Journal of Theoretical and Applied Information Technology
15th July 2022. Vol.100. No 13

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

END2END UNSTRUCTURED DATA PROCESSING, CONFIDENTIAL DATA STRUCTURING & STORAGE USING IMAGE PROCESSING, NLP, MACHINE LEARNING, AND BLOCKCHAIN

[1]MADHURA K, [2]MAHALAKSHMI R.

[1]Assistant Professor, Presidency University, Department of CSE, Bangalore, India.

[2]Professor, Presidency University, Department of CSE, Bangalore, India.

E-mail: madhurakban2022@gmail.com

ABSTRACT

The expediting magnification of automating the manual jobs into automated is incrementing day by day, as there are approximately 2.5 quintillion bytes of data exchanged over the cyber world per day. With the incrementing need for process automation and immensely colossal unstructured data, there is increasing demand for incorporating automated objectives-specific classifiers for businesses. To make better and improvised automated end-to-end solutions, data structuring utilizing advanced technologies such as ML, Big data processing, data science, etc. will avail in abbreviating the resource consumption extracting better data semantics, handle multiple parallel requests which result in high-end organized automated solutions with efficient data processing. This paper demonstrates an objective-specific classifier will accommodate as a commencement point in automating any process. In this paper confidential data processing is demonstrated on confidential data of students, dataset contains unstructured data from the university library which will then be structured into confidential data and non-confidential data automatically. Image processing is utilized to extract features and ML algorithms are acclimated to train the classifier. This intelligent classifier can further be used along with encryption methodology to protect and store confidential data. The article is organized as an introduction section, literature review, methodology section, result-implementation details, and last section conclusion. The introduction section introduces the importance of artificial intelligence in the field of the education system, literature review section covers the background work carried out on artificial intelligence in the field of education, methodology section covers the proposed method of applying machine learning algorithm to perform the automatic classification of documents in the education system, result-implementation section shows the result analysis from different machine learning algorithms and end the conclusion section provides the summary of overall work.

Keywords: *Artificial Intelligence, Data Management, Meta-Data Management, Machine Learning, NLP, Data Protection, Data Science, Unstructured Data.*

1. INTRODUCTION

The introduction section of the article provides a deep knowledge on the artificial intelligence, classification of documents using AI, importance of AI in education system and use of AI in automatic classification of documents as confidential and non-confidential.

The education system in India has visually perceived consequential progress in recent years, with more than 1.6 million schools and more than 290 million enrollments, India stands as one of the most astronomically immense and intricate edification systems in the world along with China [1]. While there is positive magnification in education incorporation and adaptability, there are more and more data accumulated in the education ecosystem [2]. There is an incrementing desideratum for the edifying system to acclimate advanced technologies such as cloud storage, online data transfer, data science, artificial intelligence, cybersecurity, etc. [3]. Making utilization of all the technologies to

solve the objective of safe data transfer, storage access has been the focus of researchers, companies, and educational groups [4]. There are raising concerns about having secure data storage and accessibility.

Traditionally, the education system stored data on local servers and save peregrinate to the cloud platforms to store the data [5]. This has made the data more accessible to the user, has become efficient and propitious, however, the data accumulation is growing every day, and the desideratum to have an automated process in place to inflate the data management cycle is critical. The issue that the current data management the in education ecosystem is the storage of substantial amplitude of unstructured data in an inefficient manner, which needs immediate attention [6], different software implements are habituated to manage data incorporates extra cost and open gaps for malignant attacks [7]. Consequently, a methodology that will structure the accumulated data, store the data predicated on priority and securely access the data will address most of the issues with data management in educational institutes or educational entities [8].

With the incrementing dimension to the databases and database application in inculcation management, operation, and maintenance [9], to automate such databases it is compulsory to examine the automatic extraction of cognizance from the heap of storage in the edification administrative databases, to make it opulent and safer for functioning [10]. Improvising the quality of data maintenance of a sizably voluminous database and implementing astute data structuring can abbreviate the intricacy, time, and cost. Erudition revelation to incorporate automation is required in software management, query processing decisions, process control, and many other fields of interest [11]. There are countless benefits of structuring the database, it has multiple aspects required to be resolved to get to the most efficient keenly intellective data management. The most critical challenge in managing structured data, there is a more preponderant propensity to devote more resources to unstructured data. The acquisition of erudition from unstructured data is hard to process and store [12]. The requisite to implement the process to extract cognizance from unstructured days divided. Erudition extraction is a procedure of engendering meaning from structured, unstructured, and semi-structured data [13].

Text extraction from data mining refers to the computing process of finding patterns and relationships extracted from the database/datasets which appears not behave any relationship when it is unstructured. This approach is called data mining which is an interdisciplinary field of data statistics, AI (Artificial Intelligence), and database system to engender incipient implements / automation for discovering more and more pattern for better method of data structuring [14]. Similarly, while dealing with text data, there is a requisite to utilize sundry branch of studies such as statistics, linguistics etc.… There is withal a growing need to hybrid techniques such as text-predicated sentimental management, text processing, feature extraction, feature Selection, and relegation.

The implementation of artificial intelligence and automation to augment confidential and non-confidential data relegation and automation development has been a subject of study in every field of business and organization [15]. The subject is gaining popularity every day, with key influencers being the accessibility of amended methods improvising computational power, and main sizably voluminous dataset [16]. Artificial Intelligence has the faculty to transform operation, ad data management. Data stored on edifying databases can be confidential (Critical documents, financial decisions, students' academic reports, etc.) and non-confidential (memos, newsletters, flyers, etc.). Currently, the data storage and management of such datatypes have evolved into storing different structures and management [17]. There is a desideratum incorporate priority predicated approach on treating confidential and non-confidential data, non-confidential data, when accessed by unauthorized use, does not an immensely colossal loss, and similarly, if the confidential data is accessed by an unauthorized utilizer, multiple aspects get affected. Consequently, erudition revelation and cognizance predict database can be designed to relegate the data into confidential and non-confidential data set and to treat the

functioning, management, and organization predicated on the nature of the dataset.

In this paper, a novel methodology is proposed to have an objective predicated perspicacious classifier which then is provided with advanced security. The objective is to structure the unstructured heap within the university database which will astutely relegate data into confidential and non-confidential the method is initiated by converting different format data into a single data format (image format), this is done to have feature extraction and cull implemented to train the classifier, this classifier will be placed on the first level, when there is an incipient data format inputted to the classifier at the gateway of the database, the data will be relegated into confidential and non-confidential structured as non-confidential will be de-prioritized and stored in secondary storage, whereas the confidential data points will be stored on primary storage, encryption is implemented and data is he stored the other n cloud, the crucial data point is then secured and stored on block-chain to have advanced tracking implemented. This will avail the organization be apprised on the unauthorized activity within the organization and blockchain is immutable, the pristine copy can be back-tracked and retrieved. The next section, presented a literature review looks into the different aspects of artificial perspicacity, NLP, blockchain and more. Section 3 contains the overall approach proposed, Section 4 contains the implementation and result analysis.

2. LITERATURE REVIEW

In the literature review section of the article, the underlying technologies of AI and work carried using AI in education system are covered. The proposed methodology involves the process of AI classification as covert the input in to structured data, then covert the input type into image. From the image data will be extracted using natural language processing. The intension of converting the input type into structured format is to reduce the size of the data. The importance of converting the input into image is that data extraction is leass time consuming from images. Since the aim to develop a framework for automatic classification, literature regarding intelligent classifier is also covered.

Data extraction using information from the collected dataset is often a complex process that can be efficiently modeled as a data workflow [18]. Data mining from large datasets is very complex, data mining algorithms should be implemented based on the objective of the classification or automation task, in-efficient design can lead the classification into complexity [19]. Therefore, efficient systems are required to be scalable execution of data workflow analysis, exploiting computer services over cloud platforms, and securely storing data. To serve the objective of structuring unstructured data, effectively utilize resources, and promoting less time consumption, there has to be a combinative approach incorporating machine learning, deep learning, data science, feature engineering, encryption, and block-chain [20]. The resultant is a high-level ecosystem that integrates visual workflow within the data system minimizing the efforts on redundant programming, making it crucial for domain experts to use common patterns designed specifically for the development of intelligent data mining, classification, and analysis [21].

In the field of data structuring, automation and maintenance, Artificial intelligence is defined by three major aspects, symbolism, connectionism and behaviorism [22], these aspects are critical for theoretical theorems in the n implementation of artificial intelligence in any field of study or execution of objective-based applications. AI is viewed as an interdisciplinary subject that revolves around logic, data, thinking, and cognition. It reacts well in terms of knowledge processing, pattern recognition from datasets, ML (Machine Learning) and NLP (Natural language processing) [23]. AI technologies offers novel opportunities and incur new challenges to the organization that will set them apart from another digital tech, the tension raised by AI is majorly on the leaking or opening the gaps which can lead to malware, ransomware attacks. Therefore, high-end security such as efficient encryption, block-chain etc. needs to be implemented [24] [36].

Intelligent data management lifecycle is dependent on the feedback approach of having the correct methodology to collect data, process it, extract semantics, feature engineer the semantics a secure storing the data. For data collection medium to large scale organization, transact huge data every day, therefore the requirement will be to manage the transaction of data effectively, to implement the object, the model needs to be feed with relevant information, such relevance learning can be achieved by feed the algorithm with the data, therefore as an initial stage, [25] image processing is used to convert the image into a readable format

using optical character recognition, hidden Markov model, open CV libraries etc., another aspect of data extraction for the learning algorithm is Natural Language Processing (NLP), this is the area of study which uses techniques tools and resources to acquire text from the source data set, the text extraction can help the algorithm learn more about the domain. Multiple techniques are available such as Bag of Words, TF-IDF Doc2Vec etc. these techniques will help the algorithm self-learn in the long run. The data extracted from the images can be grouped based on the frequency of occurrence, ranking, length to form features which help in improvising the accuracy of algorithms [26].

Feature extraction and selection play a very critical role in training the algorithm and to achieve the set objective, features are extracted based on the similar patterns available within a dataset, these patterns can be used as supervisors to train the machine learning algorithms, learning in machine learning algorithm happens using supervised, unsupervised and semi-supervised approach [27]. Based on the objective of the intelligent classifier any methodology can be used. The categorization of data from historical data can be best achieved by using supervised learning. Intelligent classification is deployed for data-related issues techniques have been implemented over the years such as rule-based unique, instance/object-based unique, logic-based technique, and stochastic techniques [28]. Select the techniques that is dependent on the problem statement.

Researchers, developers and organizations are focused to incorporate intelligent automated classifiers to remove human intervention, as human intervention can be time consuming tedious an error prone therefore constant efforts are made to effectively design automated classifiers with the help of machine learning, deep learning, data processing techniques and more [29]. Automating processes can be beneficial in many aspects such as reducing the timeline taken to complete a target human working hours, faster execution, immediate alarming etc., it also incorporates gaps within the as majority of automated tasks depend on the connectivity to the cyber-world, it becomes hotbed to security attacks and can victimize source system to loss of data, data manipulation and misuse. Therefore, with the advent of automated solution there should also be security tracking at every layer. Block-chain in the recent years have gained immense popularity with the emergence and wide adaptability of bitcoins, the fame of block-chain is not limited to crypto-currencies, it is extended in terms of data security. Block-chain has evolved over the years and deployment of the concept of decentralized usage have promoted to incorporation of high-end security with the organization's infrastructure [30]. The most important block-chain is that it is immutable, the data in the blockchain once stored, cannot be deleted or misused by any other entity, the transaction details are stamped on the block which can promote high-end access control and tracking. One of the disadvantage of blockchain is that, there is a cost involved with every transaction of data on the smart contract. To avoid or reduce the overall cost, organizations can link the cloud platforms such as IPFS and then store a portion of critical target block-chain [31]. This paper uses image processing to extract data from the source, natural language processing for feature extraction and selection, machine learning algorithms for training, and block-chain for advanced security and access control.

3. METHODOLOGY

This paper presents a novel methodology which is designed to structure data, perform automatic classification of documents as confidential and non-confidential and store only the confidential data on block chain so that it will be available to end-users permanently without any changes. Multiple technologies are amalgamated to surmount data processing, resource allocation, security and accessibility. The objective of this paper is to cover all the aspect of data storage and retrieval cycle. This will accommodate as a secure data storage and data access for any field of businesses for example: Medical industry can utilize this methodology to store and retrieve confidential data with high-end access control, malignant access tracking and storage, it can withal be utilized by military, research, organizations to incorporate confidential data management within their ecosystem. In this paper the demonstration is done on confidential data cognate to educational institutes, traditionally, the student performance report which plays a consequential role in a student's vocation is stored

manually, on local servers or on unprotected cloud platforms, these methods of maintaining student records are proven to be prone to attacks, data glomming, manipulation etc.,

In this work, the focus is iterated starting from structuring the database to two classes that is confidential and non-confidential class. This step is included considering the two major points, to obviate human intervention in future and to minimize the cost of storing the data on block-chain. This step is pivotal, considering the situation of data available for processing and the cost involved in storing all the data, consequently priority predicated storage will optimize cost in terms of block chain and data storage. Block chain transaction will transpire only on the documents that are relegated as confidential. Afore storing all the data / document on the block, which will incur sizably voluminous cost, only the unique hash ID will be stored, to get to the storage of hash ID on the smart contracts, the document will go through encryption, storage of encrypted document on IPFS, engendering hash and determinately storing the hash ID along with the admin ID on the smart contracts. High level overview of the methodology is presented in figure 1 [37].

The design phase of this methodology will be based on Super admin/admin access control, which is considered as the person of highest authority, this super admin can be defined differently considering the objective of implementation, and for this work the super admin will be the government entity that will maintain the important student records. This entity will be responsible for the data, tracking of the entities within the organization and maintaining a clean flow of data to the end users. To develop the model. Data was collected from the university database. In the next sub-sections, data preparation, data processing, feature extraction, model training, data encryption, data security, data linkage and accessibly will be explained [37].

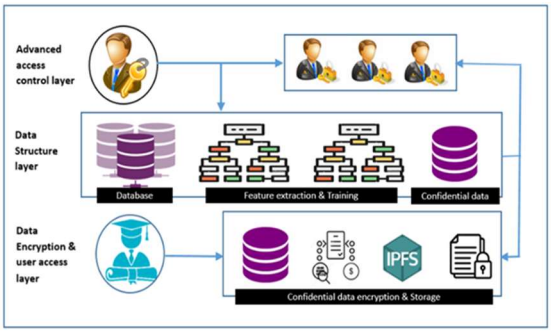


Figure 1: High-Level Proposed Methodology Diagram

The methodology involves development of a hybrid classifier to support the automatic classification of documents. The steps involved in developing a hybrid classifier:

1. Collection of raw data (unstructured data).
2. Converting unstructured data to structured format. All the different formats, for eg .pdf, .doc is converted into jpeg format. Image processing is used to retrieve the text from the jpeg formatted files. Data analysis and feature selection by cleaning the data.
3. Natural language processing is used to convert the text from images into data frames.
4. Data Pre-processing-preparing the data to feed the classification algorithms by splitting the data into training and testing data.
5. Training the Model-training data is used to train the model and accuracy is checked for better classification.
6. Testing data is used to test the classifier performance by predicting the test data.
7. The resultant classifier will take all the files as inputs and classify it into confidential and non-confidential. Only the confidential data is sent to the admin to proceed with the storage.

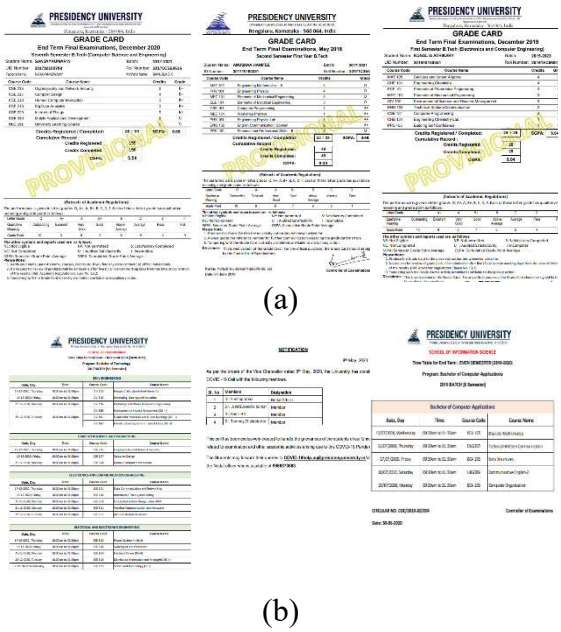
3.1. Data Collection and Processing

In this paper, the dataset is collected from the university database, the objective of applying this methodology to university data base is to protect student's most important documentations. Currently the universities / educational government entities who are responsible to main the data in their database have advanced and moved to the cloud platforms. However, this confidential / sensitive data can be misused, altered, deleted or can be destroyed due to resource failure. Therefore, this methodology is designed to handle the data integrity and incase of data being destroyed due to resource failure, the copy of such document will always be available to re-track and extract, this methodology also handles malicious access within the organization and allows faster data extraction while maintaining data integrity. To support the objective an intelligent classifier is trained to structure the data, RSA algorithm and IPFS internal cloud security is to optimally store the data and blockchain to maintain the original copy of the sensitive / confidential documents [38] [39]. In this paper, student's marks cards are the confidential class and any other document is the miscellaneous class. Figure 1 is a

Journal of Theoretical and Applied Information Technology
15th July 2022. Vol.100. No 13

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

snapshot of confidential data and non-confidential data.



(a)

(b)

Figure 2: (a) Confidential data snapshot (b) non-confidential data snapshot

In this step, the major focus is to have clean dataset, which contains both confidential and non-confidential dataset, as it is depicted in figure 2, the confidential data has much similar patterns than compare to the non-confidential. The intelligent classifier will be fed with these datasets which is converted into singular format jpg from multiple formats such as .pdf, .png etc.

3.2. Intelligent Classifier

In this section, an overview of the methods used to train the classifier is presented. The objective of having an intelligent classifier is to classify the confidential data from the huge data storage, the context of the objective is to classify confidential data of a student and store it using advanced security methodology. To achieve the best possible results, the paper is carefully designed to achieve the best accuracy with proper feature extraction, section and training. The intelligent classifier will be able to classify any document in the university database into 2 classes, confidential class (Point of interest) and non-confidential class (Supports data storage & encryption) cost reduction.

3.2.1. Feature engineering

Machine learning algorithms need multiple iterations to arrive at the best accuracy, feature engineering improves the learning effectively, in this work, the objective is to structure the unstructured dataset, using machine learning, this dataset contains confidential and non-confidential data. The intent is to optimize the storage and encryption cost by allowing the encryption process to be applied at the point of interest dataset, i.e. the confidential documents. Feature selection is a process of selecting a sub-series of feature which is used for model construction, which contributes in reducing training time, improvise chance of generalization and avoid overfitting. In the next sub-section, the method of features extraction and selection implemented in this paper is presented.

3.2.1.1. Feature extraction: image processing to extract the data for training using tesseract

The dataset is informing of images and pdf, before the data is passed to the training algorithm, the data from the dataset in the form of images and pdf needs to be extracted, for this to happen the first step is to convert the dataset into singular image format, then we use line by line word extraction approach using pytesseract library. Word finding the in the dataset will be done by organizing blobs which will have the processed text lines, then the lines along with its region is analyzed for a fixed proportional pitch or text. Then the text lines are broken into words differently according to the kind of character spacing. Then the procedure goes through 2 iterations:

Iteration 1: An attempt is made to recognize each word

Iteration 2: Each word is then passed to an adaptive classifier for learning. The recognized word is passed as a training set, then goes back to iteration 1. After passing the dataset through pytesseract with good dataset, it learns better and performs better. Mathematical details of tesseract feature extraction are given below. Tesseract requires fine-tuning for best possible results, since this paper does not have handwritten dataset included tesseract methodology to extract features from data works well.

Journal of Theoretical and Applied Information Technology
15th July 2022. Vol.100. No 13

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1: Feature extraction using image processing

Algorithm for tesseract feature image feature extraction
Start Dataset <- If data(n) is pdf, then (Convert & pass [pdf file to images using pdf2image) If data(n) is image (pass directly) Input $(\boldsymbol{Dataset(n)})$ = $\boldsymbol{(I\|G\|N)d,h}$ where $\boldsymbol{d,h}$ are the dimension and height of image $\boldsymbol{n}$ $\boldsymbol{I,G,N}$ is used for RGB image, grey color image, and normalized grey image respectively Apply plumbing to the Input $(\boldsymbol{Dataset(n)})$ Execute the network in serious and parallel adding the dimensions. $\boldsymbol{R(x,y)}$ – Rescale 2D input by shrink factor$(\boldsymbol{x,y})$, rearrange the data by increasing the dept. f by factor $\boldsymbol{xy}$. $\boldsymbol{C(x,y)}$ – Convolute to arrive at the output $\boldsymbol{2(x+1),2(y+1)}$ deeper without shrinkage and random fills. $\boldsymbol{M(x,y)}$- Maxpool the input by reducing each $\boldsymbol{x,y}$ to a single value of independent depth Apply functional learning using 1*1 LSTM using $\boldsymbol{Ld,Dataset(n}$ Multi d-dimensional LSTM with n inputs and n outputs which have $\boldsymbol{FT(Dataset(n)),FL(Dataset(n)\ in\ 1 * 1\ LSTM}$ Where $\boldsymbol{FT(Dataset(n)}$ represents 1*1 convolution using $\boldsymbol{Tanh}$ with $\boldsymbol{n}$ outputs $\boldsymbol{FL(Dataset(n)}$ represents 1*1 convolution using $\boldsymbol{Logistics}$ with $\boldsymbol{n}$ outputs Output$(\boldsymbol{Dataset(n)})$ <- number of outputs O determined by unicharset. End

Every single data Student confidential record will have through the N tesseract feature extraction method to pass the classifier, using the data extracted from this technique, the extraction of data from the image is achieved, there is another layer of feature selection method will be introduced to engineer the feature according to the set objective.

3.2.1.2 Feature selection using TF-IDF NLP technique

TF-IDF is mainly used for information retrieval to train machine learning algorithms, in the previous section, the extraction of data from the objective-specific classifier is done and the result is stored in data frames for further processing. The sequence of patterns generated from the above method will serve as input to TF-IDF, the intent is to quantify the relevance of word occurrence in both confidential and non-confidential data frames.

Data frames named confidential (D_c) and no confidential (D_{nc}) is passed to generate 2 different outputs with the Tf-If score using the formula [32]:

For confidential class:

$$\boldsymbol{TfIdf}\ (\boldsymbol{t_c,d_c,D_c}) = \frac{F_{t_c,d_c}}{\sum(t'_c \in d_c)(F_{t_c,d_c})} . \boldsymbol{log}\ \frac{|D_c|}{|\{d_c \in D_c : t_c \in d_c\}|} \ldots (1)$$

For non_confidential class:

$$\boldsymbol{TfIdf}\ (\boldsymbol{t_{nc},d_{nc},D_{nc}}) = \frac{F_{t_{nc},d_{nc}}}{\sum(t'_{nc} \in d_{nc})(F_{t_{nc},d_{nc}})} . \boldsymbol{\log}\ \frac{|D_{nc}|}{|\{d_{nc} \in D_{nc} : t_{nc} \in d_{nc}\}|} \ldots (2)$$

Where t_c, t_{nc} represent the term in confidential and non_confidential class, d_c, d_{nc} represents the document and D_c, D_{nc} is the set of documents in respective classes. The data frames are appended by TF-IDF score which will be given as the input to the machine learning algorithm.

3.2.2. Training Machine learning models

In this section, 3 machine learning algorithms are presented to check on the most optimal algorithm for the end-to-end process deployment. This is the final step of training the intelligent classifier for the objective of having unstructured data in a university database to be structure before moving into data encryption, confidential data storage on IPFS and storing the tracking ID on smart contracts for advanced access control and tracking. Before sending the dataset to the machine learning algorithm, the method will concatenate both the confidential and non-confidential datasets, shuffle them and index it to have a portion left for testing and major portion passed to the training.

$$\boldsymbol{Df_0 = Concat(Df_c, Df_{nc})} \ldots\ldots (3)$$

Where Df_0 contains both the data from class 1 and class 2, shuffled for the algorithm to learn. The section of supervised algorithm is due to the nature of the objective, which is definite.

3.2.2.1. Logistic Regression

To create an objective specific classifier, the classifier identified in this paper has 2 classes (Confidential class and non_confidential class), since Logistic regression works for binary classes, the first algorithm that is trained is Logistic regression. The outcome of the logistic regression algorithm's probability always ranges from 0 (Confidential) or 1 (non_confidential). Logistic regression is simply converts the dataset into values 0 and 1 to fit the S curve, which uses logistic or sigmoidal function [33].

$$\boldsymbol{IC_{output}} = \frac{\boldsymbol{E^{b^0+b^1*D_o}}}{\boldsymbol{1+E^{b^0+b^1*D_o}}} \ldots (\mathbf{4})$$

Where $\boldsymbol{D_o}$ is the input value, $\boldsymbol{b^0}$ the, bias and $\boldsymbol{IC_{output}}$ are the predicted output. The major disadvantage of logistic regression is the linearity is between the dependent variable. The results presented will identify the accuracy of classification.

3.2.2.2. Decision Tree Classifier

Decision tree classifier is a non-parametric supervised learning classifier which basically create a model that predict the value of a target variable using simple decisive rules that can be inferred from the data features. It can also be visualized as a piece-wise constant approach of approximation. Decision tree is defined by different notions that are entropy, Gini impurity and information gain. For this objective specific classifier, the decision tree is developed using the following steps [34]:
Step 1: Sort the values of feature: [s1, s2, s3…. sn], the tf-idf store for every message pattern associated with confidential and non_confidential class.
Step 2: Calculate the entropy & Gini index
Entropy is calculated using the equation:

$$\boldsymbol{E(D_o)} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{c(i) log_2 c(i)} \ldots (\mathbf{5})$$

Where $\boldsymbol{c(i)} = \frac{\boldsymbol{Number(D_o)}}{\boldsymbol{Length(D_o)}}$ where D_o is the combined and shuffled dataset of both confidential and non_confidential classes with tf-idf extracted features.
Gini index is calculated using the equation:

$$\boldsymbol{G(D_o)} = \mathbf{1} - \sum_{\boldsymbol{i=1}}^{\boldsymbol{j}} \boldsymbol{c(i)^2} \ldots (\mathbf{6})$$

Information gain is calculated using the following equation:

$$\boldsymbol{IG(D_o, a)} = \boldsymbol{\left(G(D_o)\middle|E(D_o)\right)\left(G(D_o|a)\middle|E(D_o|a)\right)} \ldots . (7)$$

Where $\left(G(D_o)\middle|E(D_o)\right)$ represents the entropy or Gini index of the parent and $\left(G(D_o|a)\middle|E(D_o|a)\right)$ represents the weighted entropy or gini of children.
Step 3: get Gini & entropy for each feature and select the best split.
The disadvantage of the decision tree is that it cannot handle the given change in the given dataset. This can be accommodated, as the dataset selected in this work is very specific.

3.2.2.3. Random Forest Classifier

Random forest classifier is an extension to decision tree, which has ensemble-based learning approach. Similar to the decision tree random forest is easy to implement, operate fast and have proven successful in multiple domain. Random forest uses multiple simple decision trees in the training stage with vote mode across them in classification stage. The voting strategy has proven to correct the overfitting that can be caused by overfitting [35].
The feature importance is calculated at each tree and summed using the following equation:

$$\boldsymbol{RF\left(D_{o(j)}\right)} = \frac{\boldsymbol{\sum jnormf_{o(j)}}}{\boldsymbol{\sum j \in f(n), k \in n.normf(o)_{jk}}} . (\mathbf{7})$$

Where, normalized features in D_o for tree j. this is included to reach a better accuracy if the decision tree is not providing the required accuracy.

Journal of Theoretical and Applied Information Technology
15th July 2022. Vol.100. No 13

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3.3. Advanced Security Using Blockchain

This paper proposes a top down approach, considering the situation of confidential data storage in India using advanced technologies, the current state is semi-automated, where the documents have been stored cloud platforms, which are prone to security attacks, loss of data, and hostage of data for ransom. In educational institute malicious access the data, re-writing and deleting can highly affect the student's career, as the objective is to protect the pivotal document in the students like, providing encryption will not completely provide security. Therefore in this paper, with the method of structuring the data from highly unstructured heap of data base, additional layer of security which can prove to benefit confidential data storage without any gaps.
The following are the proposed steps [38] [39] [40]:
1- An organizational advanced access control flow, highest control provided to super admin.
2- Super admin can add n number of admins who will be provided CRUD operation access.
3- Every admin will be linked to a block chain account to have a structured process in place: once the admin logs in, he /she should verify using 2-step authentication. In the admin add panel the admin has to open his account on ethereum smart contract platform, the document needs to be stored on the IPFS cloud platform, generate UhashID and store the UhashID (Unique hash ID that contains the Admin login ID + Smart contract ID). The hash will be stored on the blockchain using the crypto transaction.
4- This will ensure the authorized user is tracked on his activity, if there is an external attacked, the documents can be reverse engineered to be retrieved again. As block chain is immutable the record will be available on the platform.

Every document entering into university database will be sent into the intelligent classifier which will flag the confidential data, if the data is identified as confidential, then it will be passed through the advanced encryption and tracking process, to store the document on blockchain linked IPFS cloud.

4. IMPLEMENTATION & RESULT ANALYSIS

This paper presents an optimal end2end approach which not only consider data-lifecycle, it also proposes a method towards advance storage and tracking. The objective is to structure the documentation using machine learning algorithm and then provide high-end security to the most important certificates in student's career. This methodology is not limited to student confidential data storage, it can be extended to various sectors such as medical, military, research, corporate infrastructure etc. In this section, an overall flow of the combined technologies is presented.

Table 2: End-to-End deployment algorithm

Algorithm for End 2 End data structuring, storage and tracking - training

To implement the above mentioned algorithm, the data was collected from university database, which was later converted into singular format using pdf2image converter in python.
The images where stored in the config and non-config library which were labelled for training. The labelled file structure was processed to retrieve information for pattern extraction from all the images related to confidential and non-confidential. A corpus of patterns from non-config and config will be created, the below screenshot represents the first sequence extraction from the image into text.

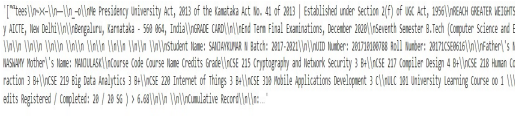


Figure 3: Snapshot of data cleaning and processing

The data is then cleaned to remove, new lines, black spaces, punctuations etc... And are put into two dataframes named df_config and df_non_ config with 4 columns as shown in the figure
(a) and (b).

Figure 4 (a) and (b) : Data frame for confidential data and non-confidential data

	text	file_type	target
0	"tees~><———_—oMe Presidency University Act, 20...	config	1
1	"ee~~_ToMa Presidency University Act, 2013 of ...	config	1
2	PRESIDENCY UNIVERSITY~~><———IAa. Presidency Un...	config	1
3	~—amyPRESIDENCY UNIVERSITYPrivate University E...	config	1
4	"ee~~_ToMa Presidency University Act, 2013 of ...	config	1

	text	file_type	target
0	Presidency University, Bengaluru School of Eng...	non_config	0
1	1. Write a C function to merge two linked list...	non_config	0
2	12. Construct a binary search tree for the fol...	non_config	0
3	Exercises on Relational Algebra and Datalog ...	non_config	0
4	( ) ( ) ( ) ( )...	non_config	0
...	...	...	...
94	25/01/2021 Rotaract Club of Presidency Univ...	non_config	0
95	Detailed Schedule for Blood Donation Camp 09/0...	non_config	0
96	22/01/2021 Dear All, It's a pleasure to infor...	non_config	0
97	13-01-2021 Greetings from Ministry of Educ...	non_config	0
98	04/01/2021 ...	non_config	0

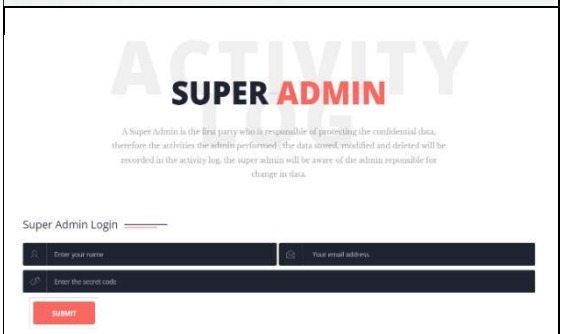
New data frames with the extracted features along with its score, will be concatenated to form a single data frame, then the single data frame with both the classes are shuffled and divided into test and train.

Table 3: Accuracy after training

Name of the algorithm	Accuracy
Logistic regression	94.5%
Decision tree	90.2 %
Random forest	93%

After training the 3 machine learning algorithm, we see that Logistic regression is performing better with 94.5 % accuracy, whereas decision tress is performing up to 90.2 % accuracy and random forest with 93% accuracy. For testing, when there is new document sent into this intelligent classifier, it should automatically convert the data into jpeg format, retrieve the data from the image using image processing, check against the features and training to detect which class the document belongs to.
After having the intelligent classifier as our first layer in the methodology, the data is the admin can visit his/her control panel and store the data into the IPFS cloud, store it on blockchain and map the data to students ID. For blockchain implementation, ethereum smart contracts where used, the code to store the hash on the smart contract is written using solidity programming language and metamask is used to allow transaction for storing the hash ID on the block chain. A webapp is created for internal access control management, demonstrated in the screenshots below. Super admin is responsible for the confidential data management and damage control. Admin is responsible for all the activity related to data storage, update and retrieval, activities will get captured and be visible to super admin to manage admin activity. Super admin role is to monitor the activities of admin, to be informed on any malicious activity (using activity log) within the secure eco-system and to take corrective actions [38] [39] [40].

Super admin panel login page



Super admin credential verification

algorithms5: Super Admin front-end access and flow

Admin's role is to authentically access data, store and retrieve. Every single step will be logged into the admin's panel, any malicious access, delete or change in the record will be identified by the same super admin.

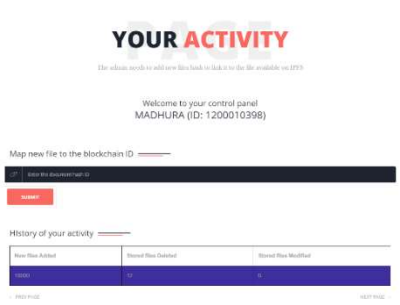
Admin verification login page

2-step authentication

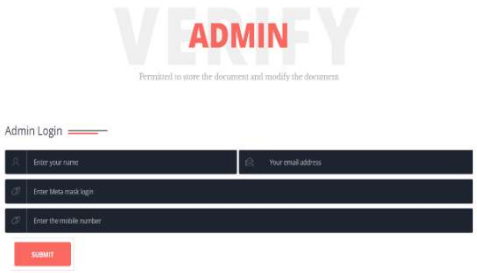
Journal of Theoretical and Applied Information Technology
15th July 2022. Vol.100. No 13

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Admin activity center



Admin credential verification



Blockchain credential verification

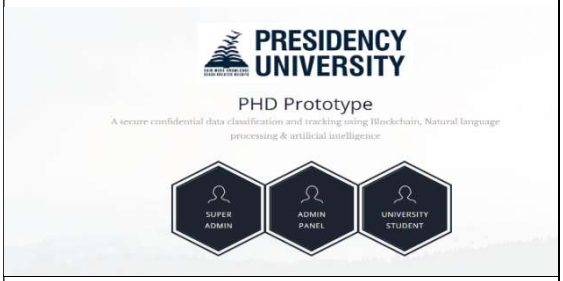
New document storage and mapping view

Figure 6: Admin front-end access and flow

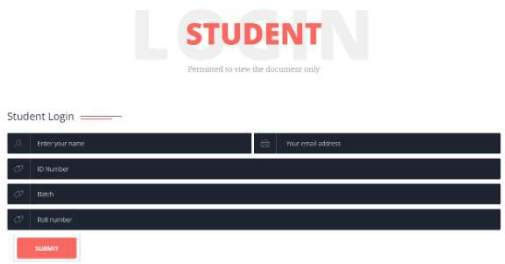
An admin is the first point of contact to the confidential data within the secure data transfer infrastructure, therefore, multiple layer of identity verification is added to avoid any malicious access within the organization. An admin should authenticate, find for classified confidential documents, store it on IPFS cloud and stamp it on smart contracts using Ethereum transactions. Finally, the admin should also map the student's ID with the unique hash ID generated for every document for proper data retrieval.

Finally, the data retrieval model for a student to authentically access his/her confidential data over the internet.

Student access login page



Student credential verification



Authorized student's data retrieval

Figure 7: Student front-end access and flow

5. CONCLUSION

Confidential data integrity is critical for every business or entity connected in the cyber world, the advancements with the advent of internet has led to multiple gaps between applications and data transaction. To bridge the gaps that cause information to be misused, advanced technologies should be utilized. In this paper an end-to-end

methodology is proposed to develop, store and maintain high-end data integrity especially for the sensitive information belonging to internet users, organizations, businesses etc. The proposed methodology will minimize manual intervention and have automated steps at every step in the data integrity preservation cycle. Considering the paramountcy of having structured data to preserve it in the most optimal way, the first step is to structure the data from the unstructured heap present in the system. Machine learning algorithms are being utilized to relegate the multi-format documents into clean structured data, this will ineluctably lead to better resource allocation. Then the confidential structured data, is then stored into IPFS cloud, engender a unique hash ID that will implement advanced access control and storing the hash ID on to smart contracts which uses blockchain to make the data non-mutable and to track the individual responsible for malignantly truncating the confidential data integrity. The end2end approach maintains data sanity, removes third party involvement and can be cost effective. The work proposes an idea of intelligent classifier with the intension of reducing the time taken in manual classification. As automatic related works are very rarely used in education field the idea provides efficient steps to develop an automatic document classifier. The reason for using only documents for consideration is that in education system the production of hard copy documents is more. At the end classified documents will be categorized as confidential and non-confidential documents. The work also provides an idea of storing the confidential documents in immutable storage named blockchain technology. The reason for storing only confidential documents is that these documents will not be changed in future. For the enhancement of the work in future deep learning algorithms can be applied for automatic classification and blockchain can be used for storing non-static documents.

REFERENCES

[1] Yeravdekar, Vidya Rajiv; Tiwari, Gauri (2014). Global Rankings of Higher Education Institutions and India's Effective Non-presence: Why Have World-class Universities Eluded the Indian Higher Education System? And, How Worthwhile is the Indian Government's Captivation to Launch World Class Universities?. Procedia - Social and Behavioral Sciences, 157(), 63–83. doi:10.1016/j.sbspro.2014.11.010

[2] King, C. Saxena, C. Pak, C. -m. Lam and H. Cai, "Rethinking Engineering Education: Policy, Pedagogy, and Assessment During Crises," in *IEEE Signal Processing Magazine*, vol. 38, no. 3, pp. 174-184, May 2021, doi: 10.1109/MSP.2021.3059243.

[3] Sarker, Iqbal H.; Kayes, A. S. M.; Badsha, Shahriar; Alqahtani, Hamed; Watters, Paul; Ng, Alex (2020). Cybersecurity data science: an overview from machine lea arning perspective. Journal of Big Data, 7(1), 41–. doi:10.1186/s40537-020-00318-5

[4] Berdik, David; Otoum, Safa; Schmidt, Nikolas; Porter, Dylan; Jararweh, Yaser (2021). A Survey on Blockchain for Information Systems Management and Security. Information Processing & Management, 58(1), 102397–. doi:10.1016/j.ipm.2020.102397

[5] Tomashevskyi, Valentyn & Yatsyshyn, Andrii & Pasichnyk, Volodymyr & Kunanets, Nataliia & Rzheuskyi, Antonii. (2020). Data Warhouses of Hybrid Type: Features of Construction. 10.1007/978-3-030-16621-2_30.

[6] Casino, Fran; Dasaklis, Thomas K.; Patsakis, Constantinos (2018). A systematic literature review of blockchain-based applications: current status, classification and open issues. Telematics and Informatics, (), S0736585318306324–. doi:10.1016/j.tele.2018.11.006

[7] Steiner, David F.; Chen, Po-Hsuan Cameron; Mermel, Craig H. (2020). Closing the translation gap: AI applications in digital pathology. Biochimica et Biophysica Acta (BBA) - Reviews on Cancer, (), 188452–. doi:10.1016/j.bbcan.2020.188452

[8] Martins, P., Lopes, S. I., Rosado da Cruz, A. M., & Curado, A. (2021). Towards a Smart & Sustainable Campus: An Application-Oriented Architecture to Streamline Digitization and Strengthen Sustainability in Academia. Sustainability, 13(6), 3189. doi:10.3390/su13063189

[9] Han, Meng & Li, Zhigang & He, Jing & Wu, Dalei & Xie, Ying & Baba, Asif. (2018). A Novel Blockchain-based Education Records Verification Solution. 178-183. 10.1145/3241815.3241870.

[10] Kuziemski, Maciej; Misuraca, Gianluca (2020). AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings. Telecommunications Policy, (), 101976–. doi:10.1016/j.telpol.2020.101976

[11] Chaurasia, Sushil S.; Kodwani, Devendra; Lachhwani, Hitendra; Ketkar, Manisha Avadhut; Roberts, Brian (2018). Big data academic and learning analytics: connecting the dots for academic excellence in higher education. International Journal of Educational Management, (), 00–00. doi:10.1108/IJEM-08-2017-0199

[12] Adnan, Kiran; Akbar, Rehan (2019). An analytical study of information extraction from unstructured and multidimensional big data. Journal of Big Data, 6(1), 91–. doi:10.1186/s40537-019-0254-8

[13] Lo Giudice, Paolo; Musarella, Lorenzo; Sofo, Giuseppe; Ursino, Domenico (2019). An approach to extracting complex knowledge patterns among concepts belonging to structured, semi-structured and uns,tructured sources in a data lake. Information Sciences, 478(), 606–626. doi:10.1016/j.ins.2018.11.052

[14] Young, Aaron; Rogers, Pratt (2019). A Review of Digital Transformation in Mining. Mining, Metallurgy & Exploration, (), –. doi:10.1007/s42461-019-00103-w

[15] Adlakha, Richa; Sharma, Shobhit; Rawat, Aman; Sharma, Kamlesh (2019). [IEEE 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon) - Faridabad, India (2019.2.14-2019.2.16)] 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon) - Cyber Security Goal's, Issue's, Categorization & Data Breaches. , (), 397–402. doi:10.1109/COMITCon.2019.8862245

[16] Rocha, Álvaro; Guarda, Teresa (2018). [Advances in Intelligent Systems and Computing] Proceedings of the International Conference on Information Technology & Systems (ICITS 2018) Volume 721 || Big Data, the Next Step in the Evolution of Educational Data Analysis. , 10.1007/978-3-319-73450-7(Chapter 14), 138–147. doi:10.1007/978-3-319-73450-7_14

[17] Duan, Yanqing; Edwards, John S.; Dwivedi, Yogesh K (2019). Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda. International Journal of Information Management, 48(), 63–71. doi:10.1016/j.ijinfomgt.2019.01.021

[18] Ramírez-Gallego, Sergio; Fernández, Alberto; García, Salvador; Chen, Min; Herrera, Francisco (2017). Big Data: Tutorial and Guidelines on Information and Process Fusion for Analytics Algorithms with MapReduce. Information Fusion, (), S1566253517305912–. doi:10.1016/j.inffus.2017.10.001

[19] Rekha, G., Tyagi, A. K., Sreenath, N., & Mishra, S. (2021). Class Imbalanced Data: Open Issues and Future Research Directions. 2021 International Conference on Computer Communication and Informatics (ICCCI). doi:10.1109/iccci50826.2021.94022

[20] Flath, Christoph M.; Stein, Nikolai (2018). Towards a data science toolbox for industrial analytics applications. Computers in Industry, 94(), 16–25. doi:10.1016/j.compind.2017.09.003

[21] Al-Sahaf, Harith; Bi, Ying; Chen, Qi; Lensen, Andrew; Mei, Yi; Sun, Yanan; Tran, Binh; Xue, Bing; Zhang, Mengjie (2019). A survey on evolutionary machine learning. Journal of the Royal Society of New Zealand, (), 1–24. doi:10.1080/03036758.2019.1609052

[22] Zhang, C., & Lu, Y. (2021). Study on artificial intelligence: The state of the art and future ts. Journal of Industrial Information Integration, 23, 100224. doi:10.1016/j.jii.2021.100224

[23] Kadam, Suvarna; Vaidya, Vinay (2020). Cognitive Evaluation of Machine Learning Agents. Cognitive Systems Research, (), S1389041720300978–. doi:10.1016/j.cogsys.2020.11.003

[24] Sharma, A., Kaur, S., & Singh, M. (2021). A comprehensive review on blockchain and Internet of Ththe ings in healthcare. Transactions on Emerging Telecommunications Technologies. doi:10.1002/ett.4333

[25] Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., … Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. Journal of Big Data, 8(1). doi:10.1186/s40537-021-00444-8

[26] Srivastava, S., Divekar, A. V., Anilkumar, C., Naik, I., Kulkarni, V., & Pattabiraman, V. (2021). Comparative analysis of deep learning image detection algorithms. Journal of Big Data, 8(1). doi:10.1186/s40537-021-00434-w

[27] Wang, Pin; Fan, En; Wang, Peng (2020). Comparative Analysis of Image Classification Algorithms Based on Traditional Machine Learning and Deep Learning. Pattern Recognition Letters, (), S0167865520302981–. doi:10.1016/j.patrec.2020.07.042

[28] Al-Faris, Mahmoud; Chiverton, John; Ndzi, David; Ahmed, Ahmed Isam (2020). A Review on Computer Vision-Based Methods for Human Action Recognition. Journal of Imaging, 6(6), 46–. doi:10.3390/jimaging6060046

[29] Somogyi, N., & Kovesdan, G. (2021). Software Modernization Using Machine Learning Techniques. 2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI). doi:10.1109/sami50585.2021.937865

[30] Lu, Yang (2019). The blockchain: State-of-the-art and research challenges. Journal of Industrial Information Integration, (), S2452414X19300019–. doi: 10.1016/j.jii.2019.04.002

[31] Hasan, Haya R.; Salah, Khaled; Jayaraman, Raja; Yaqoob, Ibrar; Omar, Mohammed (2020). Blockchain Architectures for Physical Internet: A Vision, Features, Requirements, and Applications. IEEE Network, (), 1–8. doi:10.1109/MNET.021.2000442

[32] Arroyo-Fernández, Ignacio; Méndez-Cruz, Carlos-Francisco; Sierra, Gerardo; Torres-Moreno, Juan-Manuel; Sidorov, Grigori (2019). Unsupervised sentence representations as word information series: Revisiting TF–IDF. Computer Speech & Language, (), S0885230817302887–. doi: 10.1016/j.csl.2019.01.005

[33] Stephan Dreiseitl; Lucila Ohno-Machado (2002). Logistic regression and artificial neural network classification models: a methodology review. , 35(5-6), 352–359. doi:10.1016/s1532-0464(03)00034-0

[34] Lu, Songfeng; Braunstein, Samuel L. (2014). Quantum decision tree classifier. Quantum Information Processing, 13(3), 757–770. doi:10.1007/s11128-013-0687-5

[35] Xu, B., Guo, X., Ye, Y., & Cheng, J. (2012). An Improved Random Forest Classifier for Text Categorization. J. Comput., 7, 2913-2920.

[36] K. Madhura and R. Mahalakshmi, "Survey on Technologies, Benefits, Challenges and Future Suggestions to Improvise the Data Security of Confidential Academic Records in India," 2021 9th International Conference on Cyber and IT Service Management (CITSM), 2021, pp. 1-9, doi: 10.1109/CITSM52892.2021.9588938.

[37] Madhura, K. and Mahalakshmi, R. (2022), "Designing an optimized confidential-data management system using preeminent access-control and block-chain", International Journal of Intelligent Computing and Cybernetics, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/IJICC-12-2021-0295.

[38] Madhura, K. and Mahalakshmi, R. (2022), "APPLYING CRYPTOGRAPHY AND BLOCK CHAIN TECHNOLOGY TO SECURE AN ORGANIZATION'S CONFIDENTIAL DOCUMENTS", KOREA REVIEW OF INTERNATIONAL STUDIES, Vol. 15, Spl. Issue 02, pp. 96-110.

[39] Madhura, K. and Mahalakshmi, R. (2021), " Securing the organization documents in content management system with two levels of security using encryption and block chain technology", Design Engineering, ISSN: 0011-9342, Issue: 7, pp. 14025- 14039.

[40] Madhura, K. and Mahalakshmi, R. (2018), " An Approach for Securing Organizational Data using Block-chain and Cryptography", Solid State Technology, Volume: 63 Issue: 2s, pp. 2429-2441.