A Study of Privatized Synthetic Data Generation Using Discrete Cosine Transforms

Kato Mivule
Computer Science Department
Bowie State University
Bowie Maryland, USA

***Abstract*—In order to comply with data confidentiality requirements, while meeting usability needs for researchers, entities are faced with the challenge of how to publish privatized data sets that preserve the statistical traits of the original data. One solution to this problem, is the generation of privatized synthetic data sets. However, during data privatization process, the usefulness of data, have a propensity to diminish even as privacy might be guaranteed. Furthermore, researchers have documented that finding an equilibrium between privacy and utility is intractable, often requiring trade-offs. Therefore, as a contribution, the Filtered Classification Error Gauge heuristic, is presented. The suggested heuristic is a data privacy and usability model that employs data privacy, signal processing, and machine learning techniques to generate privatized synthetic data sets with acceptable levels of usability. Preliminary results from this study show that it might be possible to generate privacy compliant synthetic data sets using a combination of data privacy, signal processing, and machine learning techniques, while preserving acceptable levels of data usability.**

Keywords—privatized synthetic data; Signal processing; Data privacy; discrete cosine transforms; Moving average filtering

I. INTRODUCTION

Realizing an equilibrium between privacy and usability needs is a challenging undertaking that organizations have to engage in, to meet the terms of privacy regulations. To implement privacy acquiescent data transactions, trade-offs have to be made between privacy and usability requirements [1][2][3][4][5]. One way to address this problem, is the generation of privatized synthetic data sets that retain the statistical traits of the original data. Therefore, as a contribution, the Filtered Classification Error Gauge (Filtered x-CEG) methodology is suggested as a heuristic for the generation of privatized synthetic data [17]. The Filtered x-CEG is a variation of the Comparative x-CEG heuristic process described in Mivule and Turner (2013) [6] and [17]. The Filtered x-CEG heuristic works as follows: (i) Data privacy is applied to the data using noise addition; (ii) in the second step, signal processing technique of discrete cosine transforms, is used to mine the coefficients; (iii) the coefficients are added back to the noisy data; (iv) new privatized synthetic data is produced with a similar formation as the original[17]; (v) the moving average filter is then applied to the privatized synthetic data to improve usability; (vi) machine learning classification is used to test the filtered synthetic data for usability, with lower classification error (high classification accuracy) as an indication of better data usability [6][17]. Initial outcome from this study indicates that privatized synthetic data could be produced with adequate usability levels. Therefore, the main focus of this study is to employ data privacy, signal processing, and machine learning classification techniques in the generation of privatized synthetic data with acceptable levels of usability. The rest of the paper is organized as follows, in Section II, background and related work is given. Section III discusses the essential terms used in this paper, while Section IV focuses on the methodology. In Section V, the experiment is outlined and results discussion is done in Section VI. Finally in Section VII, the conclusion is given.

II. BACKGROUND AND RELATED WORK

In this section, a review of related work on using signal processing techniques for data privacy applications, is given [17]. While signal processing techniques have been applied for obfuscation in image and audio applications, there is not much work on using signal processing for specifically data privacy applications, such as, privatized synthetic data generation. However, of recent, researchers have picked up interest on applying signal processing techniques for data privacy implementations. For instance, on the use of signal processing in fulfilling data privacy challenges, Sankar, Trappe, Ramchandran, Poor, and Debbah (2013), noted that the necessary optimization task between data privacy and usability is a primary signal processing issue. Sankar et al., also observed there was a possibility of privacy assurances and solutions, by employing distributed signal processing methods [7]. Furthermore, Sankar et al., (2013), suggested the U-P trade-off region data privacy and utility signal processing based measurement model, for the quantification of data privacy and utility [7]. Consequently, usability, would be a measure of the closeness between the original and privatized data [7]. However, in this study, the classification error is used as a gauge for data privacy and usability quantification [6]. On the subject of discrete cosine transforms and data privacy, studies have mostly been done in the image and audio processing areas, with focus on access control instead of confidentiality [8][9][10][11]. In this paper, discrete cosine transforms methods are employed for data privacy applications, in this case, the generation of privatized synthetic data sets. Nevertheless, applications of Fourier transforms, for example discrete cosine transforms, were suggested by Mukherjee, Chen, and Gangopadhyay (2006) for the enhancement of privacy in Euclidean distance based clustering algorithms [11]. Mukherjee et al., (2006) observed that although original data allocations can be fittingly reconstructed

in the confidential data, distance between points in the confidential data, is not conserved, thus clustering results with unsatisfactory performance [11]. At the same time Mukherjee et al., (2006) outlined advantages of employing Fourier transforms (discrete cosine transforms): (i) Conservation of Euclidean distance in the transformed data can be achieved, thus better clustering results; (ii) data compression could be attained by suppressing lesser coefficients and retaining greater coefficients; (ii) by suppressing coefficients, confidentiality of the data can be enhanced, thus making it complex for attackers to reconstruct the original data [11] [17]. In this study and in the suggested model, the suppression of coefficients as in Mukherjee et al (2006) model, is avoided. Rather, extraction of coefficients using discrete cosine transforms, and applying the coefficients in the generation of synthetic data with similar traits as the original, is done.

III. Essential Terms

While a number of data privacy and signal processing methods exist, it is beyond the span of this implementation paper to expansively survey each technique. The following are a description of some of the techniques used in this paper.

Noise addition: Random values are generated using the mean and standard deviation from the original data and added back to the original data, thus producing a confidential data set, using the following equation [12]:

$$\boldsymbol{Z = X + \varepsilon} \quad (1)$$

The symbol Z represents the confidential data, while X represents the original data, and ε symbolizes random values, chosen from a distribution of $\varepsilon \sim N(0, \sigma^2)$. The symbol ε represents an adjustable parameter, with a smaller ε producing data with traits much similar with the original, and a larger ε producing data that is much more dissimilar to the original [13]. In this paper, a normal distribution $\varepsilon \sim N(\mu = 1, \sigma = 0.2)$, is used to generate the noisy data that is then used in the signal processing, to generate coefficients which are then used to produce the privatized synthetic data set [17].

Discrete cosine transforms: Proposed by Ahmed, Natarajan, and Rao (1974), discrete cosine transform (DCT) is a process that converts a limited data sequence (real numbers) by summing up of cosine functions oscillating at different frequencies[13][14] [17]. DCT alters a set of real numbers $N: x_0, \dots, x_{n-1}$ into a set of real numbers $N: X_0, \dots, X_{n-1}$ using the following equation [14]:

$$\boldsymbol{X_k = \sum_{n=0}^{N-1} x_n \cos\left[\frac{\pi}{N}(n + \frac{1}{2})k\right], k = 0, \dots, N-1.} \quad (2)$$

The symbol X_k, represents the set of altered data as a result of the DCT computation.

Moving Average Filter: In the moving average filter, each point in the output signal is a result of averaging a number of adjacent points in the input signal using the following formula: [16].

$$y[i] = \frac{1}{M} \sum_{j=0}^{M-1} x[i + j] \quad (3)$$

The notation $x[i + j]$ symbolizes the input signal, while $y[i]$ represents the output signal, and M stands for the number of points used in the moving average [16].

IV. Methodology

In this section, the methodology used in this paper and as described in [16], is outlined. The Filtered *x*-CEG, an adaptation of the Comparative *x*-CEG heuristic model outlined in Mivule and Turner (2013), is suggested [6]. Signal processing techniques, such as, discrete cosine transforms are used in the Filtered *x*-CEG, illustrated in Figure 1, unlike the model in [6], that does not involve signal processing methods [17]. The following are the steps involved in the generation of privatized synthetic data sets.

The Filtered *x*-CEG:

- *Step 1: Data privacy:* data privacy is implemented using noise addition – noisy data with statistical traits closer to the original is generated, with a normal distribution $\varepsilon \sim N(\mu = 1, \sigma = 0.2)$.
- *Step 2: Signal processing:* discrete cosine transforms is applied on the noisy data to extract coefficients.
- *Step 3: Synthetic data generation*: the obtained coefficients from Step 2, are added to the noisy data, producing a new confidential synthetic data set. The compensation from this phase is that it would be more difficult for an attacker to rebuild the original data; furthermore, the statistical traits from the original data could be preserved by using the acquired coefficients.
- *Step 4: Filtering*: The moving average filter is used in this phase, to reduce noise that could affect the usability of the data, with the aim for better data usability.
- *Step 5: Machine learning*: Machine learning is then applied on the privatized synthetic data to gauge for usability – with less classification error as an indicator of better data usability.
- *Step 6: The threshold*: if the classification error satisfies the desired threshold, then better usability is achieved and the privatized synthetic data could be published.

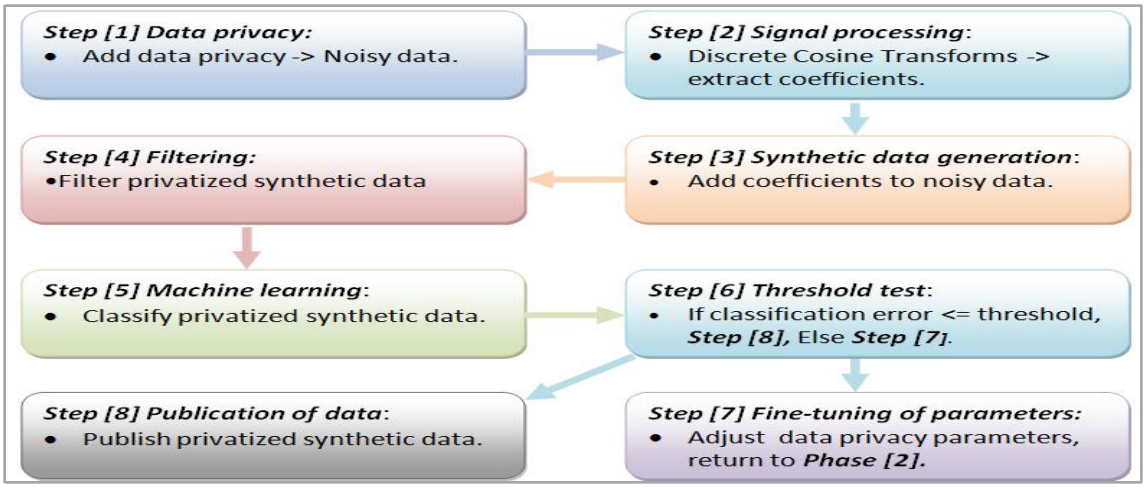

Fig. 1. The Filtered *x*-CEG process

- *Step 7: Fine tuning of parameters*: fine-tuning is done to the privacy parameters, and the signal processing is re-done if the threshold is not met. The procedure replicates *x* times until the preferred threshold is achieved, signifying improved data usability.
- *Step 8: Publication*: The privatized synthetic data with improved usability is published.

V. Experiment

The data used in this study comprised of the Fisher Iris data hosted at the UCI repository. The data contained 150 data items, four columns, the sepal length, sepal width, petal length, and petal width, with the fifth class column, representing the three classes, Setosa, Versicolar, and Virginica [15]. To produce the noisy data, the original data set was perturbed with noise addition at $N\sim(\mu = 1, \sigma = 0.2)$. This allocation of noise was selected since it mirrored statistical characteristics of the original data. After generation of the noisy data, discrete cosine transforms technique was used to obtain coefficients from the noisy data (which in this case was a close representation of the original data). The obtained coefficients were combined – added back to the noisy data, as illustrated in Figure 2, for an additional stratum of confidentiality, generating the privatized synthetic data set. The moving average filter was then used on the privatized synthetic data to remove excessive noise and thus increase usability. Machine learning classification was then applied on both the non-filtered and filtered privatized synthetic data. The following classifiers were used: Neural Networks, KNN, Naïve Bayes, Decision Trees, and AdaBoost Ensemble, employing a 10 fold cross-validation.

The threshold determination heuristic was then used by observing all classification errors and choosing data sets that met the threshold criteria. Only data sets that met the threshold criteria were published and statistical analysis performed on them.

VI. Results and Discussions

In this segment, outcome from the experiment on applying discrete cosine transforms (DCT) and filtering techniques for data privacy, is presented. Three groups of data results are observed: (i) original data, (ii) noisy data, and (iii) privatized synthetic data. A presentation of both descriptive and inference statistical results is also given.

A. Non-Filtered Privatized Sythentic DCT-based Data Results

Figures 3(a), 3(b), 3(c), and 3(d), represent results from the DCT process. In each graph of the illustrations, the lower data sequence represents the DCT coefficients, while the middle data sequence represents the noisy data, and the upper data sequence represents the generated privatized synthetic data. The DCT coefficients were mined from the noisy data and added to the same noisy data set, generating the privatized synthetic data. The noisy data was generated using very low noise addition to the original, to mimic the statistical properties of the original data. As can be seen in Figure 3(a), the privatized synthetic data sequence follows a similar pattern to the noisy data sequence, from an anecdotal view point. This could be an indication that it might be possible to generate privatized synthetic data sets that retain some statistical traits of the original data.

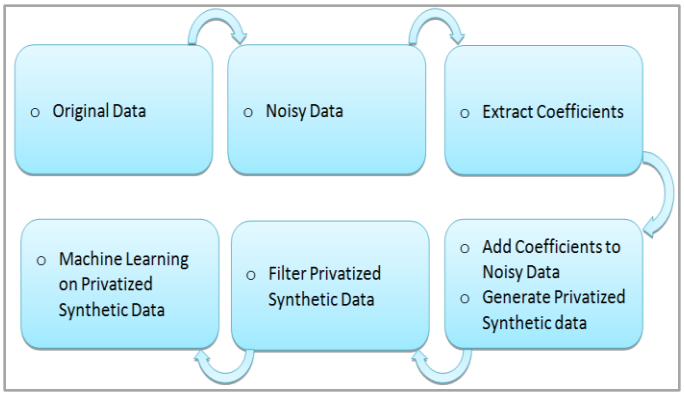

Fig. 2. Privatized synthetic data generation process

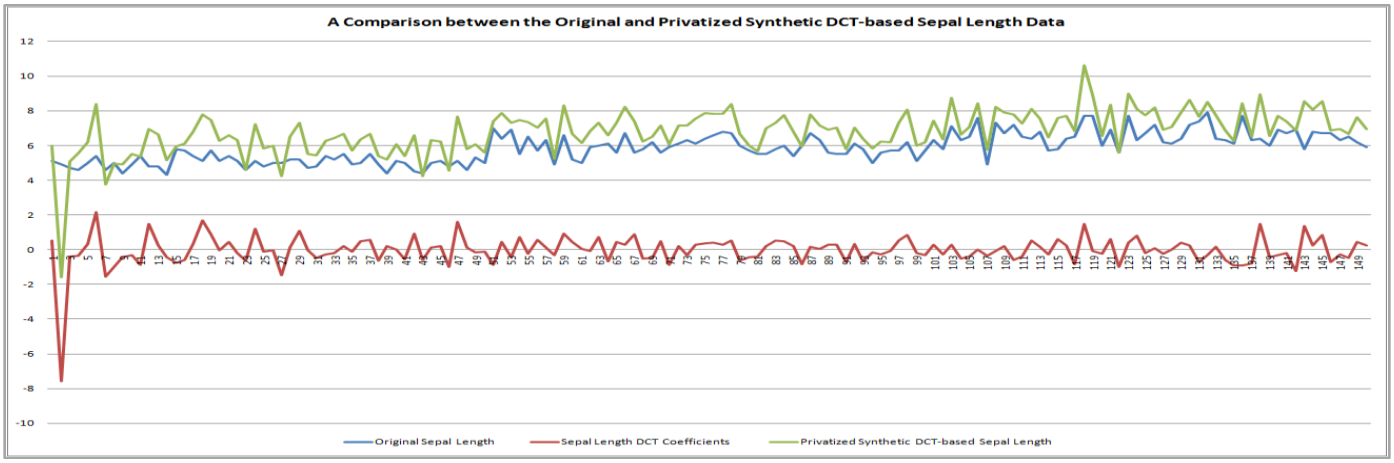

Fig. 3. (a)Privatized Synthetic Fisher-Iris data sequence – Sepal Length

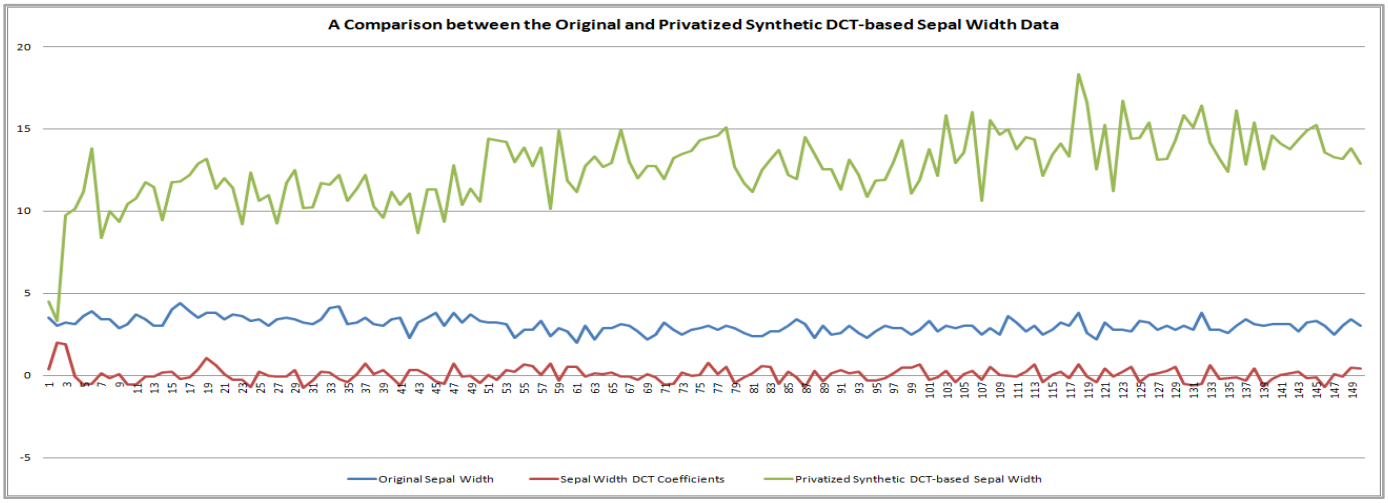

Fig. 3. (b)Privatized Synthetic data Fisher-Iris data sequence – Sepal Width

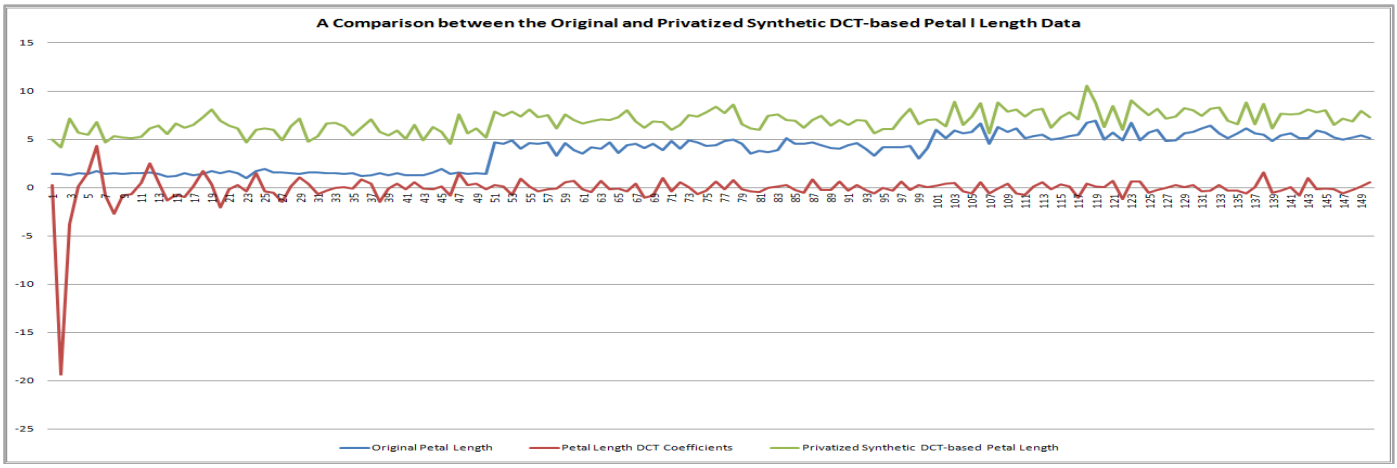

Fig. 3. (c)Privatized Synthetic data Fisher-Iris data sequence – Petal Length

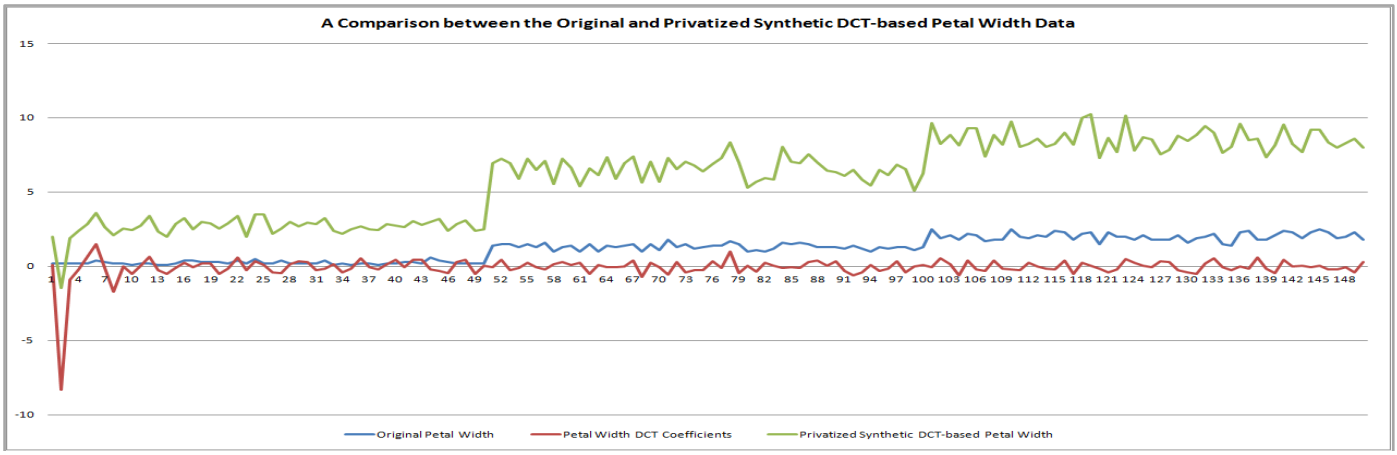

Fig. 3. (d)Privatized Synthetic data Fisher-Iris data sequence – Petal Width

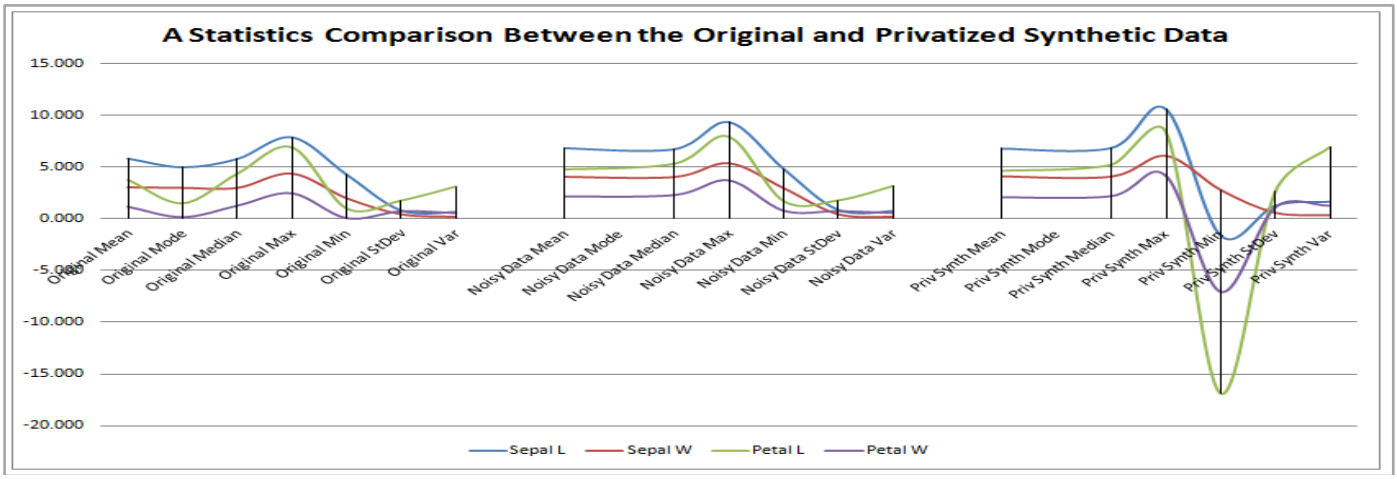

Fig. 4. Privatized Synthetic – descriptive statistics

As shown in Figures 3(b), 3(c), and 3(d), DCT-based privatized synthetic data did not automatically preserve the statistical skeletal structure of both the original and noisy data sets; and as further highlighted in Figure 4, with the descriptive statistics, a deformation of the original statistical skeletal structure occurred with the DCT-based privatized synthetic data. An anecdotal view of Figure 4 and Table I, show that the statistical skeletal structural likeness of the original data is kept in the noisy data.

TABLE I. NON-FILTERED PRIVATIZED SYNTHETIC DATA – DESCRIPTIVE STATISTICS

Statistics	Sepal L	Sepal W	Petal L	Petal W
Original Mean	5.843	3.054	3.759	1.199
Original Mode	5.000	3.000	1.500	0.200
Original Median	5.800	3.000	4.350	1.300
Original Max	7.900	4.400	6.900	2.500

Original Min	4.300	2.000	1.000	0.100
Original Stdev	0.828	0.434	1.764	0.763
Original Var	0.686	0.188	3.113	0.582
Noisy Data Mean	6.841	4.077	4.766	2.200
Noisy Data Mode	#N/A	#N/A	#N/A	#N/A
Noisy Data Median	6.744	4.060	5.323	2.333
Noisy Data Max	9.353	5.398	7.921	3.747
Noisy Data Min	4.846	2.978	1.716	0.819
Noisy Data Stdev	0.880	0.432	1.778	0.776
Noisy Data Var	0.775	0.186	3.162	0.603
Priv Synth Mean	6.801	4.124	4.632	2.125
Priv Synth Mode	#N/A	#N/A	#N/A	#N/A
Priv Synth Median	6.863	4.101	5.225	2.232
Priv Synth Max	10.608	6.115	8.356	4.173
Priv Synth Min	-1.603	2.799	-16.889	-7.010
Priv Synth Stdev	1.295	0.583	2.632	1.142
Priv Synth Var	1.677	0.340	6.926	1.305

However, the same statistical skeletal structure is deformed after applying DCT, in the privatized synthetic data. This could mean that simply adding noise addition to generate a noisy data set might not be enough, since an attacker could guess the original with a higher prospect of success. However, the statistical structure of the privatized synthetic data set is deformed when compared to the original and thus might make it more difficult for an attacker to guess the original composition while at the same time offering some usability to the end user of the privatized synthetic data set. Nevertheless, the mean of the privatized synthetic data is preserved when compared to the mean of the noisy data, as illustrated in Table I. For example, the mean of the noisy data is 6.841, whereas the mean of the privatized synthetic data is at 6.863 for the Sepal length class as recorded in Table I. Yet still, the median and max values are not preserved in the privatized synthetic data set. The covariance values between the noisy data and the privatized synthetic data sets are shown in Figure 5 and Table II. The standard deviation and covariance of the privatized synthetic data set is also not analogous to the noisy and original data. This might be good for privacy preservation in the privatized synthetic data set, while still maintaining some level of usability with the similar mean values.

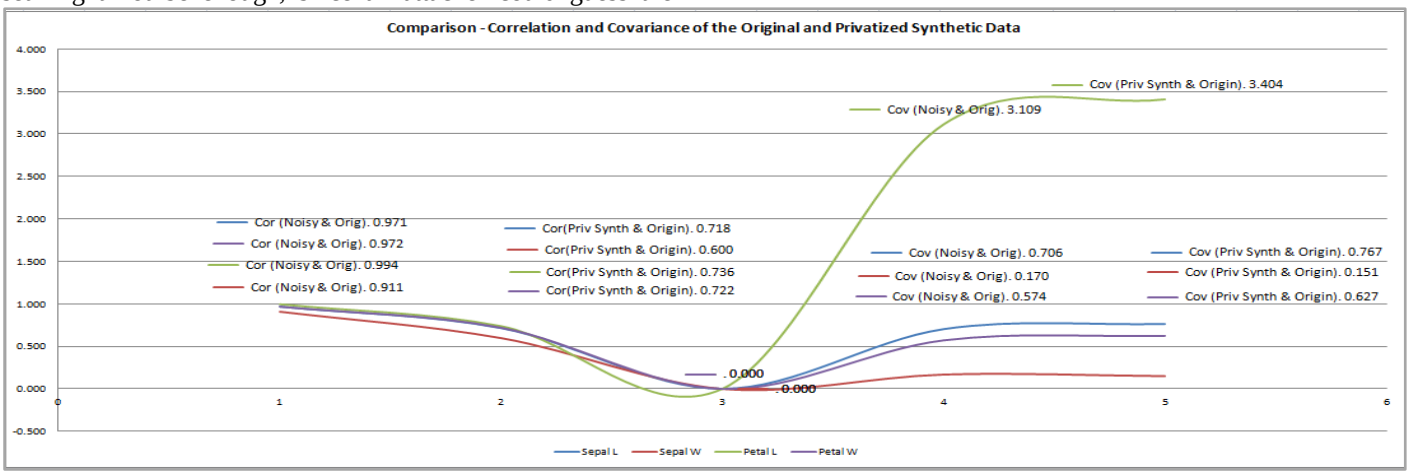

Fig. 5. Privatized Synthetic data – correlation and covariance

The results in Table II, show covariance values between 3.1 and 3.4, for the Petal length, and between 0 and 1, for the Sepal length, Sepal width, and Petal width, an indication of a diminutive inclination for the compared data to grow simultaneously.

TABLE II. NON-FILTERED PRIVATIZED SYNTHETIC DATA – CORRELATION AND COVARIANCE

Statistics	Sepal L	Sepal W	Petal L	Petal W
Correl (Noisy Data & Orig)	0.971	0.911	0.994	0.972
Correl (Synth & Orig)	0.718	0.600	0.736	0.722
Cov (Noisy Data & Orig)	0.706	0.170	3.109	0.574
Cov (Synth & Orig)	0.767	0.151	3.404	0.627

The correlation shown in Table II, between the noisy data and the original data, indicate results varying from 0.971 to 0.994, demonstrating a strong relationship. However, correlation results between the privatized synthetic and original data indicate a range of values from 0.060 to 0.74, signifying more or less a small relationship between the privatized synthetic data and the original data. Yet still, this could be good for privacy preservation even though a level of usability might be lost. Nonetheless, it might be said that DCT-based privatized synthetic data did not preserve the statistical traits of the original but did maintain the mean values. To investigate this premise further, DCT-based privatized synthetic data is passed through the filtering procedure.

B. Filtered Privatized Sythentic DCT-based Data Results

Results in Figures 6(a), 6(b), 6(c), and 6(d), represent the outcome of the experiment after applying filtering on the DCT-based privatized synthetic data. The lower sequence in each of the graphs shown in the illustrations, represents the DCT coefficients, while the middle sequence represents the noisy data, and the upper sequence represents the generated privatized synthetic data after applying filtering.

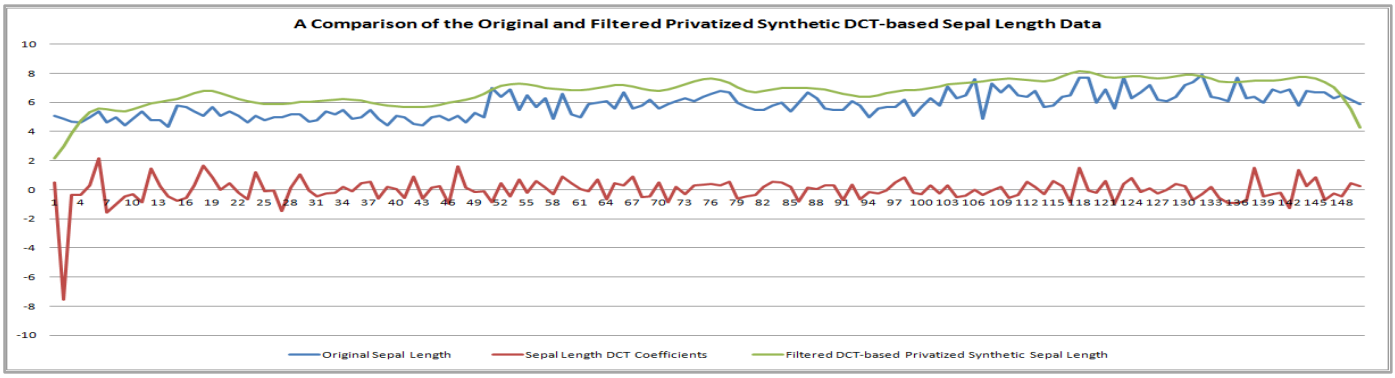

Fig. 6. (a)Filtered Privatized Synthetic Fisher-Iris data sequence – Sepal Length

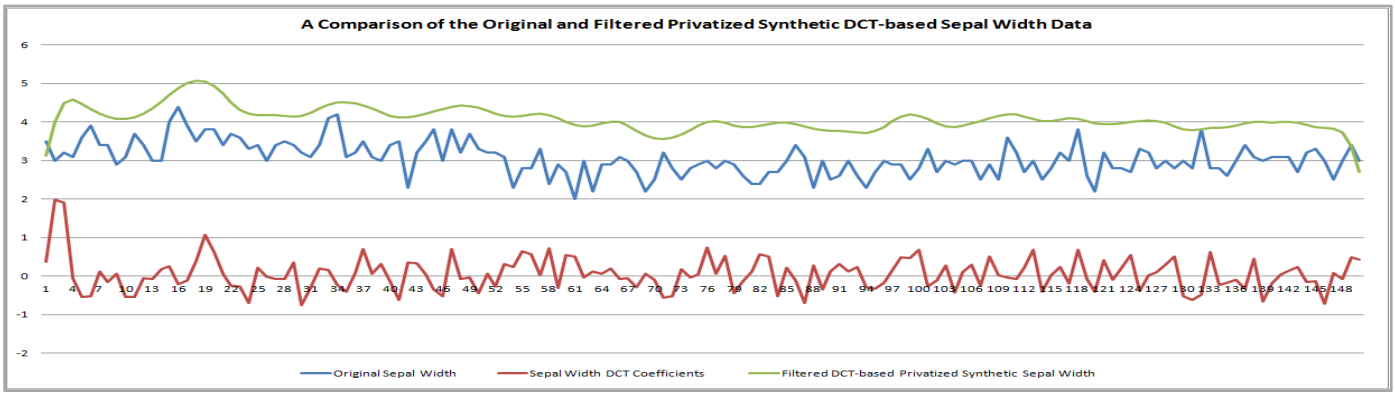

Fig. 6. (b)Filtered Privatized Synthetic Fisher-Iris data sequence – Sepal Width

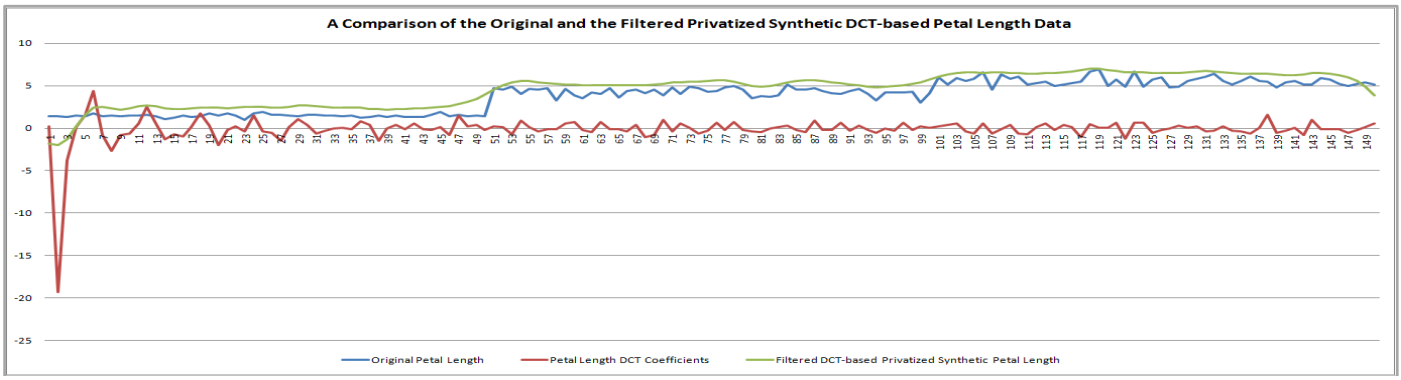

Fig. 6. (c)Filtered Privatized Synthetic Fisher-Iris data sequence – Petal Length

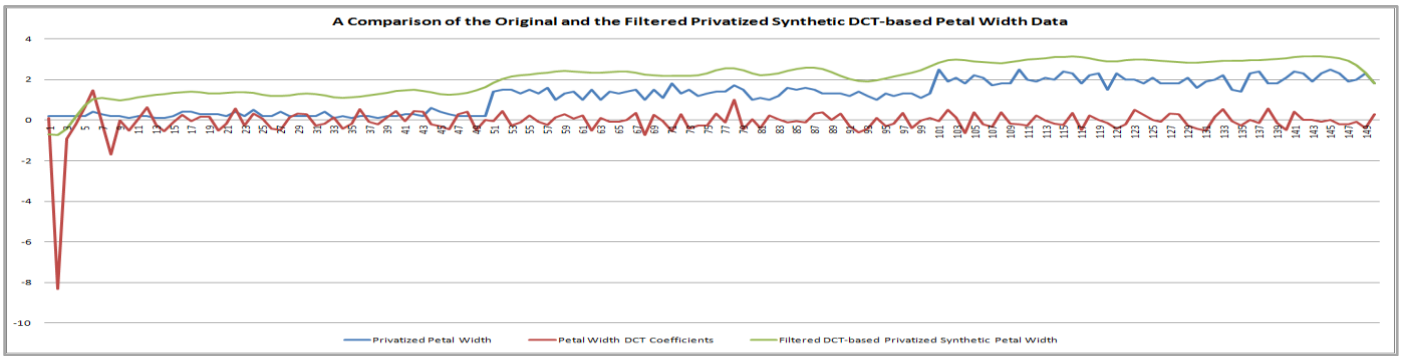

Fig. 6. (d)Filtered Privatized Synthetic Fisher-Iris data sequence – Petal Width

The moving average filtering with kernel width window of 4.0, was employed in the experiment. Regardless of the filtering process, Filtered privatized synthetic data did not preserve a good deal of the statistical traits and skeletal makeup of the noisy and original data, as illustrated in Figure 7; the results are similar to those produced for the non-filtered privatized synthetic data in Figure 4. However, the mean values were preserved in the Filtered privatized synthetic data, similar to results in the non-filtered privatized synthetic data, as shown in Table III. The outcome from this part of the study, indicates that although DCT based privatized synthetic data did not preserve some of the statistical traits, the mean values were maintained, an indication of some level of usability. Additionally, it might be possible that better privacy guarantees could be offered with DCT-based privatized synthetic data, and make it more challenging for an attacker to make precise deductions. Therefore, for the production of privatized synthetic data sets with less emphasis on data usability (utility), DCT-based privatized synthetic data sets might offer some interesting outcomes. However, there was a slight improvement in the correlation values, as shown in Figure 8. The filtered privatized synthetic data and the original data correlation values ranged from 0.5 to 0.9, compared to the 0.6 to 0.7 range of the non-filtered privatized synthetic data and the original.

TABLE III. FILTERED PRIVATIZED SYNTHETIC DATA – DESCRIPTIVE STATISTICS

Statistics	Sepal L	Sepal W	Petal L	Petal W
Original Mean	5.843	3.054	3.759	1.199
Original Mode	5.000	3.000	1.500	0.200
Original Median	5.800	3.000	4.350	1.300
Original Max	7.900	4.400	6.900	2.500
Original Min	4.300	2.000	1.000	0.100
Original StDev	0.828	0.434	1.764	0.763
Original Var	0.686	0.188	3.113	0.582
Noisy Data Mean	6.841	4.077	4.766	2.200
Noisy Data Mode	#N/A	#N/A	#N/A	#N/A
Noisy Data Median	6.744	4.060	5.323	2.333
Noisy Data Max	9.353	5.398	7.921	3.747
Noisy Data Min	4.846	2.978	1.716	0.819
Noisy Data StDev	0.880	0.432	1.778	0.776
Noisy Data Var	0.775	0.186	3.162	0.603
Priv Synthetic Mean	6.801	4.124	4.632	2.125
Priv Synthetic Mode	#N/A	#N/A	#N/A	#N/A
Priv Synthetic Median	6.863	4.101	5.225	2.232
Priv Synthetic Max	10.608	6.115	8.356	4.173
Priv Synthetic Min	-1.603	2.799	-16.889	-7.010
Priv Synthetic StDev	1.295	0.583	2.632	1.142
Priv Synthetic Var	1.677	0.340	6.926	1.305

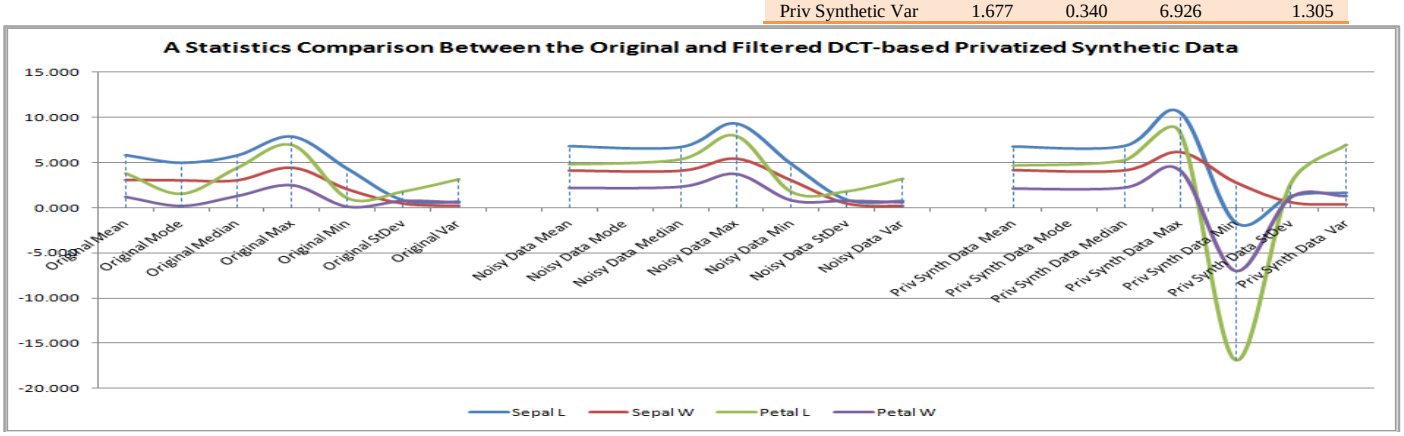

Fig. 7. Filtered Privatized Synthetic data descriptive statistics

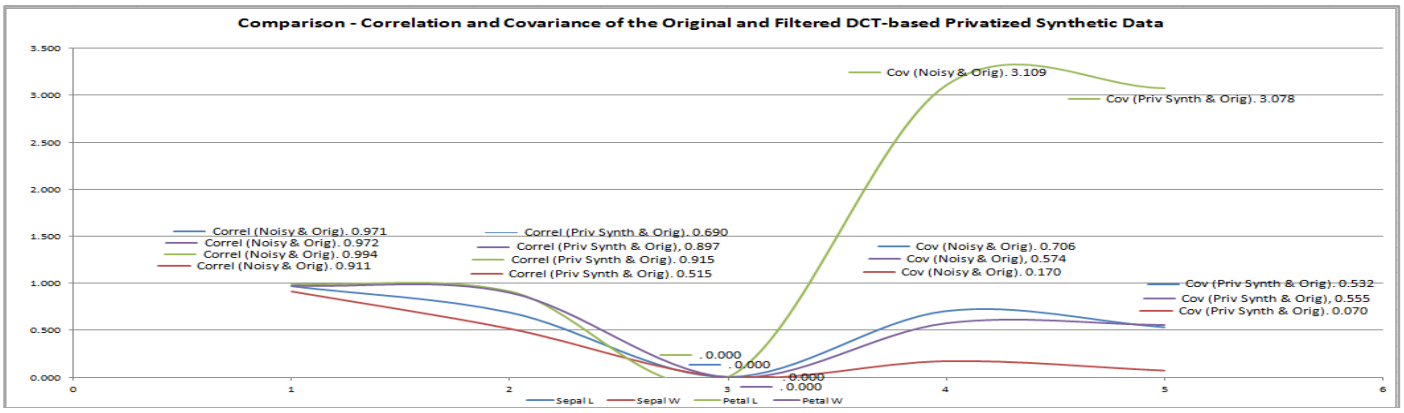

Fig. 8. Filtered Privatized Synthetic data – correlation and covariance

TABLE IV. FILTERED PRIVATIZED SYNTHETIC DATA – CORRELATION AND COVARIANCE

Statistics	Sepal L	Sepal W	Petal L	Petal W
Correl (Noisy Data & Orig)	0.971	0.911	0.994	0.972
Correlation(Priv Synth & Origin)	0.690	0.515	0.915	0.897
Cov (Noisy Data & Orig)	0.706	0.170	3.109	0.574
Cov (Priv Synth & Origin)	0.532	0.070	3.078	0.555

C. Machine Learning Classifier Results

Preliminary results from employing machine learning classification as a measure for data usability, are presented in this section. Both the non-filtered and filtered privatized synthetic data were sent through a chain of machine learning classifiers, namely, Neural Nets (NN), K-nearest Neighbor (KNN), Naïve Bayes (NB), Decision Trees (DT) – Random Forest, in this case, and AdaBoost ensemble. Each classifier returned the classification error, with a higher classification error signifying low data usability, and a low classification error representing improved data usability.

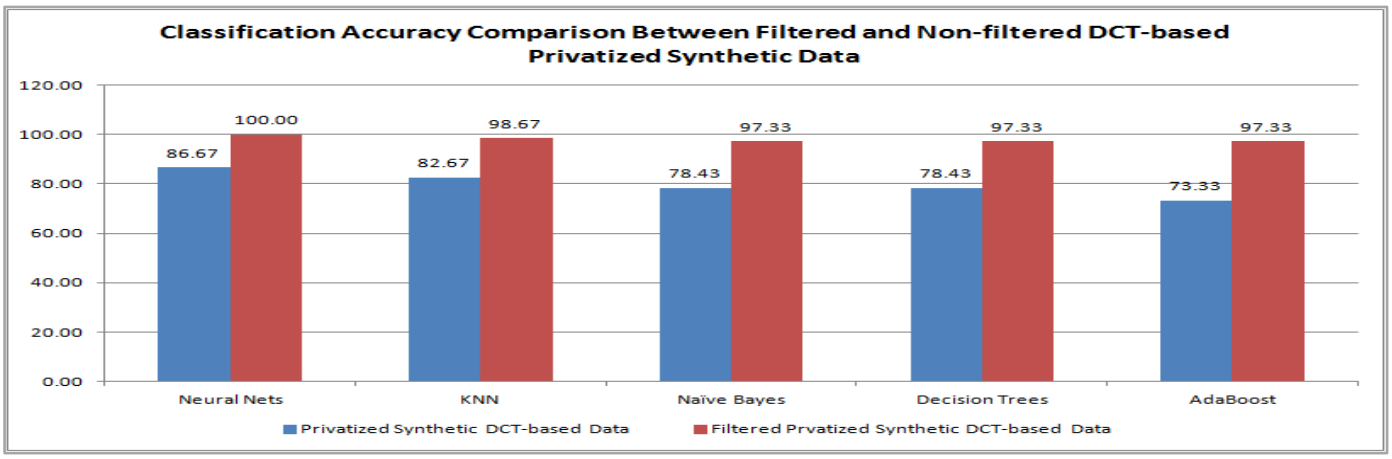

Fig. 9. Classification of Non-Filtered and Filtered data

In Figure 9 and Table V, classification accuracy results were reported – with high classification accuracy as an indication of low classification error and better data usability. However, low classification accuracy indicates higher classification error and likewise signifies low data usability.

TABLE V. CLASSIFICATION ACCURACY FOR BOTH NON-FILTERED AND FILTERED DATA

Classifier	Privatized Synthetic DCT-based Data	Filtered Privatized Synthetic DCT-based Data
NN	86.67	100.00
KNN	82.67	98.67
NB	78.43	97.33
DT	78.43	97.33
AdaBoost	73.33	97.33

Experimental results, as indicated in Figure 9, Figure 10, and Table V, show that there was better performance with filtered privatized synthetic data, with returned higher classification accuracy results, and thus lower classification error. This signifies that filtering might have a profound effect on the classification accuracy of a perturbed data set. For instance, a look at the classification accuracy results, the non-filtered privatized synthetic data, returned a classification accuracy of 86.67 for NN, 82.67 for KNN, and 73.33 for AdaBoost. However, filtered privatized synthetic data returned 100.00 for NN, 98.67 for KNN, and 95.33 for AdaBoost, an indication that filtering does have an effect. The Neural Net classifier, represented by the top sequence in Figure 10, offered the best performance in terms of resilience, among classifiers used in this experiment, on both non-filtered and filtered privatized synthetic data. In general, there was a significant improvement in the performance of all classifiers after application of filtering as illustrated in Figure 10. Consequently, our preliminary results indicate that the technique of filtering noisy data might be significant in enhancing the classification accuracy of data, as such, improving data usability for privatized synthetic data sets. However, concerns about to what degree filtering has to be employed in privatized synthetic data generation, is still challenging. Secondly, inquiries about what quantity of information might be lost at some point in the filtering process, also remain legitimate.

D. Threshold Determination Results

Results in this section, as illustrated in Figure 11, show how the threshold was determined. To find out the threshold, a heuristic was employed by first, using the average value function to compute the mid-point values, and secondly, calculating the mean values [17]. As shown in Table VI, values used in the calculation of both the mid-point and mean were selected from the classification accuracy results. After selecting the mid-point and mean values, the threshold was then selected by taking the max value between the max mid-point and max mean values as shown in Table VI. From our preliminary results, the selected threshold value was 93.34 classification accuracy or 6.66 classification error. Any privatized synthetic data set that met this threshold requirement, was selected, as offering better data usability. Once the threshold is determined and privatized synthetic data set is chosen, the Filtered *x*-CEG procedure stops; the selected data sets that meet the threshold requirement are then published. Conversely, if the threshold criteria is not satisfied, and no data sets are chosen, then the Filtered *x*-CEG algorithm would proceed to the step of adjusting data privacy parameters and going through the classifier procedure again *x*-times, until the threshold criteria is satisfied.

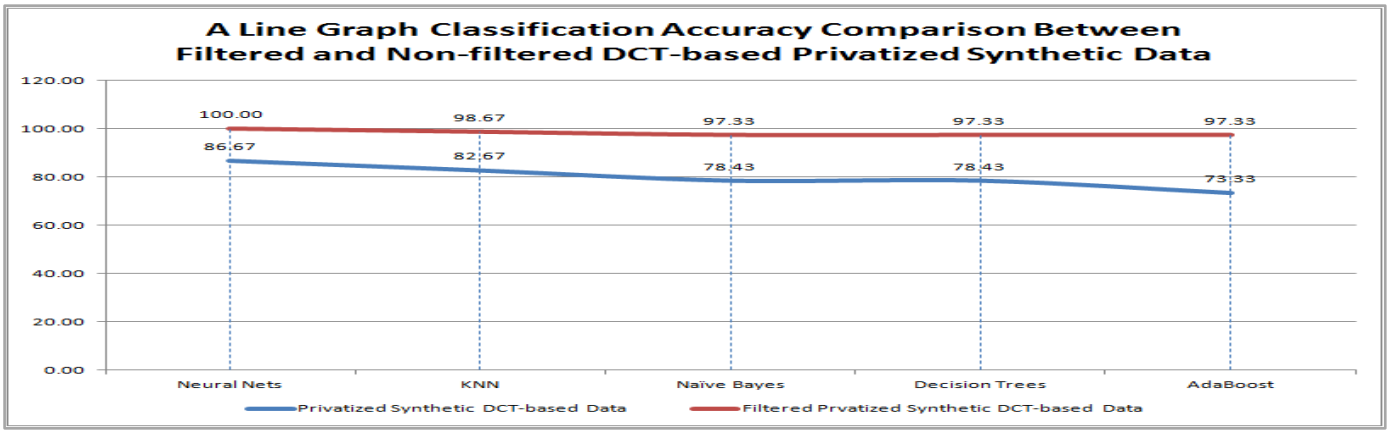

Fig. 10. Performance of classifiers on non-filtered and filtered data

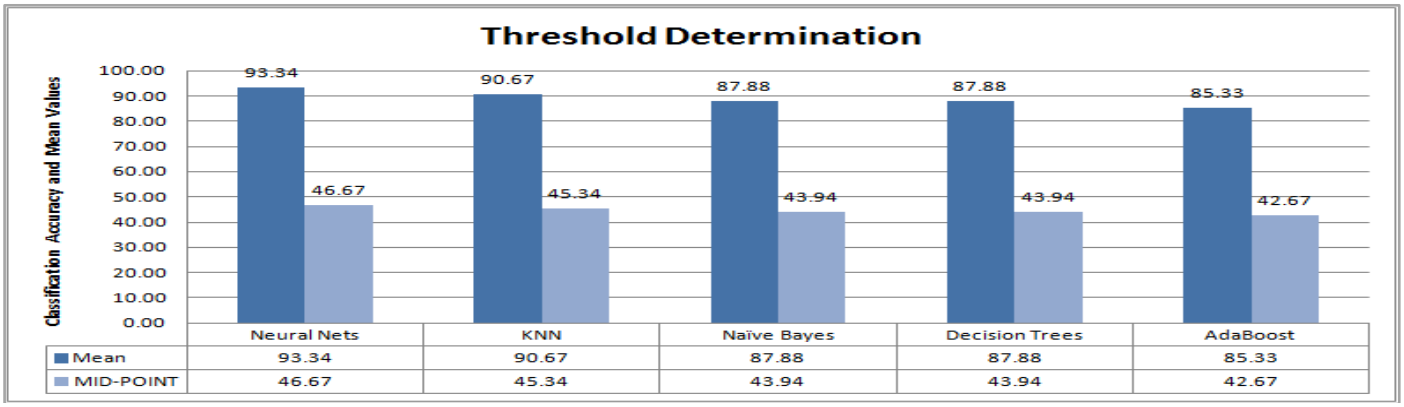

Fig. 11. The mean and mid-point values

TABLE VI. DETERMINIG THE THRESHOLD

Priv Synth Data	NN	KNN	NB	DT	AdaBoost	*Max*
Mean	93.34	90.67	87.88	87.88	85.33	*93.34*
MID-POINT	46.67	45.34	43.94	43.94	42.67	*46.67*
Max	93.34	90.67	87.88	87.88	85.33	***93.34***

VII. CONCLUSION

In this investigation, the Filtered Classification Error Gauge (Filtered *x*-CEG) heuristic was presented and tested. The suggested data privacy model, in which data privacy, signal processing, and machine learning methods are employed to generate privatized synthetic data sets with satisfactory usability levels, was implemented. Preliminary outcome from this investigation indicates that signal processing techniques, such as, discrete cosine transforms, could be used in concert with data privacy techniques to produce privatized synthetic data sets in compliance with confidentiality requirements. Additionally, initial outcome from this study, indicates that filtering might have a corollary to the usability and performance of a privatized synthetic data set when classification is applied to the data set. Filtered privatized synthetic data returned higher classification accuracy results than the non-filtered privatized synthetic data, an indication that filtering might enhance usability of privatized data sets. On the other hand, non-filtered and filtered privatized synthetic data sets did preserve the mean but not the correlation with the original data, an indication of no relationship. In addition, non-filtered and filtered privatized synthetic data sets did not maintain the skeletal structure of the original data, a further indication of dissimilarity. Yet this dissimilarity might be beneficial for improved confidentiality, and perhaps signify that it might be possible to generate confidential synthetic data sets with enhanced usability, by maintaining some statistical traits of the original data, such as, the mean. The Moving Average Filtering procedure was employed in this investigation, using a kernel width window of size 4.0. While the Filtering might have an effect on improving the classification accuracy results, as we showed in the preliminary results, experimenting with various filtering methods not used in this investigation would be worthwhile. The question of what most effective signal processing procedure one would select for executing such a privatized synthetic data generating procedure, remains a case by case proposition and open to further investigation. Yet still, a variety of algorithms could be employed in the generation of confidential synthetic data with strong privacy guarantees, such as, differential privacy. Even more, finding the right equilibrium between privacy and usability requirements, remains challenging and any proposed solution would necessitate trade-offs on a case-by-case basis.

A. Limitations and Future Work

Because of the emergent challenge of big data, the extent and complexity of data confidentiality is at the same time, growing, and as such, it is outside the reach of this investigation to tackle each subject in the data confidentiality sphere. As such, the goal of this investigation was to look at

privatized synthetic data generation, by employing data privacy, signal processing, and machine learning methods. The goal of this investigation was not focused on the type of attacks on the privatized synthetic data, a subject while important, is left for future work. The investigation was restricted to DCT transforms and the moving average filtering techniques. The Fisher-Iris data set was the only data set used in this study. Therefore, future works will comprise of testing generated privatized synthetic data against various adversary attacks, employing of various signal processing and filtering techniques, not used in this investigation, using other large data sets, finally application of various machine learning techniques not covered in this investigation.

ACKNOWLEDGMENT

Portions of this work were presented as part of the dissertation by the author in fulfillment of the requirements for the D.Sc. degree, in the Computer Science Department, at Bowie State University [17]. Special thanks and acknowledgement of Dr. Claude Turner and Dr. Soo-Yeon Ji, in the Computer Science Department, at Bowie State University, for the great and tireless efforts in making this work possible. Special thanks to the peer reviewers for their generous feedback. This study was facilitated and made possible by the HBGI Grant from the United States Department of Education.

REFERENCES

[1] R. C.-W. Wong, A. W.-C. Fu, K. Wang, and J. Pei, “Minimality Attack in Privacy Preserving Data Publishing,” in In Proceedings of the 33rd international conference on Very large data bases (VLDB ’07), 2007, pp. 543–554.

[2] A. Meyerson and R. Williams, “On the complexity of optimal K-anonymity,” in Proceedings of the twentythird ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems PODS 04, 2004, pp. 223–228.

[3] H. Park and K. Shim, “Approximate algorithms for K-anonymity,” in Proceedings of the 2007 ACM SIGMOD international conference on Management of data SIGMOD 07, 2007, pp. 67–78.

[4] A. Krause and E. Horvitz, “A Utility-Theoretic Approach to Privacy in Online Services,” J. Artif. Intell. Res., vol. 39, pp. 633–662, 2010.

[5] Y. W. Y. Wang and X. W. X. Wu, Approximate inverse frequent itemset mining: privacy, complexity, and approximation. 2005.

[6] K. Mivule and C. Turner, “A Comparative Analysis of Data Privacy and Utility Parameter Adjustment, Using Machine Learning Classification as a Gauge,” Procedia Comput. Sci., vol. 20, pp. 414–419, 2013.

[7] L. Sankar, W. Trappe, K. Ramchandran, H. V. Poor, and M. Debbah, “The Role of Signal Processing in Meeting Privacy Challenges,” IEEE Signal Process. Mag., vol. 30, no. 5, pp. 95–106, 2013.

[8] M. Diephuis, S. Voloshynovskiy, O. Koval, and F. Beekhof, “DCT sign based robust privacy preserving image copy detection for cloud-based systems,” in 2012 10th International Workshop on Content-Based Multimedia Indexing (CBMI), 2012, pp. 1–6.

[9] N. V. Lalitha, G. Suresh, and P. Telagarapu, “Audio authentication using Arnold and Discrete Cosine Transform,” in 2012 International Conference on Computing, Electronics and Electrical Technologies (ICCEET), 2012, pp. 530–532.

[10] M. Niimi, F. Masutani, and H. Noda, “Protection of privacy in JPEG files using reversible information hiding,” in 2012 International Symposium on Intelligent Signal Processing and Communications Systems (ISPACS), 2012, no. Ispacs, pp. 441–446.

[11] S. Mukherjee, Z. Chen, and A. Gangopadhyay, “A privacy-preserving technique for Euclidean distance-based mining algorithms using Fourier-related transforms,” VLDB J., vol. 15, no. 4, pp. 293–315, Aug. 2006.

[12] J. Kim, “A Method For Limiting Disclosure in Microdata Based Random Noise and Transformation,” in Proceedings of the Survey Research Methods, American Statistical Association,, 1986, vol. Jay Kim, A, no. 3, pp. 370–374.

[13] N. Ahmed, T. Natarajan, and K. R. Rao, “Discrete cosine transform,” IEEE Trans. Comput., vol. 100, no. 1, pp. 90–93, 1974.

[14] G. Strang, "The discrete cosine transform." SIAM Review, vol 41, no. 1, pp. 135-147, 1999.

[15] K. Bache and M. Lichman, “Iris Fisher Dataset - UCI Machine Learning Repository.” University of California, School of Information and Computer Science., Irvine, CA, 2013.

[16] K. Mivule and C. Turner, "Applying Moving Average Filtering for Non-interactive Differential Privacy Settings", Procedia Computer Science, (In Print), 2014, Philadephia, PA, USA

[17] K. Mivule, "An Investigation of Data Privacy and Utility Using Machine Learning as a Gauge", D.Sc. Dissertation, Computer Science Dept., Bowie State University. 2014: 262 pages; ProQuest: 3619387.