

**SWE1901 - Technical Answers to Real-World Problems (TARP)**

**Project Report**

**Title: INFORMATION LEAKAGE PROTECTION SYSTEM**

*Submitted to*

*Dr. Nisha V.M*

Submitted by:

Niranjana O (19MIS1164)

Saheli Das (19MISS164)

Khushi Kantula (19MIS1186)

**DECLARATION**

I hereby declare that the report titled “**Information Leakage Protection System”** submitted by me to VIT Chennai is a record of bonafide work undertaken by me under the supervision of **Dr. Nisha V.M** , School of Computer Science and Engineering, Vellore Institute of Technology, Chennai.

Signature of the Candidate

**ACKNOWLEDGEMENT**

We wish to express our sincere thanks and deep sense of gratitude to our project guide, Dr. Nisha V.M, School of Computer Science and Engineering, for her consistent encouragement and valuable guidance offered to us in a pleasant manner throughout the course of the project work.

We are extremely grateful to Dr. Nisha V.M (Head of the Department (HoD), Integrated M.Tech Software Engineering, SCOPE, VIT Chennai) and Dr. Ganesan R (Dean of the School of Computer Science & Engineering, VIT Chennai) for extending the facilities of the School towards our project and for his unstinting support.

We also take this opportunity to thank all the faculty of the School for their support and the wisdom imparted to us throughout the course.

We thank our parents, family, and friends for bearing with us throughout the course of our project and for the opportunity they provided us in undergoing this course in such a prestigious institution

Signature of the Candidate

**Introduction**

The modern world is based on information exchange, or the flow of data from one person to another. Since the data shared with trustworthy external parties is confidential and vital, the data transmitted by the sender must be private, confidential, and not duplicated. The data transmitted by the producer is often copied by multiple agents, causing significant damage to the organization, and this process is known as data leakage. To prevent data from becoming freely available, data leaking must be caught at an early stage. As a result, cyberspace security has become a priority. Security threats are a disadvantage for cloud computing jobbers, thus it's more important than ever to use different levels of information security in different settings.

The identification of information leakage in a data can be accomplished by applying fundamental tactics such as watermarking various types of data using cloud computing and implementing data encryption techniques that make the data, as well as the keys encoded with it, more secure. The company's Information security depends on employees learning the rules through training and awareness-building sessions. However, security must go beyond employee knowledge and cover the following areas such as a physical and logical security mechanism that is adapted to the needs of the company and to employee use than the procedure for managing updates, and finally, it needs an up-to-date documented system. Information system security is often the subject of metaphors. It is often compared to a chain in the example that a system's security level is only as strong as the security level of its weakest link. All this goes to show that the issue of security must be tackled at a global level and must comprise the following elements like making users aware of security problems then the logical security, i.e. security at the data level, notably company data, applications, and even operating systems and also products used in Telecommunications security such as network technologies, company servers, access networks, etc.

Confidential corporate information such as client or patient data, source code or design specifications, pricing lists, intellectual property and trade secrets, and spreadsheet projections and budgets are all examples of data leakage. When these are released, the firm is left defenseless and falls outside of the corporation's authority.

When cybercriminals benefit from this data, it costs our company money, hurts our competitive advantage, brand, and reputation, and undermines consumer trust. To overcome this issue, we create a model for determining an agent's guilt. In order to enhance the odds of finding a guilty agent, the distributor would intelligently send data to agents, such as adding false objects to distributed sets. At this stage, the distributor can determine if the stolen data came from one or many agents rather than being obtained independently by other ways. If the distributor has solid evidence that an agent has leaked data, they may refuse to do business with him or pursue legal action. It primarily has one limitation and one goal. The Distributor's constraint fulfills the agent by offering a sufficient quantity of objects that meet their requirements. Data should always be noticed as soon as possible.

As a result, data detection from the distributor to the agents is required. This research proposes a data leakage protection system that evaluates the chance that the leaked data came from one or more agents using various allocation algorithms. Allowing only authorized users to access sensitive data through access control policies will prevent data leakage by sharing information only with trusted parties.

**LITERATURE SURVEY**

Sushilkumar, Holambe, et al (2019) proposed a new watermarking algorithm for predicting data leakage using cloud services Watermarking techniques is used to test around 500 data points in this paper. There are certain kinds of data that cannot admit watermarks that do not detect leakage.

Bijayalakshmi et al (2020) proposed DLP technology which is used to detect the leakage and it is easy to implement which makes it easy for developers to work on. According to this paper data leakage analysis using a virtual cloud network is not implemented.

Chandu vaidya and Prashant (2019) proposed a probability-based model for data leakage. By using this model the distributor will be able to add fake objects in-order to boost the effectiveness. In order to detect leakage algorithms are used which creates issues of integrity for the users.

Priyanka, Goyal, et al (2017) world on data leakage and security which focuses on watermarking and robust watermarking techniques which can be very useful for the detection .watermarking can sometimes be destroyed if the data recipient is malicious .

Ye Tao, Peng Xu, et al (2021) used cloud-based technology it enables outsiders and insiders of the company to view the same data without sharing any confidential data this will decrease the data leakage which the use of data allocation fake objects are added to find guilty parties which create confusion to detect the correct information.

Arhan P, Surya Kiran, et al (2017) used different types of watermarking techniques like secure spread watermarking techniques which are very secure and impossible for outsiders to decode and make the data private. In this paper, wavelet-based watermarking technique is used which uses a complex algorithm which is really hard to implement for the developers and tech teams.

Tushar Agarwal et al (2018) have considered the frequency domain method as the best method to detect the data leakage because it deals with the rate of pixel change in an image instead of the image plane while detecting the data leakage through images. Better results can be provided using embedded watermarking using MATLAB but it’s not implemented in this paper.

Prashant et al (2018) use capturing data and distribution strategies to improve the detection process for data leakage and security. The overlap technique is used in data distribution which allows outsiders to guess the object by other means.

Arman Barg et al (2020) primarily focuses on browser history data to detect where the attack is happening and it can be used as evidence to detect the leakage. Privacy is invaded by the users as browser history checking is used as a technique to detect the leakage.

Ria Agarwal et al (2020) have considered the frequency domain method is the best method to detect the data leakage because it deals with the rate of pixel change in an image instead of the image plane while detecting the data leakage through images. Better results can be provided using embedded watermarking using MATLAB but it’s not implemented in this paper.

Ashish et al (2016) worked on a permutation technique which is used when the data is modified and made less sensitive before being handed to the agents. Provisions for the fake records dynamically according to the agent’s request can be implemented

Kiran et al (2017) used different types of watermarking techniques like secure spread watermarking techniques which are very secure and impossible for outsiders to decode and make the data private. In this paper, the wavelet-based watermarking technique is used which uses a complex algorithm that is really hard to implement for the developers and tech teams.

**PROBLEM STATEMENT**

Data leak prevention helps ensure that confidential data like customer information, personal employee information, trade secrets, financial data and. research and development data remain safe and secure. Data leak prevention solutions prevent confidential data by securing the data itself.

**Objective and goal of the project**

* Using steganography techniques to do data hiding for imprinting the data.
* Implementing data encryption techniques makes the data, as well as the keys encoded with it, more secure.
* Using machine learning techniques like Support vector machine, KNN classifier, and linear discriminant analysis for binary classification. Decision tree, random forest, and logistic regression for multi-class classification to understand the present situation of the amount of data leaked

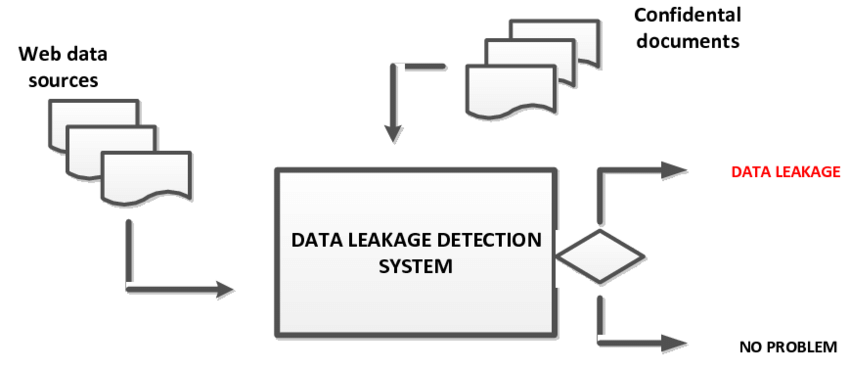
**Requirements Specification**

Software Requirements

* Python
* Kali Linux
* Sklearn
* Pandas
* Numpy
* Google Collab
* Virtual Machine (Ubuntu)

**SYSTEM DESIGN**

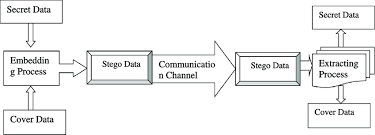
In this project, Information leakage is divided into two parts the first one is text data leakage and the second part is Image/video data leakage. For text files, the data distribution with algorithms of encryption and decryption is done which enables the users to store their data encrypted in key. Nowadays, Meta company is using encryption technique to store sensitive information. In order to provide security, the data owner should register to a cloud account and every user must register to the cloud and become an authorized user to prevent data leakage. For detecting data leakage in text, the Intrusion detection system can be used to find out the malicious activity and type of leakage in the data. In order to find out data leakage machine learning techniques are used such as SVM, KNN classifier, and linear discriminant analysts are used for binary classification data. Decision trees, random forest, and logistic regression are used for multi-class classification. And the main factor that detects better performance is accuracy. The second part is data leakage using image files. Watermarking techniques can also be used to detect information leakage in data.



**IMPLEMENTATION OF SYSTEM**

Watermarking technique is used for detecting data leakage in image files. In this type of technique, a unique code is embedded into the image files if that code is available with the third party then the leaker can be identified. And it is also believed that watermarking techniques are pretty accurate compared to other algorithms. Usage of cloud-based technology enables insiders and outsiders to view the data which will be helpful in reducing data leakage. In most cases, secure spread spectrum watermarking techniques are used which is very secure and also makes the data private.

DLP technology is used to detect the leakage and it is easy to implement in this technique distributors will be able to add fake objects which are used to boost the effectiveness. There are many drawbacks to the existing system. Certain data cannot admit watermarks that don’t detect leakage of data. Data leakage on cloud platforms such as Citrix, cloud stack, and AWS is not detected. With the use of data allocation, a few fake objects were added to find guilty parties which create confusion to detect the malicious activities. Wavelet-based watermarking techniques use a complex algorithm that is hard to implement. In order to detect leakage, various algorithms are used which creates issues of integrity. Watermarks can be sometimes destroyed if the data recipient is malicious. Overlapping techniques is really dangerous which is used in data distribution and allow outsiders to guess the object by other means.



**5.1 ​​Dataset**

The dataset used in the project is taken from a private site and it consists of sports information and CSV file used is KDDCup99.csv and it consists of 494020 rows and 41 columns. And some of the columns are protocol type, service, flag, src bytes, label, etc. The target column is label data which consists of information that is leaked and different forms of leaked data such as smurf, Neptune,land,root-kit, Perl , spy act

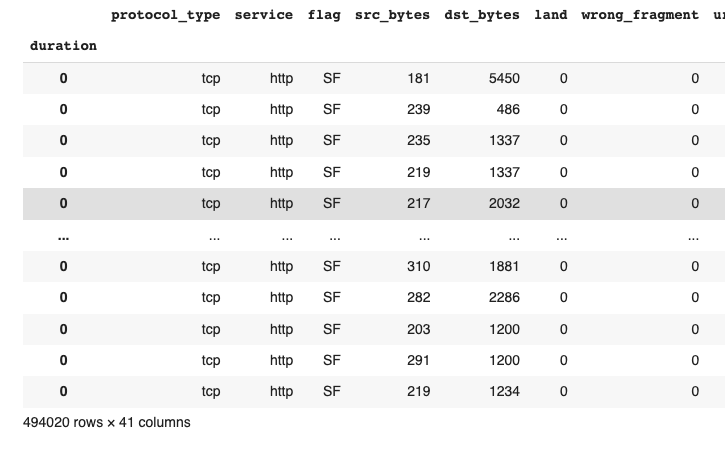


Fig 1: Dataset

**5.2 Discrete cosine wavelength**

The DCT is a method for converting a signal into its fundamental frequency components. The input signal is represented as a linear mixture of weighted basis functions that are connected to its frequency components in the DCT. The DCT does not, in general, reduce the number of bits needed to represent a block. Due to the range of coefficient values, the DCT creates an 8 8 of 11-bit coefficient for an 8 8 block of 8-bit pixels. However, because the DCT concentration, low-frequency coefficients, and remaining other coefficients are primarily zero, compression may be accomplished by transmitting the near-zero coefficients and quantizing and coding the other coefficients. The image is broken into different frequency bands, the watermark is embedded into the middle frequency to ensure the watermark will not be removed by any attacks

In this project, Discrete cosine wavelength transformation is used to detect leakage on image files by breaking the image into different frequencies and working on the different pixels and the watermarks embedded into the middle of the frequency to ensure it cannot be removed by any malicious attackers

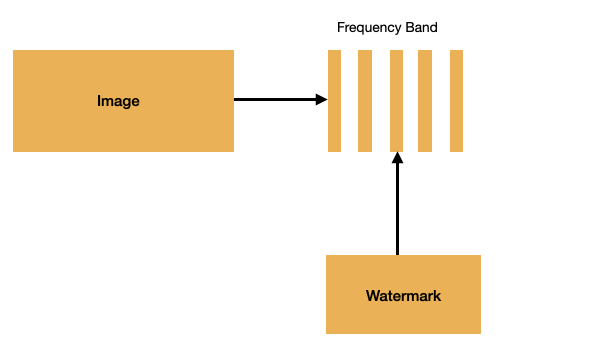


Fig 2 Discrete cosine transformation

**5.3 Discrete Wavelength Transformation**

In discrete wavelet transform, the DWR divides the signal into two halves, one for the frequency and the other for the lower frequency. Now, The higher frequency section contains information on the edge components, whereas the lower frequency section is divided into high and low-frequency sections once more and this process proceeds again. In discrete cosine wavelength, the image is broken into different frequency bands, the watermark is embedded into the middle frequency to ensure that the watermark will not be removed by any attacks. A discrete wavelet transform (DWT) decomposes a signal into a number of sets, each set including a time series of coefficients that describe the signal's time evolution in the associated frequency band. In science, engineering, mathematics, and computer science, the discrete wavelet transform have a wide range of applications. It is most famous for signal coding, which is intended to represent a discrete signal in a more redundant form and is frequently used as preconditioning for data compression.

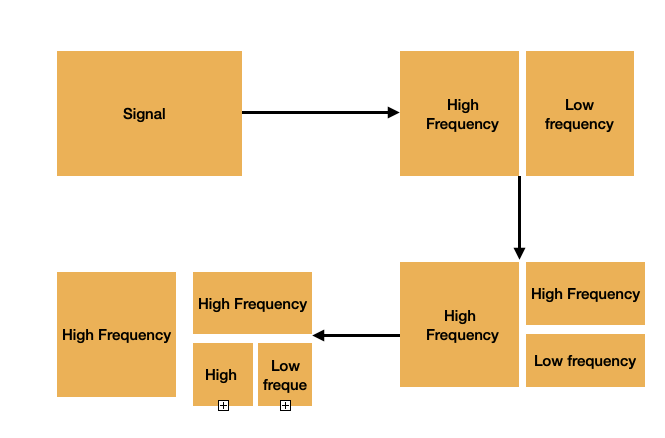


Fig 3: Discrete wavelength transformation

**5.4 Support vector machine**

SVM (Support Vector Machine) is a supervised machine learning technique that may be used to solve classification and regression problems. It is, however, mostly employed in categorization difficulties. The SVM classifier is a frontier that separates the two classes (hyper-plane/line) the most effective. A support vector machine is used to detect data leakage or intrusion detection in text files using the dataset and the accuracy obtained using that method is 98.13%

**5.5 K-Nearest neighbors**

"K-Nearest Neighbour" is the acronym for "K-Nearest Neighbour." It's a machine-learning algorithm that's supervised. Both classification and regression problem statements may be solved using the approach. The sign 'K' represents the number of nearest neighbors to a new unknown variable that must be predicted or categorized. K-nearest neighbor algorithm is used to detect data leakage or intrusion detection in text files using the dataset and the accuracy obtained using that method is 98.83%

**5.6 Linear discriminant analysis**

LDA is a classifier used to find a linear combination of features, which separates two or more classes of data. The succeeding combination can be used as a linear classifier. In LDA, the classes are expected to be normally distributed. Like PCA, LDA can be utilized for both dimension reduction and classification.LDA supports both binary and multi-class classification. Gaussian Distribution. The standard implementation of the model assumes a Gaussian distribution of the input variables.LDA algorithm is used to detect data leakage or intrusion detection in text files using the dataset and the accuracy obtained using that method is 98.13%

**5.7 Decision tree classifier**

A decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. A decision tree algorithm is used to detect data leakage or intrusion detection in text files using the dataset and the accuracy obtained using that method is 98.6%

**5.8 Logistic regression**

Logistic Regression is a Machine Learning algorithm that is used for classification problems, it is a predictive analysis algorithm and based on the concept of probability hypothesis of logistic regression tends to limit the cost function between 0 and 1.A logistic regression algorithm is used to detect data leakage or intrusion detection in text files using the dataset and the accuracy obtained using that method is 96.33%

**5.9 Random Forest**

A random forest is a machine learning technique that's used to solve regression and classification problems. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems. A random forest algorithm consists of many decision trees. A random forest algorithm is used to detect data leakage or intrusion detection in text files using the dataset and the accuracy obtained using that method is 98.83%

**6. Results and Discussion**

Data leakage for text files can be detected using machine learning models. Different types of leaked and normal data

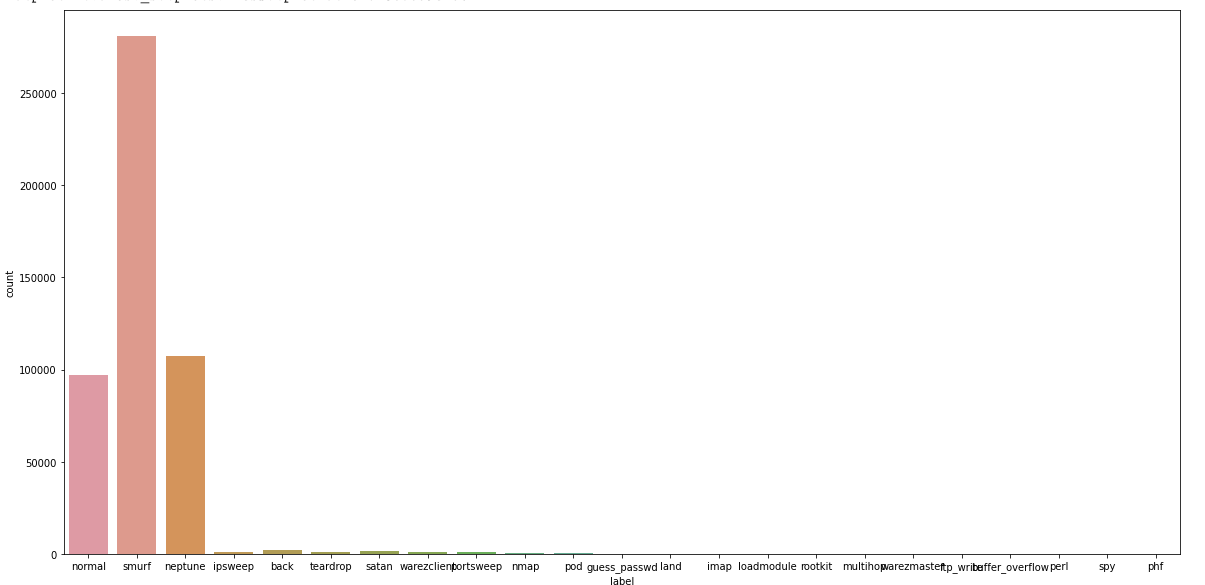
****

Fig. Different forms of malicious/leaked data

Correlation map of the dataset for data leakage detection for text dataset .

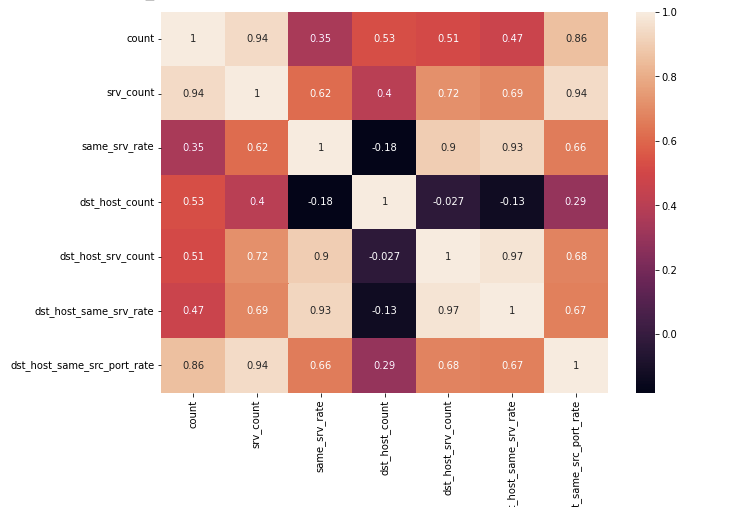
****

Fig: Correlation map for text dataset

Pie chart distribution of leaked and normal data after dividing the dataset into binary classes.

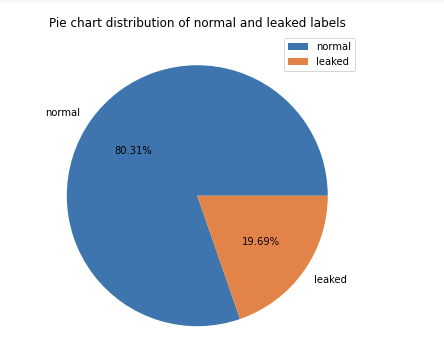
****

Fig: Pie chart distribution of normal and leaked data

Binary class classification is done using three machine learning models known as SVM, and logistic regression

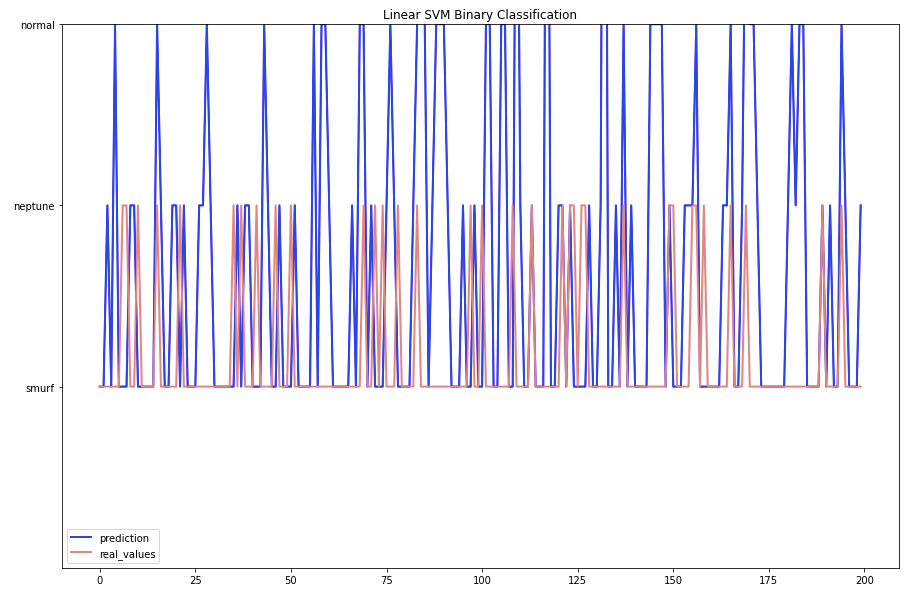
****

Fig. SVM classifer

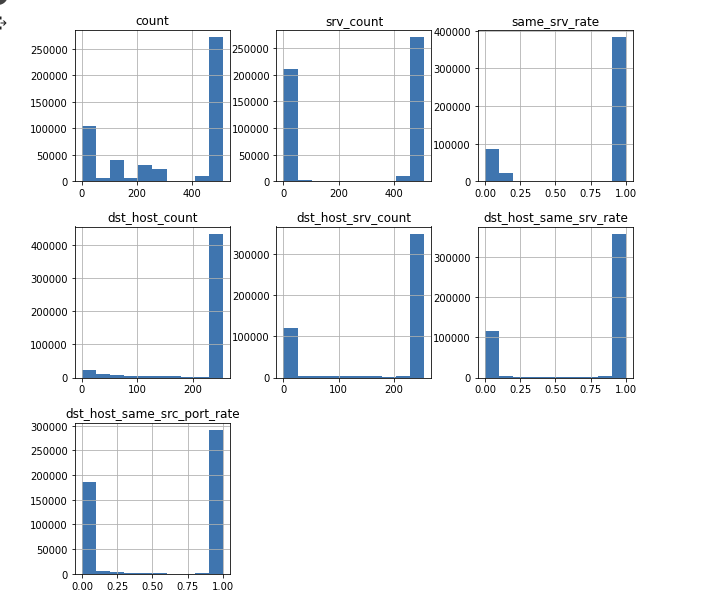
****

Fig: Histogram of all the columns

Machine learning techniques such as SVM, KNN classifier, Latent discriminant analysis, decision tree, random forest analysis, and Logistic regression models are used. The accuracy is taken as the parameter. After performing each model, the random forest has given the highest accuracy of 99, followed by a decision tree. This can be concluded by saying that the multi-class model gives better results compared to the binary class.

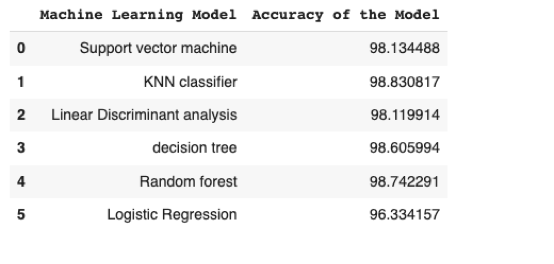
****

Fig: Performance evaluation of all the models

**7. Conclusion and Future Work**

Information leakage detection is the process that is used to provide security to sensitive, confidential, and important documents. The sensitive data is divided into two forms text data and image data. For securing the text data from getting leaked to malicious users, encryption and decryption techniques are used. With the use of encryption the data will be encoded using keys that will really hard to duplicate.

For detecting data leakage for text data machine learning techniques such as SVM, KNN, LDA, Decision tree, Random forest, and Logistic regression. Information leakage protection is the process that is used to provide security to sensitive, confidential, and important documents. The sensitive data is divided into two forms text data and image data. For securing the text data from getting leaked to malicious users encryption and decryption techniques are used. For image files, better results can be provided using MATLAB.

In future, Watermarking technique can used for detecting data leakage in image files. In this type of technique, a unique code is embedded into the image files if that code is available with the third party then the leaker can be identified. And it is also believed that watermarking techniques are pretty accurate compared to other algorithms. Usage of cloud-based technology enables insiders and outsiders to view the data which will be helpful in reducing data leakage. In most cases, secure spread spectrum watermarking techniques are used which is very secure and also makes the data private.

**References**

[1] Sandip A.Kale, Prof. Kulkarni S.V. artment Of Computer Sci. &Engg,MIT College of Engg(Dep, Dr.B.A.M.University, Aurangabad(M.S), India, Data Leakage Detection: A Survey, ( IOSR Journal of Computer Engineering org

[2] IEEE Transactions On Knowledge And Data Engineering, Vol. 22, No. 3, March 2011 Data Leakage Detection P

[3] Faith M. Heikkila, Data Leakage: What You Need to Know, Pivot Group Information Security Consultant. P.P (1-3)

[4] Rudragouda G Patil Dept Of CSE, The Oxford College Of Engg, Bangalore.International Journal Of Computer Applications In Engineering Sciences [VOL I, ISSUE II, JUNE 2011] [ISSN: 2231- 4946] P.P (1, 4) Development Of Data Leakage Detection Using Data Allocation Strategies

[5] Chun-Shien Lu, Member, IEEE, and Hong-Yuan Mark Liao, Member, IEEE MultipurposeWatermarking for Image Authentication and Protection

[6] A. Shabtai, a. Gershman, M. Kopeetsky, y. Elovici Deutsche Telekom Laboratories at Ben-Gurion University, Israel. Technical Report TR-BGU-2409-2010 24 Sept. 2010 1 A Survey of Data Leakage Detection and Prevention Solutions P.P (1-5, 24-25)

[7] Panagiotis Papadimitriou 1, Hector Garcia-Molina 2 Stanford University 353 Serra Street, Stanford, CA 94305, USA P.P (1, 4-5) A Model for Data Leakage Detection [

[8]Web-based Data Leakage Prevention Sachiko Yoshihama1, Takuya Mishina1, and Tsutomu Matsumoto2 1 IBM Research - Tokyo, Yamato, Kanagawa, Japan fsachikoy

[9] Joseph A. Rivela Senior Security Consultant P.P (4-6) Data Leakage: Affordable Data Leakage Risk Management

[10] Data Leakage Prevention: A newsletter for IT Professionals Issue 5 P.P (1-3)

[11] Panagiotis Papadimitriou, Student Member, IEEE, and Hector Garcia-Molina, Member, Data Leakage Detection IEEE P.P (2-6)IEEE transactions on knowledge and data engineering, vol. 23, no. 1, JANUARY 2011

[12] Archie Alimagno California Department of Insurance P.P (2- 7),The Who, What, When & Why of Data Leakage Prevention/Protection

[13] Faith M. Heikkila, Data Leakage: What You Need to Know, Pivot Group Information Security Consultant. P.P (1-3)

[14] Rudragouda G Patil Dept Of CSE, The Oxford College Of Engg, Bangalore.International Journal Of Computer Applications In Engineering Sciences [VOL I, ISSUE II, JUNE 2011] [ISSN: 2231- 4946] P.P (1, 4) Development Of Data Leakage Detection Using Data Allocation Strategies

[15] Chun-Shien Lu, Member, IEEE, and Hong-Yuan Mark Liao, Member, IEEE MultipurposeWatermarking for Image Authentication and Protection