A study secure multi authentication based data classification model in cloud based system

Sakshi kaushal[1], Bala buksh[2]
[1]Computer science Engineering, Career Point University, Himachal Pradesh, India
[2]Computer Science Engineering, R N Modi Engineering College, Rajasthan, India

Article Info

Article history:

Received Jul 2, 2019
Revised Apr 13, 2020
Accepted May 15, 2020

Keywords:

Bayesian technique
Cloud computing
Data mining
Internet

ABSTRACT

Abstract: Cloud computing is the most popular term among enterprises and news. The concepts come true because of fast internet bandwidth and advanced cooperation technology. Resources on the cloud can be accessed through internet without self built infrastructure. Cloud computing is effectively managing the security in the cloud applications. Data classification is a machine learning technique used to predict the class of the unclassified data. Data mining uses different tools to know the unknown, valid patterns and relationships in the dataset. These tools are mathematical algorithms, statistical models and Machine Learning (ML) algorithms. In this paper author uses improved Bayesian technique to classify the data and encrypt the sensitive data using hybrid stagnography. The encrypted and non encrypted sensitive data is sent to cloud environment and evaluate the parameters with different encryption algorithms.

Corresponding Author:

Sakshi kaushal,
Computer science Engineering,
Career point university kota Rajasthan,
Aalniya, Rajasthan 324005, India.
Email: sksakshi.kaushal@gmail.com

1. INTRODUCTION

The growth and use of internet services cloud computing becomes more and more popular in homes, academia, industry, and society [1]. Cloud computing is envisioned as the next-generation architecture of IT Enterprise, which main focus is to merge the economic service model with the evolutionary growth of many offered approaches and computing technologies, with distributed applications, information infrastructures and services consisting of pools of computers, storage resources and networks. Cloud computing has almost limitless capabilities in terms of processing power and storage [2]. Cloud computing offers a novel and promising paradigm for organization and provide information and communication technology (ICT) resources to remote users [3].

Client does not handle or control the cloud's infrastructure so far, they have power over on operating systems, applications, storage space, and probably their components selection. A Cloud System defined at the lowest level as the difficult server, which having the substantial devices, processing unit and memory [4]. To give the distribution of existing services and applications, the Cloud server is auxiliary divided in multiple virtual machines [5]. Every virtual machine is definite with various specification and characterization. The logical partition of existing memory, resources and processing capabilities is done. The separation is based on the requirement of the application, user requirement and just before achieves the quality of services. In this type of environment, there can be multiple instances of related services, products and data. When a user enters to the Cloud System, here is the main requirement of identification of useful Cloud service for Cloud user. As also there is the necessity to process the user request successfully and reliably.

Types of cloud

To given a secure Cloud computing key a main decision is to focus which type of cloud to be implemented as shown in Figure 1. There are four types of cloud deployment models are a public, community, private and hybrid cloud [6, 7].

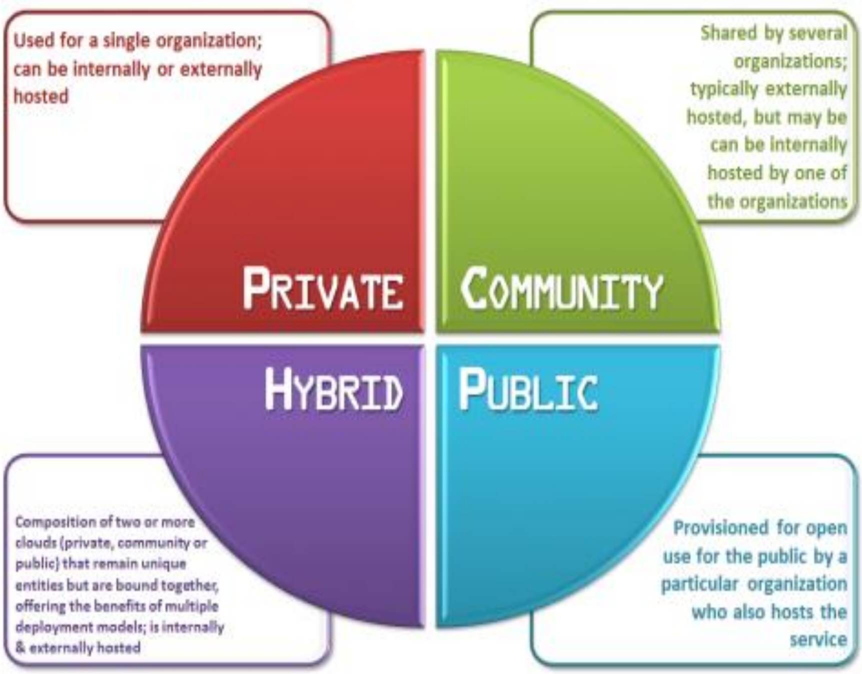

Used for a single organization; can be internally or externally hosted
Shared by several organizations; typically externally hosted, but may be can be internally hosted by one of the organizations
PRIVATE
COMMUNITY
HYBRID
PUBLIC
Composition of two or more clouds (private, community or public) that remain unique entities but are bound together, offering the benefits of multiple deployment models; is internally & externally hosted
Provisioned for open use for the public by a particular organization who also hosts the service

Figure 1. Types of cloud [2]

a. Private cloud

A private cloud is place up within an organization’s interior project datacenter. In the private cloud, virtual applications provided by cloud merchant and scalable resources are pooled together and presented for cloud users so that user share and use them. Deployment on the private cloud can be greatly more secure than that of the public cloud as of its specific internal experience only the organization and selected stakeholders have access to control on a specific Private cloud [8].

b. Hybrid cloud

A hybrid cloud is a private cloud associated to one or other external cloud services, centrally provisioned, managed as a solitary unit, and restricted by a protected network. A hybrid cloud architecture combination of a public cloud and a private cloud [9]. It is also an open architecture which allows interfaces by means of other management systems. In the cloud deployment model, storage, platform, networking and software infrastructure are defined as services to facilitate up or down depending on the demand.

c. Public cloud

A public cloud is a model which gives users to access the cloud using interfaces by web browsers. A public cloud computing most commonly used cloud computing service [10**].** It’s usually based on a pay-per-use model, like to a prepaid electricity metering system which is capable enough to supply for spikes in demand for cloud optimization. This helps client to enhance the match with their IT expenditure at operational level by declining its capital expenditure on IT infrastructure. Public clouds are not as much of secure than the other cloud models because it places an extra burden of ensuring that all data and applications accessed on the public cloud are not subjected to malevolent attacks.

d. Community cloud

The Cloud System can exist situate up particularly for a firm, organization, institution [11]. The rules for authentication, authorization and usage architecture can be defined exclusively for the organization users. The policy, requirement and access methods are defined only for the organization users. Such type of organization secret Cloud System is called Community Cloud.

Data classification

By classifying the data using supervised machine learning algorithm into sensitive data and non-sensitive data in order to reduce the data hiding time [12]. Data classification is done using improved Boosting algorithm which will classify the data according to the security requirement. Data classification is a machine learning technique used to predict the class of the unclassified data. Data mining uses different tools to know the unknown, valid patterns and relationships in the dataset. These tools are mathematical algorithms, statistical models and Machine Learning (ML) algorithms. Consequently, data mining consists on management, collection, prediction and analysis of the data. ML algorithms are described in two classes: supervised and unsupervised.

a. Supervised learning

In supervised learning, classes are already defined. For supervised learning, first, a test dataset is defined which belongs to different classes. These classes are properly labelled with a specific name. Most of the data mining algorithms are supervised learning with a specific target variable. The supervised algorithm is given many values to compare similarity or find the distance between the test dataset and the input value, so it learns which input value belongs to which class.

b. Unsupervised learning

In unsupervised learning classes are not already defined but classification of the data is performed automatically. The unsupervised algorithm looks for similarity between two items in order to find whether they can be characterized as forming a group. These groups are called clusters. In simple words, in unsupervised learning, “no target variable is identified”.

The classification of data in the context of confidentiality is the classification of data based on its sensitivity level and the impact to the organization that data be disclosed only authorized users. The data classification helps determine what baseline security requirements/controls are appropriate for safeguarding that data.

2. LITERATURE REVIEW

Singh et al. [13] clear a complete survey on different safety issues that affect communication in the Cloud environment. A complete discussion on main topics on Cloud System is given, which includes application, storage system, clustering technique. The Cloud System approval security and extra security concerns are too discussed by the authors. The module level deployment and related security impact are discussed by the authors. Authors also discussed the related faith and confidence with subject categorization. Different security threats and the future solutions are also suggested by the authors. The paper also recognized a few forensic tools used by previous researchers to path the security leakage [13].

The descriptive and relative survey is specified to recognize different security issues and threats. Several solutions set by previous researchers are too provided by the authors. The safety mixing to unusual layers of Cloud System is accessible and moderately provided the solutions and the concerns. Just the comparative and vivid review of work complete is given. No analytical explanations are provided by the authors.

Faheem Z. *et al.* [14] explored a variety of types of inner and outer attacks so as to affect the Cloud network. Authors recognized the safety requirements in Cloud System in deepness and moderately possible mitigation methods are too defined based on previous studies. Authors had known the requirement of safety in Cloud System and virtual benefits. The assault impact and assault result are provided by the authors. Authors mostly provided a result against verification attack, sql injection attack, phishing attack, XML signature wrapping attack, etc. A lesson work on special attack forms and virtual solution methods is provided in the given work. The attack solutions or protection solutions are provided as the included layer to the Cloud System different exposure and avoidance-based approaches are too defined by the authors as the included Cloud service [15]. In future, a new attack detection or preventive method is required [14]. Abuhussein, Bedi, and Shiva proposed comparison between security and privacy attributes like backup, cloud employee trust, encryption, external network storage, access control, dedicated hardware and data isolation, monitoring, access computing services so consumers can make well educated choice cloud related insider threats lay in three groups: cloud provider administrator, employee in victim organization, who uses cloud resources to carry out attacks [16].

Derbeko *et al.* [17] defined the safety aspects for Cloud System beside a variety of attacks in actual time Cloud environment. The calculation is provided for MapReduce situation with communal and confidential Cloud specification. The privacy data computation, integrity analysis and accuracy of outcome are investigated by the authors. The constraint characterization and challenges of MapReduce scheme for data safety are discussed by the authors. The security and privacy control with master procedure are defined to attain superior safety aspects for Cloud System. Different safety methods, including authentication, authorization and access control observations are also provided by the authors [17].

With reference [18, 19] Tawalbeh, Darwazeh, Al-Qassas and Aldosari propose a secure cloud computing model based on data classification. The proposed cloud model minimizes the overhead and processing time needed to secure data through using different security mechanism with variable key sizes to provide the appropriate confidentiality level required for the data. They have stored data using three levels: basic, confidential and highly confidential level and providing different encryption algorithm to each level to secure the data. This proposed model was tested with different encryption algorithms and the simulation results showed the reliability and efficiency of the proposed framework.

a. Review on the basis of authentication

Cusack and Ghazizadeh [20] recognized the service access threat with the human behavior study Authors known the best behavioral belief for Cloud server with single sign in approval the individuality management and comparatively optimization to human actions are also provided. Authors evaluated the special risk factors below security solutions recommended These acknowledged risks take in human user risk, disclosure risk and service security risk. Trust behavior-based trust analysis process is provided to manage the access behavior.

Contribution:

- Safety measure and belief observations are provided based on the client behavior analysis.
- A solo sign in approval is provided for efficient identity confirmation and management.

Scope:

- Only the way of sign in approval is provided, but authors perform not provide the real time implementation or analysis. In future, such real time execution can be applied to confirm the work.

Yang *et al.* (2016) clear a full study work on GNFS algorithm for the Cloud System. beside with method exploration and analysis, a fresh block Wiedemann algorithm is too provided by the authors. The process is based on strip and cyclic partitioning to do block encoding. The defined process works on a similar block processing to decrease the processing time. The chronological block processing can be complete to improve the information encoding by the improved strip block form as a result that the information security can be enhanced [21].

Contribution:

- A narrative widemann algorithm is providing with enhanced strip block giving out for data encoding.
- The similar block processing process has improved the processing for encoding elevated volume data.

Scope:

- The assessment of the process is provided below block size and competence parameter. No attack consideration is given. In the future, analysis relation to extra parameters such as file type or the difficulty measures can be provided.

b. Review on the basis of security assessment

Modic *et al.* (2016) provided a learn of accessible Cloud security estimation methods for actual world applications. The document also defined a fresh security assessment way called Moving Intervals Process (MIP). The quantitative analysis process is also definite with processing method based on dissimilar security or accessibility parameters. A survey is also system to control different categories and structures based on the quantitative requirement. Authors identified the least maximum and actual time requirements and then comparatively perform the price specific measure. The manage measure is here defined in the form of scores [22].

Zhang (2014) has provided a better view of Cloud security below issue examination and financial impact on industry or the organization. The dynamic body-based livelihood control is discussed on static network. The domination specific threat and fulfillment rule are evaluated to get better the power of Cloud System. The danger evaluation-based value measures are provided to put in up long word practices to the Cloud System [23].

Kalloniatis (2013) has defined a reading work to know various security intimidation in Cloud System. The solitude and security issues are recognized with related properties. This paper also definite the process to provide security over Cloud System. Unusual threats on different Cloud service models are known by the author. The necessity engineering for Cloud System is too identified with essential challenges and characterization. The admission criticality and challenges are explored with threat evaluation and impact analysis [24].

c. Review on the basis of cloud security framework

Ramachandran (2016) provided a broad descriptive study on various condition engineering ways and their management. The paper worked mostly on safety as a repair layer to get better the safety aspects and their sharing to the Cloud environment. The included service model with software development system is provided by the author. The logical research and obligation mapping are provided as an example which is being used in various models as integrated form. The safety privileges for every stakeholder is identified and provided the technique for the security requirement, method and maintenance [25].

Contribution:

- A safety as service included Cloud System development representation is provided for scheme design and distribution.
- The analytical copy is provided for the obligation of different stakeholders as well as procedure stages of Cloud System.

Scope:

- A generalized model is provided by risk rating and danger prioritization.
- The business exact model with real time configuration is not provided.
- In future, the job can be applied on on hand Cloud System application or environment.

Chang *et al.* [21] defined an original security framework for Cloud System environment. The multi layered safety protection is provided next to different attacks with far above the ground level data concerns, including volume, veracity and velocity. Authors practical model for zero knowledge Cloud System that does not contain any user information. The information sharing, dependency and the data calculation are provided to attain effective block stage. The information encoding and the safe communication are provided by the authors. Authors also maintained the private fire storage and safe key management in storage space. The information sharing and approval is also going to through the key distribution methods. The concerns are also provided against a variety of attacks practical on unusual layers of Cloud System access.

Contribution:

- A hierarchy layer safety framework is provided to attain access control, attack preservation and encoded data storage.
- The safe sharing of data is performed by means of key management and approval in Cloud file system.

Scope:

- The services can be extensive in the form of prototypes so that the use can be enhanced for different Cloud business models.

Palmieri *et al.* [26] have done an adaptive sleuth energy-based analysis to identify DoS attack in Cloud data centers. Authors provided the service level analysis under availability, operation cost and energy parameters. Authors defined the problem to identify the DoS attack in network in early stage and to provide the attack resistive communication in Cloud network. The work is able to provide the analysis under availability and visibility parameters to give the analysis under pattern specific dynamic observation. The energy impact with potential effect is here analyzed for larger infrastructure to identify the attack in Cloud network. The attack ratio analysis and computation are defined to perform the attack detection. Authors estimated the attack effect with response time violation and determines the flow analysis-based service degradation. Authors provided the power management and consumption analysis to give the component level evaluation on Cloud environment. Authors provided an energy proportional system to reduce the peak power usage in attacked Cloud network and to reduce the effect of DoS attack.

Chonka and Abawajy [27] have defined a work to detect and reduce the impact of DoS attack in web service driven Cloud network. Authors defined the security system to observe the channel communication under the common problem identification and to reduce the impact of DoS attack. A problem analysis for XML DoS attack is given as new defense system that can provide a solution with pre decision and learning based observation. The defined network attack is able to observe the network under training and testing criteria to provide the effective attack preserved solution in the DoS attack network. The scenario specific observation with specification of response 84 pattern is analyzed to generate the classification rules and to provide the attack preventive probabilistic solution.

Michelin *et al.* (2014) used the authentication API to provide the Cloud communication solution against DoS attack. An unresponsive work behavior and relative protocol specific attack mapping are provided for REST applications. Authors identified the client behavior and relatively identified the malicious client in the network with response time observation. An attack specific Cloud management system is designed to define taxonomy for DDoS attack adaptation. An automated method is defined to analyze the network features and generate the attack features. Later on, all these features are combined to identify the victim type in the Cloud network. The attack scenario specific authentication measures have defined the solution against the DoS attack. At the early stage, the communication is monitored to identify the overload conditions and later on applied the filtration stage to generate the effective Cloud communication [21].

3. RESEARCH METHOD

3.1. Phases of proposed model

The proposed display is executed as specified in the accompanying stages to meet every one of the targets portrayed that are vital piece of this contextual investigation.

a. Creating virtual environment

First stage is to make a virtual domain where there is introduction of cloud servers, information specialists, virtual machines and assignments.

b. Authentication level

Figure 2 For recovery of information, the client needs to enroll himself with organization/association to get a legitimate username and secret phrase which is additionally put away at database of the organization/association.

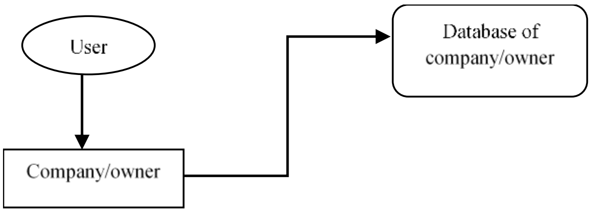


Figure 2. Process of registration

Figure 3 it portrays the transmission of client demand to cloud. The cloud checks if the client has entered amend username and secret phrase is validated into its database. In the event that yes then it then just the client is endorsed to utilize the cloud information. For confirmation process, the validated clients are coordinated with the as of now existed information put away at cloud catalogs. Client needs to give its username, secret key and answer the security question, if answers given by the client is right, at that point he is allowed to get to the cloud.

c. Secure authentication using image sequencing

To take care of the issue of security in distributed computing, we will send these two route strategies for stopping security ruptures on distributed computing. One is giving privacy at different levels of clients like proprietor, administrator and third-party utilizing image sequence base secret key that gives privacy from validation assaults at client end. Information Hiding Architecture use for safely transmitting the information over the cloud condition

This secret key depends on the groupings of a few pictures. It is more secure in light of the fact that arrangement of pictures is change without fail. Essentially that password is basically use for confirmation reason. Just authentic client will permit entering in cloud, in the event that they enter the right arrangement of picture. After verification, amid access of information tasks this interface will again tell the client arrangement, this time pictures gets rearrange, in view of succession of pictures secret word will likewise be change.

d. Proposed data classification architecture

Figure 4 is showing classification of data is a machine learning system used to anticipate the class of the unclassified data. Information mining utilizes one of kind instruments to get a handle on the obscure, genuine examples and connections inside the dataset.

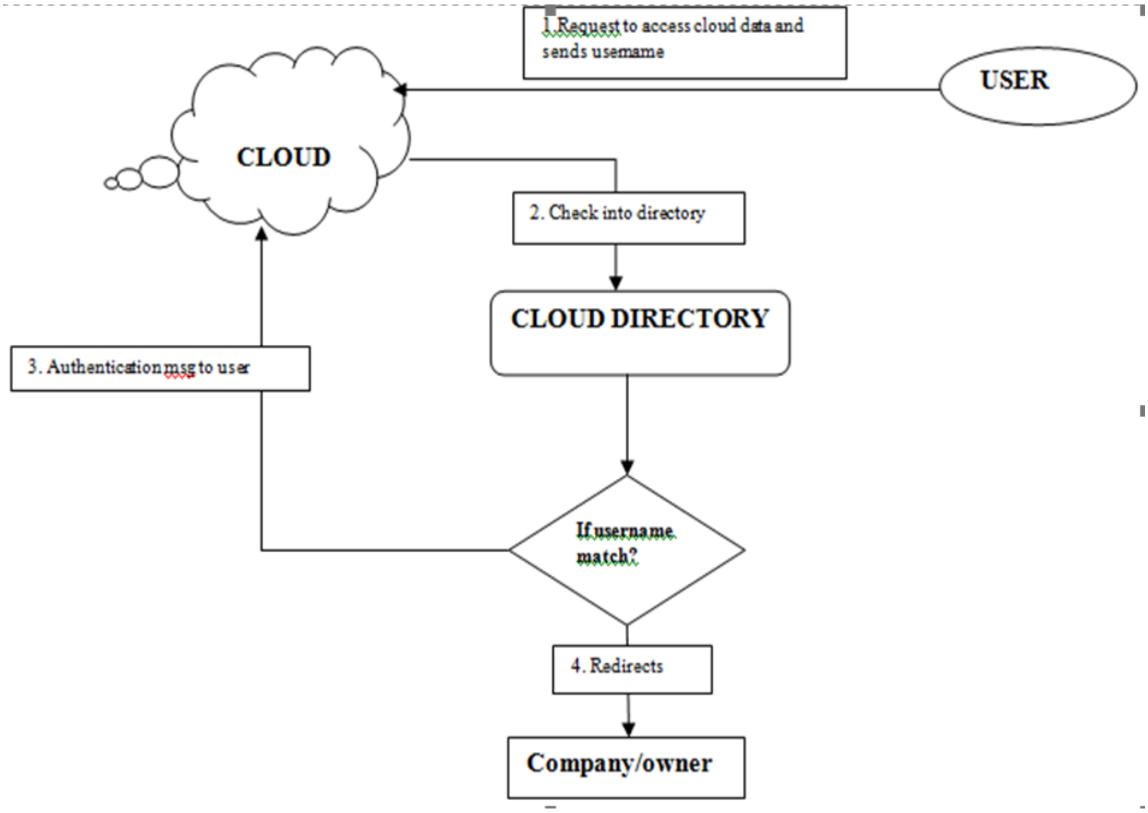


Figure 3. Request accessing process

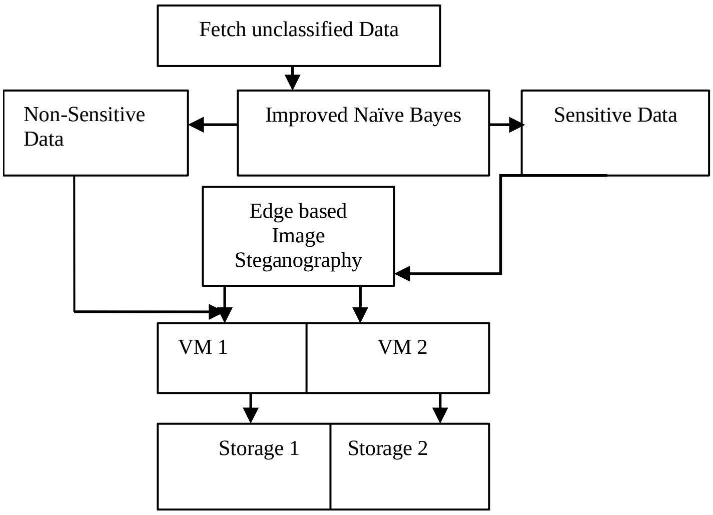


Figure 4. Data classification architecture

These devices are numerical counts, authentic models and expectation and assessment of the information. Subsequently, information mining comprises of administration, gathering, forecast and examination of the information. ML calculations are depicted in to 2 classifications: directed and unsupervised. For directed contemplating, initial, an analysis dataset is characterized which has a place with unmistakable classes. These exercises are skillfully marked with a specific recognize. Bunches of the information mining calculations are regulated concentrate with a point by point objective variable. In unsupervised learning classes are not adequately described yet rather course of action of the data is performed naturally. The unsupervised calculation searches for likeness between two contraptions keeping in mind the end goal to observe whether they can be described as shaping a group. In basic words, in unsupervised adapting, "no objective variable is distinguished". The arrangement of data inside the setting of secrecy is the characterization of learning headquartered on its affectability level and the affect to the association that information be uncovered just authorized clients. The information grouping figures out what gauge security necessities/controls are proper for protecting that information. The data is ordered into two exercises, secret and non-private (non-elite) data. The arrangement of the data relies upon the qualities of the data. The estimations of the touchy traits are delegated "classified" and estimations of the non-delicate characteristics are sorted as "non-secret".

By characterizing the information utilizing supervised machine learning calculation into delicate information and non-touchy information so as to lessen the information hiding time. Information classification is finished utilizing enhanced bayesian calculation.

Sensitive or confidential data

Private information comprises of exceptionally fundamental data of cloud clients' or affiliation. The illicit client can't get to classified/private information. Such information may incorporate the accompanying:

- Individual Information: It contains individual acknowledgment, for example, societal security ID, visa ID, CC Identification, driver's allow number.
- Monetary Accounts: Banking operational information, money related acc no.
- Trade Data: Manufactured products, future arranged information.
- Health check/Medical Data: Health concerned information of person.
- Administration Information: comprises of organization future arranging, government sane papers, organization gather papers.

Non-Sensitive or public data

This is otherwise called non-confined information. It is utilized by average citizens by means of web. The data that is considered as non-delicate or open data comprises of information which isn't crucial to the substance or affiliation. Such information contains information promoting material, squeeze explanation or preliminary information of an affiliation.

After finished with the order of information, the client became more acquainted with what dataset needs the security and what sort of dataset needn't bother with any security. To anchor the delicate or private informational collection, RSA encryption cryptographic system was utilized.

The K-Nearest Neighbor comprises of 'n' name information tests. Here 'n' is the no. of information esteems in the datasets: this can be appeared as:

$$D = \{d1, d2, d3...... d(n)\} \quad (1)$$

D= set of total number of samples. The Dmust have ‘n’ labelled value. d1, d2, d3...d (n) are diverse data samples.
The set of *n* labelled samples can be represented as:

$$D= \{d1, d2, d3...... d(n) \mid C\} \quad (2)$$

C = the data class for the target values. In this technique only one class is defined for sensitive or confidential data.

*K-NN Algorithm***:**

Step 1: To discover the arrangement of n name test-set that is D
Step 2: To discover the estimation of K
Step 3: To process the space b/w the new I/P and aggregate no of preparing dataset.
Step 4: Arrange the separations b/w the neighbour pixels and discover the K-NN which depends on the Kth remove measure.
Step 5: Determine the neighbour class.
Step 6: to discover the new class of inputted information dependent on larger part of votes.

3.2. Phase 3.data classification

In this proposed work, the improved Naïve Bayes machine learning is utilized to improve the execution of existing KNN system.

a. Combining naïve bayes with Decision table using Decision tree as Meta classifier.

b. Meta Learner is a learner scheme that combines the output of the naive bayes and decision table i.e. the base learner. The base learners' level-0 models and the meta-learner is a level-1 model. The predictions of the base learners are input to the meta-learner.

This will classify the data into: basic, confidential and highly confidential using the rules induced inthe learning algorithms which will identify which attributes of the data set are under vulnerability attacks. Troupe learning strategy includes an arrangement of various models are assemble together to enhance the forecast and soundness intensity of any model. It has two levels: base level-0 and Meta level-1. At base level a no. of calculations can run i.e, AdaBoost and packing calculation. At Meta level, otherwise called basic leadership calculation, irregular woodland tree is utilized. To use the preparation sets of information which is given by KNN display, the preparation set is registered with Euclidean separation work. To diminish the computational thickness, we enhanced the essential calculation of outfit learning.

Base level - 0: Also known as LEVEL-0. There are number of calculations running at base level in particular, Naïve Bayes and Decision Table i.e. the base students. These are parts of Ensemble Learning strategy. The individual yield of Naive Bayes and Decision Table will additionally be given to the meta level.

3.3. Working of Naïve Bayes

Naive Bayes algorithm is machine learning based approach. The fundamental requirement of machine learning based approach is a dataset that is already coded with sentiment classes. The classifier is modelled with the labelled data. For the purpose of this report, multinomial naïve Bayes is used as a baseline classifier because of its efficiency. We assume the feature words are independent and then use each occurrence to classify headlines into its appropriate sentiment class. Naive Bayes is used because this is easy to build and implement and to estimate the parameters it requires small amount of training data.

It follows from that our classifier which utilizes the maximum a posterior decision rule can be represented as:

$$c = \underset{c\epsilon C}{argmax}\ (P(c/m)) = \underset{c\epsilon C}{argmax}\ (\mathrm{P(c)}\ \prod_{1\leq k\leq n_d} P(l_k/c)\ ) \tag{3}$$

Where l_k denotes the words in each headline and C is the set of classes used in the classification, $P(c/m))$ denotes the conditional probability of class c, P(c) denotes the prior probability of a document occurring in class c and $P(l_k/c)$ denotes the conditional probability of word l_k given class c. To estimate the prior parameters, equation (1) is then reduced to

$$\mathrm{C}= (\underset{c\epsilon C}{argmax} \log \mathrm{P(c)} + \sum_{1\leq k\leq n_d} \log P(l_k/c)) \tag{4}$$

3.4. Working of decision table

A decision table is used for representing conditional logic by making a list of process which depicts the business level rules. These tables can also be used when constant numbers of conditions are there which are needed to be calculated and where an exact set of events to be used when the following conditions are to be met. These tables are very similar to decision trees except that tables will have the similar number of conditions that needed for evaluation and actions that are needed to be taken. While decision tree, contain one branch with addition of conditions that are necessary to be evaluated than other branches on the tree. The main idea of a decision table is of structuring the logic so as to generate rules that are derived from the data which have already been entered into table. A decision table consists of lists causes i.e. business rule condition and effects i.e. business rule action, which have been denoted by the matrix where each column represents a single combination. If number of rules are there inside the business which are expressed by the use of some templates and data then we can refer the decision table technique to accomplish the particular task. Individual row in the decision table collects and stores its data uniquely and then bind the data with a particular or customized template to generate a rule. It is not preferable to use decision tables if the rules are not following a set of templates.

a. Check the capability of classifier whether it can handle data or fails using getCapabilities method
b. Removing instances with missing values
c. Add each instance to decision table.
d. Each Instance I is assigned a category by finding the line in the decision table that matches the non-class values of the data item
e. Wrapper method is used to find a good subset of attributes for inclusion in the table.
f. By eliminating attributes that contribute little or nothing to a test model of the dataset, the algorithm reduces the likelihood of over-fitting and creates a smaller and condensed decision table.

g. The attribute space is searched greedily either top to bottom or bottom to top. A top-to-bottom search adds attributes at each stage; this is called forward selection. A bottom-to-top search starts with a full set of attributes and deletes attributes one at a time; this is backward elimination.

Decision table for data set D with n attributes A1, A2, ..., An is a table with schema R (A1, A2,..., An, Class, Sup, Conf). A row Ri= (a1i, a2i, ..., ani, ci,supi, confi) in table R represents a classification rule, where aij(1 j n) can be either from DOM(Ai) or aspecial value ANY, ci{ c1, c2, ..., cm}, minsupsupi1, and minconfconfi1 and minsup and minconf are predetermined thresholds. The interpretation of the rule is if (A1 = a1) and (A2 = a2) and … and (An= an) then class= ci with probability confi and having support supi, where ai ANY, 1 j n.

The decision table generated is to be used to classify unseen data samples. To classify an unseen data sample, u(a1u, a2u, ...,anu), the decision table is searched to find rows that matches u. That is, to find rows whose attribute values are either ANY or equal to the corresponding attribute values of u. Unlike decision trees where the search will follow one path from the root to one leaf node, searching for the matches in a decision table could result in none, one or more matching rows.

3.5. One matching row is found

If there is only one row, ri(a1i, a2i, ..., ani, ci, supi, confi) in the decision table that matches u (a1u, a2u, ..., anu), then the class of u is ci. More than one matching row is found: When more than one matching row found for a given sample, there are a number of alternatives to assign the class label. Assume that k matching rows are found and the class label, support and confidence for row i is ci, supi and confi respectively. The class of the sample, cu, can be assigned in one of the following ways.

a. based on confidence and support:

$$\begin{array}{l} \mathrm{k} \\ C_u = \{ c_i | conf_i = max\ \ conf_j \} \\ j=1 \end{array} \tag{5}$$

b. based on weighted confidence and support:

$$\begin{array}{l} \mathrm{k} \\ C_u = \{ c_i | conf_i * sup_i = max\ \ conf_j * sup_{ij} \\ j=1 \end{array} \tag{6}$$

The ties are treated similarly. Note that, if the decision table is sorted on (Conf, Sup), it is easy to implement the first method. We can simply assign the class of the first matching row to the sample to be classified. In fact, our experiments indicated that this simple method provides no worse performance than others.

c. No matching row is found:

In most classification applications, the training samples cannot cover the whole data space. The decision table generated by grouping and counting may not cover all possible data samples. For such samples, no matching row will be found in the decision table. To classify such samples, the simplest method is to use the default class. However, there are other alternatives. For example, we can first find a row that is the nearest neighbor (in certain distance metrics) of the sample in the decision table and then assign the same class label to the sample. The drawback of using nearest neighbor is its computational complexity.

Meta level - 1: Also known as LEVEL-1. Here we apply Decision Tree which is only a choice tree i.e. Meta classifier. The forecasts of the base students i.e., (Naïve Bayes and Decision Table) are contribution to the meta-student.

d. Decision tree: (Basic principle)

The decision tree classification algorithm is widely used classification algorithm in data mining. It operates in a divide and conquer manner, which recursively partitions the training data set based on its attributes until the stopping conditions are satisfied. The decision tree consists of nodes, edges, and leaves. A Decision Tree node has its corresponding data set this specifies the attribute to best divide the data set into its classes. Each node has several edges that specify possible values or value ranges of the selected attributes on the node.

The decision tree algorithm recursively visits each decision node, selecting the optimal split, until no further splits are possible. The basic steps of j48 algorithm for growing a decision tree are given below:

- Choose attribute for root node
- Create branch for each value of that attribute

- Split cases according to branches
- Repeat process for each branch until all cases in the branch have the same class

A question that, how an attribute is chosen as a root node? At first, we calculate of the gain ratio of each attribute. The root node will be that attribute whose gain ratio is maximum. Gain ratio is calculated by (7).

$$\text{Gain Ratio }(A) = \frac{\text{Gain}(A)}{\text{SplitInfo}(A)} \tag{7}$$

Where, A is an attribute whose gain ratio will be calculated. The attribute A with the maximum gain ratio is selected as the splitting attribute. This attribute minimizes the information needed to classify the tuples in the resulting partitions. Such an approach minimizes the expected number of tests needed to classify a given tuple and guarantees that a simple tree if found.

The data set of the node is divided into subsets according to the specifications of the edges, and the Decision Tree creates a child node for each data subset and repeats the dividing process. When the node satisfies the stopping rules because it contains homogeneous data sets or no future distinguishing attributes can be determined, the Decision Tree terminates the dividing process and the node is labeled as following the class label of the data set. This labeled node is called a leaf node. In this way, the Decision Tree recursively partitions the training data set, which creates a tree-like structure.

e. Proposed system

Flowchart of the proposed system is shown in Figure 5. In first step, create secure virtual cload environment, and continued authentification of image using image sequencing passwords. Next step is data classification with input dataset as explained in introduction section, encryption, sending and evaluation.

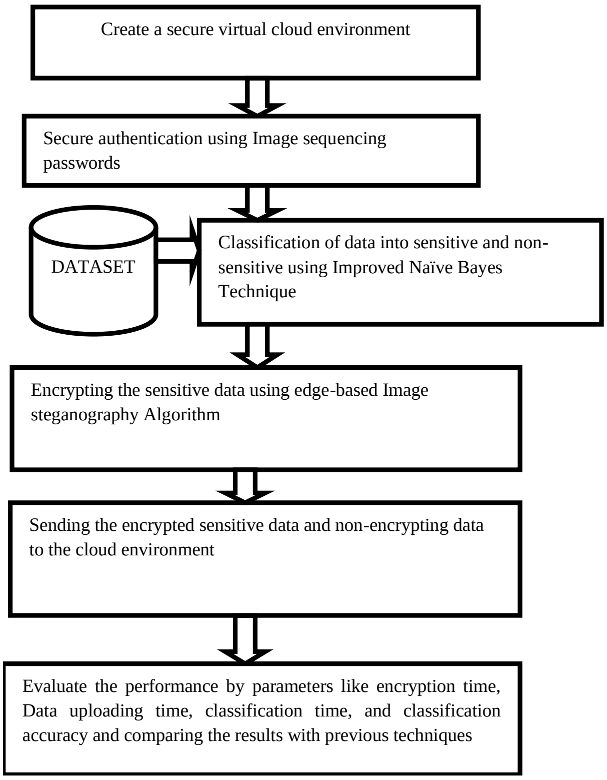

Create a secure virtual cloud environment
Secure authentication using Image sequencing passwords
DATASET
Classification of data into sensitive and non-sensitive using Improved Naïve Bayes Technique
Encrypting the sensitive data using edge-based Image steganography Algorithm
Sending the encrypted sensitive data and non-encrypting data to the cloud environment
Evaluate the performance by parameters like encryption time, Data uploading time, classification time, and classification accuracy and comparing the results with previous techniques

Figure 5. Proposed system flowchart

4. RESULTS AND DISCUSSION

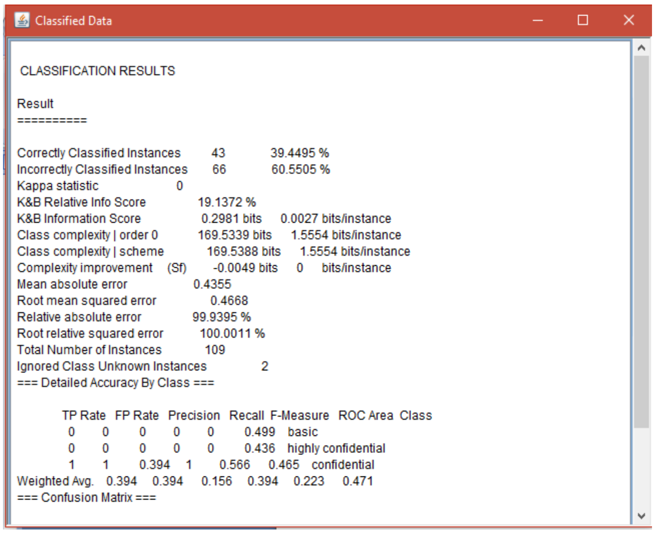
The proposed technique is executed with the assistance of CloudSim and Netbeans IDE 8.0. CloudSim is the library that gives the recreation condition of distributed computing and furthermore give essential classes portraying virtual machines, server farms, clients and applications. NetBeans is where applications are created utilizing sections called programming modules. We utilize Cloudsim test system for the examination work. Cloudsim is a system for displaying and recreation of distributed computing administrations and foundation.

4.1. Classification phase

Classification of articles is a basic subject of studies and of common-sense applications in various fields such as pattern recognition and statistics, artificial intelligence, vision analysis and medicine. An exceptionally smart method to anchor the information would be to initially arrange the information into secret and non-classified information and after that protected the delicate information as it were. This will decrease the overhead in encoding the whole information which will be especially expensive in association of both time and memory. For encoding the information numerous encryption strategies can be utilized and for ordering the information various grouping calculations are accessible in the field of information mining.

4.2. Results for KNN algorithm

Figure 6 shows the classification results of KNN algorithm with total of 109 instances out of which 43 are classified correctly. Time taken to classify the data is 329 milliseconds. By using KNN algorithm, weighted average rate for FP and Rate is 0.394, Precision is 0.156, recall is 0.394, F-measure is 0.223 and ROC Area is 0.471.



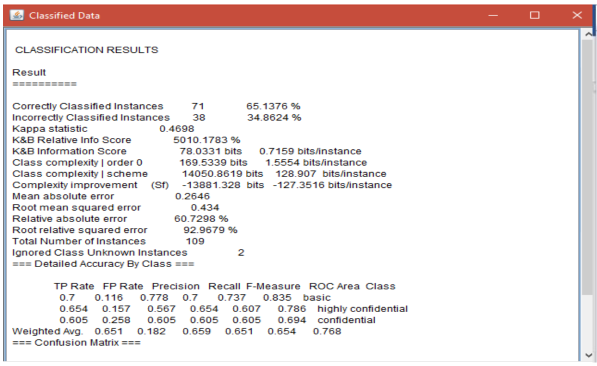
```
CLASSIFICATION RESULTS

Result
==========

Correctly Classified Instances       43           39.4495 %
Incorrectly Classified Instances     66           60.5505 %
Kappa statistic                  0
K&B Relative Info Score              19.1372 %
K&B Information Score                 0.2981 bits     0.0027 bits/instance
Class complexity | order 0           169.5339 bits     1.5554 bits/instance
Class complexity | scheme              169.5388 bits     1.5554 bits/instance
Complexity improvement    (Sf)        -0.0049 bits     0     bits/instance
Mean absolute error                 0.4355
Root mean squared error                0.4668
Relative absolute error             99.9395 %
Root relative squared error          100.0011 %
Total Number of Instances             109
Ignored Class Unknown Instances                 2
=== Detailed Accuracy By Class ===

          TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
            0        0         0         0        0         0.499   basic
            0        0         0         0        0         0.436   highly confidential
            1        1         0.394    1        0.566     0.465   confidential
Weighted Avg.   0.394    0.394     0.156    0.394    0.223     0.471
=== Confusion Matrix ===
```

Figure 6. Classification results of KNN algorithm

4.3. Results for improved Bayesian classifier

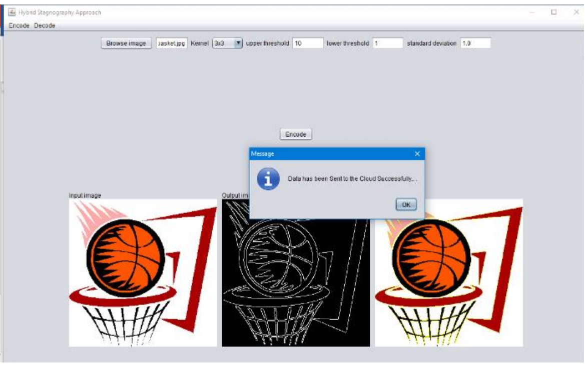
Figure 7 shows the classification results of Improved Bayesian Classifier with total of 109 instances out of which 71 are classified correctly. Time taken to classify the data is 797 milliseconds. By using Improved Bayesian Classifier, weighted average rate for FP is 0.605, Rate is 0.182, Precision is 0.659, recall is 0.651, F-measure is 0.654 and ROC Area is 0.768. From Figure 7 and Figure 8, it is shown that the improved Bayesian Classifier performs better than the KNN algorithm, i.e it will classify the data more accurate

Figure 8 shows that time taken to hide the data inside the image is 554 milliseconds which is less than the time taken by RSA algorithm by 8154 milliseconds. In this, sensitive data is hidden using hybrid steganography approach i.e. the data is first converted into binary format then this binary data is hidden inside the edges of the input image which is calculated using Canny Edge detection. From these set of edges, the randomization is applied and then edges are selected and the binary format of sensitive data is hidden in the least significant bit of that randomly selected edge. In this way whole data is hidden and the edge pixel position on which data is hidden is saved in file and sent that file to the cloud. The final encrypted image and the original image is compared using histogram equalization and it shows that both the images have same histogram i.e after hiding the data image is distorted this is because LSB method is used for hiding the data in which least significant bit is replaced with the data bit instead of most significant bit.



```
CLASSIFICATION RESULTS

Result
==========

Correctly Classified Instances        71           65.1376 %
Incorrectly Classified Instances      38           34.8624 %
Kappa statistic                    0.4698
K&B Relative Info Score              5010.1783 %
K&B Information Score                 78.0331 bits     0.7159 bits/instance
Class complexity | order 0            169.5339 bits     1.5554 bits/instance
Class complexity | scheme             14050.8619 bits    128.907  bits/instance
Complexity improvement    (Sf)    -13881.328  bits   -127.3516 bits/instance
Mean absolute error                  0.2646
Root mean squared error                0.434
Relative absolute error              60.7298 %
Root relative squared error            92.9679 %
Total Number of Instances             109
Ignored Class Unknown Instances                2
=== Detailed Accuracy By Class ===

          TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
            0.7      0.116     0.778     0.7      0.737      0.835    basic
            0.654     0.157     0.567     0.654     0.607      0.786    highly confidential
            0.605     0.258     0.605     0.605     0.605      0.694    confidential
Weighted Avg.   0.651     0.182     0.659     0.651     0.654      0.768
=== Confusion Matrix ===
```

Figure 7. Classification results of KNN algorithm



basket.jpg Kernel 3x3 upper threshold 10 lower threshold 1 standard deviation 1.0

Data has been Sent to the Cloud Successfully...

Figure 8. Sending random pixel position values to the cloud environment

In this proposed technique, with the increase in privacy no information loss is there. As results show that we got the complete sensitive data. Hence, data utilization is there.

From Tabel 1 and Tabel 2 shows the performance analysis of data classification algorithms. Table 1 and Figure 9 shows the correctly and incorrectly classified instances by KNN and Improved Bayesian. Table 2 shows the performance of the algorithms on the basis of its error rate. Both the tables show that the proposed classification algorithm performs better than the existing KNN algorithm as showing in Figure 10.

Table 1. Performance of data mining algorithms

	KNN	Improved Bayesian Classifier
Correctly classified instances	43	71
Incorrectly classified instances	66	38

Table 2. Detailed comparison of accuracy based on error

	KNN	Improved Bayesian Classifier
Mean absolute error	0.4355	0.2646
Root mean squared error	0.4668	0.434
Relative absolute error	0.9993	0.607
Root relative squared error	1.00	0.929

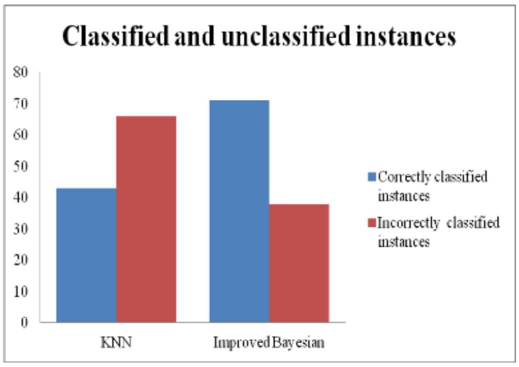


Figure 9. Correctly and Incorrectly Classified instances comparison of KNN algorithm with the proposed Algorithm

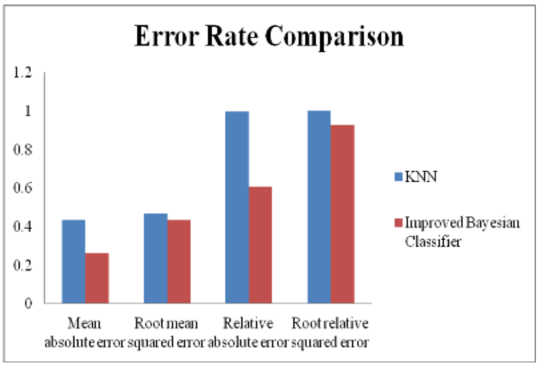
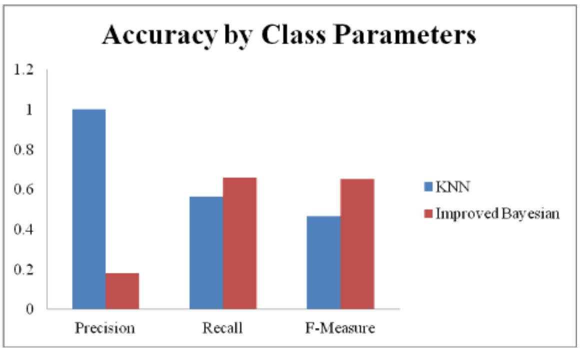


Figure 10. Detailed comparison of accuracy based on error

From Figure 11, Figure 12 and Figure 13 shows the performance analysis of the proposed methodology with the previous method. In Figure 13, encryption time of existing technique is 8154 msec and encryption time taken by proposed technique is 554msec. It is clearly analyzed from the performance graphs that the proposed technique is better than the previous approach. Figure 11 shows the accuracy comparison of data classification algorithms KNN and proposed Improved Bayesian Algorithm. KNN algorithm is having accuracy 39% and improved boosting is having 65% i.e proposed algorithm has classified data more correctly and performs better than the KNN algorithm. Time taken to classify the data is 329 milliseconds. By using KNN algorithm, weighted average rate for FP and Rate is 0.394, Precision is 0.156, recall is 0.394, F-measure is 0.223 and ROC Area is 0.471. Time taken to classify the data is 797 milliseconds. By using Improved Bayesian Classifier, weighted average rate for FP is 0.605, Rate is 0.182, Precision is 0.659, recall is 0.651, F-measure is 0.654 and ROC Area is 0.768. Similarly, Figure 12 shows the data hiding time comparison between the proposed and previous approach. Proposed Hybrid technique takes 8000 milliseconds and the existing algorithm takes 6000 milliseconds to hide the sensitive data. Figure 13 shows the time taken to encrypt the unclassified and classified data; figure clearly shows that the encryption time of classified data is less as compare to the unclassified data. From the above analysis it is shown that the proposed methodology performs betters in respect to Accuracy, classification time and data hiding time.



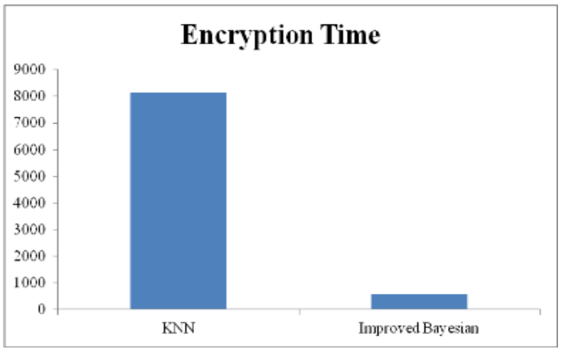


Figure 11. Accuracy comparison of KNN algorithm with the proposed Bayesian Algorithm

Figure 12. Comparison of data hiding time of existing and the proposed Hybrid algorithm

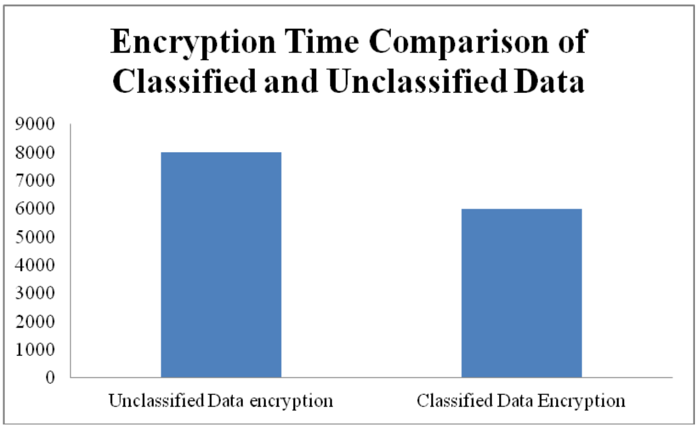


Figure 13. Comparison of encryption time of existing and proposed technique

5. CONCLUSION

Cloud computing is effectively managing the security in the cloud applications. Data classification is a machine learning technique used to predict the class of the unclassified data. Data mining uses different tools to know the unknown, valid patterns and relationships in the dataset. These tools are mathematical algorithms, statistical models and Machine Learning (ML) algorithms. In this paper author uses improved Bayesian technique to classify the data and encrypt the sensitive data using hybrid stagnography. The encrypted and non encrypted sensitive data is sent to cloud environment and evaluate the parameters with different encryption algorithms. Proposed Hybrid technique takes 8000 milliseconds and the existing algorithm takes 6000 milliseconds to hide the sensitive data. Time taken to encrypt the unclassified and classified data; It is shows that the encryption time of classified data is less as compare to the unclassified data. On conclusion we can say that the proposed methodology performs betters in respect to Accuracy, classification time and data hiding time.

REFERENCES

[1] Data, B., “For better or worse: 90% of world’s data generated over last two years,” *SCIENCE DAILY*, 2013. [Online]. Available: https://www.sciencedaily.com/releases/2013/05/130522085217.htm

[2] Botta, A., De Donato, W., Persico, V., & Pescapé, A., “Integration of cloud computing and internet of things: A survey,” *Future generation computer systems*, vol. 56, pp. 684-700, 2016.

[3] Mastelic, T., Oleksiak, A., Claussen, H., Brandic, I., Pierson, J. M., & Vasilakos, A. V., “Cloud computing: Survey on energy efficiency,” *Acm computing surveys (csur)*, vol. 47, no. 2, pp. 1-36, 2014.

[4] Messerli, A. J., Voccio, P., & Hincher, J. C. *U.S. Patent No. 9,563,480*. Washington, DC: U.S. Patent and Trademark Office, 2017.

[5] Farahnakian, *et al*, “Using ant colony system to consolidate VMs for green cloud computing,” *IEEE Transactions on Services Computing*, vol. *8*, no. 2, pp. 187-198, 2015.

[6] Botta, A., De Donato, W., Persico, V., & Pescapé, A., “Integration of cloud computing and internet of things: a survey,” *Future generation computer systems*, vol. 56, pp. 684-700, 2016.

[7] Wang, B., Zheng, Y., Lou, W., & Hou, Y. T., “DDoS attack protection in the era of cloud computing and software-defined networking,” *Computer Networks*, vol. 81, pp. 308-319, 2015.

[8] Puthal, D., Sahoo, B. P. S., Mishra, S., & Swain, S., “Cloud computing features, issues, and challenges: a big picture,” *2015 International Conference on Computational Intelligence and Networks*, pp. 116-123, 2015.

[9] Li, J., Li, Y. K., Chen, X., Lee, P. P., & Lou, W., “A hybrid cloud approach for secure authorized deduplication,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 26, no. 5, pp. 1206-1216, 2015.

[10] Chou, D. C., “Cloud computing: A value creation model,” *Com. Standards & Interfaces*, vol. 38, pp. 72-77, 2015.

[11] Ali, M., Khan, S. U., & Vasilakos, A. V., “Security in cloud computing: Opportunities and challenges,” *Information sciences*, vol. *305*, pp. 357-383, 2015.

[12] Ang, J. C., Mirzal, A., Haron, H., & Hamed, H. N. A., “Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection,” *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 13, no. 5, pp. 971-989, 2016.

[13] Saurabh Singh, Young-Sik Jeong, Jong Hyuk Park, “A survey on cloud computing security: issues, threats, and solutions,” *Journal of Network and Computer Applications*, vol. 75, pp. 200-222, 2016.

[14] Faheem Zafar, *et al*, “A survey of Cloud Computing data integrity schemes: Design challenges, taxonomy and future trends,” Computers & Security, vol. 65, pp. 29-49, 2016.

[15] A Platform Computing Whitepaper, *Enterprise Cloud Computing: Transforming IT*, Platform Computing, 2009.

[16] NITS, “Guidelines on Security and Privacy in Public Cloud Computing,” [Online]. Available http://csrc.nist.gov /publications/nistpubs/800-144/SP800-144.pdf, retrieved on Sep 29, 2012.

[17] Philip Derbeko, Shlomi Dolev, Ehud Gudes, Shantanu Sharma, “Security and privacy aspects in mapreduce on clouds: a survey,” *Computer Science Review*, vol. 20, pp. 1-28, 2016.

[18] Ogigau-Neamtiu F., “Cloud computing security issues,” *Journal of Defense Resources Management*, vol. 3, no. 2, pp. 141-148, 2012.

[19] Wu J, Ping L, Ge X, Wang Y, Fu J, “Cloud storage as the infrastructure of cloud computing,” *International Conference on Intelligent Computing and Cognitive Informatics (ICICCI*), pp. 380-383, 2010.

[20] Brian Cusack, Eghbal Ghazizadeh, “Evaluating single sign-on security failure in Cloud services,” *Business Horizons*, vol. 59, no. 6, pp. 605-614, 2016.

[21] Laurence T. Yang, *et al*, “Parallel GNFS algorithm integrated with parallel block wiedemann algorithm for RSA security in cloud computing, information sciences,” *Information Sciences*, vol. 387, pp. 254-265, 2016.

[22] Jolanda Modic, Ruben Trapero, Ahmed Taha, Jesus Luna, Miha Stopar, Neeraj Suri, “Novel efficient techniques for real-time cloud security assessment,” *Computers & Security*, vol. 62, pp. 1-18, 2016.

[23] Christos Kalloniatis, Haralambos Mouratidis, Manousakis Vassilis, Shareeful Islam, Stefanos Gritzalis, Evangelia Kavakli, “Towards the design of secure and privacy-oriented information systems in the Cloud: Identifying the major concepts,” *Computer Standards & Interfaces*, vol. 36, no. 4, pp. 759-775, 2014.

[24] Chunming Rong, Son T. Nguyen, Martin Gilje Jaatun, “Beyond lightning: A survey on security challenges in cloud computing,” *Computers & Electrical Engineering*, vol. 39, no. 1, pp. 47-54, 2013.

[25] Lofstrand M, ‘The VeriScale Architecture: Elasticity and Efficiency for Private Clouds”, *Sun Microsystems, Sun BluePrint*, 2009.

[26] Vahid Ashktoraband Seyed Reza Taghizadeh. “Security threats and countermeasures in cloud computing,” *International Journal of Application or Innovation in Engineering & Management*, vol. 1, no. 2, pp. 234-245, 2012.

[27] Brian Cusack, Eghbal Ghazizadeh, “Evaluating single sign-on security failure in cloud services,” *Business Horizons*, vol. 59, no. 6, pp. 605-614, 2016.