$$KD_{f(D)} = \left[\frac{N(f(D))_{BP} \cap N(f(D))_{AP}}{N(f(D))_{BP}} \right] \times 100 \quad (19)$$

$$IL_{f(D)} = \left[\frac{N(f(D))_{BP} - N(f(D))_{AP}}{N(f(D))_{BP}} \right] \times 100 \quad (20)$$

Knowledge discovery and information loss on the basis of seasonal diseases

$$KD_{S(D)} = \left[\frac{N(S(D))_{BP} \cap N(S(D))_{AP}}{N(S(D))_{BP}} \right] \times 100 \quad (21)$$

$$IL_{S(D)} = \left[\frac{N(S(D))_{BP} - N(S(D))_{AP}}{N(S(D))_{BP}} \right] \times 100 \quad (22)$$

Knowledge discovery and information loss on the basis of sensitive diseases

$$KD_{G(D)} = \left[\frac{N(G(D))_{BP} \cap N(G(D))_{AP}}{N(G(D))_{BP}} \right] \times 100 \quad (23)$$

$$IL_{G(D)} = \left[\frac{N(G(D))_{BP} - N(G(D))_{AP}}{N(G(D))_{BP}} \right] \times 100 \quad (24)$$

In Equations 13 and 14, KD_{CP} and IL_{CP} represent the knowledge discovery and information loss based on the process of changing the position of the diseases in the sequential rule. The calculation of KD_{CP} and IL_{CP} is shown in Equations 25 and 26 respectively, where $(S.R)_{BP}$ and $(S.R)_{AP}$ represent the number of sequential rules before and after processing and $(S.R)_{BP}$ represents the number of sequential rules before processing.

$$KD_{PC} = \left[\frac{(S.R)_{BP} \cap (S.R)_{AP}}{N(S.R)_{BP}} \right] \times 100 \quad (25)$$

$$IL_{PC} = \left[\frac{(S.R)_{BP} - (S.R)_{AP}}{N(S.R)_{BP}} \right] \times 100 \quad (26)$$

In Equations 13 and 14, KD_{RS} and IL_{RS} represent the knowledge discovery and information loss based on the process of reducing the support count value of the sequential rule. To calculate the initial values of KD_{RS} and IL_{RS}, we need to find the common rules (which may need multiple iterations on the basis of position changing and item removing). The calculations for finding the common rules are given in Equation 27 from which $\{D.R\}$ is the representation of final set of sequential rules and $\{P.R\}$ is the representation of processed rules. The result of Equation 27 produces a set of common rules represented as $\{C.R_i\} Where\ (1 \le i \ge C)$, where the value of 'C' is the representation of total number of common rules.

$$\{C.R\} = \{D.R\} \cap \{P.R\} \quad (27)$$

After calculating the common rules, we process the support count of the common rules, meaning how many rules are modified based on their support value. For this, we calculate the support similarity $Sim_S(R)$ and support dissimilarity $Diss_S(R)$ which is given in Equation 28 and 29, where $S.Cnt$ represents the similarity count and $D.Cnt$ represents the dissimilarity count, the values of $S.Cnt$ and $D.Cnt$ increase when the following condition is satisfied (Equation 30).

$$Sim_S(R) = \sum_{i=1}^{C} S.Cnt(R_i) \quad (28)$$

$$Diss_S(R) = \sum_{i=1}^{C} D.Cnt(R_i) \quad (29)$$

$$\begin{Bmatrix} if\ (S(R_i)_{BH} - S(R_i)_{AH}) = 0\ then\ S.Cnt(R_i) = 1 \\ else \qquad\qquad\qquad D.Cnt(R_i) = 1 \end{Bmatrix} \quad (30)$$

After calculating support similarity $Sim_S(R)$ and support dissimilarity $Diss_S(R)$, we use these values to calculate the values of KD_{RS} and IL_{RS}, as shown in Equations 31 and 32 respectively, where the value of $|Sim_S(R)|$ is the representation of number of rules having the same support value, $|Diss_S(R)|$ is the representation of number of rules having different support values and $|\{D.R\}|$ is the representation of number of rules in the final set of sequential rules.

$$KD_{RS} = \left[\frac{|Sim_S(R)|}{|\{D.R\}|} \right] \times 100 \quad (31)$$

$$IL_{RS} = \left[\frac{|Diss_S(R)|}{|\{D.R\}|} \right] \times 100 \quad (32)$$

In Equations 13 and 14, KD_{RC} and IL_{RC} represent the knowledge discovery and information loss based on the process of reducing the confidence value of the sequential rule. To calculate the initial

values of KD_{RS} and IL_{RS}, we need to find the common rules (which may need multiple iterations on the basis of position changing and item removing). The calculations for finding the common rules are given in Equation 27. After calculating the common rules, we process the confidence value of the common rules, meaning how many rules are modified based on their confidence value. To evaluate that, we calculate the confidence similarity $Sim_C(R)$ and confidence dissimilarity $Diss_C(R)$ as given in Equations 33 and 34, where $S.Cnt$ represents the similarity count and $D.Cnt$ represents the dissimilarity count. The values of $S.Cnt$ and $D.Cnt$ when the following condition is satisfied (Equation 35).

$$Sim_C(R) = \sum_{i=1}^{C} S.Cnt(R_i) \tag{33}$$

$$Diss_C(R) = \sum_{i=1}^{C} D.Cnt(R_i) \tag{34}$$

$$\begin{Bmatrix} if\ \left(C(R_i)_{BH} - C(R_i)_{AH}\right) = 0\ \ then\ \ S.Cnt(R_i) = 1 \\ else \qquad\qquad\qquad\qquad D.Cnt(R_i) = 1 \end{Bmatrix} \tag{35}$$

After calculating the confidence similarity $Sim_C(R)$ and confidence dissimilarity $Diss_C(R)$, we use these values to calculate the values of KD_{RC} and IL_{RC} as represented in Equations 36 and 37 respectively, where the value of $|Sim_C(R)|$ is the representation of number of rules having the same confidence value and $|Diss_C(R)|$ is the representation of number of rules having different confidence values.

$$KD_{RS} = \left[\frac{|Sim_S(R)|}{|\{D.R\}|}\right] \times 100 \tag{36}$$

$$IL_{RS} = \left[\frac{|Diss_S(R)|}{|\{D.R\}|}\right] \times 100 \tag{37}$$

5 Results and Discussion

The experimental results of the proposed technique (balanced constraint measure-based algorithm for privacy-preserved sequential rule discovery) have been described here. In this section, we evaluate our proposed algorithm in terms of running time, memory usage, and modifications on the rule in terms of disease, position, support value and confidence value. We also evaluate the set of modified rules in terms of knowledge discovery and information loss. The above measures are calculated for various values of minimum support and minimum confidence.

5.1 Experimental Design

The proposed approach is implemented using java (jdk 1.7). The experimentation of our proposed technique was carried out on a synthetic medical database using a dual core processor PC with 2 GB main memory running in 32 bit version of Windows 7 Operating System. In this paper, we have generated the synthetic medical dataset containing four attributes: patient name, place, disease name, and disease duration. The medical dataset consists of 1000 numbers of data.

5.2 Evaluation of Running Time

In this section, we evaluate running time of our proposed algorithm in terms of minimum support and minimum confidence value. Figures 2 and 3 are the representation of evaluation of running time on the basis of minimum support and minimum confidence value respectively.

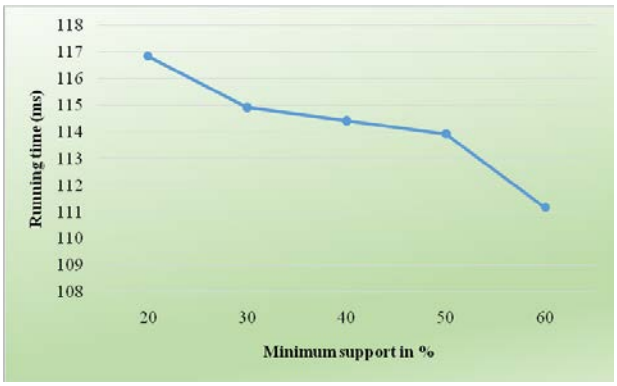

Fig.2. Evaluation of running time for various values of minimum support

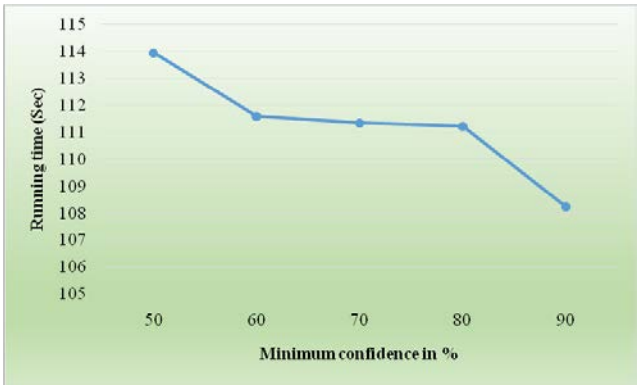

Fig.3. Evaluation of running time for various values of confidence value

Fig.2 is the representation of required running time of the proposed algorithm to apply the privacy on the sequential rules in terms of minimum support. We have maintained the value of the minimum confidence constant at 60% and evaluated the running time of our proposed algorithm for various values of minimum support. From Fig. 2, we see that when the values of minimum support increase, the running time of our proposed algorithm reduces. This is because when we increase the value of minimum support, the number of supported rules gets reduced and the processing time is also reduced.

Fig.3 shows the required running time of the proposed algorithm to apply the privacy on the sequential rules in terms of minimum confidence. Here, we have maintained the value of the minimum support constant at 25% and evaluated the running time of our proposed algorithm for various values of minimum confidence. From Fig. 3, we see that when the values of minimum confidence increase, the running time of our proposed algorithm reduces. This is because when we increase the value of minimum confidence, the number of supported rules gets reduced and the processing time is also reduced.

5.3 Evaluation of Memory Usage

Here, the memory usage of our proposed algorithm is evaluated on the basis of minimum support and minimum confidence value and represented in Figures 4 and 5 respectively.

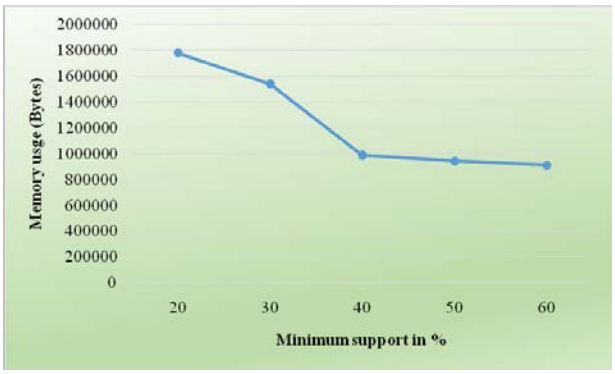

Fig.4. Evaluation of memory usage for various values of minimum support

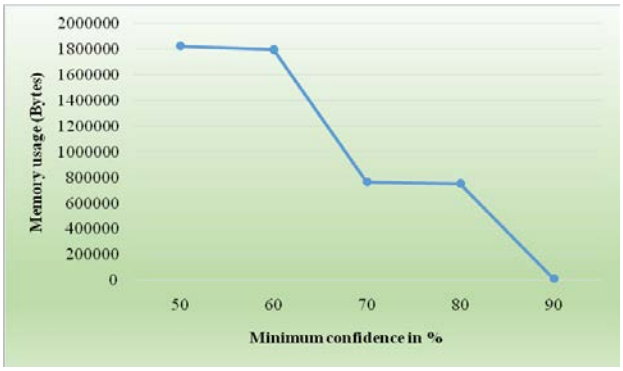

Fig.5. Evaluation of memory usage for various values of minimum confidence

Figures 4 and 5 represent the memory usage of the proposed algorithm to apply the privacy on the sequential rules in terms of minimum support and minimum confidence values respectively. From Figures 4 and 5, we see that when the values of minimum support or minimum confidence increase, the memory usage of our proposed algorithm is reduced. This is because when we increase the value of minimum support or minimum confidence, the number of supported rules gets reduced, the memory usage is also reduced.

5.4 Modification of Rules Based on Significant Diseases

In this section, the modification of the rules of our proposed algorithm on significant diseases is evaluated on the basis of minimum support and minimum confidence value. Figures 6 and 7 represent the modification on significant diseases on the basis of minimum support and minimum confidence value respectively.

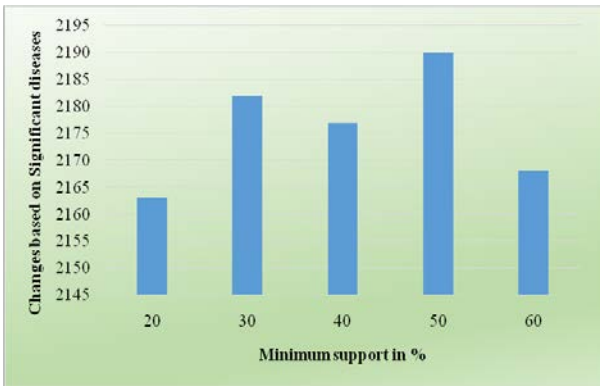

Fig.6. Modification of diseases based on minimum support

From Fig. 6, we observe that as the value of minimum support increases, the changes based on diseases vary randomly. This is because the rule is modified for any disease with random value 0 to 0.2

(and the number of rules with random value 0 to 0.2 changes with each iteration). From Fig. 6, the minimum level of disease modification is 2163 for minimum support 20 and the maximum level of disease modification is 2190 for minimum support 50.

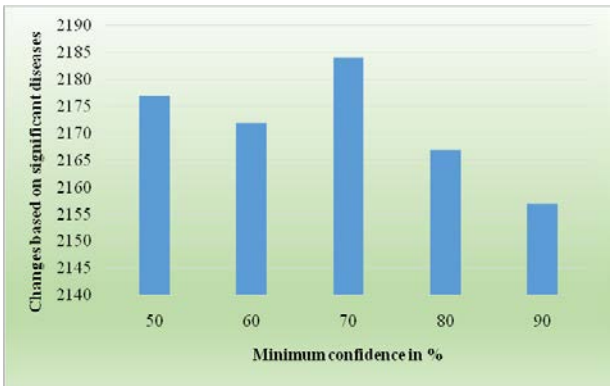

Fig.7. Modification of diseases based on minimum confidence

From Fig. 7, we observe that when the value of minimum confidence increases, the number of changing of significant diseases gradually decreases (except for the minimum confidence 70). This is because the rule is modified for any disease only for random value 0 to 0.2 (and the number of rules with random value 0 to 0.2 changes with each iteration). From Fig. 7, the minimum level of disease modification is 2157 for the minimum confidence 90 and the maximum level of disease modification is 2184 for the minimum confidence 50.

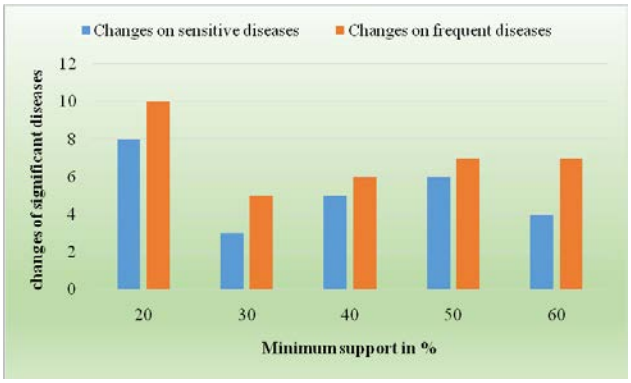

Fig.8. Modification on significant diseases based on minimum support

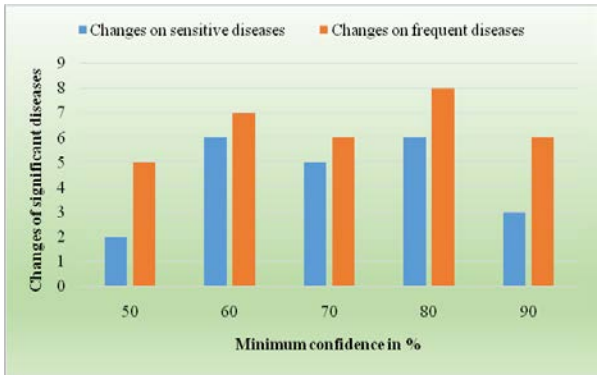

Fig.9. Modification on significant diseases based on minimum confidence

From Figures 8 and 9, we observe that when we increase the value of minimum support and minimum confidence, there is no modification in seasonal diseases and geographical diseases. This is because the database contains comparatively less seasonal and geographical diseases, so they are less likely to take part in the sequential patterns. Seasonal and geographical diseases are not included as their support value in the rules is below the minimum stipulated support value. It is observed that frequent diseases are most-used for modification because they appear more frequently than sensitive diseases.

5. 5 Modification of Rules Based on Position

In this section, the modification on position of the rules of our proposed algorithm is evaluated on the basis of minimum support and minimum confidence value. Figures 10 and 11 represent the modification on diseases on the basis of minimum support and minimum confidence respectively.

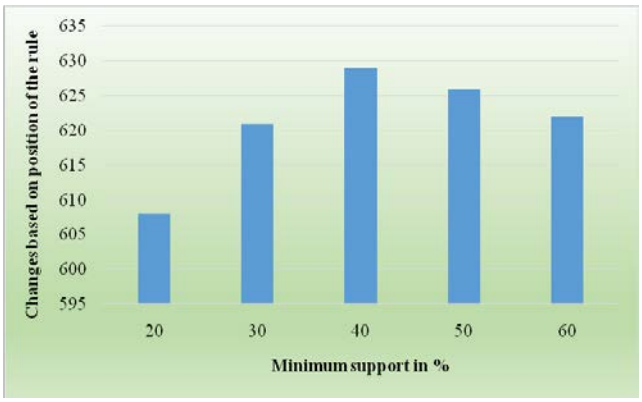

Fig.10. Modification on position of the rules based on minimum support

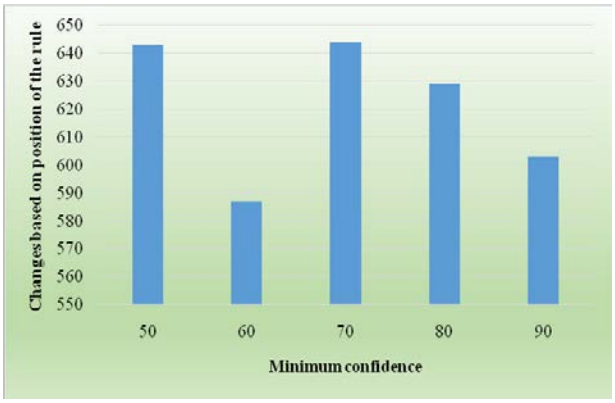

Fig.11. Modification on position of the rules based on minimum confidence

From Figures 10 and 11, we observe that when the value of minimum support and minimum confidence increases, the changes based on position vary rapidly. This is because the number of changes based on position varies at every iteration (as the rule is applied only for random values between 0.2 and 0.4). From Fig. 10, the minimum level of position modification is 608 for the minimum support 20 and the maximum level of position modification is 629 for the minimum support 40. From Fig. 11, the minimum level of position modification is 587 for the minimum confidence 60 and the maximum level of position modification is 644 for the minimum confidence 70.

5.6 Modification of Rules Based on Support Value

In this section, our proposed algorithm is evaluated on the basis of minimum support and minimum confidence value. Figures 12 and 13 represent the modification on diseases on the basis of minimum support and minimum confidence respectively.

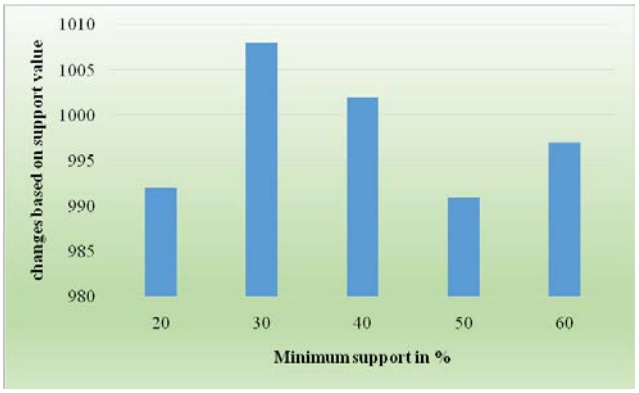

Fig.12. Reduction of support values of the rules based on minimum support

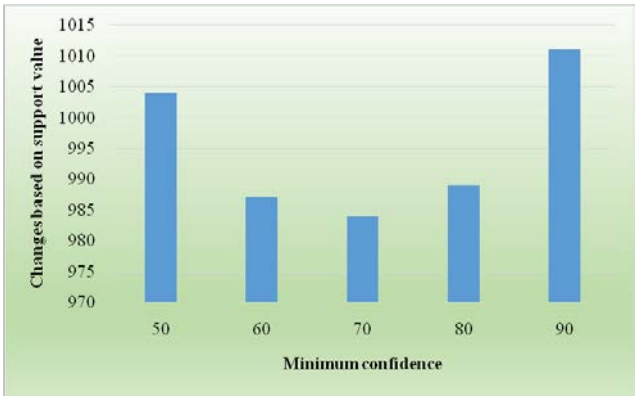

Fig.13. Reduction of support values of the rules based on minimum confidence

From Figures 12 and 13, we observe that when we increase the value of minimum support and minimum confidence, changes based on support value vary rapidly. This is because the rule is modified for any disease having random value 0.4 to 0.6 (and the number of rules with random value 0.4 to 0.6 changes with each iteration). From Figure 12, the minimum level of position modification is 991 for the minimum support 50 and the maximum level of position modification is 1008 for the minimum support 30. From Fig.13, the minimum level of position modification is 984 for the minimum confidence 70 and the maximum level of position modification is 1011 for the minimum confidence 90.

5.7 Modification of Rules Based on Confidence Value

In this section, our proposed algorithm is evaluated on the basis of minimum support and minimum confidence value. Figures 14 and 15 represent the modification of diseases on the basis of minimum support and minimum confidence respectively.

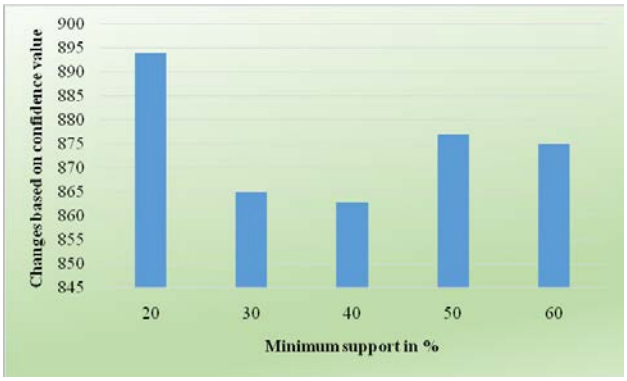

Fig.14. Reduction of confidence values of the rules based on minimum support

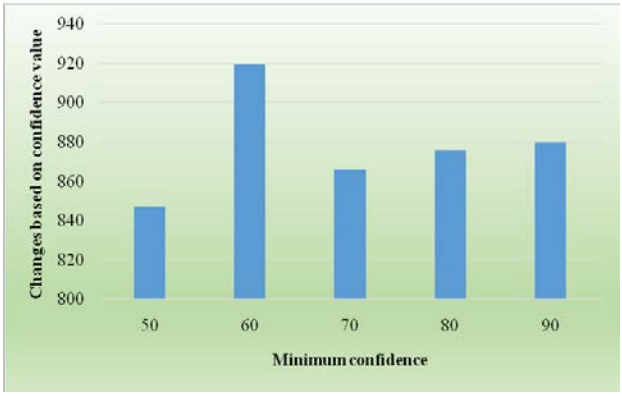

Fig.15. Reduction of confidence values of the rules based on minimum confidence

From Figures 14 and 15, we observe that when the value of minimum support and minimum confidence increase, the changes based on confidence value vary rapidly. This is because the rule is modified for any disease having random value greater than 0.6 (and the number of rules with random value greater than 0.6 changes with each iteration). From Figure 14, the minimum level of position modification is 863 for the minimum support 40 and the maximum level of position modification is 894 for the minimum support 20. From Figure 15, the minimum level of position modification is 847 for the minimum confidence 50 and the maximum level of position modification is 920 for the minimum confidence 60.

5.8 Evaluation of Knowledge Discovery and Information Loss

Here, we evaluate the knowledge discovery and information loss of our proposed algorithm on the basis of minimum support and minimum confidence value. Figures 16 and 17 represent (knowledge discovery and information loss of the processed rule) with changes in (minimum support and minimum confidence) respectively. Here, we take the knowledge discovery and information loss threshold as 70 and 30 respectively. Once the proposed algorithm evaluates the processed rules, it checks with the threshold value of knowledge discovery and information loss. The proposed algorithm releases the processed rules when the evaluated result is close to the threshold boundary value.

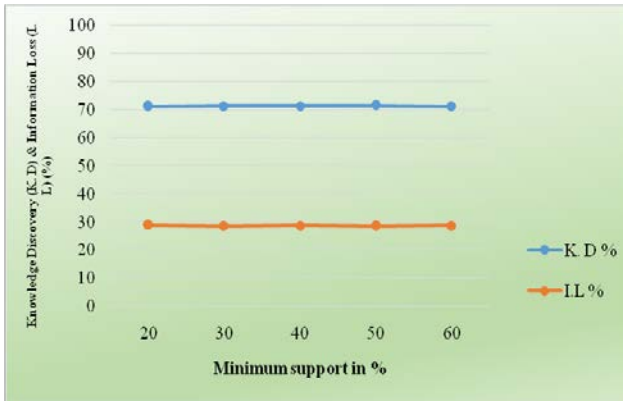

Fig.16. Knowledge discovery, information loss of the processed rules based on minimum support

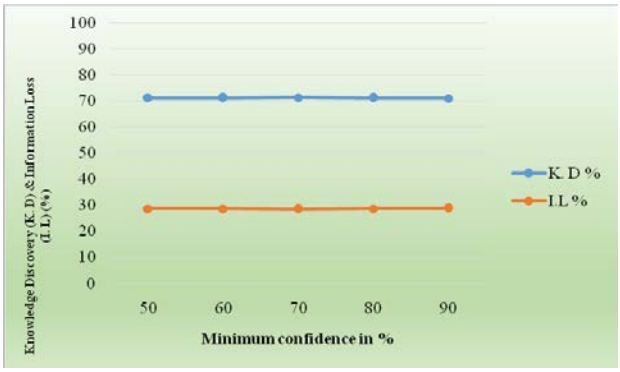

Fig.17. Knowledge discovery, information loss of the processed rules based on minimum confidence

5.9 Comparison Analysis with Existing Works

Figure 18 compares our proposed work with previous works. The earlier algorithm has much more information loss and much less knowledge discovery compared to our proposed algorithm.

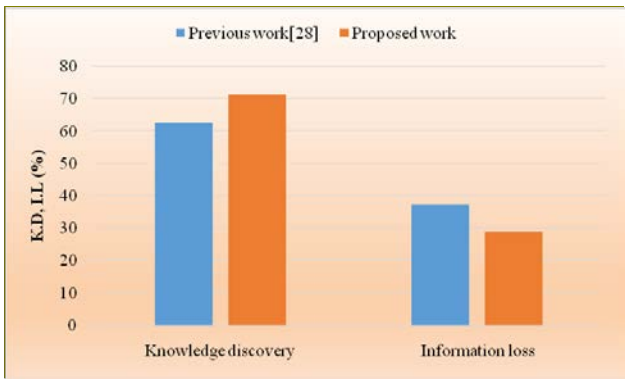

Fig.18. Comparison analysis of knowledge discovery and information loss with previous algorithm

6 Conclusion

We have presented an efficient technique for balanced constraint measure-based algorithm for privacy preserved sequential rule discovery. Initially, we generated the sequential patterns from the medical database through the prefixspan algorithm, after which the sequential patterns was converted into sequential rule. After the sequential rules were generated, we applied our proposed algorithm on sequential rules according to the random value sequential rule. Our proposed algorithm evaluated the processed rule in terms of knowledge discovery and information loss and released the sequential rule if the evaluated value satisfied the threshold values of user defined Knowledge Discovery and Information Loss, else the proposed privacy algorithm continued its modification process until the user defined threshold for the knowledge discovery and information loss was satisfied through updated random values. Finally, an experiment was carried out to evaluate the proposed algorithm on the basis of knowledge discovery and information loss.

References:

[1] Erez Shmueli, Tamir Tassa, Raz Wasserstein, Bracha Shapira, Lior Rokach, "Limiting disclosure of sensitive data in sequential releases of databases", *Journal of Information Sciences*, vol. 191, pp. 98–127, 2012.

[2] R. Agrawal, C. Faloutsos, A. Swami, "Efficient similarity search in sequence databases", *Lecture Notes in Computer Science* 730 (1993) 69–84.

[3] C. Faloutsos, M. Ranganathan, Y. Manolopoulos, "Fast subsequence matching in time-series databases", Proceedings of the *ACM SIGMOD International Conference on Management of Data*, Minneapolis, Minnesota, 1994.

[4] B. LeBaron, A. S. Weigend, "A bootstrap evaluation of the effect of data splitting on financial time series", *IEEE Transactions on Neural Networks* 9 (1) (1998) 213–220.

[5] K. Mehta, S. Bhattacharyya, "Adequacy of training data for evolutionary mining of trading rules", *Journal of Decision Support Systems* 37 (4) (2004) 461–474.

[6] C. Y. Chang, M. S. Chen, C. H. Lee, "Mining general temporal association rules for items with different exhibition periods", *IEEE International Conference on Data Mining*, Maebashi City, Japan, 2002.

[7] C.H. Lee, M. S. Chen, C. R. Lin, "Progressive partition miner: an efficient algorithm for mining general temporal association rules", *IEEE Transactions on Knowledge and Data Engineering* 15 (4) (2003) 1004–1017.

[8] Y. Li, P. Ning, X. S. Wang, S. Jajodia, "Discovering calendar based temporal association rules", *Proceedings of the 8th International Symposium on Temporal Representation and Reasoning, Cividale, Italy*, 2001, pp. 111–118.

[9] R.Srikant, R. Agrawal, "Mining sequential patterns: generalizations and performance improvements", *Research Report RJ 9994, IBM Almaden Research Center*, San Jose, California, 1995.

[10] R.Srikant, R. Agrawal, "Mining sequential patterns: generalizations and performance improvements", *Proceedings of the 5th International Conference on Extending Database Technology*, Avignon, France, 1996.

[11] RSrikant, Y. Yang, "Mining web logs to improve website organization", *Proceedings of the 10th International World Wide Web Conference*, Hong Kong, 2001.

[12] Wen-Chih Peng, Zhung-Xun Liao, "Mining sequential patterns across multiple sequence databases", *Journal of Data & Knowledge Engineering*, Vol. 68, pp. 1014–1033, 2009.

[13] X. Yan, J. Han, "CloSpan: mining closed sequential patterns in large datasets", *Proceedings of the 2003 SIAM International Conference on Data Mining (SDM'03)*, San Francisco, California, May, 2003.

[14] J.Yang, P. Yu, W. Wang, J. Han, "Mining long sequential patterns in a noisy environment", *Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data*, Madison, Wisconsin, 2002, pp. 406–417.

[15] Chung-Wen Cho, Yi-Hung Wu, Arbee L. P. Chen, "Effective database transformation and efficient support computation for mining sequential patterns", Proceedings of the 2005 *International Conference Database Systems for Advanced Applications (DASFAA)*, 2005, pp. 163–174.

[16] Jay Ayres, Jason Flannick, Johannes Gehrke, Tomi Yiu, "Sequential pattern mining using a bitmap representation", *Proceedings of the 2002 ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2002, pp. 429–435.

[17] Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, Helen Pinto, Qiming Chen, Umeshwar Dayal, Meichun Hsu, "PrefixSpan: mining sequential patterns by prefix-projected growth", *Proceedings of the 2001 IEEE International Conference on Data Engineering (ICDE), 2001*, pp. 215–224.

[18] Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, Jianyong Wang, Helen Pinto, Qiming Chen, Umeshwar Dayal, Meichun Hsu, "Mining sequential patterns by pattern-growth: the PrefixSpan approach*", IEEE Transactions on Knowledge and Data Engineering* 16 (11) (2004) 1424–1440.

[19] Rakesh Agrawal, Ramakrishnan Srikant, "Mining sequential patterns", *Proceedings of the 1995 IEEE International Conference on Data Engineering(ICDE),* 1995, pp. 3–14.

[20] Florent Masseglia, Pascal Poncelet, Maguelonne Teisseire, "Incremental mining of sequential patterns in large databases", *Data and Knowledge Engineering* 46 (1) (2003) 97–121.

[21] Hong Cheng, Xifeng Yan, Jiawei Han, Incspan, "Incremental mining of sequential patterns in large database", *Proceedings of the 2004 ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2004, pp. 527–532.

[22] Helen Pinto, Jiawei Han, Jian Pei, Ke Wang, Qiming Chen, Umeshwar Dayal, "Multi-dimensional sequential pattern mining", *Proceedings of the 2001 ACM International Conference on Information and Knowledge Management (CIKM)*, 2001, pp. 81–88.

[23] Neal Lesh, Mohammed Javeed Zaki, Mitsunori Ogihara, "Mining features for sequence classification", *Proceedings of the 1999 ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD),* 1999, pp. 342–346.

[24] Pierre-Yves Rolland, "FlExPat: flexible extraction of sequential patterns", *Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM)*, 2001, pp. 481–488.

[25] Themis P. Exarchos, Markos G. Tsipouras, Costas Papaloukas, Dimitrios I. Fotiadis, "A two-stage methodology for sequence classification based on sequential pattern mining and optimization", *journal of Data and Knowledge Engineering* 66 (3) (2008) 467–487.

[26] Osman Abul, Francesco Bonchi, and Fosca Giannotti, "Hiding Sequential and Spatiotemporal Patterns", *IEEE transactions on knowledge and data engineering*, Vol. 22, no. 12, pp. 1709-1723, 2010.

[27] Yen-Liang Chen, Ya-Han Hu, "Constraint-based sequential pattern mining: The consideration of recency and compactness", *Journal of Decision Support Systems*, Vol. 42 pp. 1203–1215, 2006.

[28] S.S. Arumugam and Dr. V. Palanisamy, "An Efficient Algorithm for Privacy Preserving Temporal Pattern Mining", *Journal of Theoretical and Applied Information Technology*, Vol. 58, December 2013.

[29] Jieh-Shan Yeh and Po-Chiang Hsu, "HHUIF and MSICF: Novel algorithms for privacy preserving utility mining", *Journal of Expert Systems with Applications*, Vol. 37, pp. 4779–4786, 2010.

[30] Tiancheng Li, Ninghui Li, Jian Zhang, and Ian Molloy, "Slicing: A New Approach for Privacy Preserving Data Publishing", *IEEE transactions on knowledge and data engineering*, Vol. 24, no. 3, 2012.

[31] En Tzu Wang and Guanling Lee, "An efficient sanitization algorithm for balancing information privacy and knowledge discovery in association patterns mining", *Journal of Data & Knowledge Engineering*, Vol. 65, pp. 463-484, 2008.

[32] Weijia Yang and Sanzheng Qiao, "A novel anonymization algorithm: Privacy protection and knowledge preservation", *Journal of expert system with application*, Vol. 37, pp. 756-766, 2010.