Journal of Computational Algorithms and Information Technology

Article

Smart Identity and Access Management Leveraging Neural Networks

Ali Raza [1,*] and Zakir Ullah[2]

[1] University of Lahore, Pakistan.
[2] University of Peshawar, Pakistan.
* Correspondence: derartushumet@gmail.com

Received: 11 July 2024; Accepted: 18 September 2024; Published: 15 October 2024.

Abstract: Cloud computing ensures easy access to data on multiple form factor devices. It provides the flexibility to connect with business anytime, anywhere. Numerous cloud benefits are accounted for, including reduced information technology costs, infrastructure costs, scalability, business continuity, collaboration efficiency, the flexibility of work practices, and automated updates. These benefits can be utilized only with a strong identity and access management of cloud applications and services. Various challenges of IAM systems include user password fatigue, failure-prone manual provision and deprovisioning, visibility of compliance, isolated user directories for each application, access management across an explosion of browsers and devices, updated application integrations, the variation of administrative models for different applications and lack of knowledge of best practices for optimal utilization of cloud services. So to overcome the challenges, this paper proposes an intelligent identity and access management model trained using neural networks. This model extracts user's file access logs, learn from them, and provides a question based intelligent identification system for user identification, authentication, and authorization of files, and ensures efficient and faster file access.

Keywords: Cloud Computing; Deep Learning; Artificial Neural Network; Identity and Access Management; Access Control

1. Introduction

Access to cloud services and resources management requires an access management interface, which causes an unauthorized user to try penetrating the system and access unauthorized resources. In the cloud, this cause is highly vulnerable to the traditional infrastructure where only a few administrators have access to functionalities and geographical security. There are four essential functions used to ensure effective identity and access management of resources on the cloud. Those are identity provisioning and de-provisioning, authentication and federation, authorization, and user profile management and support for compliance [1].

Enterprises are concerned about embedding their resources in the cloud environment due to the high-security risks involved in confidential data theft and exposure. Latest research broadly specifies extreme importance of identity and access management security [2].

Security is still a barrier for cloud users, as consolidated by International Data Corporation. Real security incidents including but not limited to outages at amazon web services, Gmail are a clear example of high-level insecurities. Principle elements of security are identity, infrastructure, and information [3].

2. Literature Review

Khandakar Entenam Unayes Ahmed and Vassil Alexandrov proposed Identity and Access Management in cloud computing [1]. International Data Corporation survey shows that 87.5% of users are reluctant to adopt cloud computing

for future project deployments. The reason behind this resistance is the security of data on the cloud. A possible solution proposed by the authors is Identity and Access Management, which ensures improved security of data on the cloud. This model includes authentication with Kerberos and authorization based on an RBAC processor that uses Java to implement authorization policies for granting user access. This model's services include authentication server service, ticket-granting server service, RBAC processor service, and edge node service. This model is prone to losing personal identity information due to man in the middle attack possibility between the cloud service provider and trusted third party.

Ruediger Schulze proposed identity and access management for cloud services used by the payment card industry [4]. Its data security standard mandates that all the cardholder data environment involved in the credit card payment process must comply with all required standards. The research aimed to identify identity and access management essential functions to control and ensure standards of PCI DSS. This model includes two methods for authenticating cloud services. This model is limited to authentication and lacks management of access authorizations. Alina Madalina Lonea, Huaglory Tianfield, and Daniela Elena Popescu proposed identity management in the cloud [5]. This model is focused on web application security and virtualization security issues. Mitigation techniques that provide adequate identity and management architecture are discussed. Further discussions are extended towards security requirements, standards of interests, and currently available identity and access management solutions.

Umme Habiba, Rahat Masood, Muhammad Awais Shibli, and Muaz A Niazi proposed assessment criteria for evaluating existing and upcoming cloud-based identity and access management systems [6]. This research includes attacks on identity and access management systems and countermeasures against the mentioned attacks and the feature mechanism relationship used to evaluate cloud-based identity and access management systems. It leads to the development of a robust identity management system. Research is evident enough to describe the pros and cons of existing architecture and functionality. It shows limitations in the context of reliability and applications. This approach lacks in mitigation of unlisted attacks identified on cloud systems. Surya Majumdar, Taous Madi, Yushun Wang, Yosr Jarraya, Makan Pourzandi, Lingyu Wang, Mourad Debbabi in "Security Compile Auditing of Identity and Access Management in the Cloud: Application to OpenStack" [7] proposed a framework for the cloud on security compliance auditing which is done by using OpenStack. Lack of security in the cloud has become a significant problem. So auditing the cloud was the main challenge. The results of experiments show that the formal method auditing is realistic in the large cloud environment. This research still needs to integrate existing techniques into their system on trusted auditing to make cloud infrastructure more trustable. Data collection is a costly part of this approach. Win-Bin Huang, Wei-Tsung Su, in "Identity-based Access Control for Digital Content based on Ciphertext Policy Attribute-Based Encryption" [8] proposed an approach to digital content. Identity-based access control is the approach, which is further based on ciphertext-policy attribute-based encryption (iDAC) [9]. If any digital content is duplicated, then also this approach will work. One copy of encrypted digital data can be shared with multiple users, reducing the content server's overhead. Performance analysis is based on considering different factors such as security, time complexity, and space complexity. It concludes that encryption-based access control approaches and the traditional access control list were outperformed by iDAC. This research still needs to apply for digital content protection. They only provide YES/NO admission control decisions as there is limited iDAC.

Samlinson.E, M.Usha in "User-Centric Trust based Identity as a Service for federated Cloud Environment" proposed a service that aims to develop trust among Cloud Service Providers (CSPs) [10]. This service is known as a user-centric trust agent identity service. The authors discussed various standards like SAML, OAuth , XACML, and SPML so that trust and secure access can be enabled to cloud services. Cloud providers can shift to an economical and centralized solution when implemented in open standards. This research paper lacks in prototype having more users. The author's proposed model does not have multiple user attribute-based trust that can be used for specific transactions. In this model, authors have not addressed user trust in various perspectives.

Jorge Werner, Carla Merkle Westphall, in "A Model for Identity Management with Privacy in the Cloud," proposed a model that will solve privacy issues related to Personally Identifiable Information (PII). A prototype was developed of dynamic scopes, federation agreements, and security policies using OpenID Connect (OIDC). This model helps to reduce the risk of breaching privacy. Other features such as data quality, user-friendly, the trust of obfuscation were developed. The authors presented the improvements in a framework that will validate this model. This research paper lacks considering or working on new approaches related to privacy in the cloud with identity management.

3. Proposed Methodology

Cloud computing is a blend of various configurable computing assets like servers, networks, services, applications, and storages that help give advantageous and on-request access to cloud services in the hands of cloud clients. Cloud

computing is generally referenced by individuals and utilized in numerous business fields. However, management of these identities and controlling access by cloud clients and applications stay a great question to date. For augmenting security in ventures, implementing a trustable identity, and access management (IAM) system in the cloud becomes necessary. According to International Data Corporation (IDC), some functionalities of a flawless IAM are User authentication/authorization, Identity management, control by privileged users, and data protection. This paper proposes an intelligent and trustable IAM system implemented through neural networks. The objective is to create a system that can achieve an intelligent identity mechanism to authenticate real users and access files on the cloud server. Access management consists of two parameters:

1. Efficient and easy file access
2. Policy management of the files

For achieving efficient access of files, those files which are used frequently by users are kept in a separate active server based on learning and association policy in order to provide faster access and to reduce the latency of those files to the IAM (Identity and Access Management) system for further processing. The remaining files will be stored in other servers according to the cloud storage servers' workload balancing techniques. For Policy management, the system will learn which files are no longer accessed by the user and ask the administrator to revoke the permissions on those files granted earlier to the respective users. For this, the machine learning model predicts files accessed by the user in the last few days, which can be best explained in the below given figure 1.

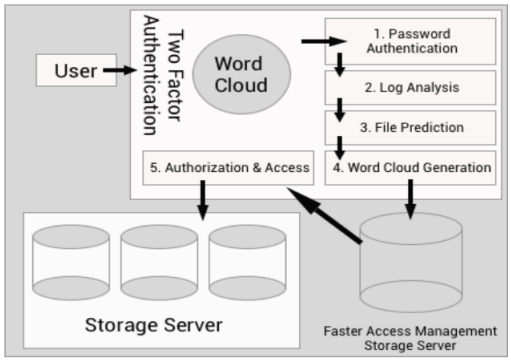

Figure 1. Machine Learning Model for File Prediction

The model will identify user logs in order to know about the recently accessed files by a user. These user file logs can be generated by setting up Audit control services on the cloud server, installed explicitly. The file logs generated will be in raw or unsupervised format. This unsupervised data obtained needs to be converted into supervised CSV (Comma Separated Values) for further processing. This conversion is accomplished through a bash script to extract essential features from the user log files. Some features stored in CSV which were extracted from the unsupervised user logs are:

- type (to identify files of type: PATH)
- msg = audit (time_stamp:ID)
- item (reference of the item among total items)
- name (Stores path of the file accessed)
- inode (Inode number associated with the file)
- dev (Device ID)
- mode (records file permissions)
- ouid (Owner's user ID)
- ogid (Owner's group ID)

These features will be identified to check which files are useful for the Identification system to authenticate genuine users. The data will be pre-processed to clean and label the string values to numeric data for building the machine learning model. After the numerical and categorical data is obtained after the pre processing phase, the data is ready for analysis. Nevertheless still, the data is unstructured. Now the question arises for clustering the files to specific user identification. So clustering will be done on the pre-processed data. The similarity in the clusters will be based on the following features:

1. OUID (for identifying the users who have accessed the file earlier)

2. NAME (to identify the names of the files the model will be using)
3. MSG (to identify the timestamp to check the last access time of that file in order to recognize recently accessed files)

After that, the random text will be extracted from the files identified for authentication of a particular user by the Identification system. The text will be used further to extract some words, which will be blended into a word cloud according to which the user has to identify the file name containing those words for gaining access to the file server.

3.1. Model Training and Evaluation Process

A machine learning model must be selected to identify the files (through filenames) based on the three essential features mentioned above. Further, these files will be used for the authentication process by the IAM (Identity and Access Management) system. The machine learning model is first directly trained and evaluated upon the pre-processed user logs stored in the CSV (Comma Separated Values) file without any clustering algorithms applied. The prediction using Random Forest Algorithm is shown in the given below figure 2. From this, it can be observed that the Random Forest Algorithm is the most suitable model that possesses:

- Accuracy score - 94%
- ROC AUC score - 99%

However, it has a training time of 7 seconds logarithmic loss is 37%, which needs to be reduced to predict the files for authentication.

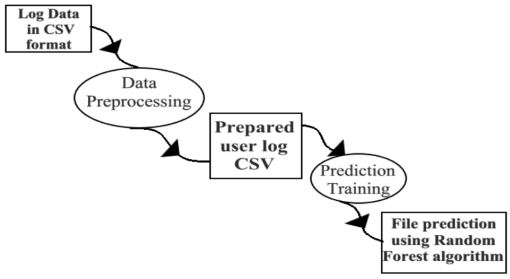

Figure 2. Prediction using Random Forest Algorithm

Another model using Artificial Neural Network (ANN) is deployed on the preprocessed CSV (Comma Separated Values) data without any clustering to identify the user files. The prediction using Artificial Neural Network is shown in the given below figure 3. The results obtained are:

- Accuracy score - 68%
- ROC AUC score - 92%

However, the training time is 0 seconds, and the logarithmic loss stands at 1.3%.

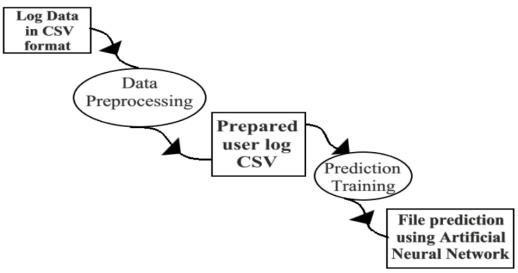

Figure 3. Prediction using Artificial Neural Network

It is observed that the accuracy of the Random Forest algorithm is much higher than the Artificial Neural Network when deployed directly without creating any clusters. However, the Random Forest algorithm's logarithmic loss is much higher in contrast to the logarithmic loss of Artificial Neural Network. So, further amendments are to be made to select

and implement a machine learning model with improved accuracy and reduced logarithmic loss to make the prediction process more efficient and effective. Artificial Neural Network is the better one with the least logarithmic loss and training time. Now, specific changes are to be made to improve the Artificial Neural Network algorithm's accuracy score.

A different approach is followed in which firstly a clustering algorithm is applied to the pre-processed data and then deploying the Artificial Neural Network algorithm using the clustered data. Several Clustering algorithms were deployed, and it is observed that the Gaussian Mixture Model (GMM) gives the best results with a Silhouette score of 0.4334 and least Inertia score in segregating the files into two clusters that is one cluster for the files which are useful for the Identification process and another cluster for the files not much useful. The GMM proved to be the best algorithm to segregate these files based on three essential features: OUID (Owner's user ID), OGID (Owner's group ID), and INODE (INODE number associated with the file) of the log files. Then a model using Artificial Neural Network is deployed upon the clustered data after Gaussian Mixture Model was applied. The below given figure 4 shows the prediction using Artificial Neural Network after clustering data through Gaussian Mixture Model.

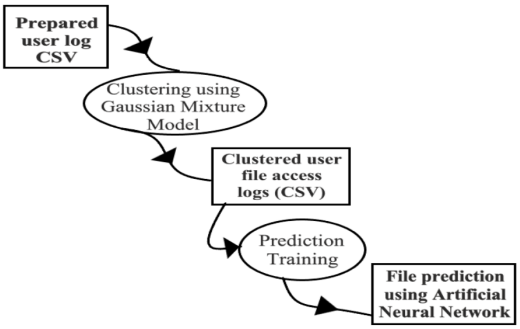

Figure 4. Prediction using Artificial Neural Network after clustering data through Gaussian Mixture Model

It can be observed that the Artificial Neural Network algorithm's accuracy score has been increased to 99% if applied after clustering the user log data using GMM (Gaussian Mixture Model). This model will help classify and predict the files used by the Intelligent Identification System to authenticate a specific user based on the user's recently accessed files. The Intelligent Identification system will then use the predicted files to extract the text, and then random words will be selected to generate a word cloud for a specific user to identify the files to which those words belong.

Figure 5. A Word Cloud

These files will be among those only which a specific user has recently accessed. This will help in enforcing the security by identifying and authenticating only the genuine users to access the file system.

4. Experimental Analysis

The model first takes audit files to begin proper data analysis, containing log records as an input and converts the raw data present into the CSV (Comma Separated Value) file. The CSV (Comma Separated Value) file will contain 16 columns in which the data will be distributed. The columns are as follows:

- Type: The type field contains the type of the record or the event type. It is specified at the beginning of every audit record.

- Msg: This field records a timestamp and a unique ID of the form audit record (time_stamp: ID). Multiple records could share the same timestamp and ID if they were generated as part of the same Audit event. Msg also records various event-specific name=value pairs provided by the kernel or user space applications.
- Arch: The arch field contains information about the CPU architecture of the system. The value which is followed by it is encoded in hexadecimal format.
- Syscall: The syscall field shows the type of system call that was sent to the kernel.
- Success: The success field records whether the system call recorded in that particular event succeeded or failed.
- Exit: The exit field contains a value that tells the exit code returned by the system call. This value of the exit is different for different system calls.
- A0-A3: These fields show the first four arguments of the system call in an event.
- Items: The items field contains the number of path records in the event.
- PPID: The PPID field will show the Parent Process ID (PPID).
- AUID: The AUID field represents the Audit user ID, that is, the login UID. This ID is assigned to a user after login and given to every process when the user's identity changes.
- UID: The UID field will show the user ID of the user who started the analyzed process.
- GID: The GID field shows the group ID of the user who started the analyzed process.
- EUID: The EUID field records the effective user ID of the user who started the analyzed process.
- SUID: The SUID field represents the set user ID of the user who started the analyzed process.

After converting our data into a CSV (Comma Separated Value) file, the model pre-processes the file to analyze it better after applying various Machine Learning algorithms to see which algorithms may provide us better performance metrics. The observations are shown in the below given table 1 which are as follows: It can be analyzed from the above table

Table 1. Performance of various algorithms on prepared data

Name	Train Time	Accuracy	Precision	Recall	Log Loss	ROC AUC
Random Forest	7s	0.76	0.37	0.44	0.62	0.98
ANN	0s	0.76	0.37	0.44	1.46	0.95
GBT	2s	0.71	0.42	0.49	1.44	0.94
Logistic Regression	9s	0.57	0.18	0.22	1.95	0.93

that Random Forest and Artificial Neural Network have high accuracy and ROC. Even though GBT also has high ROC and recall, the accuracy value for GBT is lower. Among Random Forest and Artificial Neural Network, artificial neural networks are preferred because training time for ANN is much less than Random Forest. Random forest is primarily seen to have higher ROC, so prediction is made using random forest, and the following features were discovered to be the most critical dependent variables.

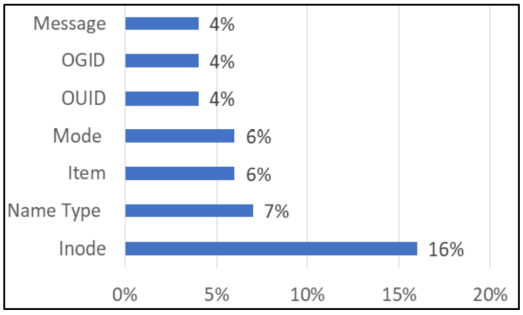

Figure 6. Variable Importance in Random Forest

As can be seen in the above figure 6, Inode, name type, mode, and item are the most influential factor in case the model uses random forest for prediction.

Table 2. Performance metrics of Random Forest

AUC	Recall	Precision	Accuracy	Log Loss
0.97	0.89	0.95	0.94	0.37

Clustering algorithms were used to make clusters of the pre - processed CSV (Comma Separated Value) file to improve algorithms' further performance. So, to do this, some algorithms were tried, and the following results were achieved.

From the table **??** given above, it is analyzed that Gaussian Mixture and Agglomerative Clustering are equally capable of clustering our data. Even though some algorithms gave a better silhouette, they created more clusters and were thus ignored. Using Gaussian Mixture, two clusters were created in the prepared dataset, and our dataset was now ready to be analyzed.

The important factors while making clusters are shown in the horizontal bar as shown in the below given figure 7. It can be seen that OUID, OGID, Inode, and Mode were the essential factors in making the clusters while performing Gaussian Mixture on the dataset.

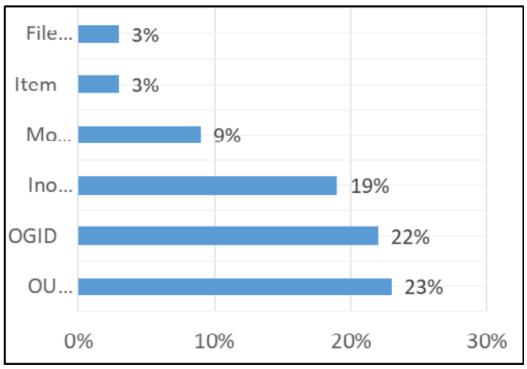

Figure 7. Variable Importance in Gaussian Mixture

Now using Gaussian Mixture, the model performed analysis using Random Forest and following confusion matrix was obtained. As shown in the below given table 3 and table 4. The following results were obtained from the confusion

Table 3. Confusion Matrix for Random Forest

	Predicted True	**Predicted False**
Actually True	57	3
Actually False	3	37

matrix. After using Random Forest, Artificial Neural Network was also used to predict the file using Gaussian Mixture

Table 4. Performance metrics of Random Forest

AUC ROC	**Recall**	**Precision**	**Accuracy**	**Log Loss**
0.94	0.95	0.95	0.94	0.41

Clustering. Following Confusion Matrix was obtained. As shown in the below given table 5 and table 6. The following

Table 5. Confusion Matrix for ANN

	Predicted True	**Predicted False**
Actually True	59	1
Actually False	0	39

results were obtained from the confusion matrix.

Table 6. Performance metrics of ANN

AUC ROC	**Recall**	**Precision**	**Accuracy**	**Log Loss**
0.98	0.99	0.989	0.99	0.13

It can be analyzed from the obtained results that Artificial Neural Network performs considerably better than Random Forest and, together with the Gaussian Mixture Clustering model, has achieved near 99% accuracy. Using

the improved artificial neural network modified with the Gaussian mixture, the model can successfully predict the word cloud generation's optimal file. Word cloud is generated using Textract. For second-factor authentication, multiple word clouds will be presented to a user out of which if an appropriate selection is identified, full file access & authorization will be granted.

5. Conclusion

In this research, an efficient model is proposed to extract user's file access logs, learn from them, and provide a word-cloud based on an intelligent identification system for user identification, authentication, and authorization of files and ensure efficient and faster file access. Model converted the access logs into CSV (Comma Separated Value) file and pre - processed it. Then various machine learning algorithms are tested on the data, and experimentally Artificial Neural Network was found to have an optimum performance metric reaching 99 percent accuracy. This model can be further improved on different file formats like images with minimal scenic texts in different languages. This file prediction model can further be extended for enterprise-level group file access and permission policy management. It ensures that optimal file access privileges will be assigned to a group user to accomplish the law of security "Principle of Least Privilege.

Author Contributions: All authors contributed equally to the writing of this paper. All authors read and approved the final manuscript.

Conflicts of Interest: "The authors declare no conflict of interest."

References

[1] Ahmed, K. E. U., Alexandrov, V. (2011). Identity and Access Management in Cloud Computing. In: Mahmood Z., Hill R. (eds) Cloud Computing for Enterprise Architectures. Computer Communications and Networks. Springer, London. https://doi.org/10.1007/978-1-4471-2236-4_6.

[2] Andrei, T. (May 21, 2011). Cloud Computing Challenges and Related Security Issues. A Survey Paper (online), http://www1.cse.wustl.edu/~jain/cse571-09/ftp/cloud/index.html.

[3] Ahmad, I., Basheri, M., Iqbal, M. J., & Rahim, A. (2018). Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection. *IEEE Access*, 33789–33795. https://doi.org/10.1109/access.2018.2841987.

[4] Anjana, P. S., Badiwal, P., Wankar, R., Kallakuri, S., & Rao, C. R. (2019). Cloud Service Provider Evaluation System Using Fuzzy Rough Set Technique. *IEEE International Conference on Service-Oriented System Engineering (SOSE)*. https://doi.org/10.1109/sose.2019.00033.

[5] Bethencourt, J., Sahai, A., & Waters, B. (2007). Ciphertext-Policy Attribute-Based Encryption. *IEEE Symposium on Security and Privacy*. https://doi.org/10.1109/sp.2007.11.

[6] Brown, K. P., Hayes, M. A., Allison, D. S., Capretz, M. A. M., & Mann, R. (2012). Fine-Grained Filtering of Data Providing Web Services with XACML. *2012 IEEE 21st International Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises*. https://doi.org/10.1109/wetice.2012.41.

[7] CSA, (October 14, 2010). Security Guidance for Critical Areas of Focus in Cloud Computing V2.1 (online) Cloud Security Alliance (2009). http://www.cloudsecurityalliance.org/csaguide.pdf.

[8] CPNI, (October 2, 2010). Information Security Briefing Cloud Computing (online) Centre for the Protection of National Infrastructure, http://www.cpni.gov.uk/Docs/cloud-computing-briefing.pdf.

[9] Chi, M.-T., Lin, S.-S., Chen, S.-Y., Lin, C.-H., & Lee, T.-Y. (2015). Morphable Word Clouds for Time-Varying Text Data Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 21(12), 1415–1426. https://doi.org/10.1109/tvcg.2015.2440241.

[10] Dwivedi, S. K., & Rawat, B. (2015). A review paper on data processing: A critical phase in web usage mining process. *2015 International Conference on Green Computing and Internet of Things (ICGCIoT)*.

[11] Fett, D., Kusters, R., & Schmitz, G. (2017). The Web SSO Standard OpenID Connect: In-depth Formal Security Analysis and Security Guidelines. *IEEE 30th Computer Security Foundations Symposium (CSF)*. https://doi.org/10.1109/csf.2017.20.

[12] Gruschka, N., Iacono, L. L. (2009). Vulnerable Cloud: SOAP Message Security Validation Revisited. In: *IEEE International Conference on Web Services, ICWS*, Los Angeles, pp. 625–631.