An Improved Machine Learning Method by applying Cloud Forensic Meta-Model to Enhance the Data Collection Process in Cloud Environments

Rafef Al-Mugern
Department of Computer Science, Faculty of Computing, Universiti Teknologi Malaysia, Malaysia | Department of Computer Science, Shaqra University, Saudi Arabia
a-20@graduate.utm.my (corresponding author)

Siti Hajar Othman
Department of Computer Science, Faculty of Computing, Universiti Teknologi Malaysia, Malaysia
hajar@utm.my

Arafat Al-Dhaqm
Computer & Information Sciences Department, Universiti Teknologi PETRONAS, Malaysia
arafat.dhaqm@utp.edu.my

Received: 12 November 2023 | Revised: 7 December 2023 | Accepted: 11 January 2024

Licensed under a CC-BY 4.0 license | Copyright (c) by the authors | DOI: https://doi.org/10.48084/etasr.6609

ABSTRACT

Cloud computing has revolutionized the way businesses operate by offering accuracy in Normalized Mutual Information (NMI). However, with the growing adoption of cloud services, ensuring the accuracy and validation of common processes through machine learning and clustering of these common concepts as well as of the processes generated by cloud forensics experts' data in cloud environments has become a paramount concern. The current paper proposes an innovative approach to enhance the data collection procedure in cloud environments by applying a Cloud Forensic Meta-Model (CFMM) and integrating it with machine learning techniques to improve the cloud forensic data. Through this approach, consistency and compatibility across different cloud environments in terms of accuracy are ensured. This research contributes to the ongoing efforts to validate the clustering process for data collection in cloud computing environments and advance the field of cloud forensics for standardizing the representation of cloud forensic data, certifying NMI and accuracy across different cloud environments.

Keywords-cloud forensics; data collection; cloud environment; cloud computing; data integrity

I. INTRODUCTION

Cloud computing has transformed the landscape of IT infrastructure, enabling businesses to harness the benefits of scalable, efficient, accurate, and flexible resources [1]. However, as organizations increasingly migrate their critical data and applications to cloud environments, concerns about data security, integrity, and the ability to effectively investigate and respond to accurate data collection incidents have grown due to redundancy during the interoperability process regarding NMI [2, 3]. Establishing the trustworthiness of accurate data collection in cloud environments is essential and it necessitates the development of innovative approaches that not only enhance data collection, but also enable intelligent analysis and proactive validation process [4]. Numerous studies introduced an improved Machine Learning (ML) method that leverages the Cloud Forensic Meta-Model (CFMM) to significantly ameliorate the data collection process in cloud environments but accuracy issues, such as classifications and NMI emerge [5, 6]. Traditional data collection practices in the cloud often lack the structured and standardized approach needed to address data collection, validation, and accuracy concerns comprehensively [7]. The integration of CFMM with ML for data collection offers a promising solution to these setbacks. The adoption of cloud computing has introduced unique challenges to data collection and validation [8, 9]. Traditional methods and tools designed for on-premises environments are often ill-suited for cloud-based systems. Cloud environments are dynamic, distributed, and virtualized, making data collection and analysis a complex and costly endeavor [10]. Additionally, the dynamic nature of accurate data collection in cloud systems makes tracking data provenance and maintaining the integrity of critical information rigorous.

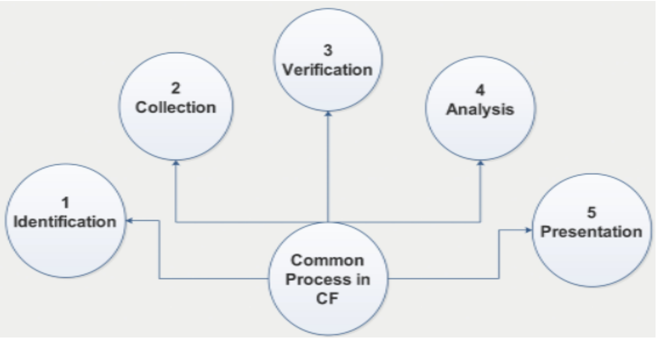


Fig. 1. Cloud forensic process.

Furthermore, Figure 1 explains the Cloud Forensic (CF) process divided into the ML techniques that are integrated into several procedures to enable intelligent data analysis, anomaly detection, and predictive analytics. ML meta-models can automatically identify patterns, anomalies, and potential validation extortions in real-time interoperability process by enhancing the proactive accuracy posture of cloud environments [11, 12]. Not only does this integration strengthen the ability to detect and manage to validate data collection promptly, but also provides valuable insights for forensic investigations [13].

In this study, we present the development of a prototype system that implements CFMM and integrates ML algorithms for data analysis to improve accuracy in terms of NMI. The system is rigorously evaluated in a simulated cloud environment employing various use cases and scenarios applied for classification. The results of this research indicate substantial improvements in data collection, accuracy, and the ability to detect and respond to NMI [14] incidents proactively. Problems related to the classification performed during the digital investigation process in cloud environments lead to a discussion of incident response strategies [15, 16]. Some of them propose new tools, procedures, and meta-models to accomplish accurate data collection investigation in the cloud [17, 18]. In summary, by introducing an improved ML method and by applying a CF meta-model, this article contributes to an enhanced data collection process in cloud environments. The critical need for upgraded data collection procedures in cloud environments is tackled by proposing an innovative method that combines the CFMM and ML techniques to ensure NMI. The main target of this paper is to propose a forensic model to organize and structure a CF field among forensic domains. The proposed abstract model combines and unifies redundant and different common CF processes and terminologies from different CF models.

II. RELATED WORK

A. *Cloud Forensics and Investigations*

Previous research has extensively explored challenges such as classification regarding NMI [19, 20]. Authors in [21] provided an overview of various forensic techniques and approaches in cloud environments, offering valuable insights into the field of establishing NMI. Digital forensics in cloud computing environments face unique difficulties related to data collection, preservation, and analysis, laying the foundation for the need to develop structured models [22-24].

Data provenance in cloud meta-model environment plays a crucial role in forensic investigations [25]. Research on data provenance in cloud environments explores methods for tracking data lineage and warranting its integrity, which aligns with the goals of the proposed CFMM approach. ML for anomaly detection in cloud data collection has been widely applied in cloud data classifications [26, 27]. Studies on anomaly classification and data collection using ML techniques [28-30] can inform the integration of such methods with CFMM to enhance data collection process and accuracy. Standardization efforts in cloud data classification bodies and organizations [31, 32] have proposed guidelines and standards for cloud classification and forensics. These standards can inform the development of CFMM and its compatibility with existing best practices on a medium and large scale.

B. *Cloud NMI and Compliance Frameworks*

Various frameworks and models, like the Cloud NMI Matrix [33], provide guidelines for the classification of cloud environments [34]. Understanding these frameworks can help in designing CFMM to align with industry best practices. Research on cloud metadata management and studies in cloud environments can provide insights into the types of metadata that are crucial for forensic investigations [35]. The proposed CFMM's reliance on forensic metadata makes these studies relevant. Cloud incident response framework research can elucidate the data collection, analysis, and response phases, which are integral to CFMM's objectives. Forensic tools and platforms have also emerged to address the unique challenges of cloud environments. Research on these methods can identify gaps that CFMM aims to fill. Although much research has been conducted on CF, a large scale systematic review on the challenges, solutions, and methodologies does not exist [36]. On the other hand, as far as cloud services are concerned, they have not been given the proper attention even though they are the most important aspects in the field, since cloud computing is based on the services offered [37]. Software designers and engineers, in many cases do not design and implement cloud services to be cloud forensic and enabled for accurate data collection. This is a major issue in CF investigation since the former cannot be conducted in a forensically data collection accurate manner [38, 39].

Authors in [40, 41] proposed various investigation models for the CF domain. However, the models suggested are limited in scope and do not encompass the entire CF field. Therefore, other researchers [42, 43] point out the need for a unified model to include all CF terms with an organized data collection process [44]. These works collectively offer a foundation for the CFMM development and integration. They underscore the significance of structured data collection, data provenance, and the application of ML in addressing the challenges posed by cloud environments, contributing to the growing body of knowledge regarding the accuracy in CF.

III. METHODOLOGY

Data collection and processing in CFMM often involves working with datasets to analyze and investigate data collection incidents or forensic cases in cloud computing environments. A step-by-step overview of how data collection and processing can be conducted using datasets within the CFMM follows.

A. Clustering Process using Datasets and Concept Datasets

To reduce the redundant complexity of the existing evidential clustering data and so improve the accuracy in terms of data collection, a new alternative version, named CFMM is proposed in this paper. CFMM is based on the following two expectations:

- For the same data set, the centers obtained in the partition and those of singleton clusters are very similar. This means that the meta-clusters can be ignored in the initial iterations because the centers of meta-clusters are defined based on the instant information of the related redundant clusters.
- Some of the objects in the dataset are difficult to be accurately assigned to specific clusters. They are then assigned to the related meta-clusters composed of only several close specific clusters. Thus, it is not necessary to expose all the objects under the power-dataset.

By the above assumptions, the CFMM method can be summarized as: (1) preliminary sorting partition, (2) partial organization data rearrangement.

B. Data sorting and Organization

The purposes of this subsection are to preliminary sort and assign each object in the dataset as the outlier, precise or imprecise, adaptively. To derive such a proposal, let us consider a query set X including n objects in p dimensions with $\Omega = \{\omega 1, . . . , \omega c \}$. The support degrees of each object belonging to different singleton (specific) cluster and the noise cluster, called the mass of beliefs in specific partition, can be minimized by an FCM-like objective function at first. There are many methods to obtain the mass of beliefs. For example, the noise clustering method can be applied for the dataset, and we have modified it as the version of the specific partial organization data arrangement to facilitate the presentation. The objective function can be expressed as:

$$JDEC - NC\ (M1, V1) = \sum n\ i = 1 \sum c\ j = 1 \quad (1)$$

$$\mathrm{m}\ \beta\ \mathrm{ij}\ \cdot\ \mathrm{d}\ 2\ \mathrm{ij} + \sum \mathrm{n\ i} = 1\ \delta\ 2\ \cdot\ \mathrm{mi}\emptyset\ \beta \quad (2)$$

$$\sum \mathrm{c\ j} = 1\ \mathrm{mij} + \mathrm{mi}\emptyset = 1, \forall \mathrm{i} = 1, \mathrm{n} \quad (3)$$

where $M1 = (m1, . . . , mn) \in R\ n\times(|\Omega|+1)$ is the mass of the belief matrix for n objects in X, and $V1 \in R\ c \times p$ is the matrix of the centers of single clusters. dij is the Euclidean distance between the object xi and the center of singleton cluster ωj. Parameters β, δ are adjustable.

This research utilized four distinct datasets, each sourced from separate investigations by the Brazilian Federal Police Department. These datasets, originally in varied formats, were converted to a consistent plain text format and underwent extensive preprocessing in order to be prepared for analysis. To assess the quality of the data partitions, each dataset was compared against a "ground truth" reference partition, established by an expert in the field. This comparison was crucial in evaluating the data's organizational structure and its relevance to the study's objectives. The datasets were quantitatively characterized using several metrics, as detailed in Table I, namely the number of documents (N), groups (G), attributes (A), singletons (S), and the distribution of documents across groups. For instance, Dataset A included 37 documents across 23 groups, with an attribute count of 1744 and 12 singletons, the largest group comprising 3 documents. Dataset B, contained 111 documents in 49 groups, with a significantly higher attribute count of 7894 and 28 singletons. This analytical approach provided in-depth insight into the structural and thematic elements of the documents, offering a comprehensive understanding of the data.

TABLE I. DATASET CHARACTERISTICS

Dataset	N	G	A	S	Largest cluster
A	37	23	1744	12	3
B	111	49	7894	28	12
C	68	40	2699	24	8
D	74	38	5095	26	17

Utilizing several meta-modeling methodologies, we were able to validate the generated datasets consistency as well as their applicability (comparison against other models, frequency-based selection). Based on the findings, it can be concluded that the built FCMM is consistent and coherent. This allows domain forensic practitioners to simply instantiate new solution models by picking and combining concept elements (attributes and operations) based on the requirements of their models. From a scientific perspective, the use of reference partitions for evaluating data clustering algorithms is considered a principled approach. In controlled experimental settings, reference partitions are usually obtained from data generated synthetically according to some probability distributions. From a practical standpoint, reference partitions are usually acquired in a different way, but they are still employed to choose a particular clustering algorithm that is more appropriate for a given application, or to calibrate its parameters. In our case, parameter partitions were constructed by a domain expert and reflect the expectations that (S) has about the clusters that should be found in the datasets. In this sense, the evaluation method that we used to assess the obtained data partitions is based on the Adjusted Rand Index, which measures the agreement between a partition P, attained from running a clustering algorithm, and the reference partition given by the expert examiner. More specifically, the greater its value, the better the agreement between P and R is.

Table II elucidates the stratification of processes into five discrete clusters using the CFMM methodology, further delineating into 30 principal categories. The dataset systematically organizes the sequences of actions within a protocol tailored for incident response and data governance. The inaugural cluster delineates the preliminary arrangements and the pinpointing of evidence. The subsequent cluster, Cluster 2, is dedicated to data verification and authentication, with a pronounced consideration for privacy. The third cluster is centered around data analysis and structuring, incorporating procedural steps for system deployment. Cluster 4 is exhaustive in its detail on the collection phase, encapsulating data acquisition and the safeguarding of digital evidence. Culminating the clusters, Cluster 5 is comprehensive of documentation and reporting mechanisms, safeguarding of records, and the presentation of findings, signifying a meticulous blueprint for incident management and regulatory compliance.

TABLE II. CLUSTERING CHARACTERISTICS

Sequence No	Processes	Cluster Number
1	1. Pre-preparation 2. Pre-incident data collection 3. Incident detection 4. Initial response 5. Identification 6. Evidence source identification 7. Log collection	1
2	1. Verifications 2. Authenticating 3. Confirmation 4. Data validation 5. Incident confirmation 6. Concurrent activities 7. Privacy	2
3	1. Data analyzing 2. Analysis 3. Implementation on open stack 4. Organization 5. Matching 6. Examination	3
4	1. Collection 2. Data acquisition 3. Data extraction 4. Collect digital evidence 5. Gathering data 6. Storage	4
5	1. Documentation 2. Reporting 3. Preservation 4. Presentation	5

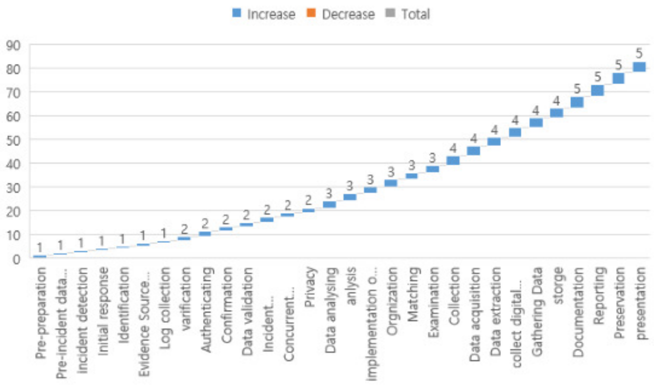


Fig. 2. CFMM partition results.

In Figure 3, the flowchart of the adaptive CFMM partitioning can be seen. By doing this, one can easily find that only a few objects need to be further reassigned. It can greatly reduce the computational complexity, with each imprecise object having a specific dynamic edited forensic meta-model. The diagram outlines a structured framework for data classification within a CFMM designed to optimize the sorting of data elements. The process commences with the introduction of the data set into the system, which then undergoes evaluation using a clustering function. The mass of belief (*Bel*) for the object i is assigned to the cluster $\emptyset$ or ω. This function determines a Bel for each data element, reflecting its affinity to a particular cluster. Data elements whose *Bel* strongly surpasses the threshold are confidently allocated to a cluster, while those with less definitive scores are marked as imprecise and set aside for further analysis.

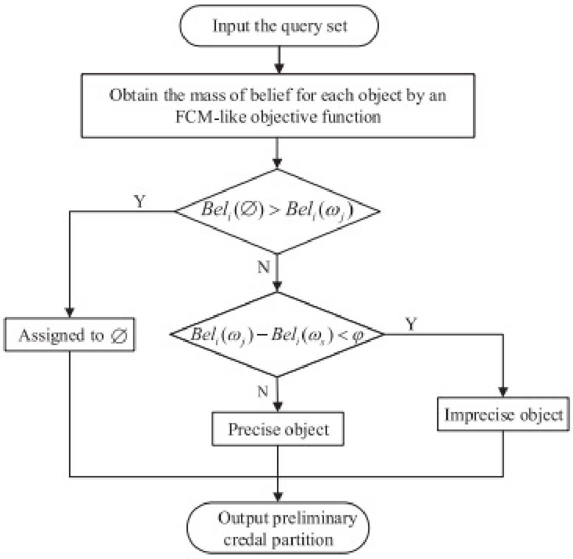

Input the query set
Obtain the mass of belief for each object by an FCM-like objective function
$Bel_i(\varnothing) > Bel_i(\omega_j)$
Y
N
Assigned to $\varnothing$
$Bel_i(\omega_j) - Bel_i(\omega_s) < \varphi$
Y
N
Precise object
Imprecise object
Output preliminary credal partition

Fig. 3. CFMM partitioning process.

This initial sorting results in a preliminary categorization, separating data elements into precise and imprecise groups. Precise elements are those with clear cluster alignment, whereas imprecise elements undergo a more detailed review. A specialized dynamic edited forensic meta-model is then employed to reassess the imprecise elements, considering their individual characteristics to assign them to the appropriate cluster. This method greatly reduces computational load by quickly classifying elements with clear cluster associations, thereby focusing resources on elements with more ambiguous status. Such an approach is particularly advantageous in cloud forensics, where the vast amounts of data make identifying relevant information a challenging and resource-intensive task. The CFMM Algorithm is shown in Figure 4. The CFMM Clustering Algorithm is a methodical process for organizing data specifically for forensic analysis. Required inputs include a dataset labeled dataset to cluster: $X = \{x1, ..., xn\}$ in R^p. Parameters: c, β, δ, ϕ ensure the cluster decision results. An object is constructed without meta-clusters, containing n entities, each within p dimensions, alongside c, β, δ, ϕ, which guide the clustering. The procedure is twofold. It begins by establishing objects devoid of meta-clusters and updating each object's belief mass through iterative calculations. Based on these beliefs, objects are first sorted as outliers, or as belonging distinctly or ambiguously to clusters. This establishes an initial separation of the data. Subsequently, the method recalculates meta-cluster centers and modifies the CFMM for a more equitable distribution of belief. It then focuses on the previously ambiguous objects, re-evaluating and reassigning them to the most fitting clusters. The algorithm concludes with a dynamic partition that offers a detailed and nuanced categorization of the data, tailored for forensic applications.

Algorithm 1: CFMM Clustering Algorithm

```
Require: Dataset to cluster: $X = \{x1, ..., xn\}$ in $R^p$; Parameters: $c$, $\beta$, $\delta$, $\phi$
Ensure: Cluster decision results
Step 1
Construct the object without meta-clusters using (1)–(2); Iterate the mass of beliefs for each object by (2)–(3);
For
    The 1th to $n$th query object preliminary assign the object as the outlier, either precise or imprecise using (3);
End
Sub-return: Preliminary partition.
Step 2 Calculate the centers of meta-cluster using (1);
Reconstruct the objective method CFMM for credal redistribution using (3);
Reiterate the mass of beliefs for the $q$ imprecise objects using (1), (2);
For the 1th to $q$th imprecise object reassign the object to a singleton cluster or meta-cluster.
    End
Sub return: Partial specific redistribution.
Return: Dynamic partition
```

Fig. 4. Pseudo code of CFMM.

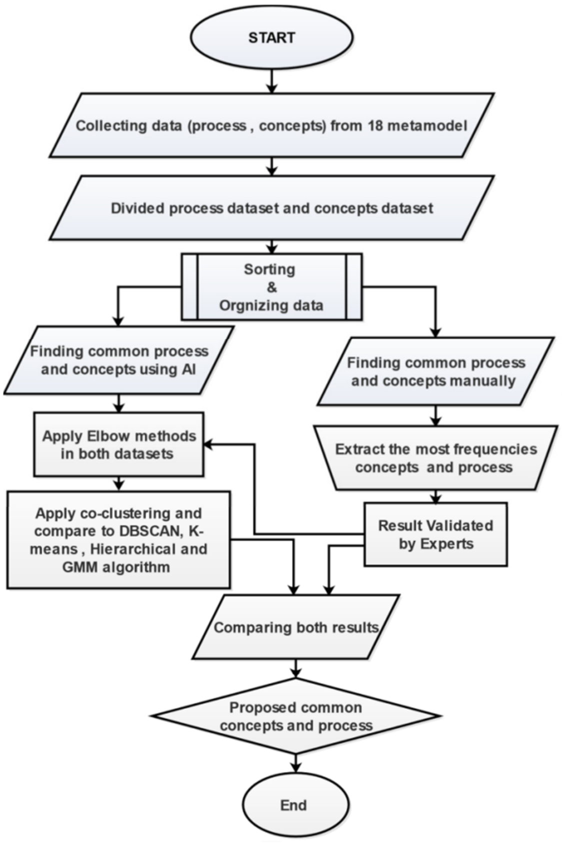

START
Collecting data (process , concepts) from 18 metamodel
Divided process dataset and concepts dataset
Sorting & Orgnizing data
Finding common process and concepts using AI
Finding common process and concepts manually
Apply Elbow methods in both datasets
Extract the most frequencies concepts and process
Apply co-clustering and compare to DBSCAN, K-means , Hierarchical and GMM algorithm
Result Validated by Experts
Comparing both results
Proposed common concepts and process
End

Fig. 5. CFMM clustering process.

Figure 5 explains the CFMM clustering process, which is divided into 2 major portions to sort and organize the clustered data. The flowchart illustrates a structured method for identifying shared processes and concepts within a dataset collected from 18 different metamodels. Initially, the data are gathered and then segregated into two distinct datasets—one for processes and the other for concepts. These are then methodically sorted and organized. Subsequently, two simultaneous approaches are undertaken to determine the commonalities: an algorithmic approach, which incorporates methods such as the Elbow method, co-clustering, DBSCAN, K-means, Hierarchical, and GMM algorithms, along with a manual approach. The latter involves extracting the most frequently occurring concepts and processes. Following the identification stage, the results procured from both approaches undergo expert validation. The final step involves a comparative analysis of the findings from the algorithmic and manual methods. The culmination of this procedure is the determination of the agreed-upon common processes and concepts, signifying the end of the process. This flowchart depicts a comprehensive approach, melding algorithmic computation with manual scrutiny to validate the outcomes through expert review, ensuring the findings are reliable and accurate.

IV. RESULTS AND DISCUSSION

Cloud environments can vary significantly based on the service providers, configurations, and technologies used. Creating a unified metamodel helps standardize the representation of cloud forensic data, ensuring consistency and compatibility across different cloud environments. A unified metamodel enables different tools and systems to exchange and share cloud forensic data seamlessly. Usually, interoperability is crucial for the collaboration between multiple organizations or teams involved in cloud forensic investigations. Cloud forensics often requires data from various sources, such as logs, network traffic, virtual machine snapshots, and configuration data. A unified metamodel allows investigators to integrate and analyze diverse data types more efficiently, leading to a more comprehensive understanding of the incident to improve accuracy and NMI. The results of the proposed CFMM are explained below.

A. Artifact-wise Clustering Results

Table III explains the Artifact-wise Clustering Results which are selected from datasets and are the most meaningful in terms of accuracy to NMI. Table IV contains clusters no, 4, 5, 6, 7, 8, 9, 10, and 11 from the datasets mentioned in Table I. Figure 6 explains the results of Table III. The Elbow method is utilized to determine the optimal number of clusters in a dataset. The frequency values indicate how often each cluster number was identified as the optimal choice in different iterations. Notably, cluster number 5 emerges as the most frequent optimal choice, being identified 3 times, which suggests it might be the most suitable number of clusters for the dataset. Cluster numbers 4 and 6 also appear as potentially optimal, but less frequent, each with a frequency of 2. In contrast, the rest cluster numbers are considered suboptimal for the given data. Table IV illustrates the selected clusters and compares them with existing approaches in terms of accuracy.

A total of 8 iterations were performed for comparison. We can see how the proposed CFMM approach is significantly more accurate than the existing techniques. Figure 7 exhibits the superb accuracy of the proposed CFMM approach. Figure 7(i) states the overall average % results between the proposed CFMM and existing approaches.

TABLE III. RESULTS OF THE CFMM METHOD

Cluster No.	4	5	6	7	8	9	10	11
Frequency	2	3	2	0	1	0	2	1

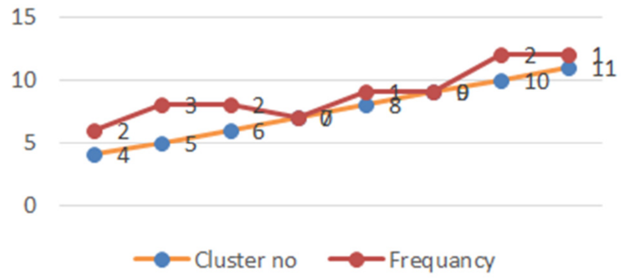


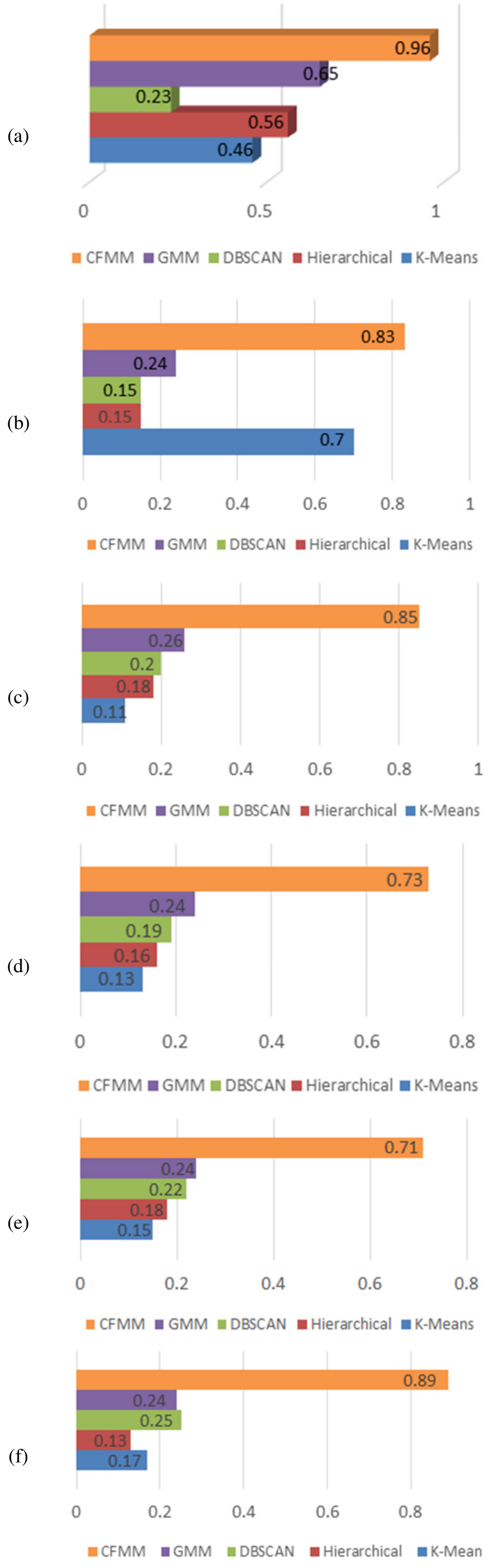
Fig. 6. CFMM clustering results.

TABLE IV. RESULT COMPARISON

	K-Means [28]	**Hierarchical [24]**	**DBSCAN [24]**	**GMM [45]**	**CFMM**
Cluster No 4	0.46	0.56	0.23	0.65	0.96
Cluster No 5	0.7	0.15	0.15	0.24	0.83
Cluster No 6	0.11	0.18	0.2	0.26	0.85
Cluster No 7	0.13	0.16	0.19	0.24	0.73
Cluster No 8	0.15	0.18	0.22	0.24	0.71
Cluster No 9	0.17	0.13	0.25	0.24	0.89
Cluster No 10	0.19	0.11	0.26	0.24	0.79
Cluster No 11	0.21	0.14	0.23	0.19	0.81

Figure 7(a) illustrates the comparison for cluster no 4. Clustering quality metric accuracy is used. It rates the performance on a scale from 0 to 1, where 1 signifies the ideal clustering arrangement. The CFMM algorithm emerges as the most effective, with a near-perfect score of 0.96, indicating its superior ability to cluster the given dataset. The GMM method registers a moderate effectiveness with a score of 0.6. The other algorithms, have a lower clustering performance. The DBSCAN algorithm's performance is not depicted, suggesting a negligible score.

Figure 7(b) compares the same set of algorithms, however, the performance scores vary from Figure 7(a). The proposed CFMM maintains its lead with a score of 0.83, affirming its robustness in effectively clustering the dataset. K-means clustering is still a competitive method, with a score of 0.7, denoting a high level of effectiveness. Conversely, DBSCAN, GMM, and Hierarchical show considerably lower scores, all either at or below 0.26, which implies these methods are less suitable for the dataset based on the metric used. The CFMM algorithm's consistent high performance across both figures indicates its potential as the preferred method for the dataset. For the rest of the cluster numbers mentioned in Figure 7, the accuracy of the proposed approach is always better than that of the compared methods.



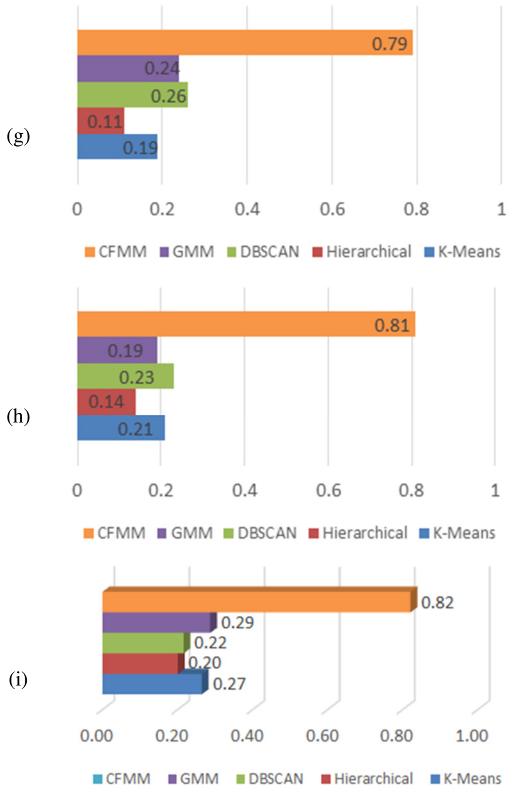


Fig. 7. (a) Cluster 4, (b) Cluster 5, (c) Cluster 6, (d) Cluster 7, (e) Cluster 8, (f) Cluster 9, (g) Cluster 10, (h) Cluster 11, (i) overall average ratio.

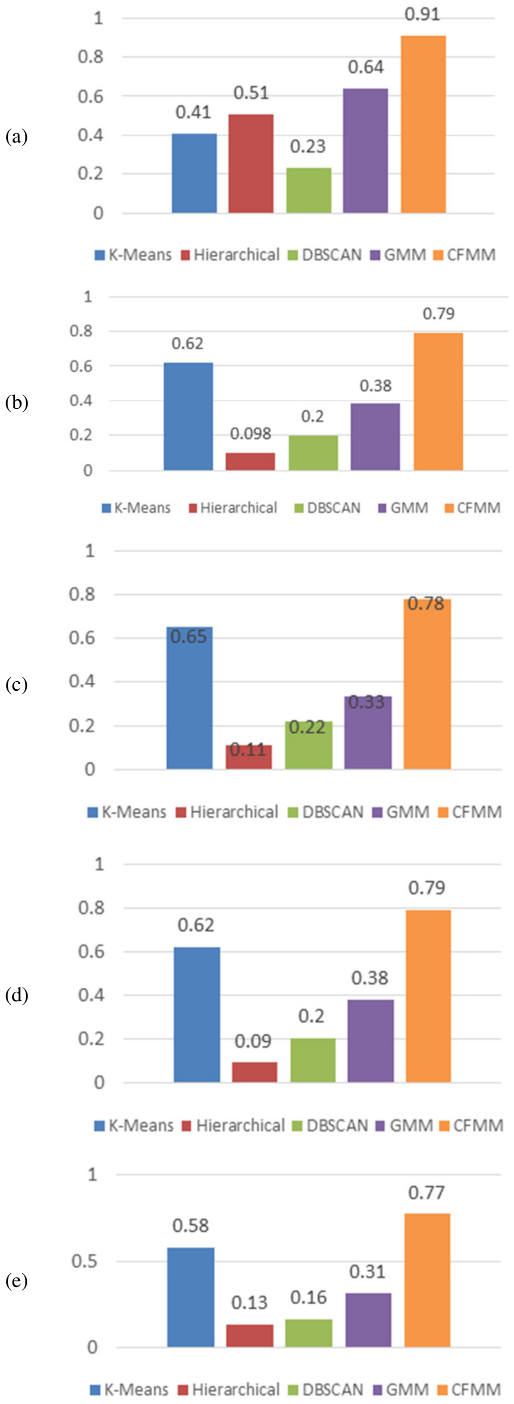
B. *Clustering of Digital Forensics Results*

Table IV illustrates the selected cluster results of the proposed CFMM and compares them with the ones of known approaches such as K-Means, Hierarchical, DBSCAN in terms of NMI. A total of 8 iterations have been performed. We can see that the proposed CFMM approach had significantly improved NMI in comparison to the other techniques. Figure 8 shows the result comparison (NMI).

TABLE V. NMI RESULTS OF SELECTED CLUSTERS

Algorithms/ Clusters	K-Means	Hierarchical	DBSCAN	GMM	CFMM
Cluster No 4	0.41	0.51	0.23	0.64	0.91
Cluster No 5	0.62	0.09	0.2	0.38	0.79
Cluster No 6	0.65	0.11	0.22	0.33	0.78
Cluster No 7	0.62	0.09	0.26	0.38	0.79
Cluster No 8	0.58	0.13	0.16	0.31	0.77
Cluster No 9	0.72	0.14	0.31	0.39	0.85
Cluster No 10	0.74	0.16	0.29	0.41	0.88
Cluster No 11	0.75	0.17	0.27	0.43	0.89

Considering the performance scores across the approaches, the average effectiveness of each algorithm can be estimated. CFMM appears to be the most efficient on average, given its leading scores in both evaluations. K-Means, despite its lower score, shows a moderate average performance, potentially indicating its versatility across different experiment settings. Hierarchical and GMM algorithms present a consistent, though not leading, performance, suggesting their utility in certain contexts. DBSCAN appears less effective on average, which could be due to its sensitivity to density parameters or the data nature. These averages provide insight into the general applicability of each algorithm, though the specific choice would depend on the particularities of the dataset and the clustering objectives.



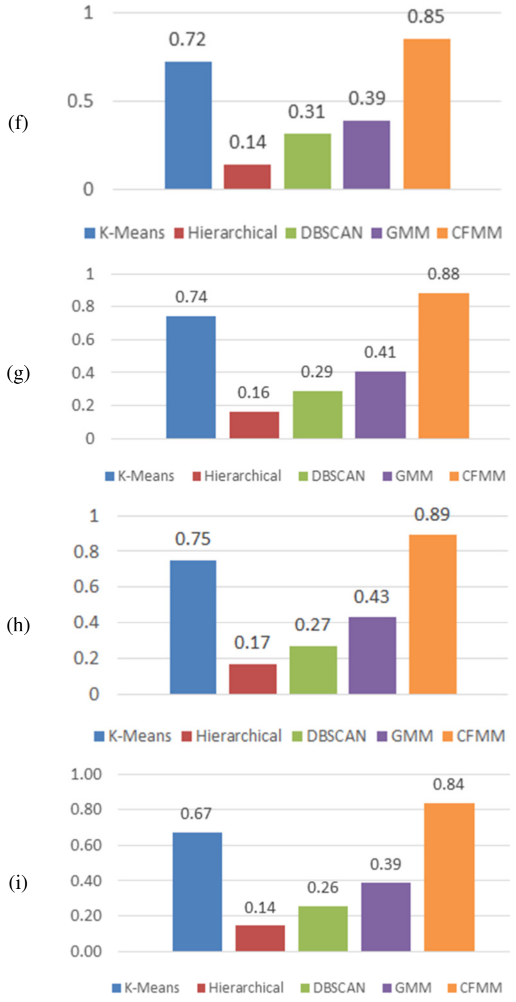


Fig. 8. (a) Cluster 4, (b) Cluster 5, (c) Cluster 6, (d) Cluster 7, (e) Cluster 8, (f) Cluster 9, (g) Cluster 10, (h) Cluster 11, (i) overall average ratio.

V. CONCLUSION

In this paper, we introduced the innovative Cloud Forensic Meta-Model (CFMM), seamlessly integrated with advanced machine learning algorithms, to significantly enhance data collection and analytical capabilities within cloud computing environments. Not only does this approach set a new standard for cloud forensics in terms of data processing and reliability, but also distinctly differentiates itself from the existing theoretical models. A key contribution of our research is the standardization of CFMM processes, which guarantees consistent and precise forensic data analysis across diverse cloud platforms. Furthermore, we outlined the common investigation processes within the CFMM framework, identifying five core stages: pre-preparation, verification, data analysis, collection, and documentation. Our study also effectively categorized and extracted several pivotal concepts from the dataset-CFMM. In addition, this study identified several concepts that were gathered from the dataset-based CFMM. On the other hand, these processes differ and have diverse meanings and synonyms. After that, the frequency feature was employed to choose the most prevalent concepts for each category. As a result, a total of 5 common concepts were chosen based on machine learning approach and the results were compared with the ones of known approaches. Several concept definitions were reconciled using the CFMM method.

ACKNOWLEDGMENT

The authors appreciate the help of the research team and the staff support of the School of Computing at the University Technology Malaysia. We would like also to thank the Deanship of Scientific Research at Shaqra University for supporting this work.

REFERENCES

[1] X. Wu, Z. Jin, J. Zhou, and C. Duan, "Quantum walks-based classification model with resistance for cloud computing attacks," *Expert Systems with Applications*, vol. 232, Dec. 2023, Art. no. 120894, https://doi.org/10.1016/j.eswa.2023.120894.

[2] Y. Liu, Z. Liu, S. Li, Y. Guo, Q. Liu, and G. Wang, "Cloud-Cluster: An uncertainty clustering algorithm based on cloud model," *Knowledge-Based Systems*, vol. 263, Mar. 2023, Art. no. 110261, https://doi.org/10.1016/j.knosys.2023.110261.

[3] K. G. Maheswari, C. Siva, and G. Nalinipriya, "Optimal cluster based feature selection for intrusion detection system in web and cloud computing environment using hybrid teacher learning optimization enables deep recurrent neural network," *Computer Communications*, vol. 202, pp. 145–153, Mar. 2023, https://doi.org/10.1016/j.comcom.2023.02.003.

[4] D. Wang, "Internet of things sports information collection and sports action simulation based on cloud computing data platform," *Preventive Medicine*, vol. 173, Aug. 2023, Art. no. 107579, https://doi.org/10.1016/j.ypmed.2023.107579.

[5] X. Wu, "Big data classification of remote sensing image based on cloud computing and convolutional neural network," *Soft Computing*, Jan. 2022, https://doi.org/10.1007/s00500-021-06562-y.

[6] K. Singh and J. Malhotra, "IoT and cloud computing based automatic epileptic seizure detection using HOS features based random forest classification," *Journal of Ambient Intelligence and Humanized Computing*, Dec. 2019, https://doi.org/10.1007/s12652-019-01613-7.

[7] J. Wang, J. Yu, Y. Song, X. He, and Y. Song, "An efficient energy-aware and service quality improvement strategy applied in cloud computing," *Cluster Computing*, vol. 26, no. 6, pp. 4031–4049, Dec. 2023, https://doi.org/10.1007/s10586-022-03795-w.

[8] D. Thakur, J. K. Saini, and S. Srinivasan, "DeepThink IoT: The Strength of Deep Learning in Internet of Things," *Artificial Intelligence Review*, vol. 56, no. 12, pp. 14663–14730, Dec. 2023, https://doi.org/10.1007/s10462-023-10513-4.

[9] A. Alshammari, "A Novel Security Framework to Mitigate and Avoid Unexpected Security Threats in Saudi Arabia," *Engineering, Technology & Applied Science Research*, vol. 13, no. 4, pp. 11445–11450, Aug. 2023, https://doi.org/10.48084/etasr.6091.

[10] A. Penuelas-Angulo, C. Feregrino-Uribe, and M. Morales-Sandoval, "Revocation in attribute-based encryption for fog-enabled internet of things: A systematic survey," *Internet of Things*, vol. 23, Oct. 2023, Art. no. 100827, https://doi.org/10.1016/j.iot.2023.100827.

[11] S. Krishnaveni and S. Prabakaran, "Ensemble approach for network threat detection and classification on cloud computing," *Concurrency and Computation: Practice and Experience*, vol. 33, no. 3, 2021, Art. no. e5272, https://doi.org/10.1002/cpe.5272.

[12] B. Alghamdi, L. E. Potter, and S. Drew, "Validation of Architectural Requirements for Tackling Cloud Computing Barriers: Cloud Provider Perspective," *Procedia Computer Science*, vol. 181, pp. 477–486, Jan. 2021, https://doi.org/10.1016/j.procs.2021.01.193.

[13] M. Awad, "Google Earth Engine (GEE) cloud computing based crop classification using radar, optical images and Support Vector Machine Algorithm (SVM)," in *3rd International Multidisciplinary Conference*

on Engineering Technology, Beirut, Lebanon, Dec. 2021, pp. 71–76, https://doi.org/10.1109/IMCET53404.2021.9665519.

[14] S. Mian Qaisar and S. F. Hussain, "An effective arrhythmia classification via ECG signal subsampling and mutual information based subbands statistical features selection," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 3, pp. 1473–1487, Mar. 2023, https://doi.org/10.1007/s12652-021-03275-w.

[15] S. Husnain and R. Abdulkader, "Fractional Order Modeling and Control of an Articulated Robotic Arm," *Engineering, Technology & Applied Science Research*, vol. 13, no. 6, pp. 12026–12032, Dec. 2023, https://doi.org/10.48084/etasr.6270.

[16] H. A. Almashhadani, X. Deng, O. R. Al-Hwaidi, S. T. Abdul-Samad, M. M. Ibrahm, and S. N. A. Latif, "Design of A new Algorithm by Using Standard Deviation Techniques in Multi Edge Computing with IoT Application," *KSII Transactions on Internet and Information Systems*, vol. 17, no. 4, pp. 1147–1161, Apr. 2023.

[17] T. Zhao, L. Wu, D. Wu, J. Li, and Z. Cui, "Multi-factor Evolution for Large-scale Multi-objective Cloud Task Scheduling," *KSII Transactions on Internet and Information Systems*, vol. 17, no. 4, pp. 1100–1122, Apr. 2023.

[18] A. Ghosh, D. De, and K. Majumder, "A Systematic Review of Log-Based Cloud Forensics," in *Inventive Computation and Information Technologies*, S. Smys, V. E. Balas, K. A. Kamel, and P. Lafata, Eds. New York, NY, USA: Springer, 2021, pp. 333–347.

[19] V. S. Bai and T. Sudha, "A Systematic Literature Review on Cloud Forensics in Cloud Environment," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 11, no. 4s, pp. 565–578, Feb. 2023.

[20] G. Surange and P. Khatri, "IoT Forensics: A Review on Current Trends, Approaches and Foreseen Challenges," in *8th International Conference on Computing for Sustainable Global Development*, New Delhi, India, Mar. 2021, pp. 909–913.

[21] E. Saiti and T. Theoharis, "An application independent review of multimodal 3D registration methods," *Computers & Graphics*, vol. 91, pp. 153–178, Oct. 2020, https://doi.org/10.1016/j.cag.2020.07.012.

[22] F. Alotaibi, A. Al-Dhaqm, and Y. D. Al-Otaibi, "A Conceptual Digital Forensic Investigation Model Applicable to the Drone Forensics Field," *Engineering, Technology & Applied Science Research*, vol. 13, no. 5, pp. 11608–11615, Oct. 2023, https://doi.org/10.48084/etasr.6195.

[23] O. Ameerbakhsh, F. M. Ghabban, I. M. Alfadli, A. N. AbuAli, A. Al-Dhaqm, and M. A. Al-Khasawneh, "Digital Forensics Domain and Metamodeling Development Approaches," in *2nd International Conference on Smart Computing and Electronic Enterprise*, Cameron Highlands, Malaysia, Jun. 2021, pp. 67–71, https://doi.org/10.1109/ICSCEE50312.2021.9497935.

[24] K. Tange, M. De Donno, X. Fafoutis, and N. Dragoni, "A Systematic Survey of Industrial Internet of Things Security: Requirements and Fog Computing Opportunities," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 4, pp. 2489–2520, 2020, https://doi.org/10.1109/COMST.2020.3011208.

[25] G. Adamo, C. Ghidini, and C. Di Francescomarino, "What is a process model composed of? A systematic literature review of meta-models in BPM," *Software and Systems Modeling*, vol. 20, no. 4, pp. 1215–1243, Dec. 2021, https://doi.org/10.1007/s10270-020-00847-w.

[26] A. Diab, R. Kashef, and A. Shaker, "Deep Learning for LiDAR Point Cloud Classification in Remote Sensing," *Sensors*, vol. 22, no. 20, Jan. 2022, Art. no. 7868, https://doi.org/10.3390/s22207868.

[27] H. Zhang *et al.*, "Deep learning-based 3D point cloud classification: A systematic survey and outlook," *Displays*, vol. 79, Sep. 2023, Art. no. 102456, https://doi.org/10.1016/j.displa.2023.102456.

[28] H. Daghigh, D. D. Tannant, V. Daghigh, D. D. Lichti, and R. Lindenbergh, "A critical review of discontinuity plane extraction from 3D point cloud data of rock mass surfaces," *Computers & Geosciences*, vol. 169, Dec. 2022, Art. no. 105241, https://doi.org/10.1016/j.cageo.2022.105241.

[29] H. Shukur *et al.*, "A State of Art Survey for Concurrent Computation and Clustering of Parallel Computing for Distributed Systems," *Journal of Applied Science and Technology Trends*, vol. 1, no. 4, pp. 148–154, Dec. 2020, https://doi.org/10.38094/jastt1466.

[30] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Information Sciences*, vol. 622, pp. 178–210, Apr. 2023, https://doi.org/10.1016/j.ins.2022.11.139.

[31] H. Yu and X. Hou, "Hierarchical clustering in astronomy," *Astronomy and Computing*, vol. 41, Oct. 2022, Art. no. 100662, https://doi.org/10.1016/j.ascom.2022.100662.

[32] N. Tissir, S. El Kafhali, and N. Aboutabit, "Cybersecurity management in cloud computing: semantic literature review and conceptual framework proposal," *Journal of Reliable Intelligent Environments*, vol. 7, no. 2, pp. 69–84, Jun. 2021, https://doi.org/10.1007/s40860-020-00115-0.

[33] W. Wang and S. Yongchareon, "Security-as-a-service: a literature review," *International Journal of Web Information Systems*, vol. 16, no. 5, pp. 493–517, Jan. 2020, https://doi.org/10.1108/IJWIS-06-2020-0031.

[34] A. Ometov, O. L. Molua, M. Komarov, and J. Nurmi, "A Survey of Security in Cloud, Edge, and Fog Computing," *Sensors*, vol. 22, no. 3, Jan. 2022, Art. no. 927, https://doi.org/10.3390/s22030927.

[35] G. Li, H. Dong, and C. Zhang, "Cloud databases: new techniques, challenges, and opportunities," *Proceedings of the VLDB Endowment*, vol. 15, no. 12, pp. 3758–3761, Dec. 2022, https://doi.org/10.14778/3554821.3554893.

[36] M. Stoyanova, Y. Nikoloudakis, S. Panagiotakis, E. Pallis, and E. K. Markakis, "A Survey on the Internet of Things (IoT) Forensics: Challenges, Approaches, and Open Issues," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 2, pp. 1191–1221, 2020, https://doi.org/10.1109/COMST.2019.2962586.

[37] S. Shiju George and R. Suji Pramila, "A review of different techniques in cloud computing," *Materials Today: Proceedings*, vol. 46, pp. 8002–8008, Jan. 2021, https://doi.org/10.1016/j.matpr.2021.02.748.

[38] M. Rahimi, N. Jafari Navimipour, M. Hosseinzadeh, M. H. Moattar, and A. Darwesh, "Toward the efficient service selection approaches in cloud computing," *Kybernetes*, vol. 51, no. 4, pp. 1388–1412, Jan. 2021, https://doi.org/10.1108/K-02-2021-0129.

[39] M. Masdari and H. Khezri, "Efficient VM migrations using forecasting techniques in cloud computing: a comprehensive review," *Cluster Computing*, vol. 23, no. 4, pp. 2629–2658, Dec. 2020, https://doi.org/10.1007/s10586-019-03032-x.

[40] R. Al-Mugerrn, A. Al-Dhaqm, and S. H. Othman, "A Metamodeling Approach for Structuring and Organizing Cloud Forensics Domain," in *International Conference on Smart Computing and Application*, Hail, Saudi Arabia, Feb. 2023, pp. 1–5, https://doi.org/10.1109/ICSCA57840.2023.10087425.

[41] S. Sureshkumar, N. Kirthiga, T. A. Kumar, P. N. Kumar, Y. P. Kumar Reddy, and R. S. Reddy, "Dual Access Control for Cloud-Based Data Storage and Sharing," in *2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies*, Vellore, India, Dec. 2023, pp. 1–6, https://doi.org/10.1109/ViTECoN58111.2023.10157156.

[42] A. Hakiri, A. Gokhale, S. Ben Yahia, and N. Mellouli, "A Comprehensive Survey on Digital Twin for Future Networks and Emerging Iot Industry." Rochester, NY, USA, Aug. 09, 2023, https://doi.org/10.2139/ssrn.4535810.

[43] A. Naghib, N. Jafari Navimipour, M. Hosseinzadeh, and A. Sharifi, "A comprehensive and systematic literature review on the big data management techniques in the internet of things," *Wireless Networks*, vol. 29, no. 3, pp. 1085–1144, Apr. 2023, https://doi.org/10.1007/s11276-022-03177-5.

[44] Y. Jing, X. Lu, and S. Gao, "3D face recognition: A comprehensive survey in 2022," *Computational Visual Media*, vol. 9, no. 4, pp. 657–685, Dec. 2023, https://doi.org/10.1007/s41095-022-0317-1.

[45] L. Li, R. Wang, and X. Zhang, "A Tutorial Review on Point Cloud Registrations: Principle, Classification, Comparison, and Technology Challenges," *Mathematical Problems in Engineering*, vol. 2021, Jul. 2021, Art. no. e9953910, https://doi.org/10.1155/2021/9953910.