APPLICATION OF AI AND BLOCKCHAIN TECHNOLOGY IN DCMS FOR THE AUTOMATIC DOCUMENT CLASSIFICATION AND IMPROVE THE SECURITY

Abstract

Every day, the volume of information carried back and forth across the internet grows at a rate of around 2.3 quintillion bytes. Accelerated technical advancement, increased user interaction, and very efficient application development are just a few of the various advantages that may be realized by increasing speed. This methodology uses artificial intelligence to arrange unstructured yet sensitive data. This prevents unauthorized third parties from accessing the information and ensures that data flows correctly so that the programmer may continue to function effectively. To accomplish this, cutting-edge access control technology, encryption, and block chain technology are used to protect sensitive data.

An automatic classifier that will assist in the classification of both secret and non-confidential data has been implemented. To prevent unauthorized access to such sensitive documents, a solution that does not involve human interaction is required. This technique handles all parts of data processing, such as recognizing the type of data, assessing whether it is critical, and determining the next encryption step to safeguard the crucial data. This technology has the potential to be applied in a wide range of sectors, including the security of sensitive data in military and medical research, the protection of confidential data in enterprises, and other areas of interest. Analysis is used to extract content from photos, and a trained classifier is then used to determine whether or not the input data is secret. An automated approach

Author

Dr. Madhura K
Assistant Professor-Senior Scale
IT Department
MIT, MAHE Bengaluru.
Bangalore, India
Madhura.k@manipal.edu

known as machine learning is used to train the classifier. To prevent unauthorized access to critical information, papers placed in the institution's principal storage area are first encrypted using an RSA algorithm before being stored.

This is done on IPFS cloud to avoid storing a large amount of data on the block chain, which will incur a large cost, because every single storage on the block chain requires a cryptocurrency transaction. In this work, Ethereum is used to transact the data and store on the smart contracts, and the data is stored on IPFS cloud. This is done on the IPFS cloud to avoid keeping a significant amount of data on the block chain, which would incur a large cost. Third when block chain is integrated with other parts of block chain, a high-end access control system can be built in which authenticated users are traced at every step and held accountable for their actions. Because of the automated end-to-end data structuring, processing, securing, and storing system that was built on high-end internal access control and tracking in order to have a secure eco-system with authorized users and administrators, the company will be able to maintain its competitive advantage.

Keywords: Data structuring, data extraction, image processing, optical character recognition, artificial intelligence, machine learning, encryption, RSA, IPFS, Ethereum, smart contracts, cloud storage, security, blockchain

I. INTRODUCTION

The collection of enormous volumes of data from various sources is continually expanding in this current era of the internet. This need is increasingly becoming a crucial feature for all small, medium, and big size organizations / businesses that store sensitive internal information or sensitive data from their users, making it critical to minimize any security risk in order to keep end user trust. The increase in data volume is increasing the demand for a secure web content management system, particularly for organizations with a proactive presence in the cyber world. Since their debut, when they depended exclusively on HTML editors, web content management systems have evolved significantly. These systems have advanced to the point where they use highly optimized tools for content creation, administration, publishing, and maintenance. These tools are developed and implemented by a variety of parties and cross-functional organizations, and their number is increasing by the day. This work analyses the Web Content Management System (WCMS) as well as the supporting hierarchy and performs a comprehensive lifecycle analysis on the WCMs in order to develop an ecosystem that is both highly useful and safe for sensitive web content management systems [1].

It is apparent that online content management systems are growing in popularity since they are an important component that serves as a platform that leads the organization to functional success. This is accomplished by creating a secure environment in which corporate websites and intranets can operate. Web content management's major purpose is to provide end users with an abstraction layer that hides the complexities of the technology stack. This allows end users to concentrate on the most important component of the WCMS, which is giving secure access to the web portal's content management [2]. The efforts of researchers and organizations have mostly concentrated on the goal of building and developing Web-oriented data access, with one click away comfort, fully created using influential technical combinations such as ASP.net, sql, Django, php, and so on.

The security management of variety of web content management companies as it scales pose greater security risk on the companies. The most worrying issue is that 80% of the data is unstructured and is stored in different local and cloud locations. However, the unstructured data accumulated is very important as it contain sensitive information, innovative finding and more that drive the authenticity of data management within the organization. As these risks posed by unstructured data is driving research and cost allocation from the companies, security implementation within the web content management system with highly influential technology is the need of every company [3]. The market for Web content management is booming and increasing exponentially based on the rate of adoption. The web content management Infrastructure has become more and more easy and executable with multiple releases.

However, a single definition of web content management and a set of risks associated with it cannot be rightly identified or limitedly allocated. For eg. Considering that the organizations web site has completely got corrupted, there could not be one or a series of definite reasons for the failure. There could be a specific reason deployed in multiple ways, for eg: to steal the confidential data, the cybercriminal can deviate the organizations focus from security issues by making the site slow, or not functional. As a result, automated scripts must be deployed at each trigger point in order to correctly determine the cause of the website's crash. The main focus of this study is on deploying systems to avoid such occurrences [4].

In the traditional methodology, static data was stored in the local storage, devices and would have limited access to external parties, this had various limitations such as long data access time, costly, and prone to data loss due to device corruption. In the current state, the dynamic operation of web content management is included which on the fly works on assembling the databases from the backend and third party application servers, holds and executes programs written in multiple scripting programming languages and the functionality is distributed across the globe [5]. The number of end-users has increased drastically and the content is evolving every transaction. Therefore, the traditional approach is in the transition phase of businesses revolution and not just a place of data repository for extraction, storage and usage. A link that is connected to the web content management can be easily accessed and can be broken by an external party that can hugely impact the bottom line of the organization. The website must be tracked and kept up to date without sleeping time, as the users around the globe given in scope can access the website at any given point of time from any device, therefore, IoT devices that are used to access information can also lead to multiple holes causing threat to the confidentiality of the data. Therefore, higher access control management should approve changes, access devices before they post the site assuring needed tolerance control is implemented. Due to the vast expansion of the web content access, it become more complex and critical to have automated tools involved as a part of web content management system [6].

The goal of this project is to cover the data flow cycle and data protection cycle for the university database that contains student-specific important information work. This section provides an overview of WCMs and how they are becoming increasingly important for businesses, organizations, and government bodies. It begins with an overview of the history of WCMs and moves on to describe the many types, characteristics, and advantages and downsides of WCMS. Organizations may easily create and manage digital material on their website using a web content management system (WCMS), which eliminates the need for in-depth knowledge of markup or web programming in the creation and maintenance processes involved. In addition to HTML pages, photographs, and other types of media, the system manages a dynamic collection of web content. Using this tool allows to keep track of the documents, audit them, change them, and manage the timeline. Therefore, web content management can be valuable in the corporation, providing insights for decision-making and delivering outcomes, as well as value, by properly managing web content [7].

1. **History of WCMS:** Consider the mid-1990s, when the World Wide Web became increasingly popular and the requirement for regular updates on websites grew as a result of this rising popularity. This represents a departure from the early days, when the information was static and pamphlet- like [10]. Several FileNet WCMS systems, notably Vignette's Story Builder and Documentum, have been launched as a result of this. All of these goods are closed-source private products that were widespread at the time of their development. Open source software is evolving along the same trend line as content management systems, which is driven by the way information is consumed on the Internet. New technologies are emerging to fulfil new needs, as is the case with the evolution of content management systems. In reality, there appears to be a mutual reliance between content management systems, the World Wide Web, and open source [8].

   End-Users are not likely to see a reduction in the requirement to manage an ever-increasing amount of content in the near future. There are a variety of reasons to expect an increase in the use of open source software in the near future. However, in the early 2000s, open source content management systems (CMS) such as WordPress, Drupal, and Joomla [9] began to arise. A plugin architecture that allows users to develop websites without having to know HTML or CSS is built into WordPress. These templates are available through an extensible plugin architecture. WordPress content management system software is deployed on a web server and is usually used in conjunction with a MySQL or Maria DB database (of course, both open source).

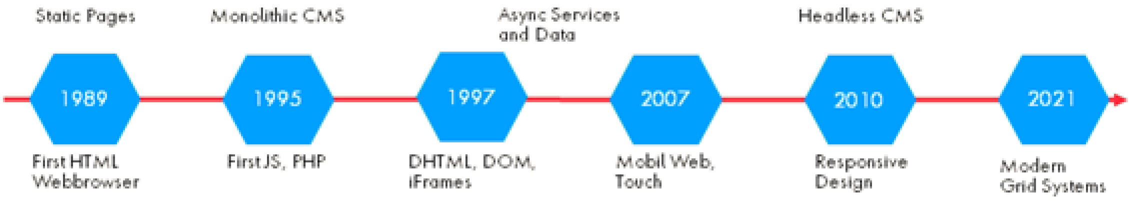

Figure 1: Evolution of Web Content Management

System vulnerabilities are faults in the design and implementation of a system; the WCMS is not an exception. Bugs and errors are both types of software difficulties that exhibit themselves at different levels of the software system hierarchy (implementation and design). In any event, they are extremely harmful and should be thoroughly studied. To put it another way, an attack comprises two phases. To put it another way, more time-consuming detection and exploitation are required immediately after the discovery of a WCMS vulnerability. Secure applications provide assurances in the areas of authentication, confidentiality, integrity, and availability. Depending on the service the concentration is on some of these security features (such as confidentiality). However, web content management systems (WCMS) are a vital component of the Internet nowadays and are a favorite target for hackers. WCMS will be rendered inaccessible in the event of a security breach, which could result in negative consequences. The following attack sequences are frequently encountered: sensitive data corruption, data tampering, web server misuse for unlawful activities, and denial of service (DoS) [11].

As a result of the significant role that web content management systems (WCMS) play on the Internet today, as well as the potential security risks that might develop if they are utilized or set incorrectly, this article serves two functions. By first demonstrating how to operate WCMS from an instructional standpoint and then explaining what can accomplish with it based on the expansion of the previous work. WordPress and Joomla! were selected as the most popular and effective web content management systems (WCMSs) for this purpose. Drupal is included as well. WCMS was used to build only 23.6 percent of websites in January 2011, but by January 2018, that proportion had climbed to 48.8 percent. WordPress was used to create 13.1 percent of websites in 2011, while Joomla! was used to create 2.6 percent of websites. Aside from that, Drupal was used by 1.4 percent of the websites surveyed. WordPress was the most widely used content management system in 2018 (29.3 percent), Joomla was the second most often used (3.2 percent), and Drupal was the third most widely used (3.2 percent) (2.3 percent). As a result, as of January 2018, WordPress has a market share of 60 percent, followed by Joomla! which is not far behind. Drupal has a market share of 6.5 percent, whereas WordPress has a market share of 4.6 percent. In the research of, which is compared to the performance of many web content management systems, and in both studies, WordPress, Joomla! In this study, Drupal was found to be the most efficient because it lowered load times and static content while increasing instalments and improving document support [12] [13].

2. **Working Model of WCMS:** When it comes to managing digital information on a website, a web content management system (WCMS) is a platform that allows businesses to manage various types of digital information on their website by creating and managing dynamic content without the need for programming or markup language expertise [15]. By deploying dynamically managed online content, businesses may make educated decisions based on the insights provided by the platform and the platform itself. In order to effectively promote the organization, websites are an absolute must-have marketing platform. Indeed, websites are critical to the success of modern businesses since they serve as an important route for generating inbound attention [16]. Updating the content of a website on a regular basis is an important part of keeping it running smoothly. This is not something that can be handled manually. A web content management system (WCMS) is therefore essential as a result of this. Upon closer inspection of the numerous digital marketing activities, such as emails, social media posts, print advertisements, and other forms of advertisement and promotion, it can observe that they all direct visitors to the company's website [17]. Consequently, it is vital for businesses to maintain their websites in order to have a strong online presence that is effective.

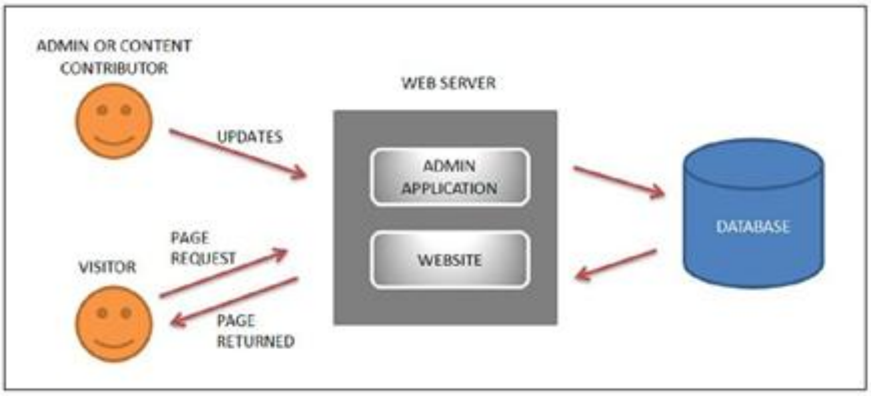

Figure 2: Overall working of Web Content Management System

II. LITERATURE REVIEW

1. **WCMS Lifecycle:** The ability to comprehend Web Content Management (WCM) protocols enables DECISIVE to manage Web Content Management principles. The search statement for the direction of the organization's growth stretches the conventional life cycle from four to six stages or steps, depending on how the organization is structured. After the primary sources have been altered, they can be utilized to change the average of procedures or life cycles, and the actual content of each stage is retained in their life cycles even after the primary sources have been altered. As demonstrated by the arrows, the continuous arrangement of steps that comprise the straight WCM process is depicted in these search sources. Continual deployment and transmission of essential depletion times, as well as the association's internal WCM, are required for the creator's long-term prospects. According to Hong and Lee's Iterative Information Management,

lifecycle information is accessible, created, shared, and used in a consistent manner across the entire process [18]. In the control and organization work process, this phase is concerned with the administration and maintenance of the iteration period for the purpose of content development and organization.

Two iterative phases comprise the life cycle, and each period will continue to be active for as long as the web remains active in the organization. This comprises the gathering and organizing of content, as well as the mediating and dissemination of that content through the internet. When one can write or create new material, are referring to the process of creating new content or getting old content from current sources. If a big number of employees from different departments or from different parts of an organization are active in content development, this is conceivable. Each individual may be given a specific assignment to complete [19]. Modify or evaluate reports in the case of specialized offices or features, as Appropriate. Non-professionals are the majority of customers, and they want a standard interface that is simple to use and understand. While it is possible to incorporate a large number of clients from various geographical regions in the content selection process, the variety of the framework's application is limited when only a small number of clients are present. It is true that program less bespoke programming eventually limits the capacity to employ frameworks to a substantial extent, as the author points out in his article. The process of distributing material to a website, also known as delivery, broadcast, or circulation, can be made more accessible to web clients by removing the content storage component of the website and creating an explicit website page for the purpose of accessing the content. It is included in the package as number [20].

Control and organizational actions are important components of the overall efficacyof the WCM life cycle. Identified and organized customer workloads are identified and organized, and customer clusters with bundle-level capabilities (for example, gathering approvals for promotional materials) are separated from other customers. In addition, activities such as content information storage management, data vault categorization, and system security are performed. This service comprises both on-the-record and off-the-record assessments of the requirements. Workers must be given approval for their work processes, and bottlenecks in the workplace must be eliminated. Providing content and delivery support is the responsibility of the workflow component of the Web Content Management lifecycle. Using events and confirmation forms, which are employed in this work process, this work process encourages collaborative efforts in WCM activities and allows for the development of material to be completed quickly and effectively. Members of this group may be involved in a one-tiered, one-way synchronous venture, and dynamic alerts may be required to advise them of the required activity process. They may also be involved in the elimination of bottlenecks that occur during the distribution cycle. It is probable that a large-scale investigation will be required. Depending on the organization, the workflow can be quite slow. There are no clearly defined duties, verification steps, or expected response periods from clients in this project. To be maintained, work procedures that are casual or obscure must be described and described as a function of the content management framework's implementation, rather than being maintained by the content management framework [22].

Components of WCMS: Many web-based content management systems (WCMS) are built in programming languages such as Java and PHP, and they are hosted on a web server to allow users to manage information on the internet. The web content management system (WCMS) may incorporate additional components in addition to the web server, including a workflow engine, a search Engine, and an email integration module. MySQL (an open source database) and Oracle (a proprietary database) are two examples of data stores or databases used to store web content and data (commercial). It is possible to integrate text or images from earlier publications. The site administrator can also examine older versions of web pages for a given site thathave been kept in a database that has been made available to him. Draught web pages are rarely posted straight to the production web server, in contrast to final products. Instead, the user saves an offline copy of the drafting page until the publication is approved, after which the copy is deleted. Once the final page has been approved and released, the file transfer application is automatically activated, and the final page is posted to the production web server and linked to it [23]. Online content management systems (WCMS) are simply web applications that are backed by back-end databases and feature additional capabilities such as search engine integration and, in some cases, connectivity with translation engines.

2. **Types of Security attacks on WCMS:** Web content management systems are confronted with the following types of security concerns on a daily basis and at varying degrees of resource deployment in the operational ecosystem:

- **Brute force Attack:** Brute force attacks on the WCMS are one of the most popular types of attacks. Attempting to guess a legitimate login and password in order to obtain access to the WCMS administration panel is the goal of this type of attack. A common tactic used byattackers is to take advantage of the user's choice of ill-conceived credentials [25]. Successful brute force websites are typically used to host command and control (C&C) servers, frauds, and drive-by assaults in order to disseminate malware over the internet.

- **DDoS Attack:** Denial of service (DoS) attacks are launched against the web server, causing it to become overwhelmed, slowing it down, and eventually forcing it to crash. A large number of requests are being directed to the server, leading it to become overloaded and preventing the intended user from accessing the service. DDoS (Distributed Denial of Service) attacks are a more advanced form of the DoS assault that is used against websites. In contrast to distributed denial of service (DDoS) attacks, which are launched from a network of computers known as botnets, denial of service (DoS) attacks are launched from a single computer. Make it difficult to determine where the traffic is coming from and increase the amount of requests that are going out at once. Because of this, efforts must be made to build botnet C and C traffic detection technologies, which will be based on data mining and DNS traffic, in order to prevent botnet proliferation and proliferation [26].

- **SQL Injection:** The SQL database backend is used by the vast majority of web content management systems (WCMS) available today. These back-end databases employ application-specific authentication rather than user-level credentials to safeguard critical information. Data breaches occur throughout the database as a result of malicious code being introduced into the web layer via SQL injection. SQL injection attacks, like other types of code injection attacks, have the capability of sending any SQL code directly to the database layer without being detected. The most typical cause of this type of vulnerability is a lack of parameter sanitization, which allows an attacker to send database instructions directly to the database and alter the database directly from the attacker's computer. In spite of the fact that SQL injection has been around for quite some time, it is still one of the most common vulnerabilities in content management systems. The passage of time has resulted in users discovering additional injection sites. For the purpose of preventing SQL injection attacks, it is usual practice to do parameter value sanitization before performing input value processing [27].

- **Directory Traversal:** In order to get access to restricted files, folders, and commands stored outside of the web server's root directory, attackers must traverse the directory structure of the web server. ...../....../ (dot slash) Backtracking is synonymous with attack, as is directory climbing, path traversal, and other names that refer to the same thing. This vulnerability allows an attacker to obtain access to the web server's root directory and other parts of the file system to which the web server has read access as a result of a successful exploit. Using distributed denial of service (DDoS) assaults, attackers can cause a website's service to be unavailable for an extended period of time or indefinitely, resulting in huge financial losses. Accordingly, under this scenario, an adversary would be able to view the restricted files and gain the information necessary to further hack the system [28].

- **File Inclusion Exploits:** Using the file include vulnerability, an attacker can obtain access to a file on the victim's computer. The vulnerable confinement technology, which is implemented in the application, must first be exploited by a malicious attacker in order to read sensitive files, get access to more sensitive information, or execute arbitrary instructions. Hackers can exploit this vulnerability by including malicious files in online applications that are vulnerable to it, allowing the malicious files to be executed by the web application. This has the potential to severely demolish computer system. A vulnerability known as file inclusion vulnerability is frequent on websites that have been written incorrectly. Local file includes (LFI) and remote file include (RFI) vulnerabilities are the two most common forms of file include vulnerabilities, with local file inclusion (LFI) being the most prevalent (RFI). A vulnerability in LFI allows an attacker to read and possibly execute a file on the victim's computer as a result of a security flaw. If a web server is not properly configured or is running with elevated privileges, an attacker may be able to get access to sensitive information on a computer. In order to obtain access to a configuration file on the server, an attacker would need to take use of this vulnerability. An attacker, on the other hand, may use his machine to execute RFI code instead of reading a file on a local web server to commit his assault [29].

3. **Advanced tech for WCMS:** For many years, beginning with the introduction of the first document management systems in 1990, one of the most difficult aspects of managing web content was determining the meaning of unstructured information, which was regarded as one of the organization's most valuable assets. Determining the meaning of

unstructured information was regarded as one of the most difficult aspects of managing web content. For example, among their current objectives are the automatic understanding and comprehension of unstructured text, as well as the transformation of such information into structured data that can be connected to multiple systems and processes. Data integrity and security are two further issues that must be addressed through the application of current technologies. It is discussed in this section how to deal with challenges such as unstructured, meaningless data storage, how to generate meaning and structure data issues, and how to construct and extract complex access control and trace back functions from an immutable record system.

Encryption Techniques

Table 1: Features of different encryption algorithms [32]

Features	DES	Triple DES	RSA
Key Used	Same key is used for encryption and decryption	Same key is used for encryption and decryption	Different key is used for encryption and decryption
Scalability	Scalable due to varying key size	Scalable due to varying key size and block size	Not scalable
Confidentiality	Low	Medium	High

Cryptography is the process of transforming plaintext (normal text or plain text) into cipher text, and then back to plaintext (normal text or plaintext) again (called decryption). Primary role is to perform encryption and decryption procedures utilizing two unique keys, a public key and a private key, both of which are known only to the parties engaged in the transaction, in order to protect sensitive information. It is not possible to extract the private key from the public key. The use of public keys significantly increases the security of cryptographic systems. Although the overall method differs between symmetric and asymmetric encryption, this is primarily owing to the significant differences between the two. Although this is a slight distinction, it is large enough to have an impact on overall security. However, because it requires private maintenance, symmetric encryption is considered to be less safe than asymmetric encryption. Symmetric encryption is also considered to be more efficient, lighter, and better suited for applications that transmit a significant volume of data. It is commonly recognized that they are more exposed to a broader number of attack vectors than other types of organizations.

The Risvest, Shamir, and Adleman (RSA) algorithm was invented in 1977 and is an asymmetric encryption technique. Two keys are generated, one for use in encrypting the communication and another for use in decrypting the message [31]. RSA's algorithm is broken into three stages, each of which has three pieces. A key is generated in the first stage, and this key will be used to encrypt and decrypt data in the following steps. It is necessary to turn plaintext into cipher text in order to proceed to the third stage, which is to decrypt the cipher text. The first phase, often known as the key generation phase, is the most important. Using an encryption method, it is turned into cipher text, which is then converted back into plaintext for further processing. One of the most common types of encryption, symmetric encryption is used for both encryption and decryption, with a single key being used for both operations. DES (Data Encryption Standard) and Advanced Encryption Standard (AES) are the two methods of encryption that are available (AES). Asymmetric algorithms encrypt and decrypt data using a multitude of keys in order to provide maximum security. The RSA cryptographic method is widely used in electronic commerce protocols, and it is also utilized in many other applications. It is deemed secure if the key is sufficiently long and if the currently available implementation is used. The RSA algorithm is commonly regarded as being absolutely safe.

A hash function is simply a function that accepts an input value and generates an output value that is deterministic of the input value based on the input value. When the hash function is run, the outcome is always the same and consists of the same y value regardless of the x value supplied. A hash function is anything that takes an input (which can be any data such as integers, files, etc.) and returns a hash. Numbers and files can be used as input. A hash value is frequently expressed as a hexadecimal number. MD and SHA are the most often used hash functions, but there are other additional hash algorithms. Because hash functions are often irreversible, sometimes known as "one-way," it is impossible to determine the input if only the output is known – unless every conceivable input is tried (which is called a brute-force attack). Hash functions have many applications, but one of the most common is validating the integrity of data. It is used in the generation of checksums for data files. Through the usage of this tool, the user gains trust in the accuracy of the data. Using a digital signature is one way to ensure that the information contained in a message has not been altered while the message is being transmitted. When a document is digitally signed by the sender, the server, it also encrypts the message content with a one-way hash using the sender's public and private keys. This keeps the message content secure during transmission. However, the process will produce a "signature" that can only be decrypted with the server's public key. Organization’s client will still be able to read it. By utilizing the server's public key, the client is thus able to authenticate not just the authenticity of the message's sender but also the completeness and accuracy of the message's contents. The following are the major aspects of the algorithm that will assist organizations in determining how to use it:

III. ARTIFICIAL INTELLIGENCE

Artificial intelligence is the ability to replicate human-like abilities in a variety of fields, including as reading, learning, memory, and reasoning, and it is used to describe a wide range of technologies. Automating even the tiniest of repetitive tasks can have a significant positive influence on both organizations and the lives of their employees. Using artificial intelligence in the disciplines of data management and security, this paper explores the potential applications of this technology. For a computer, software, or methodology to be considered intelligent, it must possess the following fundamental capabilities: It is the study and application of natural language that is known as Natural Language Processing (NLP). The use of these strategies in conjunction with other technologies can be beneficial in overcoming any difficulties that may arise in the process of teaching a computer to understand human-created languages. Knowledge representation (KR) is a technique for representing data in a structured manner. It does this in order to keep what it learns or hears safe and secure.

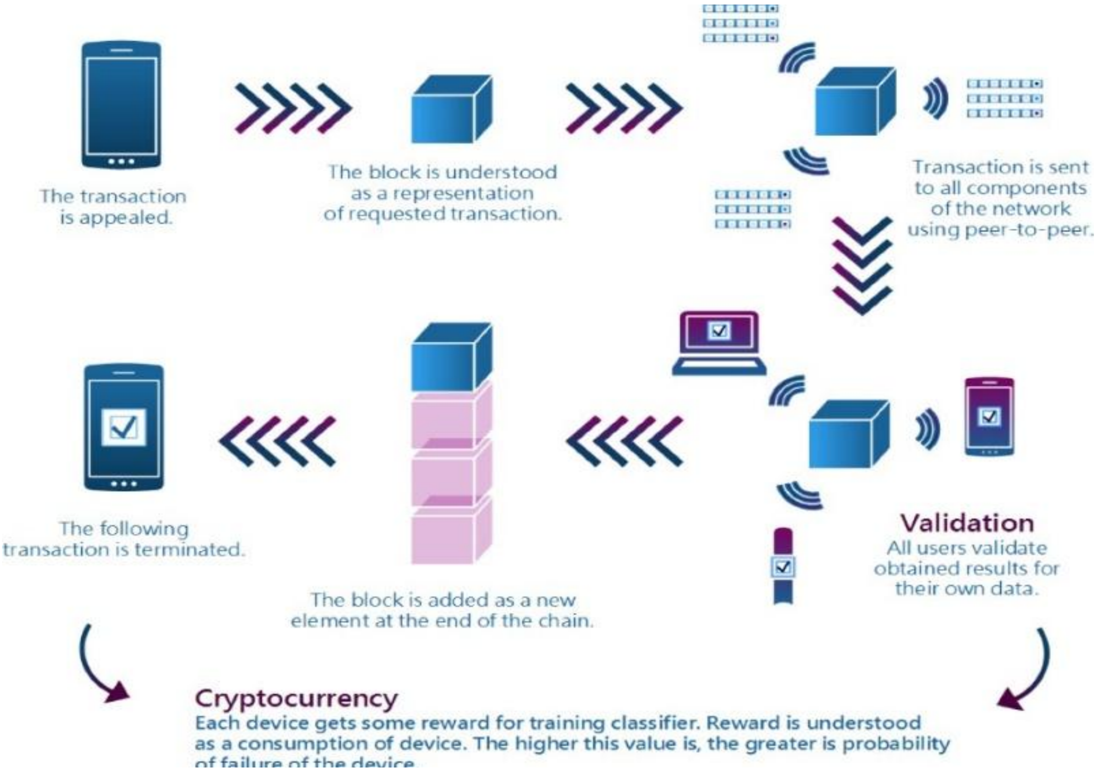
The transaction
is appealed.
The block is understood
as a representation
of requested transaction.
Transaction is sent
to all components
of the network
using peer-to-peer.
The following
transaction is terminated.
The block is added as a new
element at the end of the chain.
Validation
All users validate
obtained results for
their own data.
Cryptocurrency
Each device gets some reward for training classifier. Reward is understood
as a consumption of device. The higher this value is, the greater is probability
of failure of the device

Figure 2: Overview on the working of blockchain using cryptocurrency transaction and block creation

1. **Benefits of Artificial Intelligence:** Some of the benefits / advantages of utilizing artificial intelligence to maintain and secure academic records are as follows: Using artificial intelligence to maintain and secure academic records has the following advantages:

   - The ability to learn how to correct human errors or defects that are produced by system design or implementation.
   - Making massive amounts of data manageable through the use of a variety of artificial intelligence approaches is becoming increasingly popular.
   - Increased effectiveness in recognizing threats and exploiting weaknesses, among other things.
   - Utilizing logic and reasoning skills to reduce unnecessary storage data from the system.
   - The design and management of the complete security system, which makes use of models that have been efficiently trained on previously collected data.

2. **Block Chain:** Following the debut of blockchain technology and the development of the cryptocurrency bitcoin, blockchain technology has emerged as a disruptive force in the financial industry. Due to its decentralized structure, as well as its ability to store immutable and secure ledger data, blockchain technology is used as the backend technology for cryptocurrencies [33]. Blockchain technology, also known as distributed ledger technology, is becoming increasingly popular. Blockchain is a relatively new technology with enormous potential for use in a wide range of businesses and

organizations. It is often viewed as little more than a backend technology for Bitcoin (digital currency). However, the Blockchain is much more than that: This technology, a one-of-a-kind answer to the age-old human problem of trust, will profoundly change the way online transactions are conducted by determining whether or not two unknown persons can be trusted. Furthermore, they ensure immutability, which means that information input cannot be changed; disintermediation, which means that no third party is necessary to validate transactions; and, last, they result in cheaper costs because third party fees are reduced. When these and other benefits are properly utilized, they have the potential to cause disruptive changes, encouraging the creation and adoption of a wide range of applications.

Data security, safe access control, and backtracking are just a few of the applications for bitcoin and blockchain technology that have been demonstrated in a range of fields. Blockchain are divided into four types: public blockchain, private blockchain, permission- based block chains, and permission less blockchain. Public blockchain are those that are accessible to the public, while private blockchain are those that are only accessible to the private blockchain community. The public blockchain is the most commonly seen type of blockchain. The blockchain is an open platform for storing information that is both free and safe. The blockchain technology is still a relative newcomer to the market, especially when compared to other notable technologies. In situations when there is a requirement for hidden data security and access control, blockchain technology may be a feasible option [34]. Blockchain as a Service (BaaS) is the most essential component of blockchain in the market right now (BaaS). The exponential growth of data has created an ever-increasing demand for adequate security as well as a reliance on it over the long run. It is possible for corporations and businesses to benefit from Blockchain as a Service, which is cloud-based collaboration that gives a solution for running security-based procedures.

Benefits of Blockchain Technology: Despite the fact that Blockchain is currently considered an emerging technology [35], there is still room for advancement and research. Its characteristics, on the other hand, may be clearly identified and debated as a result of its sturdy and specified technological basis. In their most recent work, searcher and his colleague investigated and discovered key components of Blockchain technology, which was organized and submitted for peer review [36]. Blockchain technology provided some advantages over older systems. However, among the attributes stated, trust and decentralization stand out the most. Nonetheless, these two fundamental characteristics permit the generation of other significant benefits. Blockchain technology has a number of advantages, especially in this day and age of digital technology, which is particularly advantageous.

- For starters, blockchain eliminates the requirement for third-party interaction, hence eliminating the risks associated with third-party participation. A frequent citation of this as one of the most major advantages of blockchain technology is found.

- For the second time, because a copy of every record is stored in an unknowable location that can be traced back to the block's hash, blockchain records/contracts are impossible to delete. The block hash is used to facilitate transactional backtracking in this manner.

The bigger the number of records that can be saved and connected together, the greater the level of trust that can be placed in a particular transaction.

- Single public ledger – By utilizing blockchain technology, the problems associated with a centralized tiered system can be avoided.

- Distributed ledger – Blockchain connects every record to a shared public ledger, avoiding the need for a third-party mediator in the transactional process. An exchange of commercial information is taking place between two individuals. The building of trust occurs when a transaction is locked in a block due to the fact that this information is inaccessible to anyone, whether in the private or public sector.

- Reliability: Mining, validating, and confirming a transaction are the three steps necessary before adding it to the public ledger. After that, it is copied across the network using the decentralization storage technique of the blockchain, also known as distributed ledgers, which makes the data more robust and resistant to power failures, data loss, and other types of loss. The dependability of blockchain systems is ensured by the fact that transactions are not added to the ledger until after they have been mined, inspected, and confirmed. [37]

- Privacy: It is a benefit as well as a challenge at the same time. On the plus side, participant identities are concealed which makes it possible for participants who take part in the research to keep a high level of confidentiality throughout the process. In addition, the public-key encryption method is used for any and all communications that take place between network nodes. Whereas in the systems that are in place now, every single participant's identity as well as the particulars of their transactions are kept and managed by third parties. [38]

- Transparency: Blockchain technology provides complete transparency to everyone on the network because all transactions are accessible to all computers connected to the network and are not subject to the authority of a third party. However, before any changes to the Blockchain can be done, the majority of these machines must provide their permission. This prevents the potential of transactions being tampered with or disguised. Blockchain-based systems offer considerable transparency advantages over traditional centralized ledgers. Because changes to the ledger are accessible to everyone on the network, and transactions cannot be altered or erased once they are recorded on the blockchain, the amount of friction experienced during the transaction is reduced. [39]

- Versatility: Blockchain technology has far-reaching ramifications and may function as a public record for any form of transaction, not simply those involving digital currency. The Blockchain concept was initially designed for use with the digital currency Bitcoin. Despite the fact that the Blockchain technology was initially developed for use with bitcoin, its scope of use has now significantly expanded. For instance, numerous academics and industry professionals believe that this technology will have a significant impact on the fields of medicine, the music industry, and the energy industry. As a result of the fact that this technology is based on the same fundamental notion, it will have an effect on all of the other industries. One of the many benefits that Blockchain possesses is adaptability, which is demonstrated by the fact that it can be utilized in a wide variety of contexts. [40]

- Immutability and data integrity: The integrity of the data is one of the most important requirements that must be fulfilled, and it is also one of the most important components of the Blockchain technology. The implementation of a challenging cryptography-based procedure known as the consensus mechanism is required in order to guarantee the legitimacy of the data.

IV. OVERVIEW OF THE PROPOSED METHODOLY

The methodology of this research deems it necessary to complete a variety of tasks in a systematic and repetitive manner. In the subject of artificial intelligence, a thorough investigation and review of the literature reveals that the Artificial Intelligence domain is divided into several subdomains, the most prominent of which are Natural Language Processing (NLP), Machine Learning (ML), and Machine Vision, among others is combatively used, and the encryption method, as well as blockchain technology, have been employed to ensure high-end security

In this work, multiple technologies have been used to address issues identified at every aspect of web content management system. This work is an objective oriented development which develops an End2End sensitive data development in the university. The Figure below present an overview of the overall methodology and the expected outcomes. The objective of this work is to deploy a highly automated data management system with high-end access control and tracking. The dataset in most of the educational universities in India is unstructured, involving human intervention, will risk the authenticity of data by internal malicious access and human errors. Therefore, step by step design is proposed to cancel out all the risks. This section presents an overview on the technologies used, the application of the technologies for objective specific Student data protection and to have high-end tracking to ensure no internal malicious access is executed and as the end result have an alarming system if there is a malicious behavior seen.

This study is concerned with a web-based content management system (WCMS) that transacts and saves confidential / sensitive data. Security is required in every part of the WCMS in order for the storage system to function properly. When it comes to any application, security is a concern. One of the most important steps that many organizations overlook when designing operational apps and deploying them on the internet is doing a risk assessment of the areas that could be vulnerable to security breaches. This research is focused on the organizational aspect of a WCMS, which serves as its foundation. A carefully designed security implementation internal to the operating organization can make the WCMS immune to external-party attempts at malicious access while also reducing the attempts at malicious access by internal-party employees.).

As seen in the image, academic data request and response models are being changed. As illustrated in the image, new and improved procedures are being implemented. Many Indian state governments continue to employ the methods mentioned in (a), and a few have also implemented additional access mechanisms, such as those outlined in (b) (b). However, additional adjustments, analyses, and organizational structures must still be implemented. This section will examine the problems of using the manual process, the issues to be considered when moving from the manual to the digital processes, the flexibility of the digital process, and the obstacles associated with it.

Figure (a) displays the traditional data flow that was followed until 1995, and more crucially, it was in the late 1990s that the educational system began to adopt and comprehend the internet. Many people have reported that their academic records, which were thought to be the most important documents for them, had been misplaced or were otherwise unrecoverable. Several methods were proposed to address this issue, but the true solution was identified in 2006, when Google developed network-based storage, the first to be commercially viable. Many firms were working on network-based storage, which is now known as cloud storage / computing. Many businesses, including education, were working on incorporating cloud storage / computing.

A number of critical concerns were recognized when internet communication and network- based storages revolutionized the way academic documents were saved, accessed, and retrieved. The majority of the difficulties were raised by those involved in the education / academic data flow. The definitions of these entities, which are based on the present methodology of the Indian educational system, are as follows:

The implementation closes with the creation of a confidential data tracker. A major benefit of the overall system implementation was that it enabled the system's super administrator to track and control authorized user activity, as well as to ensure data immutability based on the permissions granted to users. The usage of IPFS has resulted in a reduction in the overall costs of the system. Transaction storage costs have been decreased thanks to efficient data storage in the IPFS cloud and the use of a Hash ID produced by an Ethereum smart contract in the Meta mask account. Given below is the end to end flow of when the data flows into the system, the data acceptance can be for both structured and unstructured. In the subsection, the method of choice and implementation is explained. The methodology is divided into 3 major aspects: Intelligent data classifier, protective storage with high-end with encryption & blockchain on IPFS cloud and access tracking.

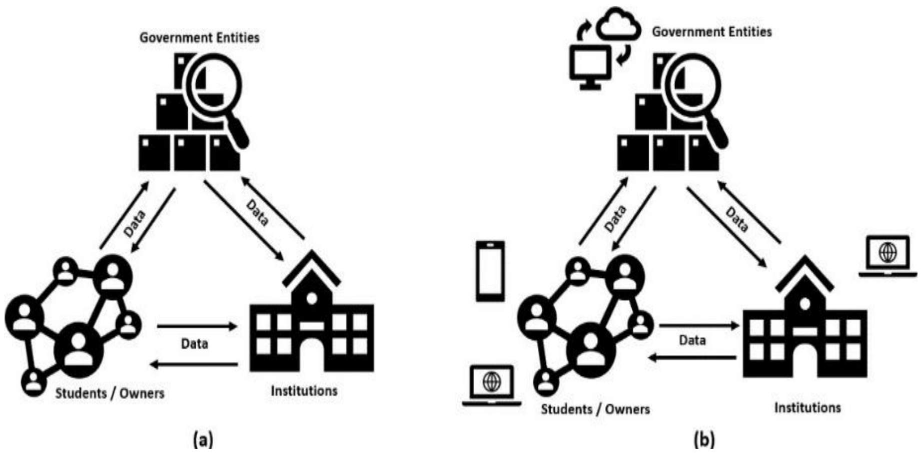

Figure 3: Organization interaction ecosystem

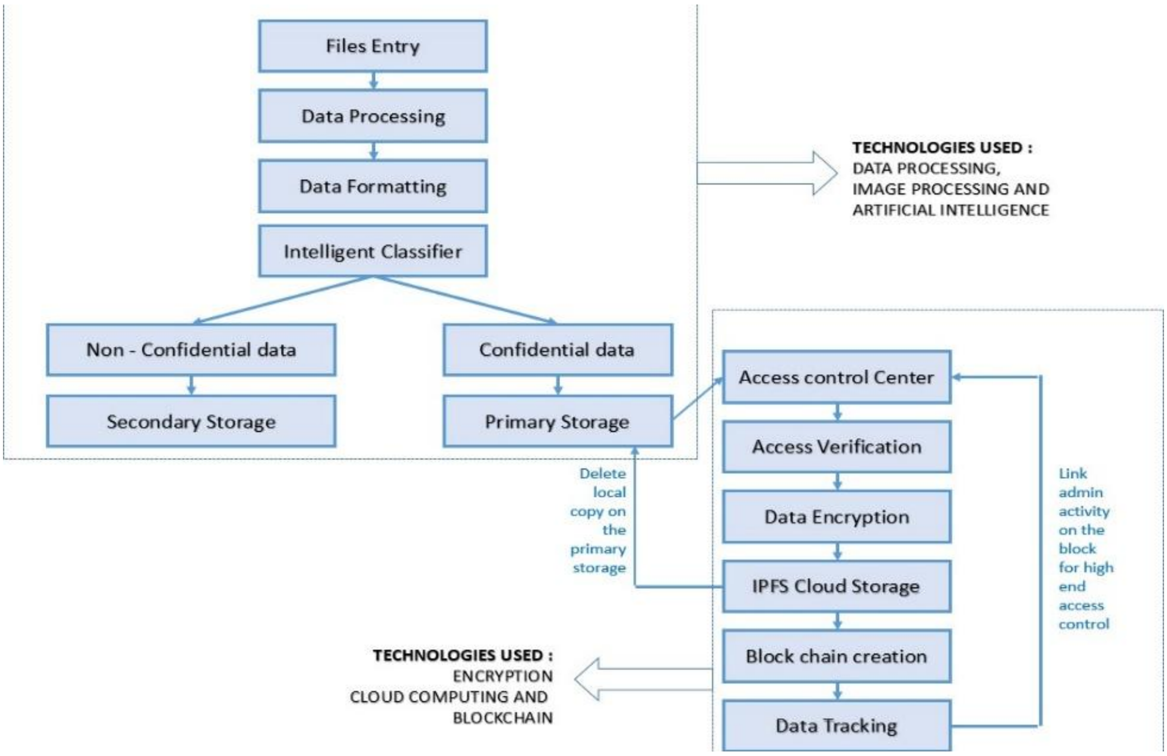

Figure 3: Overall Methodology Flow chart

V. CONCLUSION

A large increase in online user activity has occurred as a result of the accessibility of the internet, resulting in a massive amount of user-produced material in the form of structured, semi-structured, and unstructured data being generated. More information on a specific product, brand, or event can be obtained by viewing this type of content (or anything else for that matter). This has resulted in an increased level of interest in the safety of the WCMS by organizations in both the corporate and scientific sectors over the last decade. This is advantageous to both enterprises and service providers on an equal footing. Customers may find it easier to make decisions when using premium security, while service providers may find it easier to improve their offerings when using premium security. In order to achieve this goal, it is important to conduct research and development in the fields of artificial intelligence-assisted automation and security implementation using Ethereum smart contracts.

A WCMS security and job automation strategy has been effectively implemented across a wide range of industries and has seen significant improvement over time; nevertheless, there is still much more work and exploration to be done and explored in terms of detailed and fine-grained approaches.

Using cutting-edge technology implementations and application analytic methodologies, the book suggests end-to-end data transfer within the source system, tracking internal users, and alerting the WCMS service owner in the event of fraudulent activity. The book also addresses task automation (to organize data) and top-tier organizational security, which are both addressed in the book. It is necessary to employ real-world university library datasets acquired from a range of university databases to validate the models proposed in this book. Future scholars and legislators will be able to quickly implement the remedies that have been proposed because they yield promising results.

The overall conclusion derived from this work is the following:

- Data breaches can cause huge damage in terms of time, resources, money etc. for an organization or businesses.
- The current situation of data security is getting affected by multiple parameters, such as data accumulation and manual sorting which is infusing a gap into the system for malicious access, encryption alone cannot handle huge threats, cloud computing can serve as a hotbed for multiple data breaches.
- To ensure all the issues are taken care, this end to end framework is proposed.
- The first step of the framework deals with automatically sorting the confidential data and non-confidential data. Since the area of interest of this study is to secure the confidential data, the other data which is labelled non-confidential is de-prioritized and the primary resource is not used to store such data, this assures optimal resource usage and reduce in time consumption.
- Confidential and non-confidential data classifier in this work is implemented on 10th and 12th marks card which serve as a critical document in every individual's career. The documents other than that is classified as non-confidential, i.e. this class of data does not require security.
- The automatic classifier is designed to learn from the image by extracting text line-by-line and associating the pattern between the confidential class a non-confidential class.
- For the confidential class, RSA algorithm is applied at the local host, store the encrypted document on the IPFS cloud service and store immutable info on the dataset in the Ethereum blockchain smart contract.

REFERENCES

[1] Maican, Catalin, and Radu Lixandroiu. "A System Architecture Based on Open Source Enterprise Content Management Systems for Supporting Educational Institutions." *International Journal of Information Management*, vol. 36, no. 2, Apr. 2016, pp. 207–214.

[2] Martinez-Caro, Jose-Manuel, et al. "A Comparative Study of Web Content Management Systems." *Information*, vol. 9, no. 2, 27 Jan. 2018, p. 27.

[3] Viduka, Dejan, et al. "Analyzing the Potential Mechanism for Measurements - the Most Popular Open Source Web Content Management System." *Proceedings of the International Scientific Conference - Sinteza 2017*, 2017.

[4] Attanayake, D.N.P, and R.G.S. Thilanka. "Design and Implementation of Web-Based Management System instead of Manual Process Efficiently in ATI, Galle." *International Journal of Scientific and Research Publications (IJSRP)*, vol. 11, no. 1, 6 Dec. 2020.

[5] Berdik, David, et al. "A Survey on Blockchain for Information Systems Management and Security." *Information Processing & Management*, Jan. 2021, p. 102397.

[6] Robles, Rocio Garcia, et al. "Accessibility via Metadata in a Semantic Web-Driven Content Management System." *International Journal of Metadata, Semantics and Ontologies*, vol. 1, no. 3, 2006, p. 224.

[7] Kashmar, Nadine, et al. "From Access Control Models to Access Control Metamodels: A Survey." *Lecture Notes in Networks and Systems*, 2 Feb. 2019.

[8] Xu, Qiaoyun, et al. "A Visually Secure Asymmetric Image Encryption Scheme Based on RSA Algorithm and Hyperchaotic Map." *Physica Scripta*, vol. 95, no. 3, 7 Feb. 2020.

[9] Wei, Xin, et al. "Secure Data Sharing: Blockchain Enabled Data Access Control Framework for IoT." *IEEE Internet of Things Journal*, 2021, pp. 1–1.

[10] Ouaddah, Aafaf, et al. “FairAccess: A New Blockchain-Based Access Control Framework for the Internet of Things.” *Security and Communication Networks*, vol. 9, no. 18, Dec. 2016, pp. 5943–5964.

[11] Varfolomeev, Alexander A. “analysis of the change of the concept “financial cryptography” on the basis of 20 years subjects of the international conference “financial cryptography and data security.”” *statistics and economics*, no. 4, 1 jan. 2016, pp. 12–15.

[12] Gajmal, Yogesh M, and R. Udayakumar. “Blockchain-Based Access Control and Data Sharing Mechanism in Cloud Decentralized Storage System.” *Journal of Web Engineering*, 30 Aug. 2021.

[13] Rana, Saurabh, and Dheerendra Mishra. “Efficient and Secure Attribute Based Access Control Architecture for Smart Healthcare.” *Journal of Medical Systems*, vol. 44, no. 5, 30 Mar. 2020.

[14] Andziulienė, Beatričė, and Povilas Narbutas. “Open Source Web Content Management Systems Analysis: The Use of Server Resources.” *Lietuvos Matematikos Rinkinys*, vol. 51, 21 Dec. 2010.

[15] Kashmar, Nadine, et al. “A Review of Access Control Metamodels.” *Procedia Computer Science*, vol. 184, 2021.

[16] Kumar, Naveen, et al. “Security Attacks in Named Data Networking: A Review and Research Directions.” *Journal of Computer Science and Technology*, vol. 34, no. 6, Nov. 2019.

[17] Arslan, Suayb S., and Turguy Goker. “Compress-Store on Blockchain: A Decentralized Data Processing and Immutable Storage for Multimedia Streaming.” *Cluster Computing*, 25 Mar. 2022.

[18] Li Li-Xiang, et al. “Parameter Estimation for Lorenz Chaotic Systems Based on Chaotic Ant Swarm Algorithm.” *Acta Physica Sinica*, vol. 56, no. 1, 2007.

[19] Dai, Mingjun, et al. “A Low Storage Room Requirement Framework for Distributed Ledger in Blockchain.” *IEEE Access*, vol. 6, 2018.

[20] Hammi, Mohamed Tahar, et al. “Bubbles of Trust: A Decentralized Blockchain-Based Authentication System for IoT.” *Computers & Security*, vol. 78, Sept. 2018. Accessed 2 Dec. 2019.

[21] Meike, Michael, et al. “Security in Open Source Web Content Management Systems.” *IEEE Security & Privacy Magazine,* vol. 7, no. 4, July 2009.

[22] Bodkhe, Umesh, et al. “Blockchain for Industry 4.0: A Comprehensive Review.” *IEEE Access*, vol. 8, 2020.

[23] Shamshad, Salman, et al. “A Secure Blockchain-Based E-Health Records Storage and Sharing Scheme.” *Journal of Information Security and Applications*, vol. 55, Dec. 2020.

[24] Wang, Zeli, et al. “Ethereum Smart Contract Security Research: Survey and Future Research Opportunities.” *Frontiers of Computer Science*, vol. 15, no. 2, 2 Oct. 2020.

[25] Wang, Yilei, et al. “Randomness Invalidates Criminal Smart Contracts.” *Information Sciences*, vol. 477, Mar. 2019.

[26] Zhu, Lie-Huang, et al. “Data Security and Privacy in Bitcoin System: A Survey.” *Journal of Computer Science and Technology*, vol. 35, no. 4, July 2020.

[27] Zheng, Bao-Kun, et al. “Scalable and Privacy-Preserving Data Sharing Based on Blockchain.” *Journal of Computer Science and Technology*, vol. 33, no. 3, May 2018.

[28] Augustus Devarajan, A, and T SudalaiMuthu. “Cloud Storage Monitoring System Analyzing through File Access Pattern.” *2019 International Conference on Computational Intelligence in Data Science (ICCIDS)*, Feb. 2019.

[29] Mohamed, Tamara Saad. “Analytical View of Web Security and Sophisticated Ways to Improve Web Security.” *Journal of Physics: Conference Series*, vol. 1530, May 2020.

[30] Thakur, Anusha. “A Comprehensive Study of the Trends and Analysis of Distributed Ledger Technology and Blockchain Technology in the Healthcare Industry.” *Frontiersin Blockchain*, vol. 5, 3 Mar. 2022.

[31] Sukhodolskiy, Ilya, and Sergey Zapechnikov. “A Blockchain-Based Access Control System for Cloud Storage.” *2018 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, Jan. 2018.

[32] Ferrag, Mohamed Amine, et al. “Security and Privacy for Green IoT-Based Agriculture: Review, Blockchain Solutions, and Challenges.” *IEEE Access*, 13 Feb. 2020.

[33] Cheng, Xu, et al. “Design of a Secure Medical Data Sharing Scheme Based on Blockchain.” *Journal of Medical Systems*, vol. 44, no. 2, 8 Jan. 2020.

[34] Casino, Fran, et al. “A Systematic Literature Review of Blockchain-Based Applications: Current Status, Classification and Open Issues.” *Telematics and Informatics*, vol. 36, no. 36, Mar. 2019.

[35] Zheng, Qiuhong, et al. “An Innovative IPFS-Based Storage Model for Blockchain.” *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, Dec. 2018.

[36] Jadhav, Maruti R., et al. “Impact Assessment of Outcome Based Approach in Engineering Education in India.” *Procedia Computer Science*, vol. 172, 1 Jan. 2020.

[37] Shukla, Shubhendu. “Management Education in India Issues and Concerns.” *International Journal of Education and Learning*, vol. 2, no. 2, 30 Sept. 2013.

[38] Farber, Julie, and Sara Munson. “Strengthening the Child Welfare Workforce: Lessons from Litigation.” *Journal of Public Child Welfare*, vol. 4, no. 2, 4 June 2010.

[39] Jiangbo Shu, et al. “Exploration on College Education Big Data Open Service Platform.” *2017 IEEE 2nd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, Apr. 2017.

[40] Chandra, Deka Ganesh, and Dutta Borah Malaya. “Role of Cloud Computing in Education.” *IEEE Xplore*, 1 Mar. 2012.

[41] Bhatiasevi, Veera, and Michael Naglis. “Investigating the Structural Relationship for the Determinants of Cloud Computing Adoption in Education.” *Education and Information Technologies*, vol. 21, no. 5, 25 Jan. 2015.

[42] Foster, Derek, et al. “Cloud Computing: Developing Contemporary Computer Science Curriculum for a Cloud-First Future.” *Proceedings of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*, 2 July 2018.