shown in figure 2. More precisely, the below steps of the algorithm is followed

Step 1: Fit the model once for all perceptions in the present hub by assessing ˆθ through minimization of the target work Ψ.

Step 2: Evaluate whether the parameter measures are fixed regarding each requesting Z1, Z`. On the off chance that there is some general control, select the variable Zj related with the highest parameter insecurity, if not it will stop.

Step 3: Find the split point(s) that locally streamline Ψ, either for a settled or adaptively selected some parts.

Step 4: Split the data node into daughter nodes hubs and repeat the strategy.

The details for steps 1 to 3 are determined in the accompanying figure 2. To keep the notation simple, the dependency on the current fragment is suppressed and the signs instituted for the universal model are utilize, i.e., n for the number of explanations in the current node, ˆθ furthermore parameter estimate and B for the number of daughter nodes chosen [9].

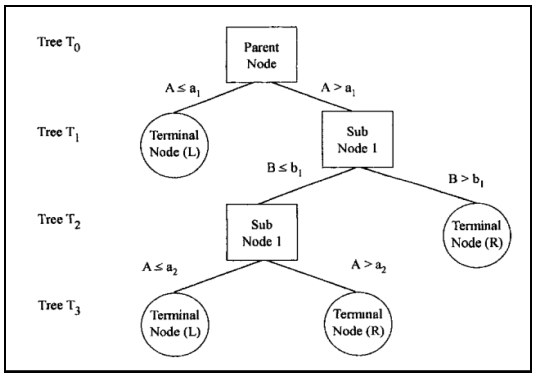

Fig. 2.Recursive Partitioning Algorithm Tree (Addict Behav., 2007)

III. EXPERIMENTAL EVALUATION

To comprehend the Recursive Partitioning process, a couple of fundamental ideas should initially be comprehended. The first of these is the idea of splitting (partitioning) the data set which the main training was set. Within the procedure of decision tree enlistment, we need to consider what questions we will request that all together direct the client down the suitable sides. For straightforwardness, how about we consider that every potential inquiry can have a genuine or wrong answer, hence, a specific set will have at most two ways from it to the following node(s) in the side. Each conceivable estimation of each conceivable component inside the training set will be a potential split that should be possible. The outcome is that we will have the capacity to go down the right side or the left side in view of the information and we will adequately part the information at every hub into two autonomous gatherings – this is apportioning.

For instance, let us consider a training set which has simply numerical information which is our data set that was taken for experiment. The highlights will be called Xn and the conceivable values for those highlights will be called Ym. Subsequently, every inquiry that could be asked can take the shape, "Is Xn not exactly or equivalent to Ym?" The subsequent answer will guide us down the proper side, e.g., if Xn is not exactly or equivalent to Ym, at that point go left, generally, go right. When we have two new sets (youngsters'sets) connected to a past set (the parent set), we can rehash the procedure for every child set autonomously utilizing just the perceptions present in that child – this is the step for recursive process.

The following idea is identified with how we pick which thing to ask. Keep in mind that we could make an inquiry of the above shape for each conceivable estimation of each conceivable element inside our training set. To choose the most proper inquiry, we build up a measure called the perfection measure that we can use to choose which split is the most ideal split from our decisions. There are numerous perfection estimates accessible. The most straightforward (however certainly not the best) is the outright perfection of classes spoke to as a percent virtue that we can accomplish in the sets, and this will serve to show the procedure. The virtue measure answers the inquiry, "In light of a specific split, how great of work did we do of isolating the two classes from one another?" We figure this perfection measure for each conceivable split and pick the one that gives the most astounding conceivable esteem.

At long last, we should have some kind of ending criteria. If we somehow happened to enable the part procedure to proceed until each leaf just had one perception, we would have a decision tree! This circumstance, be that it may be exceptionally dangerous in the event that we need to utilize our subsequent decision tree with the end goal of prediction. Doubtlessly a tree of this sort is over fit for the training information and won't perform well on new information. To get out this procedure, we acquaint with ending criteria with end the recursive partitioning process. Much like perfection measures, halting criteria come in a wide range of structures, including

- A most extreme number of sets in the tree. When this maximum number is achieved, the procedure is ended.
- A minimum number of perceptions in a specific set can be set with the end goal that if the quantity of perceptions in a set is not exactly or equivalent to a base esteem, dividing of that set won't be endeavored, and it turns into a leaf.
- An edge for the perfection measure can be forced with the end goal that if a set has virtue esteem higher than the edge, no apportioning will be endeavored paying little heed to the quantity of perceptions. (A set which is sufficiently complete in view of a predefined edge does not require any further part paying little mind to what number of perceptions are available.)When we have confirmed that we will stop the procedure, we have viably achieved a leaf in our tree. In view of the perceptions that have endured the tree to that leaf, we can appoint it class esteem.

One normal approach to do this is to figure out which class is in higher numbers and utilize that class as the leaf's class esteem. This is only a dominant part controls voting technique and, similar to every other part of tree acceptance, various strategies exist by which to decide a leaf's class mark.

While the above system for decision tree acceptance can be adjusted and tweaked for the job needed to be done, the general technique continues as before paying little respect to the decision of perfection measures or ending criteria. Moreover, while this is a technique that is utilized broadly all through the field of example acknowledgment, a potential defect exists. At every set which we have chosen should be part, we take a look at the quick consequences for the youngsters. As specified above, we glance through every last conceivable approach to partition this present set's perceptions and pick the one which gives us the best virtue in the prompt child. It is possible, in any case, that in the event that we pick a part criterion which isn't ideal, we may have better options considerably later in the tree which we can't find in the quick youngsters. At the end of the day, an imperfect split may prompt an expansion in optimality of decisions for later parts.

The model was implemented by using the python language including the crypto, random, requests, etc.., as we did for the bounded MAC-IP addresses for all the users who are connected to the network which belong to the organization that are stored in the file that was placed in the host server, and also the log file which record each movement of data of its users [6]. For user to access the data, IP and MAC address are linked and stored as a file in server. When there is an unapproved record of move(s) in the convention, that record is distinguished. Utilizing the MAC-IP address connected to the specific log document and the time stamps distributed by these records, the blame operator can be recognized when he makes a move without the authorization of the super-client to an unapproved client.

When all the user's data requests, responses, moving records are stored in the server as a log record files then after, severe distinguish that miss matches demands and check all the log records of it conclude the information status. All those things can be done in small amount of dataset and that too limited users only. Coming to large dataset, this model will take time to get notified the miss match request and then its late of getting alerts are also the huge loss to an organization so, for this issue we are applying the recursive partitioning mechanism to get more effective results. This recursive partitioning has been creating a decision tree; means spilled large dataset into the sub-partition and does the easiest way to get alerts of miss match request in the large dataset.

Data Set that was taken for the exploratory is NUIX which utilized the EDRM ENRON PST informational collection that contains: 1.3 million email, messages and connections from previous ENRON staff having 168 Microsoft standpoint .PST documents that was right around 40 GB of information [10]. This contained numerous examples of private, wellbeing and monetary information and individual data each one of those sends was sent and got by staff of Enron in the advancement of everyday business.

IV. RESULT AND DISCUSSION

After getting the recorded file for data moves for the entire users in the organization then the model was applied on it, then its split according to the categories. Performance measures Sensitivity and Specificity are the two familiar meters of the natural statistical validity of a leakage test are the probabilities of detecting by test among the true leakages nodes (L+) and true non- leakages nodes (L-). For the two responses, the results regarding true positive (T+) and true negative (T-) can be summarized in a 2×2 contingency Table 1. The columns stand for the two categories of true leaked status and rows stand for the test results. Sensitivity(S) or True Positive Rate (TPR) is conditional probability of in the approved manner recognizing the data leakage by test: SN = P (T+/L+) = TP/ (TP + FN) and specificity or True Negative Rate (TNR) is conditional probability of correctly identifying the non-leakage by test: SP = P (T-/L-) = TN/(TN +FP). False positive rate (FPR) and false negative rate (FNR) are the two other common terms, which are conditional probability of positive test in non-leakage: P (T+/L-) = FP/ (FP + TN); and conditional probability of negative test in leaked: P (T-/L+) = FN/ (TP + FN), respectively. Calculation of sensitivity and specificity of the used data set and the same data have been used to draw a ROC curve.

Table- I: A confusion matrix

		Actual	
		Sensitive	***Non- Sensitive***
Predicted	***Sensitive***	**TP**	**FP**
	Non-Sensitive	**FN**	**TN**

Few available sensitive datasets are accessible to internet, data set which are taken for the experimentation was the source data at a glance for this exercise, NUIX used the EDRM ENRON PST data set, which comprises of 1.3 million email, messages and attachments of documents from former ENRON staff having 168 Microsoft outlook .PST files that were almost 40 GB of data (10) contained many instances of private, health and financial data and personal information all those emails was sent and received by the staff of Enron in the progress of the day-to-day business.

The dataset is a sequence of items shows on EDRM ENRON, consisting of 12,892 trained files and 5,240 testing files in 130 categories, every group in single folder. Also a multi-label dataset indicates every file might fit in to numerous groups. It is identified that the data leak identification is dissimilar from the chore of files categorization. The data leakage method assesses the sensitivity of files accordingly to sensitivity semantic.

Therefore, it relates the precise group of files, whereas the multi-label cases shall be overlooked by data leakage method. While many groups in the essential dataset remain excessively minute (e.g., it's having one or two files) to execute additional assessment, the chosen group whose numeral of files is further than hundred, for both trained and test. Seven capable groups (the figure of files varies from hundreds to thousands) are exhibited as of the 130 groups as the crucial dataset.

The fundamental dataset involves of 870 training files and 130 testing files. To enhance simulation of the real-world information leak situations, specific sensitive information by deducing single group as sensitive and the remaining are non-sensitive. For further complete assessment, that does unite single group as sensitive and another as non-sensitive accidentally from the fundamental dataset every occasion. Lastly, 48 sets of information are created for the subsequent numerical trials, respectively with training and a test set encompassing thousands of files.

The consequence of recognition is calculated by accuracy, recall by using the metrics of Receiver Operating Characteristic (ROC) curve. The confusion matrix in Table3.1 shows the complete study constraints of detection consequences. The Accuracy (Acc) methods the percentage of appropriately identified cases, measured by Accuracy (TP+TN)/ total. The recall measures the percentage of sensitive files that remain the identified as sensitive, deliberate by Recall TP/ (TP+FN), which replicates the capability to identify sensitive information leaks. The ROC curve is drawn with TP compared to FP at numerous thresholds, demonstrating the diagnostic capability of our projected method.

Table- II: Performance Measures of the Model

Number of Groups	*Accuracy*	*Recall*
39	100%	100%
4	96.94%	99.91%
3	96.91%	99.91%
1	96.78%	98.99%

The experimentation engages 42 sets of information ranging in volume from a limited hundred to numerous thousand. Situations of identification original information and identified transmission information are trailed, with dissimilar volume of dataset. The files of similar group in trial set are dissimilar from the exercising set, so for trial set is employed as the customized or new information to be identified. The subsequent trials in this chapter aspire to projected method deals with huge quantity of information capably, tolerate the transmitted or new data in identification and compare to the similar dataset in conditions of accuracy, recall, ROC are shown in figures 3 to figure 6.

Sensitivity and Specificity the specificity or True Negative Rate (TNR) is defined as the percentage of clients who are correctly identified as being leaked the data or not: Specificity = TN / (TN + FP). The quantity 1-specificity is the false positive rate and is the percentage of clients that are incorrectly identified as having the probability of leaked information. The sensitivity or true positive rate (TPR) is defined as the percentage of clients who are correctly identified as having the probability of leaked information. Sensitivity = TP / (TP + FN).

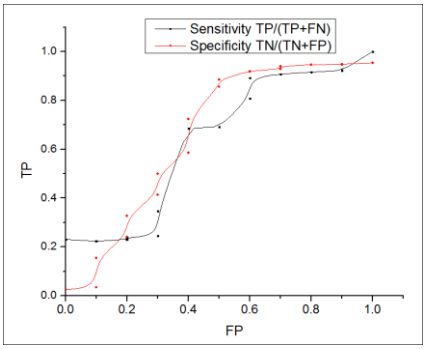

Fig. 3.Performance Measures of the test data set Sensitivity and Specificity

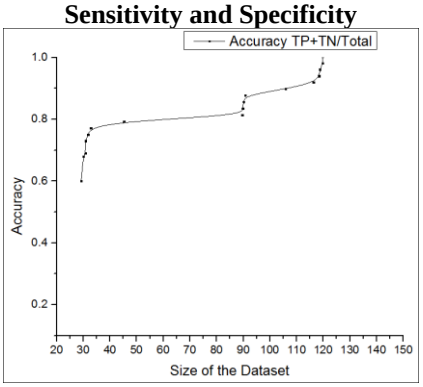

Fig. 4.The accuracy is plotted for the various range of testing data set

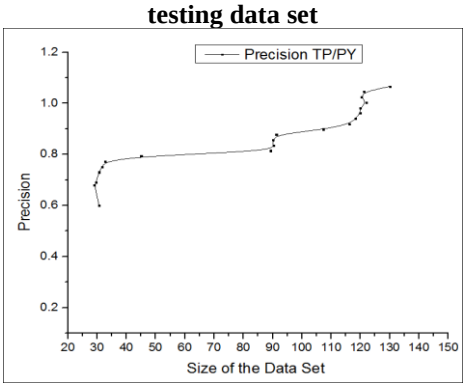

Fig. 5.The Precision is plotted for the various range of testing data set

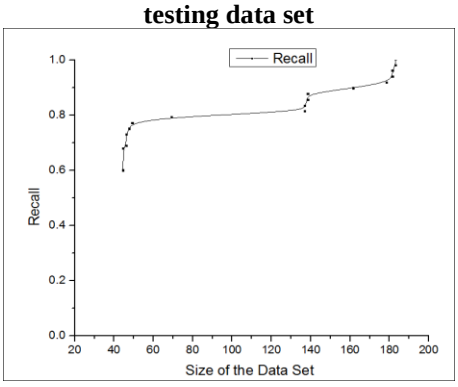

Fig. 6.The Recall is plotted for the various range of testing data set

V. CONCLUSION

In this article we proposed the data-leak detection model and its identification of guilt agent. As to determine Sensitive data issue by utilizing of MAC-IP authoritative along with encryption strategy for ensuring leaked information.

Using especially recursive partitioning algorithm to the exposure of the sensitive data is kept to a minimum during the detection and identify where data was got leaked so that an organization will take immediate action for leakage and also to control advance leakage problem. We have conducted extensive ex-pediments to validate the accuracy, privacy, and efficiency of our solution to the model. For future work, we intend to center around outlining a helped component for the entire information leak detection or recognition for extensive organizations.

REFERENCES

1. Xiaokui Shu and Danfeng (Daphne) Yao, "Data Leak Detection as a Service", Department of Computer Science Virginia Tech Blacksburg VA, USA.
2. B. Raja Koti, Dr. G.V.S. Raj Kumar, Dr. Y. Srinivas, "A Com-prehensive Study and Comparison of Various Methods on Data Leakages", International Journal of Advanced Research in Com- puter Science, Volume 8, No.7, July – August 2017, pp-627-631.
3. "Management of Data Breaches Involving Sensitive Personal Information (SPI)". Va.gov. Washington, DC: Department OF Veterans Affairs. 6 January 2012. Archived from the original on 26 May 2015. Retrieved 25 May 2015.
4. Balachander Krishnamurthy, Craig E. Wills, "On the Leakage of Personally Identifiable Information Via Online Social Networks", WOSN'09, August 17, 2009, Barcelona, Spain. 2009 ACM 978-1-60558-445-4/09/08.
5. P. Papadimitriouand H. Garcia-Molina, "Data leakage detection", Technical report, Stanford University, 2008.
6. B Raja Koti, GVS Raj Kumar, Y Srinivas, "Identification of Guilt Agent and Leaked Data by Using MAC-IP", International Journal of Applied Engineering Research, 2017, Volume 12, Issue 22, pp 12237-12245.
7. B Raja Koti, G V S Raj Kumar, "Information leakage detection and protection of leaked information by using the MAC-IP binding technique",International Journal of Engineering &Technology, Volume 7, Issue 1.7, 2018, pp230-235.
8. Torsten Hothorn Kurt Hornik Achim Zeileis "Unbiased Recursive Partitioning: A Conditional Inference Framework" Department of Statistics and Mathematics Wirtschafts universität Wien, Research Report Series, Report 8, July 2004, http://statistik.wu-wien.ac.at/
9. Achim Zeileis, Torsten Hothorn, Kurt Hornik "Model-based Recursive Partitioning", Journal of Computational and Graphical Statistics, Volume 17(2), 492–514.
10. B. Raja Koti, Dr. G V S Raj Kumar, Dr.Y. Srinivas, Dr. K. Naveen Kumar, "A Methodology for Identification of the Guilt Agent based on IP Binding with MAC using Bivariate Gaussian Model", Journal International Journal of Advanced Computer Science and Applications(IJACSA), Volume 9, Issue 11, November 2018, pp. 293-299.
11. https://www.nuix.com/edrm-enron-data-set/edrm-enron-data-set.

AUTHORS PROFILE

Mr. B. Raja Koti is a Ph.D. student at GITAM Deemed to be University, Andhra Pradesh. Department of Computer Science and Engineering. His research interests include computer and network security, information security, Block Chain, Internet of Things, Log Analysis, and Network Security.

Dr. G.V.S. Raj Kumar received his Ph.D. in Computer Science and Engineering from the Andhra University. He is currently working as a Professor in the Department of Computer Science and Engineering at GITAM Deemed to be University, Andhra Pradesh. He is a member of the Computer Society of India, the Cryptology Research Society of India and the International Association of Engineers. He guided 3 research scholars in computer science Engineering and Information Technology. He published more than 35 publications, articles in refereed journals and conference proceedings. His research interests include Computer Networks and Network Security, Computer Forensics, Mobile Security, Block Chain technology and Image Processing.

Dr. K. Naveen Kumar is an Assistant Professor and has been on the faculty at the GITAM Deemed to be University. He teaches in the Department of Computer Science and Engineering, primarily in the areas of information security and Image processing, mobile security and computing. He has published numerous articles in magazine, refereed journals and conference proceedings, His concentrates with in the areas of Information Security, Block Chain, Internet of Things, E-Commerce and Technological Innovation.