International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v11i8.8005
***Article Received**: 16 May 2023 **Revised**: 20 July 2023 **Accepted**: 12 August 2023*

Advancing Data Privacy: A Novel K-Anonymity Algorithm with Dissimilarity Tree-Based Clustering and Minimal Information Loss

Abhiram Patil, Prof. Binghui Wang
Department of CS
IIT, CHICAGO
apatil57@hawk.iit.edu, bwang70@iit.edu

Abstract: Anonymization serves as a crucial privacy protection technique employed across various technology domains, including cloud storage, machine learning, data mining and big data to safeguard sensitive information from unauthorized third-party access. As the significance and volume of data grow exponentially, comprehensive data protection against all threats is of utmost importance. The main objective of this paper is to provide a brief summary of techniques for data anonymization and differential privacy.A new k-anonymity method, which deviates from conventional k-anonymity approaches, is proposed by us to address privacy protection concerns. Our paper presents a new algorithm designed to achieve k-anonymity through more efficient clustering. The processing of data by most clustering algorithms requires substantial computation. However, by identifying initial centers that align with the data structure, a superior cluster arrangement can be obtained.Our study presents a Dissimilarity Tree-based strategy for selecting optimal starting centroids and generating more accurate clusters with reduced computing time and Normalised Certainty Penalty (NCP). This method also has the added benefit of reducing the Normalised Certainty Penalty (NCP). When compared to other methods, the graphical performance analysis shows that this one reduces the amount of overall information lost in the dataset being anonymized by around 20% on average. In addition, the method that we have designed is capable of properly handling both numerical and category characteristics.

I. INTRODUCTION

In today's environment, businesses are given access to customers' personal information in order to improve their customer service and decision-making processes; nevertheless, a significant portion of the data's potential is still unrealized. It is the goal of many organisations to disseminate this information while still protecting the privacy of individuals, as it may be of use to independent researchers and analysts in addressing a variety of concerns, ranging from those pertaining to community planning to cancer research. However, ensuring that the data continue to serve its intended purpose is essential in order to generate reliable analytical results.

The owners of the data are looking for ways to convert extremely sensitive datasets into low-risk, privacy-protecting datasets that may be shared with a variety of audiences. These audiences can range from researchers to business partners. It is unfortunate that organizations are now releasing databases that are claimed to have undergone anonymization, but it turns out that a significant portion of the records can still be used to re-identify the original subject of the record. It is vital to have a solid understanding of the inner workings of anonymization methods, as well as their appropriate applications, benefits, and downsides.

The focus of this article is on k-anonymity, which is a widely used privacy technique for protecting the privacy of data subjects in data sharing scenarios. Additionally, the article explores the advantages of implementing k-anonymity for data anonymization. The basic goal of many different privacy-preserving systems is to provide data subjects with the ability to remain anonymous. At first look, anonymity seems to mean not being able to be identified, but upon further inspection, it becomes clear that simply eliminating names from a dataset is not enough to guarantee anonymity. When anonymised data are compared with another dataset, there is a risk of re-identification occurring. It is possible for data to contain quasi-identifiers, which are tidbits of information that on their own do not constitute a unique identifier but which, when paired with other datasets, can be used to identify specific individuals.

The basic concept behind k-anonymity is to prevent re-identification of anonymized data by linking it to other datasets. K-anonymity was developed to prevent this. K-anonymization is a powerful tool, provided that it is used appropriately and is accompanied by the required safeguards, such as access control and contractual protections. It is an essential part of the armoury of privacy-enhancing technologies, along with other approaches like differentially private algorithms, and it is currently being developed. As the use of big data grows more widespread, there has been an increase in the data dimensionality as well as a growing number of datasets that are available to the public and can help with re-identification.

The study's results indicate that most standard k-anonymity techniques rely on generalization and suppression methods. Because of their substantial reliance on ordering relations derived from preset generalisation layers on attribute domains, these techniques result in a large loss of information and suffer from this problem. As a consequence of this, the outcomes of anonymization frequently result in a significant loss of information and a decreased level of utility. In addition, the majority of current anonymization algorithms are geared towards protecting private information at the expense of considering the applicability of anonymized data, which has led to a dearth of this data in situations that take place in the real world.

We propose a clustering-based enhanced k-anonymity privacy protection system and seek to improve the clustering process for improved data confidentiality scenarios. Our approach is based on k-anonymity and clustering. When performing clustering, it is possible to obtain better cluster sets by locating initial centroids that are consistent with the data structure. During the clustering process, the ideal cluster set can be obtained if the initial centroids that are found are those that align with the dataset. We recommend an approach that is based on the Dissimilarity Tree for determining the best initial centroids and more accurate clusters in a shorter amount of time using the computer. The next sections will cover important ideas related to k-anonymity.

1.1 K-anonymity

K-anonymity is a concept related to database privacy, where a database is considered K-anonymous if its attributes have been generalized or suppressed to the extent that each row cannot be distinguished from at least k-1 other rows.The use of k-anonymity ensures that specific database linkages are not established, and the accuracy of the released data is maintained. The foundation of k-anonymity lies in the use of two essential techniques: generalization and suppression.To ensure the privacy of the respondents' identities, data holders generally eliminate explicit identifiers like names and social security numbers from microdata [8] before sharing it.While de-identifying data may seem like a way to ensure anonymity, it does not guarantee it, as certain pieces of information such as birth dates, sex, and ZIP codes can still be linked to public databases and used to re-identify respondents. However, the concept of k-anonymity has emerged as a solution to protect microdata tables from such risks. This involves connecting each row in the published table to at least k respondents in an indistinguishable manner. K-anonymity is a critical component of data protection protocols that aim to maintain the integrity of data. Originally, data mining techniques relied on altering input data to safeguard user privacy.The previous approach of altering input data lacked a structured framework to demonstrate the extent of privacy protection it provided. Another subfield of data mining that prioritized user privacy emerged, utilizing cryptographic techniques. However, neither method proved entirely effective in addressing the challenge of data mining while safeguarding individuals' privacy. As a result, there has been a growing recognition among policymakers and businesses that k-anonymity [9] is a more practical standard for defining privacy, and it has been widely accepted as such.

1.2 Attack On Data (re-identification attack)

The data presented in the rightmost circle of Figure 1 includes information such as the name, address, ZIP code, birth date, and gender of each voter. By using the ZIP code, birth date, and gender, it is possible to link this information to medical records, allowing for the association of diagnoses, procedures, and medications with specific individuals. One notable example of this is William Weld, who was serving as the Governor of Massachusetts when his medical records were included in the GIC data. Based on the Cambridge Voter List, there were only three males who shared his birth date, and he was the only person living within his 5-digit ZIP code.

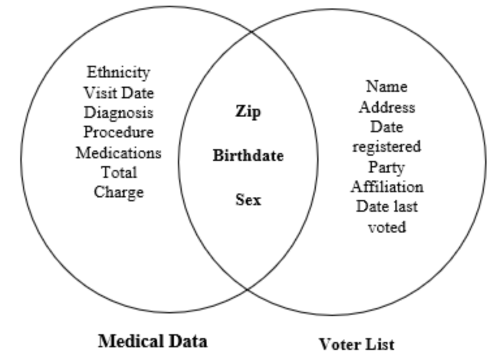

Figure 1: Linking to data with quassi-identifier

The re-identification process can be demonstrated by explicitly linking (or "matching") on shared attributes, as seen in the example that was just presented. The research that was provided in this paper demonstrates that one way to protect oneself from an assault of this kind is to modify the information that was disclosed so that it can be mapped to a wide variety of potential individuals. When there are more options to choose from, the process of linking them becomes murkier, which ultimately results in the data being less identifiable.

1.3 Proposed Contribution

The contributions of this work include:

- Releasing a substantial amount of data that can be utilized by various organizations for business or research purposes.
- Implementing the Dissimilarity Tree algorithm for selecting centroids in clustering processes.

- Protecting the privacy of individuals involves safeguarding the released data against inference and potential attacks, ensuring that personal information is not compromised.

II. LITERATURE SURVEY

In recent years, the confidentiality and privacy of healthcare information have become increasingly important due to the rapid development of digital healthcare systems and the growth of data being collected and stored. Various studies have focused on addressing these concerns and implementing strategies to ensure the protection of sensitive health data.

In their study, Rizwan et al. [1] introduced a risk monitoring strategy aimed at preserving the confidentiality of healthcare information. The authors developed a framework that utilized multi-objective optimization algorithms to identify potential threats and establish countermeasures for mitigating risks associated with data breaches in healthcare systems.

El Zarif and Haraty [2] focused on information preservation in healthcare systems, exploring various techniques to safeguard sensitive data. The authors highlighted the importance of data integrity, availability, and confidentiality as key factors in establishing a secure healthcare environment, emphasizing the need for robust cybersecurity measures to protect patient information from unauthorized access.

In their study, Andrew et al. [3] presented an approach to the publication of large data that protects privacy by utilising Mondrian anonymization techniques and deep neural networks. The proposed method sought to secure the privacy of patients by transforming sensitive information into a more general format, thereby making it difficult for adversaries to re-identify individuals from the published dataset.

Dhasarathan et al. [4] investigated the analysis of COVID-19 health data and the preservation of personal information by implementing a homomorphic privacy enforcement approach. The authors proposed a method that enabled data analysis without revealing the underlying sensitive information, ensuring that privacy was maintained throughout the process.

In their research, Haraty et al. [5] proposed a hash-based assessment and recovery algorithm for healthcare systems. This innovative approach was designed to improve data security and reliability in healthcare settings. The proposed method employed cryptographic hash functions to secure patient information and ensure data integrity, while also providing a recovery mechanism in case of data corruption or loss.

Liu et al. [6] tackled privacy-preserving raw data collection in the context of the IoT without the need for a trusted authority. The authors proposed a scheme that maintained data privacy during the collection process, ensuring that sensitive information remained secure even when transmitted across untrusted networks.

Daries et al. [10] discovered that removing student identifiers from the 2013-2014 edX dataset improved the profiles of students and grade distributions. In their 2015 study of the edX dataset, According to Angiuli et al. [14], anonymization techniques that rely heavily on suppression can distort column values, whereas techniques that rely heavily on generalization can distort column correlations. High-achieving students have distinctive, quasi-identifying characteristics, making it easier to conceal their academic records.

Ghinita, Tao, and Kalni [12] consider l-diversity, where k records share a quasi-identifier. Generalisation and permutation-based l-diversity methods are noted. Generalisation substitutes quasi-identifier values with semantically coherent values [13]. Generalisation loses correlations between characteristics, reducing data value and information loss.

Samarati and Sweeney [15] propose the k-anonymity concept, a strategy to release person-specific data while preserving privacy. They demonstrate how to achieve k-anonymity using generalization and suppression techniques, while minimizing the degree of data distortion. Sweeney et al. [16] also suggest the use of k-anonymization, emphasizing the importance of ensuring each person is connected to a set of at least k records.

Meyerson and Williams [17] investigated k-dimensional perfect matching, demonstrating that it is NP-hard and developing two approximation algorithms. Li et al. [18] examined k-anonymization and proposed a differential privacy method using random sampling. Dankar et al. [19] explored practical challenges of applying differential privacy to health data, suggesting a model for its application.

Friedman and Schuster [20] enhanced the Data Mining-Decision Tree algorithm for better privacy and data accuracy. Simi, Nayaki, and Elayidom [21] addressed business and research-oriented fields, discussing the use of k-anonymization to protect privacy. They proposed three effective algorithms based on a series of studies and systematic comparisons.

III. PROPOSED SYSTEM

1.4 Proposed System Architecture

The suggested architecture for the system is displayed below in Figure 2. The data preprocessing step is followed by the selection of k-anonymity attributes, and if no attacks are detected, the dataset is deemed valid. However, in the event of an attack, preventive measures must be implemented before dissimilarity tree centroid selection takes place. Subsequently, the clustering and K-anonymity-based data production steps are executed, leading to the presentation of evaluated graphs.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v11i8.8005
Article Received: 16 May 2023 Revised: 20 July 2023 Accepted: 12 August 2023

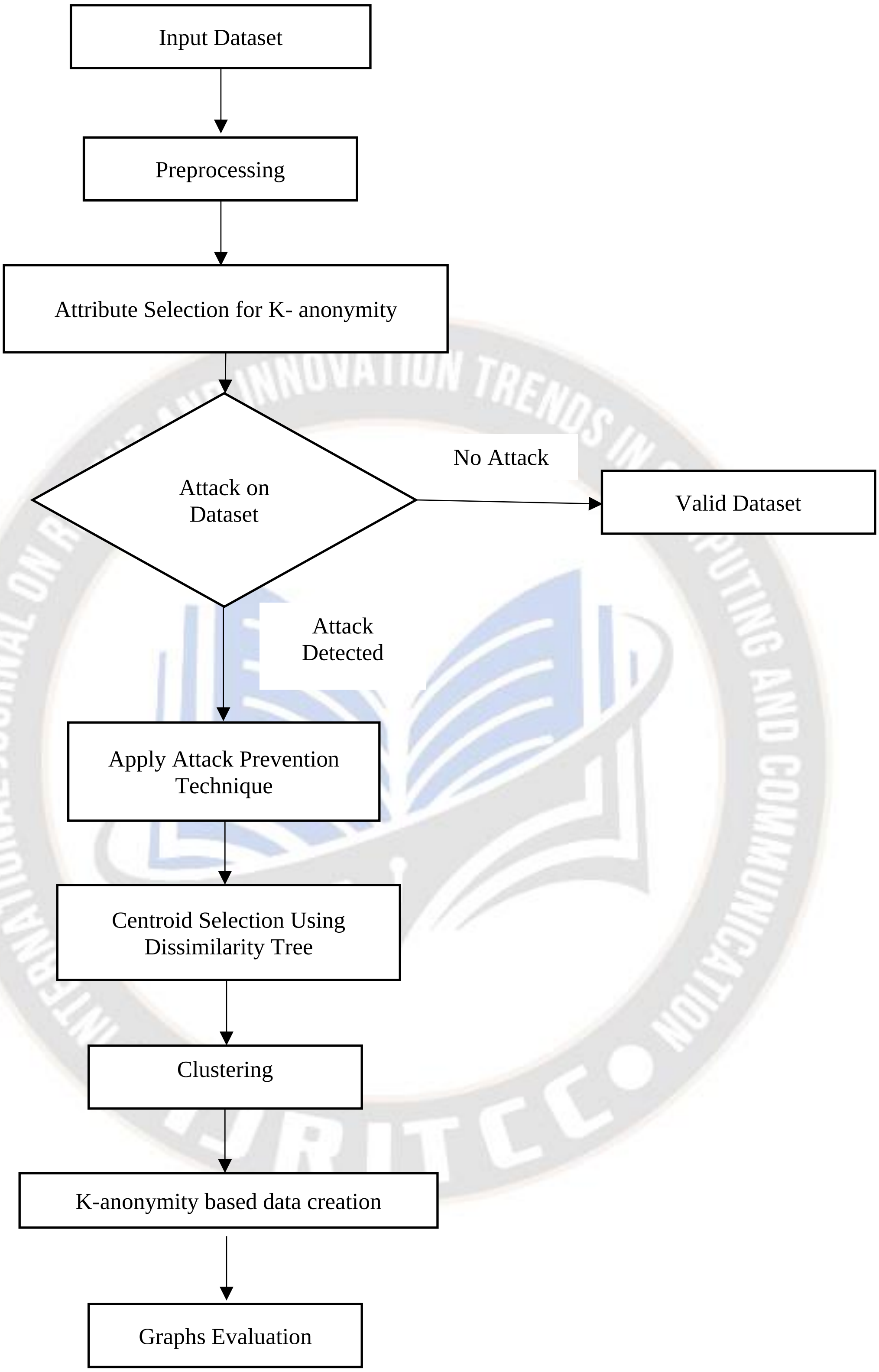
Input Dataset
Preprocessing
Attribute Selection for K- anonymity
Attack on Dataset
No Attack
Valid Dataset
Attack Detected
Apply Attack Prevention Technique
Centroid Selection Using Dissimilarity Tree
Clustering
K-anonymity based data creation
Graphs Evaluation

Figure 2: Proposed System Architecture

DOI: https://doi.org/10.17762/ijritcc.v11i8.8005
***Article Received**: 16 May 2023 **Revised**: 20 July 2023 **Accepted**: 12 August 2023*

1.5 Algorithm

Algorithm 1: Improved K-Anonymity Algorithm Based on Clustering

Input: The initial dataset is denoted by S, while the degree of generalization is represented by the parameter K.

Output: Anonymized table AT with K

Steps:

```
1: Initialize clusters as empty, and randomly pick a record index from S
2: While the length of S is greater than or equal to K, do the following:
3: If the number of clusters is less than 1, then
4: Select the next record as the furthest record from the current record
5: Else
6: Employ the Dissimilarity Tree algorithm to determine the appropriate degree of generalization for each attribute in the dataset.
7: End if
8: While the number of clusters is less than K, do the following:
9: Choose the best record and insert it into the cluster
10: End while
11: While the length of S is greater than 0, do the following:
12: Get a record, choose the best cluster, and insert the record into it
13: End while
14: End while
```

Figure 3 illustrates the steps involved in the clustering algorithm. Initially, a centroid is selected from a given number of clusters. Next, the distance between each object and the centroid is calculated. Based on these distances, objects are grouped together. If no objects change groups, the process ends; otherwise, a new centroid is selected, and the steps are repeated.

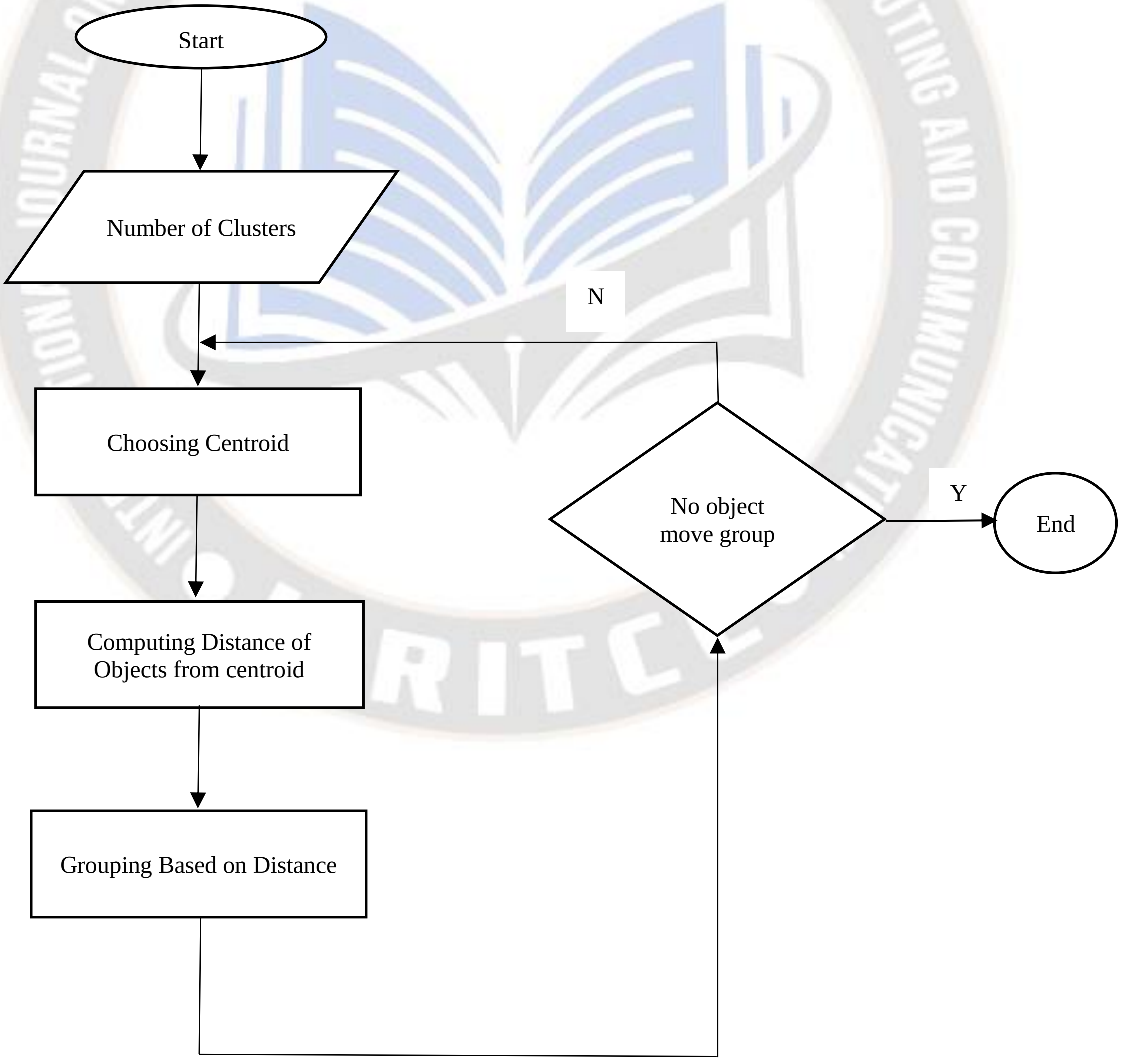
Start
Number of Clusters
Choosing Centroid
Computing Distance of Objects from centroid
Grouping Based on Distance
No object move group
N
Y
End

Figure 3: Centroid Selection Technique

***Article Received**: 16 May 2023 **Revised**: 20 July 2023 **Accepted**: 12 August 2023*

IV. RESULT ANALYSIS

1.6 Dataset Description

Adult dataset:

The United States Census Bureau's Data Extraction System was utilized to gather the dataset, which contains 32,561 entries and incorporates 15 distinct attributes. These attributes include age, education, work class, final weight, marital status, education number, occupation, race, relationship, sex, hours per week, capital gain, native country, capital loss, and salary class. Our work specifically addresses the protection of location privacy by focusing on the location-based attributes present in the dataset.

1.7 Performance Parameters

To establish a consistent evaluation criterion and enhance data quality measurement, we utilize NCP as the data quality assessment metric. As previously discussed, NCP is defined based on attributes, aligning with the distance definition in clustering mentioned earlier. To determine the information loss for both individual generalized records and the entire dataset, it is necessary to normalize NCP for both. This can be achieved by following the below method:

$$NCP(dataset) = \frac{\sum_{i=1}^{n} NCP(record_i)}{n} \quad ...(1)$$

$$NCP(record) = \frac{\sum_{i=1}^{d} NCP(attribute_i)}{d} \quad ...(2)$$

In the aforementioned equations, I represent the no. of attributes in a record, whereas n represents the total number of records in a dataset. By taking the average of the NCP for both records and datasets, the normalization process simplifies the assessment of data quality, enabling easier comparisons between different algorithms.

1.8 Result

Figure 4 shows the NCP comparison graph of proposed system and existing system for different value of K.

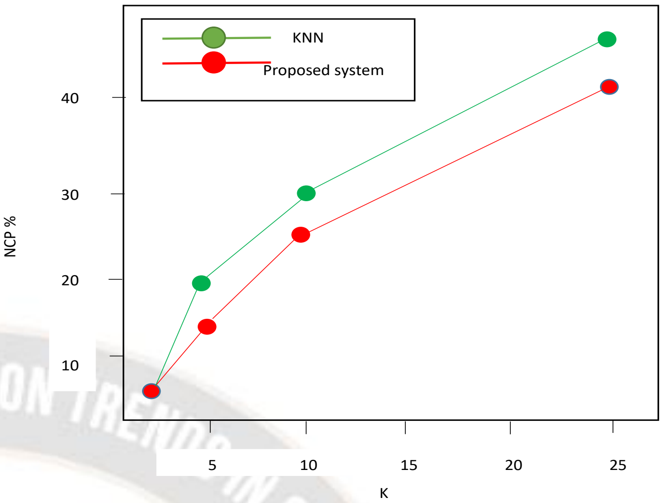

Figure 4: NCP Comparison Graph

First, we are going to keep the same amount of records in the dataset (n), but we are going to experiment with different values for the anonymity parameter (k). As can be seen in Figure 4, when n equals 5000, our method performs noticeably better than kNN in terms of the data loss measures, which ultimately results in superior data quality. Furthermore, as k value increases, the equivalence class includes more records, leading to a greater loss of information for the anonymization techniques currently in use.Furthermore, we conducted another test on a larger dataset, where we used nearly the entire Adult dataset as the focus of our study. When compared to other algorithms, even when k is set to 25, the information loss caused by our approach is still kept below 50%, which is approximately 15% less than that caused by the other techniques. The steady performance is another indication of the smooth curve.

Figure 5 presents a time comparison graph, which is used to demonstrate the execution time of each algorithm. As depicted in the figure, the proposed algorithm takes longer to execute than the existing clustering algorithm. In the graph, the x-axis represents the algorithms, while the y-axis indicates the time (in seconds) required for algorithm execution.

***Article Received**: 16 May 2023 **Revised**: 20 July 2023 **Accepted**: 12 August 2023*

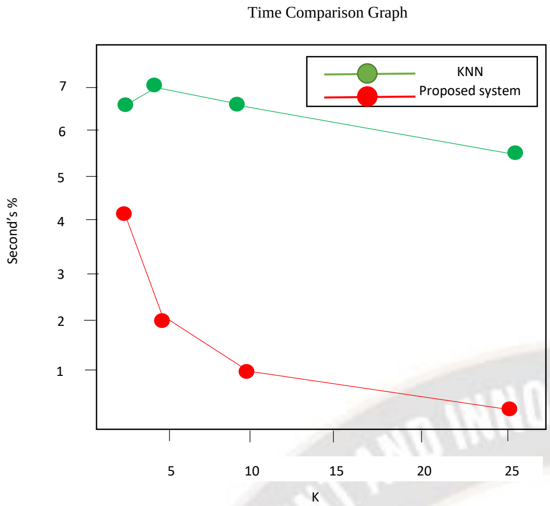

Figure 5: Graph Comparing Time Required to run algorithms for different K values

V. CONCLUSION

This paper presents a novel cluster-based k-anonymity algorithm for data privacy publishing, addressing the limitations of traditional k-anonymity methods. The proposed method reduces information loss, enhances dataset accuracy, and improves data quality for data mining applications. By utilizing a dissimilarity tree to select the optimal initial centroid, the approach increases system efficiency and accuracy while reducing computational time.

In addition to its current benefits, the proposed algorithm has the potential for further enhancements and applications. Future work may focus on refining the algorithm to handle large-scale datasets more efficiently, improving scalability. Additionally, integrating the algorithm with other privacy-preserving techniques, such as differential privacy and l-diversity, could lead to more robust privacy protection mechanisms. Furthermore, exploring the algorithm's applicability across various domains, such as finance, social networks, can broaden its utility and impact on data privacy. By continually advancing and expanding the scope of this cluster-based k-anonymity algorithm, researchers and practitioners can better address the growing challenges of data privacy and security in an increasingly data-driven world.

REFERENCES

[1] Rizwan M, Shabbir A, Javed AR, Srivastava G, Gadekallu TR, Shabir M, et al. Risk monitoring strategy for confidentiality of healthcare information. Comput Electr Eng. (2022) 100:107833. doi: 10.1016/j.compeleceng.2022.107833

[2] El Zarif O, Haraty RA. Toward information preservation in healthcare systems. Innov Heal Informat A Smart Healthc Prim. (2020) 163–85. doi: 10.1016/B978-0-12-819043-2.00007-1

[3] Andrew J, Karthikeyan J, Jebastin J. Privacy preserving big data publication on cloud using mondrian anonymization techniques and deep neural networks. In: 2019 5th International Conference on Advanced Computing and Communication Systems. (2019). p. 722–7.

[4] Dhasarathan C, Hasan MK, Islam S, Abdullah S, Mokhtar UA, Javed AR, et al. COVID-19 health data analysis and personal data preserving: a homomorphic privacy enforcement approach. Comput Commun. (2023) 199:87–97. doi: 10.1016/j.comcom.2022.12.004

[5] Haraty RA, Boukhari B, Kaddoura S. An effective hash-based assessment and recovery algorithm for healthcare systems. Arab J Sci Eng. (2022) 47:1523–36. doi: 10.1007/s13369-021-06009-4

[6] Liu YN, Wang YP, Wang XF, Xia Z, Xu JF. Privacy-preserving raw data collection without a trusted authority for IoT. Comput Networks. (2019) 148:340–8. doi: 10.1016/j.comnet.2018.11.028

[7] Tiancheng Li, Ninghui Li, Senior Member, IEEE, Jia Zhang, Member, IEEE, and Ian Molloy "Slicing: A New Approach for Privacy Preserving Data Publishing" Proc. IEEE Transactions On Knowledge and Data Engineering, Vol. 24, No. 3, March 2012.

[8] Sampurnanand Dwivedi, Vipul Singhal. (2023). A Study of Responsive Image Denoising Algorithm . International Journal of Intelligent Systems and Applications in Engineering, 11(3s), 286–291. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/2686

[9] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati On K-Anonymity. In Springer US, Advances in Information Security (2007).

[10] Latanya Sweeney. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5):557–570, 2002.[1] Swagatika Devi, K-ANONYMITY: The History of an IDEA International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-1, March 2011.

[11] Daries, J. P., Reich, J., Waldo, J., Young, E. M., Whittinghill, J., Ho, A.D., Seaton, D. T., Chuang, I. Privacy, anonymity, and big data in the social sciences. Communications of the ACM 57(9): 56-63,2014.

[12] Angiuli, O., Blitzstein, J., and Waldo, J. How to de-identify your data. Queue 13, 8 Sept. 2015.

[13] G.Ghinita, Y. Tao, and P. Kalnis, "On the Anonymization of Sparse High-Dimensional Data," Proc. IEEE 24th Int'l Conf. Data Eng. (ICDE), pp. 715-724, 2008.

[14] Ghinita, Member IEEE, Panos Kalnis, Yufei Tao," Anonymous Publication of Sensitive Transactional Data" in Proc. Of IEEE Transactions on Knowledge and Data Engineering February 2011 (vol. 23 no. 2) pp. 161-174.

[15] Ana Oliveira, Yosef Ben-David, Susan Smit, Elena Popova, Milica Milić. Improving Decision Quality through Machine Learning Techniques. Kuwait Journal of Machine Learning, 2(3). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/202

[16] Zhao FeiFei, Dong LiFeng, Wang Kun, Li Yang, Study on Privacy Protection Algorithm Based on K-Anonymity, Physics Procedia, Volume 33, ISSN 1875-3892, 2012

[17] Samarati, Pierangela; Sweeney, Latanya, “Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression.” Carnegie Mellon University. Journal contribution, 2018.

[18] Sweeney L, k-anonymity: a model for protecting privacy. Int J Uncertain Fuzzy Knowledge Based System 10 (5):557–570,2002.

[19] Meyerson A, Williams R, On the complexity of optimal k-anonymity. In: PODS ’04: proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on principles of database systems, pp 223–228,2004.

[20] Mr. Rahul Sharma. (2013). Modified Golomb-Rice Algorithm for Color Image Compression. International Journal of New Practices in Management and Engineering, 2(01), 17 - 21. Retrieved from http://ijnpme.org/index.php/IJNPME/article/view/13

[21] Li, N., Qardaji, W. H., & Su, D. Provably private data anonymization: Or, k-anonymity meets differential privacy. 2011.

[22] Dankar, F. K., & El Emam, K. (2012, March). The application of differential privacy to health data. In Proceedings of the 2012 Joint EDBT/ICDT Workshops (pp. 158-166). ACM.

[23] Friedman, A., & Schuster, A. (2010, July). Data mining with differential privacy. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 493-502). ACM

[24] Simi, Ms & Nayaki, K & Elayidom, An Extensive Study on Data Anonymization Algorithms Based on K-Anonymity. IOP Conference Series: Materials Science and Engineering. 225. 012279. 10.1088/1757-899X/225/1/012279,2017

[25] Isabella Rossi, Reinforcement Learning for Resource Allocation in Cloud Computing , Machine Learning Applications Conference Proceedings, Vol 1 2021.

[26] Feng Bo, Hao Wenning, Chen Gang, Jin Dawei, Zhao Shuining, “An Improved PAM Algorithm for Optimizing Initial Cluster Centre,” IEEE, 2012, 978-1-4673-2008- 5/12.