iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** V **Month of publication:** May 2024

DOI: https://doi.org/10.22214/ijraset.2024.62303

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue V May 2024- Available at www.ijraset.com

Documentation Management System with PDF Chat

Vaibhav Devkate[1], Atharv Divekar[2], Prasad Hulmukhe[3], Om Randhir[4], Dr. Rajesh Chowdhary[5], Prachi Nilekar[6], Sushrut Joshi[7]

[1, 2, 3, 4]*B.E Student (Information Technology) International Institute of Information Technology (I2IT) Pune, India*
[5]*Head - Research & Development, Consultancy and Collaboration | International Relations Research Center, International Institute of Information Technology (I2IT) Pune, India*
[6]*Professor (Information Technology) International Institute of Information Technology(I2IT) Pune, India*
[7]*Senior Research Associate Pralhad P. Chhabria Research Center, International Institute of Information Technology (I2IT) Pune, India*

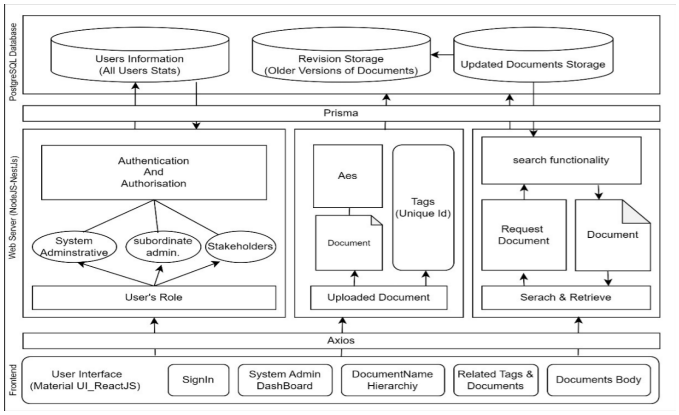
Abstract:DMS represents the principle of software and hardware solutions that provide many capabilities including data storage, version control, metadata tagging, management and advanced search. These systems are essential tools for businesses, governm ent agencies, schools, doctors, professionals, and individuals looking to improve their information management strategies. By fa cilitating collaboration in the business environment to ensures data integrity and public compliance management, DMS has prov en to be relevant in many ways.DMS represents the principle of software and hardware solutions that provide many capabilities i ncluding data storage, version control, metadata tagging, management and advanced search. These are essential tools for busine sses, government agencies, schools, doctors, professionals, and individuals looking to improve their datas management strategies . To Build this system we have used ReactJS for frontend, NestJS for Backend, for Database we used PostgreSQL and for the PDF chat we used Langchain in Python.

Keywords: Documentaion Managemnet system, Access Control, Version Control, PDF Chat, Langchain, LLM, OpenAI, Embeddings

I. INTRODUCTION

A Documentation management system (DMS) in the industry is an important tool for the effective use of the crucial data. The system provides central storage of all project information to ensure security and auditability of electronic information. It is designed to manage big documents, thereby reducing the administrative effort associated with data collection management. DMS allows rapid retrieval of information using specified content and makes this information relevant. engineers, geology Scientists, stakeholders, and analysts can easily access data for remote operations. It is also the first of the assets to provide acces control and updating of information to ensure that only authorized people are allowed to see it, reducing overall data loss. For an industry, DMS can improve coordination, improve management information, provide instant access to information, and reduce risk and liability.

A. Basic Architecture Diagram OF DMS



II. ACCESS CONTROL METHODS

A. *Discreationary Access Control (DAC)*

Discretionary Access Control is a type of security control that allows or reject access to an object through access rights determined b y the group owner and/or administrator of the object. The DAC's system is defined by the credentials (such as username and passwo rd) provided during user identification and authentication. DAC is arbitrary because the content (owner) can send credentials to othe r users to access items decides.

B. *Mandatory Access Control (MAC)*

Mandatory access control is a network security policy that accepts or denies access to an organization's private and crucial information. This distribution of the access rights depends upon the hierarchy of employees and employees in the organization. When the user try to access the resource, the system automatically checks whether he or she is granted to access that resource and which category the resource is assigned to. To access information, users must meet security and classification requirements.

For example, some employees need some information and permission from the organization to do their jobs. To effectively track and share this information, a system designed to allow multiple users to access the resources they need must be implemented.

To use the Mac, organizations must first invest time and effort in understanding and monitoring data flows. This includes viewing users and processes, accessing resources and policies, and properties (subscribers, groups). The system is easy to use because it usually needs to be done once and then only needs to be adjusted as the job or process changes.

C. *Role-Based Access Control (RBAC)*

Role-based Access control is a method of restricting access to a network based upon a user's job role. Organizations use RBAC, also known as role, based security, to resolve access levels based on employee roles and responsibilities.

Restricting access to the network is important for organizations that have large numbers of employees, employees, or authorized third parties (e.g., customers and suppliers) access to the network because Monitoring network access will be difficult. Companies that rely on RBAC can protect their sensitive data.

An employee's role in the organization determines what permissions a person is allowed to ensure that low-level employees cannot access paper-sensitive or high-performance data.

RBAC is centered on the definition of responsibility and authority. Access rights are based on factors such as authority, capacity, and responsibility. Employees can restrict access to the network and other services (such as access to certain files or programs). For example, certain files may be read only, but access to certain files or programs may be allowed temporarily to complete a task. Organizations can designate users as end users, administrators, or experts. These roles may also overlap or provide different permission levels for specific roles.

D. *Attribute-Based Access Control (ABAC)*

Attribute-based acces control is an authorization model that evaluates attributes (or properties) rather than roles to determine access control. The purpose of ABAC is to protect objects such as data, network equipment, and IT equipment from unauthorized users and operations without "approved" features as specified in the organization's security policy. ABAC has become important over the last decade as a form of access control developed from simple, role-based control lists. as control (RBAC). The Federal Board of Informatics Commissioners approved ABAC in 2011 as part of a plan to help federal agencies improve their access control systems. They recommend ABAC as the standard for organizations to share information securely.

III. VERSION CONTROL SYSTEMS

Version control is an Important component of a Documentation Management System (DMS). It is a systematic approach to managing changes and revisions in documents, ensuring that the most up-to-date and accurate information is always available. This also make sure that the pre-existing file and the revision both should be there to keep up the record of changes.

A. *Types of Version Control Systems*

There are Mainly two different version control systems as discussed in [3].

1) *Centralized version-control system (CVCS) [3]:*

There are few key points about Centralized version control system:

a) *Repository*: There is only one central repository which is the server.

b) *Repository Access*: Every user who needs to access the repository must be connected via network

c) In a CVCS, files are typically locked by a developer when they need to make changes. This prevents concurrent edits and potential conflicts.

d) CVCS systems maintain a complete history of all changes made to files and directories, including who made the changes and when.

2) Distributed Version-Control System (DVCS) [3]:

There are few key points about Distributed version control system:

a) *Repository:* Every user has a complete repository which is called local repository on their local computer

b) *Repository Access:* repository must be connected via network. DVCS allows every user to work completely offline. But user need a network to share their repositories with other users.

c) *Version History:* Every local copy in a DVCS contains the full version history of the project. This ensures that even if a central server is lost, individual copies can still be used to recover the project's history.

d) *Flexibility:* DVCS systems offer flexibility in defining workflows. Teams can choose a workflow that best suits their development process, whether it's centralized, feature-based, or topic-based branching.

B. Importance of Version-Control in DMS [2,3]:

a) *Accuracy and Consistency:* Version control helps maintain the accuracy and consistency of documents within an organization. It ensures that everyone is working with the same, current version of a document, reducing the risk of errors due to outdated information.

b) *Traceability:* Version-control provides a complete history of document revisions. This traceability is valuable for auditing, compliance, and accountability purposes, allowing organizations to track who made changes and when.

c) *Recovery from the Disaster:* Accidental deletions or data corruption can be disastrous for a DMS. Version control offers a safety net by allowing you to restore previous versions of documents, ensuring data integrity and continuity in the event of data loss.

C. Benefits

a) *Security:* Many industries require strict adherence to compliance regulations and data security standards. Version control helps organizations maintain compliance by ensuring that document revisions are tracked, audited, and securely stored.

b) *Helps in Decision Making:* Access to historical versions of documents aids in decision-making processes. Teams can review past iterations to understand the evolution of ideas, strategies, and decisions, leading to more informed choices.

c) *Improved productivity:* Version control eliminates the confusion and inefficiency associated with multiple document copies, email attachments, and conflicting edits. This results in increased productivity as team members can focus on their tasks instead of managing document versions.

IV. SECURITY

Advanced Encryption Standard (AES) is a cryptographic algorithm and encryption standard proposed by NIST to replace DES in 2001. Use of the AES algorithm supports a combination of data (128 bits) and a key length of 128 bits (AES 128), 192-bit (AES 192), and 256-bit (AES 256) depending on the length of the key. Apart from that there are various rounds used by AES which include 10 rounds, 12 rounds and 14 rounds.

The AES algorithm, a symmetrical block cipher, operates in multiple stages. It can encrypt and decrypt data efficiently. Initially, data (plain-text) is converted into cipher-text during encryption, and decryption reverses this process, restoring the plain text. This algorithm employs 128, 192, or 256-bit cryptographic keys and processes data in 128-bit blocks.

There are five modes of operation recommended by NIST that

can be used. Each mode of operation has its own parameters which are important to provide the necessary security of the algorithm.

The five modes of operation: Electronic Codebook (ECB), Cipher Block Chaining (CBC), Cipher Feedback (CFB), Output Feedback (OFB) and Counter (CTR).

AES (Advanced Encryption Standard) is a widely used symmetric key encryption algorithm that ensures the confidentiality and security of data. It uses the same key for both encryption and decryption. Here is an explanation of how AES encryption and decryption work:

A. AES ENCRYPTION:

1) *Key expansion:* The process begins by expanding the encryption key into some set of round keys. The number of rounds depends on size of the key: 128-bit keys for 10 rounds, 192-bit keys for 12 rounds, and 256-bit keys for 14 rounds.
2) *Initial Round:* In the first round, the plaintext (the data to be encrypted) is combined with the initial round key using a process called XOR (Exclusive OR).
3) *Rounds:* AES consists of a series of rounds, with each round applying specific mathematical operations to the data. The primary operations in each round are:
   - *a)* SubBytes: Substitutes each byte of data with another byte from a fixed substitution table called the S-box.
   - *b)* ShiftRows: Rearranges the bytes within each row of the data block.
   - *c)* MixColumns: Mixes the columns of the data block using a mathematical transformation.
   - *d)* AddRoundKey: XORs the data with the round key for that specific round.
   - *e)* Final Round: The final round omits the MixColumns operation but includes all other operations.
   - *f)* Output: After all rounds are completed, the result is the ciphertext, which is the encrypted form of the original plaintext

B. AES DECRYPTION

Decryption using AES is a reversible process that uses the same algorithm but in reverse order. Here's how AES decryption works:

1) *Key expansion:* Similar to encryption, the decryption process starts by expanding the encryption key into a set of round keys.
2) *Initial Round:* In the first decryption round, the ciphertext is combined with the initial round key using XOR.
3) *Rounds:* Just like in encryption, AES decryption consists of several rounds (10, 12, or 14, depending on the key size). These rounds apply the inverse of the operations used in encryption:
   - *a)* Inverse SubBytes: Reverses the substitution made during encryption.
   - *b)* Inverse ShiftRows: Restores the original arrangement of bytes within rows.
   - *c)* Inverse MixColumns: Undoes the mixing of columns.
   - *d)* AddRoundKey: XORs the data with the round key in reverse order.
   - *e)* Final Round: The final decryption round includes all operations except MixColumns.
   - *f)* Output: After all decryption rounds are completed, the result is the original plaintext, which is the data as it was before encryption.

V. PDF CHAT

PDF chat is designed to extract and interactively retrieve information from PDF documents. The application leverages several state-of-the-art natural language processing (NLP) and document processing techniques to enable users to ask questions and receive answers directly from PDF content.

A. Text Extraction and splitting:

The whole pdf is kind of a large data to handle and process at one go to minimize this and increase the accuracy we can divide the whole pdf into multiple chunks so the processing part will be easier, and the accuracy will be more.
For the first step we extract all the text present in the pdf.
For the second we fix the size of a chunk which is nothing but a smaller text paragraph.
We divide the pdf by giving the chunk size and some overlapping to not lose any data in the pdf file.

B. OpenAI Embeddings:

OpenAI embeddings typically refer to word embeddings, which are numerical representations of words in a format that can be used by machine learning models, such as neural networks. Word embeddings are a fundamental concept in natural language processing (NLP) and are used to convert words or text into vectors of real numbers. These vector representations capture semantic relationships between words and are useful in various NLP tasks, including text classification, language modeling, machine translation, and sentiment analysis.
These word embeddings are employed to transform the raw text data into numerical vectors, making it easier to perform semantic search and similarity calculations.

OpenAI embeddings play a critical role in performing semantic search within the document corpus. When a user submits a query, OpenAI embeddings are used to transform the query into a vector, and then a similarity search is conducted in the vector space to identify relevant data.

C. FAISS

FAISS (Facebook AI Similarity Search) is an open-source library formed by Facebook AI Research for precise similarity search and clustering of large datasets. It is primarily designed for searching for vectors in high-dimensional vector spaces, making it useful in various applications, including machine learning, computer vision, and natural language processing. FAISS is known for its speed and scalability in similarity search tasks.

The vectors generated by the embeddings are then indexed using FAISS. FAISS creates an index structure that allows for fast and efficient similarity searches. It organizes the vector data in a way that makes it optimized for nearest-neighbor searches, which is crucial for matching user queries with relevant document segments.

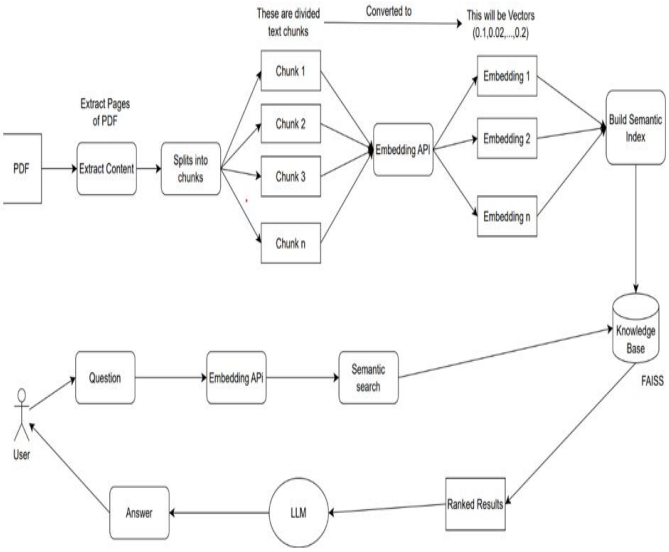
When a user submits a query through the web interface, the application uses FAISS to perform a similarity search. FAISS finds the most relevant text segments (chunks) based on the vector representations of the user's query. This allows the system to quickly identify which documents contain content relevant to the user's question.

After retrieving the most relevant text segments, the question answering chain is used to generate answers to the user's query. These answers are then presented to the user through the web interface.

D. Langchain:

Langchain is a newer framework for the applications which are powered by language models. Large-language models (LLMs) can be said as core component of the Langchain framework. In this PDF Chat Model used the langchain’s large language model.

E. Architecture Diagram Of PDF CHAT:

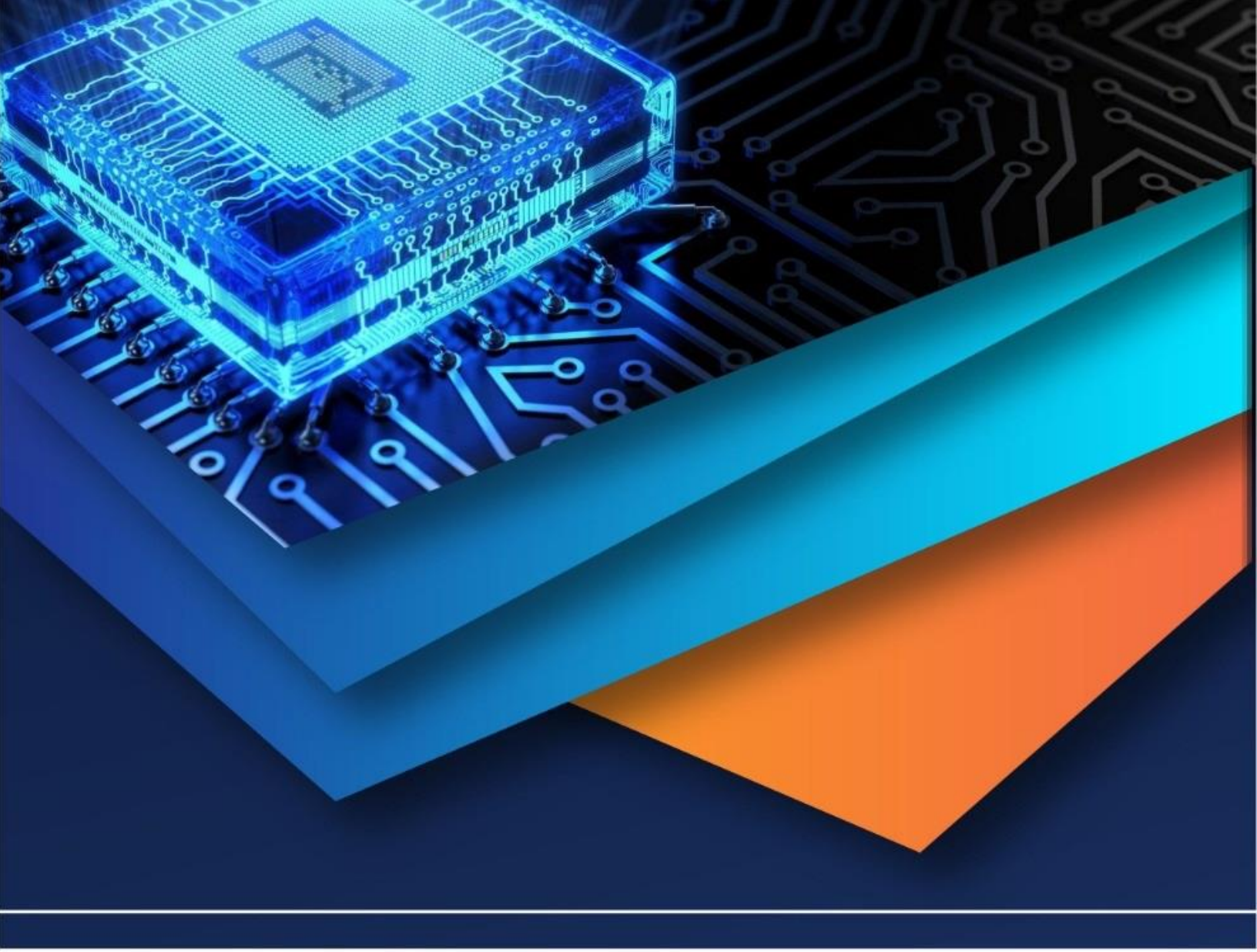


VI. CONCLUSION

In the above paper we’ve mentioned various methodologies about the document-management system like how we are able to practice access control to the system, how can we add version control to the document, to store the revisions and how the security algorithm we can implement. We have also seen the pdf chat model where if we want some particular information about something in the document, then we are able to simply ask the question and the model will give the answer correctly. As we are using langchain's LLM the accuracy of the Langchain-model is nearly 92.5% so we simply can clearly say that our model is also correct in giving the right solution to the question asked.

REFERENCES

[1] *I.*S. Jacobs and C.P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.
[2] Dmitriev, Sviatoslav O., Dmitrii A. Valter, and Artemii M. Kontsov. "System for Efficient Storage and Version Control of Arbitrary File Collections." 2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus). IEEE, 2020.
[3] Zolkifli, Nazatul Nurlisa, Amir Ngah, and Aziz Deraman. "Version control system: A review." Procedia Computer Science 135 (2018): 408-415.
[4] Xu, D. and Zhang, Y., 2014, June. Specification and analysis of attribute-based access control policies:An overview. In 2014 IEEE Eighth International Conference on Software Security and Reliability- Companion (pp. 41-49). IEEE.Zolkifli, Nazatul Nurlisa, Amir Ngah, and Aziz Deraman. "Version control system: A review." Procedia Computer Science 135 (2018): 408-415.
[5] Liddell Henry George, Scott Robert, Jones Henry Stuart, McKenzie Roderick (1984). A Greek-English Lexicon. Oxford University Press. page 827.
[6] Agrawal, Monika (2012). A Comparative Survey on Symmetric Key Encryption Techniques. International Journal on Computer Science and Engineering. 4: pages 877–882. CiteSeerX 10.1.1.433.2037
[7] El Sibai, R., Gemayel, N., Bou Abdo, J., & Demerjian, J. (2020). A survey on access control mechanisms for cloud computing. Transactions on Emerging Telecommunications Technologies, 31(2), e3720.
[8] Cho, V. (2008). A study of the effectiveness of electronic document management systems. International journal of information technology and management, 7(3), 327-352.
[9] Alade, S. "Design and Implementation of a Web-based Document Management System." Information Technology and Computer Science 2 (2023): 35-53.
[10] Livina, I. S. I., Chukwuemeka, A. E., Nnanna, E., Obi, N., Ikechukwu, I. S., & Ogonnaya, A. B. (2020). A Web Based Document Encryption Application Software for Information Security in Tertiary Institutions. Journal of Cybersecurity and Information Management (JCIM) Vol, 4(1), 26-35.
[11] Simarmata, J., Limbong, T., Ginting, M.B., Damanik, R., Nasution, M.I.P., Hasugian, A.H., Mesran, M., Sembiring, A.S., Hutahaean, H.D., Taufik, I. and Hasugian, P.M., 2018. Implementation of AES Algorithm for information security of web-based application. Int. J. Eng. Technol, 7(3.4).
[12] Sambetbayeva, Madina, Inkarzhan Kuspanova, Aigerim Yerimbetova, Sandugash Serikbayeva, and Shynar Bauyrzhanova. "Development of Intelligent Electronic Document Management System Model Based on Machine Learning Methods." Eastern-European Journal of Enterprise Technologies 1, no. 2 (2022): 115.
[13] Abidin, S.S.Z. and Husin, M.H., 2018. Improving accessibility and security on document management system: A Malaysian case study. Applied Computing and Informatics, 16 (1–2), 137–154.
[14] Ismael, Arkan, and Ibrahim Okumus. "Design and implementation of an electronic document management system." Mehmet Akif Ersoy Üniversitesi Uygulamalı Bilimler Dergisi 1.1 (2017): 9-17.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY