

Cost-effective internet of things privacy-aware data storage and real-time analysis

Femi Abiodun Elegbeleye, Munienge Mbodila, Omobayo Ayokunle Esan, Ife Elegbeleye

Department of Information Technology Systems, Faculty and Information Technology Systems, Walter Sisulu University, Queenstown, South Africa

Article Info

Article history:

Received Jul 26, 2022

Revised Jan 31, 2023

Accepted Feb 9, 2023

Keywords:

Data anonymity

Data privacy

Differential privacy

Internet of things

Machine learning

ABSTRACT

It has been estimated that about 20 billion internet of things (IoT) devices are currently connected to the Internet. This has led to voluminous data generation which makes storing, managing, and decision making on data to be challenging. Hence, exposes users' privacy to be vulnerable to unauthorized people. To address these issues, this research proposed cost-effective storage for keeping and processing the IoT data in real-time. The proposed Fframework utilized a reliable hybridised data privacy model to protect the personal information of users. An empirically evaluation was done to identify the best models using data k-anonymity (KA), l-diversity (LD), t-closeness (TC), and differential privacy (DP). The performance evaluation of cloud computing and fog computing was done through simulations. The results obtained show that the combination of two data privacy models: differential privacy and k-anonymity models performed better than any individual model and any other combined models in the protection of users' personal information. Lastly, fog computing was found to perform better than the cloud in terms of latency, energy consumption, network usage and execution time. In conclusion, the current study strongly recommends the use of hybridised privacy model of differential privacy (DP) and k-anonymity (KA) for the protection of IoT generated data privacy.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Femi Abiodun Elegbeleye

Department of Information Technology Systems, Walter Sisuu University

Eastern Cape, South Africa

Email: felegbeleye@wsu.ac.za

1. INTRODUCTION

In recent years, advances in wireless sensor networks (WSNs) has given birth to a new computing paradigm known as the internet of things (IoT) [1]. IoT is currently gaining momentum and is one of the emerging 21st-century technologies. It is used to enhance the connection of people and things at any time, any place, with anything and anyone, typically with the use of some path/network and any service. The importance of IoT is to build "a better world for human beings [2]," where the objects around us understand our wants, likes and needs. IoT eases the accessibility to information. Today, the Internet has created a link for the exchange of data, information's, opinions and news among over 100 countries [3]. The primary drivers of IoT are large organizations and industries that greatly benefit from the predictability and foresight afforded by the ability to monitor all objects through the service chains in which they are embedded. The applications of IoT have many remarkable applications in our day to day lives, including smart cars, home appliances and security, health-tracking wearable devices and weather monitors.

With the advent of IoT, there has been a tremendous proliferation of smart devices and applications which generate massive data called "Big data" on a daily or weekly basis depending on the application. They include sensors, devices, social media, temperature sensors, and health care applications and so on. They

constantly generate a huge amount of data characterized by structured, unstructured, or semi-structured [4] outputs which are deemed insufficient for the traditional databases in terms of storage, processing and analysis. However, this generated data is very useful to the organizations that own them, and data analysts are playing critical roles in improving their usefulness to further improve the growth of companies and to enhance decision making, day to day communication, relationship and building a good network among various customers. Currently, several organizations have benefited greatly from the development of IoT technology and large volumes of data generated are being assembled and transmitted from one device to another, device to business systems, and seldom from device to humans.

In spite of the benefits accrued to IoT, handling this data has become a major challenge [5] and the technology is faced with several challenges which include security, privacy, scalability and so on. These challenges are in terms of storage, analysis and processing of large volumes of data emanating from numerous data resources or heterogeneous IoT devices [6]. Moreover, these generated data are greatly prone to the risk of data theft, identity, manipulation of devices, falsification of data, and manipulation of server/network owing to inappropriate privacy models put in place in securing users' personal information. Therefore, to properly manage the high volume of data generated and to avoid the violation of data or misuses, proactive and data privacy-preserving measures must be taken to store, process and publish data to prevent breaches of sensitive information and other types of privacy and security incidents. With advancements in information and communications technology (ICT) in the past few decades [7], various computing models or paradigms such as cloud computing have come to the limelight. Cloud computing facilities are centred on the "data centre" procedure, where networks of hundreds of thousands of servers are assembled to provide services.

In addition to several dedicated servers positioned in data centres, there are also billions of seldom used personal computers (PCs) belonging to private owners and organizations worldwide, usually used for a few hours per day [8]. Their massive unused compute and storage capabilities can be combined as a substitute cloud fabric for the provision of extensive cloud services and predominantly infrastructure services. Cloud computing (CC) is like a conveyor that carries information and data for various users and offers services that can be utilized at a low cost. Nowadays, the growth of large and small scale companies largely depend on their data, maintaining these data requires a lot of money and resources [9]. Most of these organisations cannot afford the huge cost and maintenance of in-house built IT infrastructure and backup support services. Thus, cloud computing stands as a cheaper and best alternative to store their generated data due to data storing efficiency, low maintenance and computational cost which has attracted most individuals, organisations or even governments in recent years. CC plays a widespread measure of data accessibility where various users can store information via the cloud and pay to get it to reproduce for further use when needed [10]. However, CC has its challenges which are enormous, and most consumers and establishments are uninformed about the third-party vulnerabilities of their stored data into the cloud.

Considering the above background and the nature of IoT generated data, the generated data should be managed properly using cost-effective storage, processed, and analysed in real-time and personal information kept secure. To strive to achieve the stated instances is, therefore, the intention of this research. IoT devices generate a high volume of data on a daily or weekly basis and thus, handling this data has become a major challenge. The data generated require huge storage space and real-time data analysis for dynamic decision making. The data are characterised by structured, unstructured, or semi-structured [11] information which is considered insufficient for the traditional databases in terms of storage, processing and analysis. That is, the data contains useful and meaningful hidden information whose behavioural patterns are very hard to detect. Thus, providing appropriate storage architecture to store the generated data and algorithm for real-time data analysis is highly important to discover the hidden knowledge and aid dynamic decision making.

Moreover, the data generated contains important personal information of users and this information is not protected. Such data can easily be collected, and personal information exploited to endanger the privacy of the owners. As data has become a valuable asset used in promoting businesses and an effective source of decision making [12], security breaches, data leakage and cybercrime have also risen sharply globally due to ubiquitous modes of access. For IoT generated data, the intuition is that, though data cannot be completely secured, the privacy of the data owners should always be protected. Though several privacy models and security approaches used to protect data from unauthorized access exist, each has its strengths and limitations which can easily be exploited. In particular, "k-anonymity (KA) fails to prevent the background knowledge and homogeneity attacks, suffers from attribute linkage and record linkage and long processing time [13], l-diversity is prone to skewness and similarity attacks while t-closeness (TC) loses the correlation between changed attributes since each attribute is generalised separately. In this case, the data utility is damaged when it is very small. Lastly, in differential privacy, data utility may be reduced, a data miner is only allowed to pose aggregate queries and the probability of attacking both the database by an adversary is not taken into account". Consequently, there is the need for a secured and effective privacy model to protect personal information in published data. This paper, therefore, uses data privacy model which is the combined cost-effective storage

architecture for data management and data privacy. The proposed model assists in ensuring that the voluminous data is effectively managed, ubiquitously accessed, and personal information is well-protected.

2. RELATED WORKS

Different works have been done in literature on internet of things and data privacy. Table 1 shows the summary of existing work with their remarks, solutions, models used, attacks and data utility. From our studies and previous research, it is evident that the differential data privacy model has proven to be more secure. Additionally, information loss was observed across the four data privacy models utilized in this investigation, but differential data privacy model outperformed the others.

Table 1. Summary of related works on privacy models

Ref.	Remarks	Solutions	Models Used	Attacks	Data Utility
[14]	The study Proposed the use of multiple differential privacy model, on real-time analysis.	The approach helps to offer better and stronger data privacy protection.	Multiple DP	N/A	Reduced Information loss
[15]	a. Studied show that increase data utility has to be parallel with data privacy. b. It was also suggested in the study that the SMR layer show a low loss of information. c. Lightweight encryption is used in the model in other to protect the data. d. Issues with scalability was settled in the study.	From the study, it was noted from the results that CPU consumption, RAM usage and lastly information loss was reduced.	SMR model Randomization Perturbation	N/A	Information loss
[16]	EHRs system is prone to privacy violations, especially when stored in healthcare medical servers.	This study provides a discussion on several anonymity techniques designed for preserving the privacy of microdata	TC, LD and KA	N/A	Information loss
[17]	From the study data utility was little and the model cannot be recommended in many areas.	From the study, a new novel model of protecting data was presented.	Slicing model KA, and Anatomy model	Skewness attack, Sensitivity attack, Similarity attack	Information loss
[18]	The study presented a personalized approach or method of preserving the data using (α, ω) -anonymity model. Exploring the use of QI attribute and sensitive attribute.	From the study, the core solution provided was that privacy is based on the measure from the individuals' needs and requests and this was fully achieved in the study.	Anonymity model (α, ω)	Similarity attack	Information loss
[19]	The study showed the various data privacy and security issues and possible solutions.	Homomorphic encryption, Storage path encryption and Attribute-based encryption access control were used in the study.	KA, LD, TC	Background knowledge attack, similarity attack and reconstructions attack	Information loss
[20]	Data privacy, security, and data management of published data was the fulcrum of the study.	Anonymization technique was proposed as an efficient method to realize privacy-preserving.	Suppression, perturbation, and DP	Reconstruction's attack, tracing attack and similarity attack	Information loss
[21]	Privacy issues in published stored data with major challenges discussed in the study.	The study suggested that the scalability and efficiency of these models are improved to provide a suitable solution	Encryption-based, anonymization and DP models	Probabilistic attack, reconstructions attack tracing attack and similarity attack	Information loss
[22]	The proposed method has helped in decreasing the average re-identification risks between 100% and 2.33%.	The study result shows that re-identification risks are far less ranging from 100% to 2.33%.	δ -Presence, TC, LD, and KA	Background knowledge attack and similarity attack	Information loss
[23]	The need to apply suitable privacy models to the published data becomes very necessary.	Semantic anonymization approach methods were proposed.	KA, LD	Background knowledge attack	Decreases the data utility
[24]	The need to use micro aggregation leading to adding and deleting some of the data and records is updated.	Dynamics micro aggregation method. MDAV	KA	Background knowledge attack Preventing Identity disclosure	Information loss

3. METHOD

The method of data collection was a secondary data approach, with datasets being analysed, respectively. The researchers adopted qualitative research methods. Qualitative research is used to understand and explain phenomena on how to better interpret the data. Also, an in-depth literature review was carried in other to identify the problem under study and to have a better background knowledge of the data to analyse and a better approach in solving the identify problem. Quantitative research deals with the numerical analysis of collected data for decision making. In this research quantitative data were collected from my empirical analysis and simulations.

3.1. Tools and technologies

Three different software packages were used in our analysis. The three were as follows. ARX open-source software, Orange3 open-source software and iFogSim open-source software, these three tools help in the presenting of our research results/findings in a more meaningful way. ARX open-source software was used for the experiment and the software supports the transformation of the dataset in a way that ensures the data conforms to user-specific privacy models and risk thresholds that hinder attacks that may result in privacy breaches. ARX can be utilized to eliminate direct identifiers (e.g., names) from datasets and to put additional restrictions on indirect identifiers. Indirect identifiers (or quasi-identifiers, or keys) are attributes that do not directly classify a person but may combine with other indirect identifiers to produce an identifier that can be utilised for connection attacks. There is a usual assumption that data identifiers are accessible to a third party (in some form of background knowledge), and it is difficult for them to be removed from the dataset (e.g., because they are required later for analyses). Lastly, the ARX software supports methods for the protection of sensitive attributes and sensitive disclosure attacks using and semantic privacy models [25].

3.1.1. Orange3 software

Orange3 tool was utilised. It is an open-source software implemented in python and C++ Programming languages. It is a visual programming front-end for explorative information examination and perception. It underpins documents in .csv. It is a segment based visual programming for information mining, ML, and information investigation. Its parts are called gadgets and range from information perception subset choice and pre-preparing to exact assessment of learning calculations and prescient displaying.

3.1.2. iFogSim software

iFogSim is an open-source software that was used in performing the simulation. iFogSim has different types of physical entities such as device or node, sensor, and actuator. The logical entities used in modelling applications include AppModule models used for IoT services, the AppEdge model for data dependency among services, and the Tuple models which oversee entities communication. The simulations and results are presented in result session.

3.2. Data privacy and analytics mode choice

This chapter is aimed at selecting the best performing data privacy model. ARX software was used to analyse the data. ARX software provides a platform where the data privacy models can be used and to test the performance and evaluate the data. For effective proof of concepts, this research used IoT data generated from healthcare as a case study. This is because about 60% of the global healthcare organizations have incorporated IoT technology into their daily use to better enhance the overall healthcare working environment. These IoT devices are effective in helping healthcare practitioners and patients to monitor, track, trace medical reports of patients, analyse the hospital details, record patient's health status in a consistent manner which would otherwise be difficult for physicians alone to do. Accordingly, this greatly reduces the cost of healthcare and helps to minimise the chances of errors in patients' health records.

Moreover, for the data privacy model, different data privacy models were employed based on existing models such as k-anonymity (KA), l-diversity (LD), t-closeness (TC), and differential privacy (DP), for the test. For cost-effective data storage, the fog and the cloud data centres were used while empirical analysis of some ML algorithms was conducted to select the best performing algorithm for usage in the real-time data analysis to help in effective and reliable decision making in terms of classification accuracy and time efficiency. The idea is to automate the building of a data analytics model that uses the algorithm to learn from data interactively. By choosing the best model, decision making can be improved over time with less human intervention this is as shown in Table 2.

In this research implementation, the data analysis was performed qualitatively conducted on the collected data using defined metrics in Table 3. These experimental metrics are used to show the performance of the proposed model on the collected data. The detailed experiments utilizing these metrics are in sections 4.1. to 4.3. respectively.

Table 2. Considered privacy models

Model	Motivation
KA	Implementation is easy and fewer chances of data identification.
LD	It summarizes data and prevents data attribute disclosure.
TC	it promotes sensitive value variation with a group, disclosure of attributes and skewness attacks prevention.
DP	Most effective privacy model, add noise without loss of information and minimize data utility.

Table 3. Data privacy parameters

Parameter	Description
Receiver operating characteristic (ROC) curve	It is a graphical plot of TP and FP that shows a classification model's performance at all thresholds of the classification.
Area under the curve (AUC)	Shows the classification model's ability to differentiate between classes. It is a measure of the model's performance. Higher AUC signifies better model performance.
Re-Identification Risks Analysis	It is a view for quantifying the risks associated with attacks on the privacy models.
Brier skill score	It quantifies relative accuracy against reference accuracy of a classification model
Prosecutor Attacker Method	The first stage of the re-identification risks model and measures the thresholds of the attacker model and provides a record of risks, the highest risks level and the success rate of the anonymization process.
Journalist Attacker Method	The second stage of the re-identification risks model and measures the thresholds of the attacker model and provides a record of risks, the highest risks level and the success rate of the anonymization.
Marketer Attacker Method	The final stage of the re-identification risks model measures the thresholds of the attacker model and provides a record of risks, the highest risks level and the success rate of the anonymization.
True positive (TP)	The classification model correctly classified risky class as truly risky.
True negative (TN)	The classification model correctly classified as not a risky class as truly not risky.
False-positive (FP)	The classification model incorrectly classified a risky class as not risky
False-negative (FN)	The classification model incorrectly classified a not risky class as risky.

4. RESULTS AND DISCUSSION

This subsection presents the results of the analysis for the four selected data privacy models. The analysis was performed using defined metrics in Table 3. Accordingly, Table 4 shows the results of the AUC, Brier skill score and risk analysis of the KA, LD, TC and DP data privacy models. The results show that ARX performed substantial extensive measurements and attacks were predicted from the four attributes of Id, age, gender, and income from a diabetic dataset.

4.1. BSS

Table 4 shows the relative accuracy of the anonymization model where BSS achieved 0.00037 for KA, 0.00931 for LD, -0.43760 for TC and 0.0506 for DP. The BSS ranges between -0.43760 and 0.0506. The indication is that all the models provided a high degree of protection for the given dataset or record. However, based on the results, TC is not recommended due to its inability to handle large scale datasets as seen from the literature. The resulting privacy-preserving models of KA, LD and DP exhibited high protection power. Accordingly, from all the values obtained, the DP privacy model performed better in terms of accuracy with a value of 0.0506 obtained for its BSS which was closest to 1, this suggests that the DP model performed better in terms of accuracy. To obtain a more efficient privacy model, DP can be combined with KA [13].

Table 4. Summary of BSS, AUC, and risk analysis

Model	KA	LD	TC	DP
AUC	53.61%	50.11%	46.62%	45.73%
BSS	0.00037	0.00931	-0.43760	0.0506
Risk Analysis	0.52125	0.16084	0.10866	0.08065

4.2. Receiver operating characteristics curves

In the context of the experiment conducted, the data privacy models trained on unmodified data attained a ROC AUC of about 53.61% for KA, 50.11% for LD, 46.62% for TC and 43.73% for DP. Compared to the initial performance the relative ROC AUC was between 45.73% and 53.61%. This is shown in Figures 1-4 for each of the models considered in this study. Accordingly, KA appears to be the best performing with 53.61% which was the highest accuracy obtained. The implication is that KA offers effective data protection in terms of anonymization than the other models considered. Thus, KA can be combined with DP to form a hybrid model that can offer a high degree of protection. This is because DP is the most accurate in terms of the BSS and re-identification risk while KA has a good threshold in terms of the ROC AUC. Thus, combining the two privacy models could go a long way to offer a high degree and more efficient privacy protection.

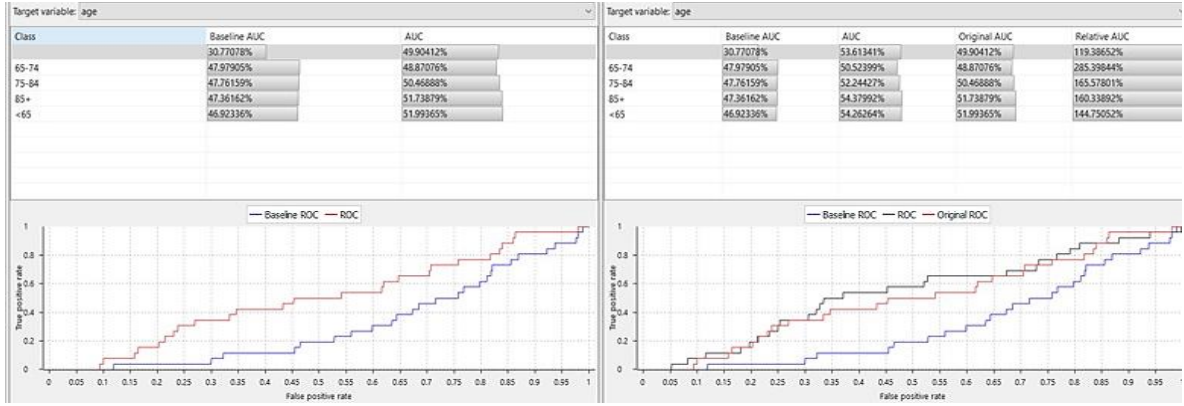


Figure 1. KA AUC performance

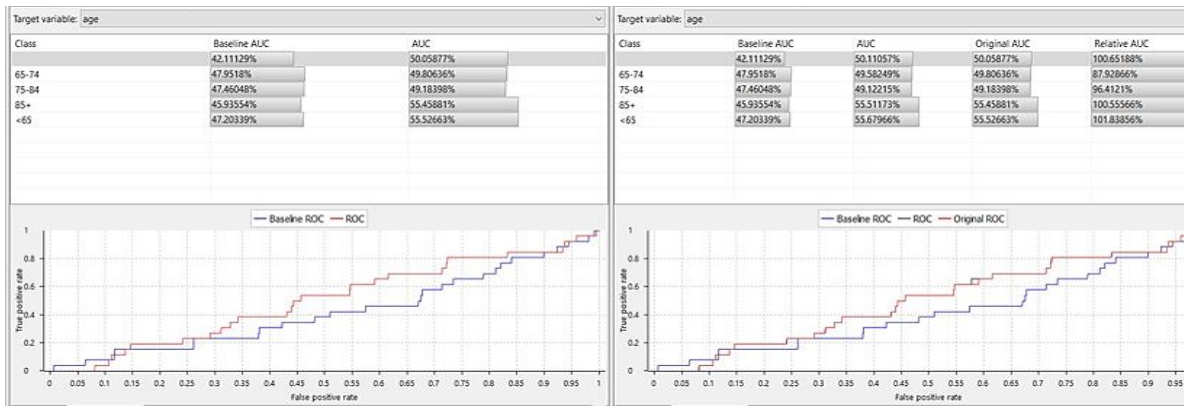


Figure 2. LD AUC performance

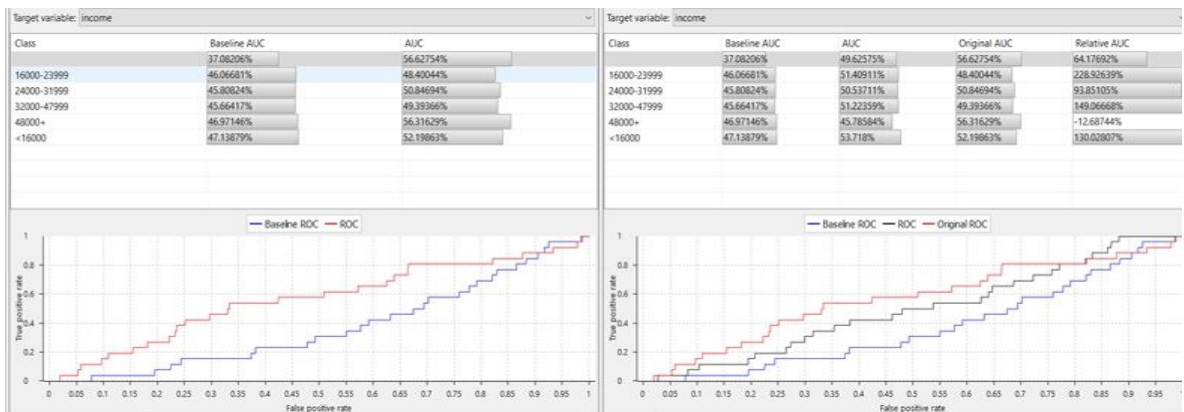


Figure 3. TC AUC performance

4.3. Re-identification risk

This summarizes the risks of all records in a dataset in terms of each possible risk level and the number of affected records is shown in Table 5. Based on the experiment conducted, the re-identification risk obtained was 0.52125 for KA, 0.16084 for LD, 0.10866 for TC and 0.08065 for DP. The summary is shown in Table 5. In Table 5, the re-identification risks value for the DP privacy model is 0.08065. DP value is smaller than values obtained for other models, signifying its suitability in protecting the privacy of our data. As shown in Table 5 also are records of risk for the prosecutor attacker model, journalist attacker model and marketer

attacker model. Accordingly, KA has value 0 as the highest risk value for the prosecutor attacker model, while both journalist attacker model and marketer attacker models have 20 while the success rate is 0.521. The success rate of 0.521 is the highest value obtained with the indication that using the KA privacy model to anonymize data makes it vulnerable to the attacker. Thus, KA cannot provide efficient privacy for the data, and this corroborates with what is in the literature that KA fails to prevent background knowledge. KA is vulnerable to matching, temporal, homogeneity, and complementary release attack.

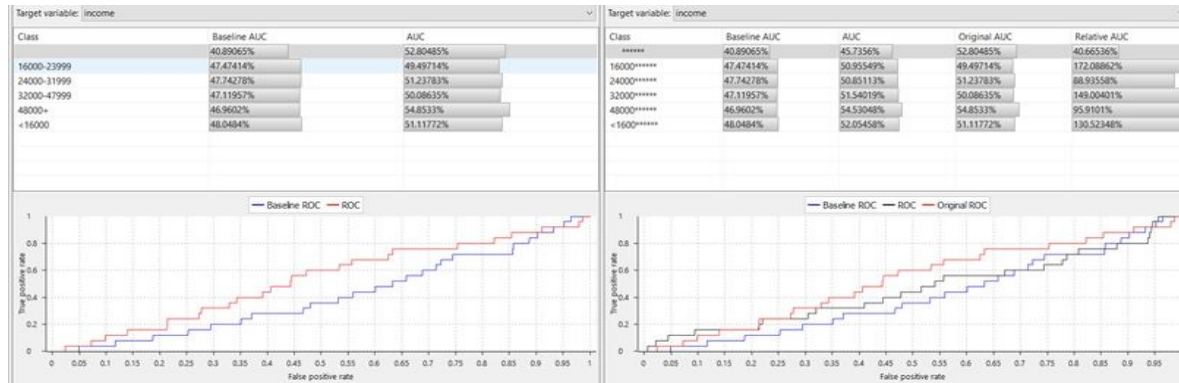


Figure 4. DP AUC performance

Table 5. Attacker method risk analysis

Model	Method	Record at Risk	Highest Risk	Success Rate
KA	Prosecutor Attacker Method	0	20	0.52125
	Journalist Attacker Method	0	20	0.52125
	Marketer Attacker Method	-	-	0.52125
LD	Prosecutor Attacker Method	0	0.34965	0.16084
	Journalist Attacker Method	0	0.34965	0.16084
	Marketer Attacker Method	-	-	0.16084
TC	Prosecutor Attacker Method	0	0.33866	0.10866
	Journalist Attacker Method	0	0.33866	0.10045
	Marketer Attacker Method	-	-	0.10866
DP	Prosecutor Attacker Method	0	0.17921	0.09385
	Journalist Attacker Method	0	0.15267	0.08065
	Marketer Attacker Method	-	-	0.08065

Accordingly, LD had 0.349 as the highest risk value for the three attack models while the success rate for the models was 0.160. The success rate of 0.160 is lower than that of the KA’s attack models suggesting the LD privacy model can provide more efficient data privacy when compared to KA. However, the pitfall of LD is that it is subject to both skewness and similarity attacks, cannot prevent attribute disclosure and is susceptible to both homogeneity and background knowledge attacks. Moreover, the 3 attack models have a value of 0 for TC record at risks, the highest risk of 0.338 and a success rate of 0.100. The low success rate of 0.100 obtained suggests that using the TC privacy model on the anonymized data would provide a more efficient privacy mode when compared to KA and LD. However, TC is limited by the fact that, as the size and variety of the data increases, the chances of re-identification of data also increase.

In the same vein, the record at risk for DP for the 3 attack models is 0 while the highest risks for the prosecutor attacker model and the journalist attacker model are 0.179 and 0.153 respectively with a success rate of 0.094, 0.081 and 0.081 respectively for the 3 attack models. The low success rate value achieved indicates that using the DP privacy model to anonymize data would provide a more efficient privacy mode when compared to KA, LD, and TC. Thus, DP could be the most suitable model and most appropriate for preserving IoT data. The essence is that DP does not allow the degradation of the system’s speed compared to other models. Privacy is preserved by making it cumbersome for an attacker to deduce any person involved regardless of the attack knowing the precise information of all the persons present in the dataset. Based on the result, one can see that combination of DP and KA can provide a more stronger data privacy model that can be used to secure the data. This is because they can offer more efficient privacy as seen from their re-identification risk, BSS for DP, and AUC ROC analysis for KA [13], as shown in Table 5.

5. CONCLUSION

Conclusively, the combination of differential privacy and k-anonymity as showed in our results to protect the data more, the two data privacy model algorithms (DP and KA) which were used to design a hybrid privacy model proposed in this paper provide a stronger data privacy model which therefore enhance the protection of the personal information of users. It is recommended that a novel data privacy model should be developed that can do both the real-time analysis and as well protect the data from attack in any form. Furthermore, it is suggested that more of the currently used data privacy model be combined to see what effect it would have on the dataset protection and to see if information loss is reduced.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the resources and financial support made available by department of Information Technology Systems, Walter Sisulu University, South Africa.





REFERENCES

- [1] D. N. Murthy and B. V. Kumar, "Internet of things (IoT): Is IoT a disruptive technology or a disruptive business model?," *Indian Journal of Marketing*, vol. 45, no. 8, pp. 18–27, Aug. 2015, doi: 10.17010/ijom/2015/v45/i8/79915.
- [2] P. Radanliev *et al.*, "Cyber risk from IoT technologies in the supply chain-discussion on supply chains decision support system for the digital economy," *University of Oxford combined working papers and project reports prepared for the PETRAS National Centre of Excellence and the Cisco Research Centre*. pp. 1–9, Mar. 2019, doi: 10.13140/RG.2.2.17286.22080.
- [3] H. Kaur and A. S. Kushwaha, "A review on integration of big data and IoT," in *Proceedings-4th International Conference on Computing Sciences, ICCS 2018*, Aug. 2019, pp. 200–203, doi: 10.1109/ICCS.2018.00040.
- [4] M. Stoll, "A data privacy governance model," *International Journal of IT/Business Alignment and Governance*, vol. 10, no. 1, pp. 74–93, Jan. 2019, doi: 10.4018/ijitbag.2019010105.
- [5] F. E. Mdarbi, N. Affi, I. Hilal, and H. Belhadaoui, "An approach for selecting cloud service adequate to big data," *International Journal of Computer Science and Information Security (IJCSIS)*, vol. 16, no. 8, pp. 20–32, Mar. 2018.
- [6] I. Lee and K. Lee, "The internet of things (IoT): Applications, investments, and challenges for enterprises," *Business Horizons*, vol. 58, no. 4, pp. 431–440, Jul. 2015, doi: 10.1016/j.bushor.2015.03.008.
- [7] M. Schunter, "Data security and privacy in 2025?," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8425 LNCS, pp. 37–41, 2014, doi: 10.1007/978-3-319-06811-4_8.
- [8] S. Supriya and S. Padaki, "Data security and privacy challenges in adopting solutions for IoT," in *Proceedings-2016 IEEE International Conference on Internet of Things; IEEE Green Computing and Communications; IEEE Cyber, Physical, and Social Computing; IEEE Smart Data, iThings-GreenCom-CPSCom-Smart Data 2016*, Dec. 2017, pp. 410–415, doi: 10.1109/iThings-GreenCom-CPSCom-SmartData.2016.97.
- [9] M. Bahrami and M. Singhal, "The role of cloud computing architecture in big data," in *Studies in Big Data*, vol. 8, Springer International Publishing, 2015, pp. 275–295.
- [10] P. D. Pise and N. J. Uke, "Efficient security framework for sensitive data sharing and privacy preserving on big-data and cloud platforms," in *ACM International Conference Proceeding Series*, Mar. 2016, vol. 22-23-Marc, doi: 10.1145/2896387.2896423.
- [11] D. Kumar and M. N. Mohanty, "A survey: Classification of big data," in *Advances in Intelligent Systems and Computing*, vol. 768, Springer Singapore, 2019, pp. 299–306.
- [12] D. Drewer and V. Miladinova, "The big data challenge: Impact and opportunity of large quantities of information under the Europol regulation," *Computer Law and Security Review*, vol. 33, no. 3, pp. 298–308, Jun. 2017, doi: 10.1016/j.clsr.2017.03.006.
- [13] F. A. Elegbeleye, M. Mbodila, A. Mabovana, and O. A. Esan, "Data privacy on using four models- a review," Jul. 2022, doi: 10.1109/ICECET55527.2022.9872999.
- [14] J. Wang, H. Han, H. Li, S. He, P. K. Sharma, and L. Chen, "Multiple strategies differential privacy on sparse tensor factorization for network traffic analysis in 5G," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 3, pp. 1939–1948, Mar. 2022, doi: 10.1109/TII.2021.3082576.
- [15] Z. Guan, Z. Lv, X. Du, L. Wu, and M. Guizani, "Achieving data utility-privacy tradeoff in internet of medical things: a machine learning approach," *Future Generation Computer Systems*, vol. 98, pp. 60–68, Sep. 2019, doi: 10.1016/j.future.2019.01.058.
- [16] R. Keerthana, M. Jayabalan, and M. E. Rana, "A study on k-anonymity, i-diversity, and t-closeness techniques focusing medical data," *IJCSNS International Journal of Computer Science and Network Security*, vol. 17, no. 12, pp. 172–177, Dec. 2017.
- [17] S. Kim and Y. D. Chung, "An anonymization protocol for continuous and dynamic privacy-preserving data collection," *Future Generation Computer Systems*, vol. 93, pp. 1065–1073, Apr. 2019, doi: 10.1016/j.future.2017.09.009.
- [18] J. Le, X. Liao, and B. Yang, "Full autonomy: A novel individualized anonymity model for privacy preserving," *Computers and Security*, vol. 66, pp. 204–217, May 2017, doi: 10.1016/j.cose.2016.12.010.
- [19] Y. Yang, L. Wu, G. Yin, L. Li, and H. Zhao, "A survey on security and privacy issues in internet-of-things," *IEEE Internet of Things Journal*, vol. 4, no. 5, pp. 1250–1258, Oct. 2017, doi: 10.1109/JIOT.2017.2694844.
- [20] M. Jayabalan and J. Asare-Frempong, "Exploring the impact of big data in healthcare and techniques in preserving patients' privacy," *IJCSNS International Journal of Computer Science and Network Security*, vol. 17, no. 8, pp. 143–149, Aug. 2017.
- [21] A. Mehmood, I. Natgunanathan, Y. Xiang, G. Hua, and S. Guo, "Protection of big data privacy," *IEEE Access*, vol. 4, pp. 1821–1834, 2016, doi: 10.1109/ACCESS.2016.2558446.
- [22] H. Taneja, Kapil, and A. K. Singh, "Preserving privacy of patients based on re-identification risk," *Procedia Computer Science*, vol. 70, pp. 448–454, 2015, doi: 10.1016/j.procs.2015.10.073.
- [23] W. Fang, X. Z. Wen, Y. Zheng, and M. Zhou, "A survey of big data security and privacy preserving," *IETE Technical Review (Institution of Electronics and Telecommunication Engineers, India)*, vol. 34, no. 5, pp. 544–560, Sep. 2017, doi: 10.1080/02564602.2016.1215269.
- [24] J. Salas and V. Torra, "A general algorithm for k-anonymity on dynamic databases," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11025 LNCS, Springer International Publishing, 2018, pp. 407–414.





- [25] F. Prasser, J. Eicher, H. Spengler, R. Bild, and K. A. Kuhn, "Flexible data anonymization using ARX-current status and challenges ahead," *Software-Practice and Experience*, vol. 50, no. 7, pp. 1277–1304, Feb. 2020, doi: 10.1002/spe.2812.

BIOGRAPHIES OF AUTHORS







Femi Abiodun Elegbeleye     received his M.Sc. in computer science and information systems from North-West University. He is presently a PhD candidate in computer science and information systems at North-West University South Africa. I also hold dual B.Sc. honors degrees in applied physics and computer science and information systems from the Universities of Venda in South Africa and Ladoke Akintola University of Technology in Ogbomosho, Nigeria, respectively, in 2010 and 2015. Since 2021, he has been a lecturer at the Walter Sisulu University in the department information systems, South Africa. His research interests include technical programming, blockchain, artificial intelligence, machine learning, and data privacy. email: felegbeleye@wsu.ac.za or phernet123@gmail.com.







Munienge Mbodila     is completing his PhD in Computer Science from Northwest University South Africa. He is the Acting-HoD in the department of Information Technology Systems and Manager of Data Analytics (student tracking and monitoring) at Walter Sisulu University, South Africa. His research interests are SDN, wireless networks, computer networks, ICT in education, and the use of ICT in teaching student engagement. He can be contacted at email: mmbodila@wsu.ac.za.



Omobayo Ayokunle Esan     obtained Master of Technology (M. Tech) in Computer Systems Engineering from Tshwane University of Technology, South Africa. He received Bachelor of Technology (B. Tech Honors) in Computer Engineering from Ladoke Akintola University of Technology (LAUTECH), Ogbomosho, Nigeria between 2003 to 2008. He is currently pursuing his PhD in Computer Science from School of Computing in College of Science, Engineering and Technology (CSET), University of South Africa, South Africa. Mr Esan is a lecturer at Information Technology Systems in Walter Sisulu University (WSU), Eastern Cape, South Africa. He has published many articles both in international and local conferences. His research areas include image processing, computer vision, security, machine learning, and artificial intelligence (AI). He can be contacted at email: oesan@wsu.ac.za.



Ife Fortunate Elegbeleye     received her PhD in Physics from University of Venda, South Africa 2019. holds a Bachelor Honors degree in Science (B.Sc.) in Applied Physics from Ladoke Akintola University of Technology, Ogbomosho, Nigeria. 2011 and had her Master. of Science (M.Sc.) degree in Physics, from University of Ibadan, Nigeria. 2014. She is currently a postdoctoral scholar. Her current research interests are computational physics, energy resources, and renewable energy. She has published many articles in both international and local conference at email: ifelove778@gmail.com.