IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 12, Issue 9, September 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 9, September 2023 ||

| DOI:10.15680/IJIRSET.2023.1209110 |

Protecting Cloud Data: A Machine Learning Approach for Data Classification in Cloud Computing

Prof. Nidhi Pateriya[1], Prof. Gulafsha Anjum[2], Prof. Neha Thakre[3], Vaishnavi Shrivaas[4], Vibhanshu Soni[5]

Department of Computer Science & Engineering, Baderia Global Institute of Engineering & Management, Jabalpur, Madhya Pradesh, India

ABSTRACT: Cloud computing (CC) is a modern framework that enables users to store data on remote servers accessible via the internet. This model facilitates easy access and transfer of personal and critical data, leading to increased demand. Users can store various types of data, including financial transactions, documents, and multimedia content. Additionally, CC reduces reliance on local storage and lowers operational and maintenance costs. However, existing systems typically encrypt all data with the same key size, irrespective of its confidentiality level, resulting in higher processing costs and time. Moreover, these methods often classify data with low accuracy and fail to provide adequate confidentiality.

This research introduces a cloud computing approach that employs automated data classification to assess data sensitivity. The proposed model categorizes data into three sensitivity levels: basic, confidential, and highly confidential. It utilizes Random Forest (RF), Naïve Bayes (NB), k-nearest neighbor (KNN), and Support Vector Machine (SVM) classifiers, incorporating automated feature extraction. The model achieved an accuracy of 92%, as demonstrated in simulation results. The findings indicate that RF, NB, and KNN outperform SVM. The research also offers valuable guidelines for cloud service providers (e.g., Dropbox and Google Drive) and researchers.

I. INTRODUCTION

Cloud computing (CC) is a highly sought-after platform that offers advanced technological applications for securely storing and retrieving personal documents. Users can access cloud services and applications through various devices, including mobile phones, laptops, and Android devices. In terms of document security, cloud computing is considered one of the most reliable and secure platforms. In contrast, other storage devices like mobile phones and laptops may not provide the same level of security due to limitations in battery life, performance, and storage capacity [1, 3]. Cloud storage services are widely utilized for storing and backing up data in a cost-effective, user-friendly, and quickly accessible manner [1]. They also facilitate easy data sharing and device synchronization.

There is no universally accepted set of attributes for cloud storage architecture, and different cloud storage systems utilize various architectural schemes. Typically, cloud storage systems consist of numerous storage devices clustered together and connected through a distributed file system, network, and other middleware. Cloud storage architecture generally includes a storage resource pool, service level agreements (SLAs), a distributed file system, and service interfaces. The primary objective of cloud storage design is to offer scalable, on-demand storage in a multi-tenant environment.

A standard cloud storage design features a front end with an API for storage access. While traditional storage systems use the SCSI protocol for this API, cloud storage often employs various APIs, including file service front ends, web service front ends, and even older protocols like iSCSI and internet SCSI. The storage logic layer, a form of middleware, operates behind the front end and handles functionalities such as data minimization and replication. The back end of cloud storage architecture, which may involve specific internet protocols or traditional storage drives, is responsible for the physical storage of data.

|e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 9, September 2023 ||

| DOI:10.15680/IJIRSET.2023.1209110 |

The cloud model is characterized by five key attributes: on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service. The National Institute of Standards and Technology (NIST) categorizes service models into three types: Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). Additionally, cloud services are available through various deployment options, including private, public, hybrid, and community clouds [1, 2, 3].

Figure 1 illustrates the high-level architecture of NIST's cloud computing standard [3]. The architecture defines five key actors: the cloud customer, cloud supplier, cloud carrier, cloud auditor, and cloud broker. In cloud computing, an actor is an individual or organization involved in transactions or processes and/or fulfilling certain responsibilities. The roles of these actors, as outlined in the NIST cloud computing standard, are summarized in Table 1 [3]. In this framework, security is considered a part of the Cloud Provider's responsibilities.

Table 1
Actors in NIST Cloud Computing reference architecture adopted from [26]

Actor	Definition
Cloud Consumer	A person or organization that maintains a business relationship with, and uses service from, Cloud Provider.
Cloud Provider	A Person, organization, or entity responsible for making a service available to interested parties.
Cloud Auditor	A party that can conduct independent assessment of cloud services, information system operations, performance and security of the cloud implementation.
Cloud Broker	An entity that manages the use, performance and delivery of cloud services, and negotiates relationships between Cloud Providers and Cloud Consumers.
Cloud Carrier	An intermediary that provide connectivity and transport of cloud services from Cloud Providers to Cloud Consumers.

Data in the cloud is often stored in a random manner, making it challenging for users to locate specific information as the volume of data grows. In contrast, organizing data systematically simplifies access. Therefore, there is a need for a model that enables data to be stored in an organized manner on the cloud. The advantages of data classification include.

- **Improved Accuracy:** Ensures precise classification with fewer false positives through the use of compound word searches.
- **Efficient Indexing:** Provides an index for searching sensitive phrases without the need to re-crawl data storage.
- **Customizable Taxonomy:** Features a flexible taxonomy manager that allows customization of classification parameters.
- **Automated Processes:** Includes procedures for automating tasks such as transferring sensitive data from public sharing.
- **Broad Support:** Accommodates on-premises and cloud content sources, as well as structured and unstructured data.

Users are often reluctant to upload personal and confidential files to online storage due to concerns that service providers might misuse their data. They also worry about data breaches resulting from prevalent cloud storage attacks [2]. Current cloud-based systems use a uniform key length for encrypting all data, which may not be ideal as it does not account for varying levels of data sensitivity. Treating all data equally, regardless of its confidentiality level, leads to unnecessary overhead and slows down processing.

To address these issues, this study focuses on three key aspects of cloud computing: data sensitivity, automatic classification, and high accuracy. We propose an effective system designed to ensure automatic data classification using machine learning algorithms, thereby maintaining confidentiality and integrity in cloud storage. This approach aims to reduce manual classification efforts while achieving high accuracy.

|e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 9, September 2023 ||

| DOI:10.15680/IJIRSET.2023.1209110 |

The proposed model categorizes data into three sensitivity levels: Basic Class, Confidential Class, and Highly Confidential Class, as illustrated in Fig: 2. The model development involves three phases: pre-processing text datasets, extracting features using the Python sk learn library, and training the model based on these features. The final phase involves classifying text data into the three sensitivity classes using various classifiers to achieve optimal accuracy.

Information classification helps determine the appropriate security standards and policies needed to protect data. Data is categorized into personal and non-exclusive (non-distinct) types based on its features. Sensitive data is classified as "confidential" or "highly confidential," while other data is categorized as "basic." The proposed cloud data classification framework utilizes RF, NB, KNN, and SVM algorithms to achieve high accuracy and automatic classification.

This study focuses on three critical aspects of cloud computing: data sensitivity, automatic classification, and high accuracy. The proposed system ensures data confidentiality and integrity during both transmission and storage, highlighting the importance of data confidentiality in cloud environments. Additionally, the framework aims to minimize manual classification efforts and achieve high accuracy.

The structure of the remaining sections of this paper is as follows: Related work is discussed in the second section, the proposed methodology is outlined in the third section, experiments are detailed in the fourth section, results and discussions are covered in the fifth section, and future directions are explored in the sixth section. The paper provides evidence and analysis for future research directions.

II. RELATED WORKS

In [4], digital signatures were enhanced to improve security, with the RSA method employed to protect the confidentiality aspect. The encryption process is executed in five stages: key generation, digital signature implementation, encryption, decryption, and signature verification.

In [5], an architecture was proposed that integrates digital signatures and Diffie-Hellman key exchange with the Advanced Encryption Standard (AES) to secure data confidentiality in the cloud. The Diffie-Hellman key exchange ensures that the key in transit remains ineffective if compromised, as it cannot be used without the user's private key, which is only accessible to authorized users. This three-tiered approach significantly strengthens the security of cloud-stored data against potential breaches.

Sinha et al. [6] provide a comprehensive overview of cloud computing, covering its benefits, architecture, implementation, and potential issues. The study reviews various security, data, and performance challenges associated with cloud computing. Meanwhile, [7] investigates different cryptographic algorithms to ensure data confidentiality in the cloud. This study evaluates several cryptographic methods based on criteria such as block size, key length, and characteristics, examining their effectiveness in securing cloud data.

The KNN technique is utilized for classifying data to maintain confidentiality, with the primary goal of ensuring security. By using KNN, data is categorized into sensitive and non-sensitive groups. Sensitive data is then encrypted to enhance security. Classification simplifies the process of selecting the appropriate security level based on the data's requirements, thereby improving overall security.

In [9], another critical technique for enhancing cloud data security is presented. This approach involves classifying data based on various parameters—such as access control, content, and storage—and then applying encryption to the classified data. This method aims to improve both security and efficiency by first classifying the data according to these properties and then encrypting it.

The K-NN and Improved Naïve Bayes algorithms have been employed to identify and ensure data confidentiality. The main objective is to protect confidential information by classifying data into sensitive and non-sensitive categories. Using K-NN and Improved Naïve Bayes, sensitive data is encrypted for protection. The Improved Naïve Bayes method achieved a 72% accuracy rate, facilitating better protection of the data.

Decision trees, used for classification, recursively partition the training dataset into smaller subsets at each node based on a set of test criteria [20]. Each node in the tree tests a feature from the dataset, with branches representing possible

|e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 9, September 2023 ||

| DOI:10.15680/IJIRSET.2023.1209110 |

values of that feature. Classification begins at the root node and follows down each branch based on the feature values, with this process being repeated recursively [17].

The discussion indicates that existing methods often treat all data with the same level of confidentiality, using a uniform key size for encryption, which leads to increased processing costs and time. Additionally, these methods typically involve manual data classification with low accuracy and insufficient security.

The proposed approach aims to address these issues by reducing the need for manual classification in cloud computing. It focuses on achieving high accuracy and providing appropriate security levels by automatically classifying data using machine learning algorithms. The proposed method evaluates results across different classifiers, categorizing data into three sensitivity levels: Basic, Confidential, and Highly Confidential. This approach employs four machine learning algorithms to significantly enhance the accuracy and efficiency of data classification.

III. PROPOSED WORK

This research article examined the effectiveness of different machine learning data categorization techniques, including NB, RF, SVM, and KNN algorithms. The study focused on classifying data based on sensitivity levels in the cloud. The proposed model features three categories: Basic Class, Confidential Class, and Highly Confidential Class, as illustrated in Figure 2.

3.1 BASIC CLASS

In our proposed model, the Basic Class includes common data types like text documents that have a low level of confidentiality. Examples of such data are advertisements, announcements, and notices. This level offers minimal data security and does not necessitate client-side encryption. However, when transmitted, the data will be encrypted on the server side using the backup service's key.

3.2 CONFIDENTIAL CLASS

This class encompasses personal files such as private accounts, web accounts, and professional details. Designed for data with a medium level of confidentiality, the Confidential Class requires robust security measures to protect sensitive and private information. To ensure protection, encryption methods like AES can be applied. In this class, encryption will be performed on the client side.

3.3 HIGHLY CONFIDENTIAL CLASS

This class handles highly sensitive data such as financial transactions, confidential organizational documents, and military information. Given the high level of confidentiality, users might be cautious about new services. To address these concerns, security will be ensured using two widely recommended algorithms. The AES-256 encryption, endorsed by the US National Security Agency (NSA), will protect against unauthorized access to top-secret information. Additionally, the SHA-2 algorithm will be used to maintain data integrity by generating a hash value before any modification or transfer. This hash value will be used to verify data during retrieval, ensuring that it remains unaltered.

3.4 DATASET

We obtained the Reuters-21578 text categorization dataset from the UCI ML repository. For testing our proposed system, we also gathered confidential and highly confidential data from the CIA public library. This dataset includes various types of content such as advertisements, announcements, news articles, organizational account information, and military data, sourced from the mentioned international repositories. The datasets used in this study are available in the [UCI, CIA] repositories, with access links provided in Table 2.

|e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 9, September 2023 ||

| DOI:10.15680/IJIRSET.2023.1209110 |

Table 2
Dataset with corresponding repositories

S.No	Dataset Type	Number of documents	Dataset Link
1	Public and Confidential Data	4000	https://miguelmalvarez.com/2015/03/20/classifying-reuters-21578-collection-with-python-representing-the-data/
2	Highly Confidential Data	2010	https://www.archives.gov/research/intelligence/cia

3.5 DATA PROCESSING

Natural Language Processing (NLP) involves analysing, manipulating, and extracting meaning from human language in a way that computers can comprehend. The NLTK library is used to pre-process text data before it is fed into the algorithm. This transforms unstructured text into a structured format. Processing is a crucial element in many machine learning techniques and significantly affects the classification process as well [15].

IV. RESULTS AND DISCUSSION

The goal of the proposal is to validate the application by creating relevant modules. Python 3.7 is used on the backend for effective data analysis with appropriate tools and frameworks. The process starts with the training phase of document classification, which includes NLP preprocessing, feature extraction, and vector creation, followed by the testing phase. The prediction module applies various classification algorithms for comparison. Figures 5 and 6 illustrate the performance of the RF, NB, KNN, and SVM algorithms using different methodologies. The comparison shows that the proposed approach outperforms existing methods. Performance graphs reveal that the new technique exceeds the old one. The average precision, recall, and F-measure of each classifier are shown, with Figure 5 displaying these metrics for the three class labels. Figure 8 compares the accuracy of RF, NB, KNN, and SVM algorithms, showing SVM with 43% accuracy, KNN with 83%, NB with 72%, and RF with 92%. This indicates that the proposed algorithm achieves superior classification accuracy. Additionally, Figure 6 compares the precision, recall, and F1-measure for these methods. The proposed methodology, which uses machine learning algorithms to categorize cloud data based on security needs, reduces encryption time and demonstrates improved accuracy, precision, recall, and F1-measure.

V. CONCLUSION AND FUTURE WORK

This study proposes a method for ensuring data confidentiality and automatic text document classification within a cloud environment. The primary objective is to define data sensitivity levels and achieve high accuracy in classification. The research focuses on using machine learning algorithms to categorize data into three levels: basic, confidential, and highly confidential. The main contribution of this security model is the integration of data confidentiality and classification through a machine learning approach. The methodology aims to develop application modules that can be validated effectively. Python 3.7 was chosen for the system's backend due to its suitability for data analysis when paired with the right tools and libraries. The results indicate that the proposed method is more effective than simply storing data without assessing its security needs. Additionally, the random forest algorithm surpasses the NB, SVM, and KNN techniques in terms of accuracy, precision, recall, and F1-measure. Future plans include transmitting the classified data using TLS, AES, and SHA, and expanding the system by addressing additional research gaps with other cryptographic methods to enhance reliability and security.

REFERENCES

1. Sun, X., Wang, Z., Wu, Y. et al. A price-aware congestion control protocol for cloud services. J Cloud Comp 10, 55 (2021). https://doi.org/10.1186/s13677-021-00271-5
2. Al-Said Ahmad, A., Andras, P. Scalability resilience framework using application-level fault injection for cloud-based software services. J Cloud Comp 11, 1 (2022). https://doi.org/10.1186/s13677-021- 00277-z.
3. Song, D., E. Shi, I. Fischer and U. Shankar, “Cloud data protection for the masses”, IEEE Computer. Soc., Vol. 45, Issue 1, pp.39-45, 2012.

|e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 9, September 2023 ||

| DOI:10.15680/IJIRSET.2023.1209110 |

4. Somani U, Lakhani K, Mundra M. Implementing digital signature with RSA encryption algorithm to enhance the Data Security of cloud in Cloud Computing. In Parallel Distributed and Grid Computing (PDGC), 2010 1st International Conference on 2010 Oct 28 (pp. 211-216). IEEE.
5. Rewagad P, Pawar Y. Use of digital signature with Diffie Hellman key exchange and AES encryption algorithm to enhance data security in cloud computing. In Communication Systems and Network Technologies (CSNT), 2013 International Conference on 2013 Apr 6 (pp. 437-439). IEEE.
6. Sinha N, Khreisat L. Cloud computing security, data, and performance issues. In2014 23rd Wireless and Optical Communication Conference (WOCC) 2014 May 9 (pp. 1-6). IEEE.
7. Diwan V, Malhotra S, Jain R. Cloud security solutions: Comparison among various cryptographic algorithms. IJARCSSE, April. 2014 Apr.
8. Dubey AK, Dubey AK, Namdev M, Shrivastava SS. Cloud-user security based on RSA and MD5 algorithm for resource attestation and sharing in java environment. In Software Engineering (CONSEG), 2012 CSI Sixth International Conference on 2012 Sep 5 (pp. 1-8). IEEE.
9. Yellamma P, Narasimham C, Sreenivas V. Data security in cloud using RSA. In Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on 2013 Jul 4 (pp. 1-6). IEEE.
10. Lo'ai Tawalbeh NS, Raad S. Al-Qassas and Fahd Al Dosari,"A Secure Cloud Computing Model based on Data Classification". In First International Workshop On Mobile Cloud Computing Systems, Management and Security (MCSMS-2015) 2015 (Vol. 52, pp. 1153-1158).
11. Jacovi A, Shalom OS, Goldberg Y. Understanding convolutional neural networks for text classification. arXiv. 2018.
12. Pennington J, Socher R. Manning CD. Glove: Global Vectors for Word Representation; 2014. p. 1532– 43.
13. Thiyagarajan, D., Shanthi, N.: A modified multi objective heuristic for effective feature selection in text classification. Cluster Comput. 22(5), 10625–10635 (2019).

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.423

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details