

2022

## Exploring the Existing and Unknown Side Effects of Privacy Preserving Data Mining Algorithms

Hima Bindu Sadashiva Reddy

Follow this and additional works at: [https://nsuworks.nova.edu/gscis\\_etd](https://nsuworks.nova.edu/gscis_etd)



Part of the [Communication Technology and New Media Commons](#), [Computer Sciences Commons](#),  
and the [Library and Information Science Commons](#)

## Share Feedback About This Item

---

This Dissertation is brought to you by the College of Computing and Engineering at NSUWorks. It has been accepted for inclusion in CCE Theses and Dissertations by an authorized administrator of NSUWorks. For more information, please contact [nsuworks@nova.edu](mailto:nsuworks@nova.edu).

**Exploring the Existing and Unknown Side Effects of Privacy Preserving Data Mining  
Algorithms**

by

Hima Bindu Sadashiva Reddy

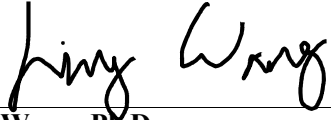
A dissertation submitted in partial fulfillment of the  
requirements for the degree of Doctor of Philosophy  
in  
Information Systems

College of Computing and Engineering  
Nova Southeastern University

2022



We hereby certify that this dissertation, submitted by Hima Bindu Sadashiva Reddy conforms to acceptable standards and is fully adequate in scope and quality to fulfill the dissertation requirements for the degree of Doctor of Philosophy.



Ling Wang, Ph.D.  
Chairperson of Dissertation Committee

9/9/22  
Date



Ajoy Kumar, Ph.D.  
Dissertation Committee Member

9/9/22  
Date



Martha M. Snyder, Ph.D.  
Dissertation Committee Member

9/9/22  
Date

Approved:



Meline Kevorkian, Ed.D.  
Dean, College of Computing and Engineering

9/9/22  
Date

College of Computing and Engineering Nova  
Southeastern University

2022

An Abstract of a Dissertation Submitted to Nova Southeastern University  
in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

Exploring the Existing and Unknown Side Effects of Privacy Preserving Data Mining  
Algorithms

by  
Hima Bindu Sadashiva Reddy  
July 2022

The data mining sanitization process involves converting the data by masking the sensitive data and then releasing it to public domain. During the sanitization process, side effects such as hiding failure, missing cost and artificial cost of the data were observed. Privacy Preserving Data Mining (PPDM) algorithms were developed for the sanitization process to overcome information loss and yet maintain data integrity. While these PPDM algorithms did provide benefits for privacy preservation, they also made sure to solve the side effects that occurred during the sanitization process. Many PPDM algorithms were developed to reduce these side effects. There are several PPDM algorithms created based on different PPDM techniques. However, previous studies have not explored or justified why non-traditional side effects were not given much importance.

This study reported the findings of the side effects for the PPDM algorithms in a newly created web repository. The research methodology adopted for this study was Design Science Research (DSR). This research was conducted in four phases, which were as follows. The first phase addressed the characteristics, similarities, differences, and relationships of existing side effects. The next phase found the characteristics of non-traditional side effects. The third phase used the Privacy Preservation and Security Framework (PPSF) tool to test if non-traditional side effects occur in PPDM algorithms. This phase also attempted to find additional unknown side effects which have not been found in prior studies. PPDM algorithms considered were Greedy, POS2DT, SIF\_IDF, cpGA2DT, pGA2DT, sGA2DT. PPDM techniques associated were anonymization, perturbation, randomization, condensation, heuristic, reconstruction, and cryptography. The final phase involved creating a new online web repository to report all the side effects found for the PPDM algorithms. A Web repository was created using full stack web development. AngularJS, Spring, Spring Boot and Hibernate frameworks were used to build the web application. The results of the study implied various PPDM algorithms and their side effects. Additionally, the relationship and impact that hiding failure, missing cost, and artificial cost have on each other was also understood. Interestingly, the side effects and their relationship with the type of data (sensitive or non-sensitive or new) was observed. As the web repository acts as a quick reference domain for PPDM algorithms. Developing, improving, inventing, and reporting PPDM algorithms is necessary. This study will influence researchers or organizations to report, use, reuse, or develop better PPDM algorithms.

## Acknowledgments

A special thanks to my mother, the late Nagamani Gandluri, for sacrificing her life to see her children happy and successful. Amma, your important words will always be remembered:

*"No matter what, always be calm, happy, brave, fearless, truthful, loyal, and independent."*

This study was completed successfully only because of the super supportive dissertation committee chair and members, Dr. Ling Wang, Dr. Ajoy Kumar, and Dr. Marti Snyder. Their valuable time, guidance, and feedback helped me gain more knowledge.

I'm thankful to my brothers Roopesh Reddy and Ratnaditya Jonnalagadda for helping financially to finish my doctorate degree.

My thanks to Avinash Gogineni, Andrea Green, John Papavaritis, Jamaica Jordan, Evangeline Faraldo, Henry Faraldo, Rosalie Faraldo, Elizabeth Faraldo, Ellie, the late Dr. Ravi, and everyone else who helped, guided, and supported me during my doctoral studies.

May all beings live in happiness, peace, and harmony.

## Table of Contents

<b>Abstract</b>	<b>ii</b>
<b>List of Tables</b>	<b>iv</b>
<b>List of Figures</b>	<b>v</b>

### Chapters

<b>1. Introduction</b>	<b>1</b>
Background	1
Problem Statement	2
Dissertation Goal	5
Research Questions	6
Relevance and Significance	6
Barriers and Issues	9
Assumptions, Limitations and Delimitations	9
Definition of Terms	10
Acronyms Used in this Dissertation	10
Summary	13
<b>2. Review of the Literature</b>	<b>14</b>
Introduction	14
Critical Review of Articles	14
Privacy Preserving Data Mining Algorithms	14
Performance or Evaluation Criteria	15
Limitations	19
Side Effects	19
Design Science Research	22
Summary	25
<b>3. Methodology</b>	<b>27</b>
Overview of Research Methodology	27
Research Methods	27
Phase One	35
Phase Two	35
Phase Three	36
Phase Four	36
Instrument Development and Validation	38
Sampling	35
Data Analysis	41
Formats for Presenting Results	43
Resources	44
Summary	44
<b>4. Results</b>	<b>46</b>
Data Collection	46

Phase One	51
Phase Two	53
Phase Three	55
Phase Four	57
Findings	68
Phase One	68
Phase Two	76
Phase Three	82
Phase Four	97
Summary	104
<b>5. Conclusions, Implications, Recommendations and Summary</b>	<b>105</b>
Conclusions	105
Implications	114
Recommendations	115
Summary	116
<b>References</b>	<b>111</b>



## **List of Tables**

### **Tables**

1. Research Questions and the Artifacts Implementation in Phases 31
2. PPDM Algorithms Selected from PPSF 39
3. Datasets Selected from SPMF 40
4. Expected Data in Each Phase 42
5. Maven Project Related Details 59
6. REST Endpoint for PPDM Algorithms Application 62
7. Identified Common Side Effects 71
8. Relevant Information Retrieved from the Research Studies 73
9. Non-traditional Side Effects 77
10. PPDM Algorithms and Related Side effects Evaluated 81
11. Small Datasets and Output Files used for Each Algorithm 83
12. Small Dataset File's Details for each Row 83
13. Sensitive File Details Showing Sensitive Itemset Selected from each Row in Small Dataset 83
14. REST Endpoint Results 97

## **List of Figures**

### **Figures**

1. DSRM Approach Demonstrating Different Phases of this Study 30
2. The Three-step Literature Review Process for Information Systems Research 32
3. Five Stages of Systematic Literature Review 33
4. JabRef Tool Used to Convert BibTeX to CSV format 45
5. Finding Duplicate Values 48
6. Displaying the Duplicate Values in Red 48
7. Selecting the “Remove Duplicate” option in “Data Tools” 49
8. Remove Duplicates Based on “Title” 49
9. Number of Duplicate Values Removed, and Unique Value Remained are Shown 50
10. Flowchart for Phase One Systematic Review 51
11. Flowchart for Phase Two Systematic Review 54
12. Greedy Algorithm Running in PPSF Tool 57
13. Created PpdmAlgorithms Application Using Spring Initializer 63
14. Spring Tool Suite Project Directory Structure 60
15. PostgreSQL Dependency Details in pom.xml 61
16. PostgreSQL Connection Details in Application Properties File 62
17. PPDm Domain Implementation 63
18. JpaRepository Dependency in pom.xml 64
19. Spring Tool Suite’s Console Showing the Status as Started 65
20. Table Ppdm\_algorithms Auto Created by JPA’s @Entity and @Table Annotations 66
21. AngularJS and BootStrap Dependencies in pom.xml 67

22. AngularJS Route Defined as /main for Homepage	68
23. Single Page Application Directory Structure	70
24. Data Sanitization Process	71
25. PPSF Tool Showing sGA2DT Algorithm is Running	88
26. PPSF Tool Showing pGA2DT Algorithm is Running	89
27. PPSF Tool Showing cpGA2DT Algorithm is Running	90
28. PPSF Tool Showing cpGA2DT Algorithm Testing was Interrupted due to Index Out of Bound Error	91
29. PPSF Tool Showing pGA2DT Algorithm is Running and Finished	92
30. PPSF Tool Showing Greedy Algorithm is Running and Stats after Finishing	93
31. PPSF Tool Showing PSO2DT Algorithm Testing was Interrupted due to Index Out of Bound Error	94
32. Greedy Algorithm Testing for Customized Small Dataset	95
33. PSO2DT Algorithm Testing for Customized Small Dataset and Index Out of Bound Error Encountered	96
34. cpGA2DT Algorithm Testing for Customized Small Dataset and Index Out of Bound Error Encountered	97
35. pGA2DT Algorithm Testing for Customized Small Dataset with After Execution Stats	98
36. sGA2DT Algorithm Testing for Customized Small Dataset with After Execution Stats	99
37. REST Endpoint Created New PPDM Algorithm Successfully	101
38. REST Endpoint Successfully Displaying All PPDM Algorithms	101

- 39. REST Endpoint Successfully Updated cpGA2DT PPDM Algorithm 102
- 40. REST Endpoint Successfully Requesting to Delete cpGA2DT PPDM Algorithm by id  
“1” 102
- 41. REST Endpoint Successfully Deleted cpGA2DT PPDM Algorithm 103
- 42. The Main Page of PPDM Algorithm Repository 103
- 43. Report Algorithm Page of PPDM Algorithm Repository 104
- 44. List Algorithms Page of PPDM Algorithms Repository 104
- 45. Report Algorithm Page with Details Before Clicking Report Button 105
- 46. List Algorithms Page with Details Showing PPDM Algorithms Reported 105
- 47. Edit in List Algorithms Page Redirects to Update Algorithms Page 106
- 48. Email is Successfully Updated as Shown in List Algorithms Page 106

## **Chapter 1**

### **Introduction**

#### **Background**

According to Bélanger and Crossler (2011), information privacy is a subset of overall concepts of privacy related to four dimensions: privacy of a person, personal behavior privacy, personal communication privacy, and personal data privacy. Definitions of privacy are ambiguous. Lampinen et al. (2013) considered privacy in the social media domain as “an interpersonal boundary process by which a person or group regulates interaction with others” (p. 57). Smith et al. (2011) discussed that there is no single concept for privacy and defined it as limited access to information. During the process of data sharing, the preservation of privacy mainly deals with protecting confidential information from being stolen and misused by fraudsters (Menzies et al., 2014). Confidential information includes SSN details, user transactions, date of birth, contact details, medical details, purchase history, credit history, passwords, and bank account information. Hence, preserving privacy has become an important topic for researchers due to the pervasiveness of computer systems, data, Internet users, transactions, data collections, and data analysis.

Protecting sensitive information is given importance during the privacy preservation process (Aggarwal & Philip, 2008). This is done by using data distortion, data reconstruction, and data encryption technology (Sharma et al., 2013). Several types of privacy preserving techniques are heuristic-based, reconstruction, and cryptography-based (Patel, 2016). In recent years, the importance of privacy preserving has increased due to extensive growth of data extraction. Therefore, during the knowledge-mining process Privacy Preserving Data Mining (PPDM) was incorporated to ensure there is no leakage of the sensitive information (Chaudhary

et al., 2013). Different methods of PPDM used were association rule mining, association rule hiding, downgrading classifier effectiveness, query auditing, and inference control (Mendes & Vilela, 2017). The algorithms developed so far were unable to cope with the enormous increase in data collection, data transfer, and database size (Aggarwal et al., 2015). Aggarwal et al. (2015) compared the MapReduce algorithm with MapReduce Top-Down Specialization (MRTDS) and integrating models such as k-anonymity and l-diversity. The integrating models were used to tackle information loss during privacy preservation in big data. The results showed significant degradation in performance and an increase in privacy preservation iterations in the MapReduce algorithm as compared to MRTDS; however, side effects were not discussed. Side effects determine the authenticity of protecting confidential information in PPDM algorithms. They are important to evaluate the characteristics, performance, and data quality of the PPDM algorithm during the sanitization process. Few research studies apply association rule hiding algorithms such as Hiding-Missing-Artificial Utility (HMAU) (Shah et al., 2012; Gayathiri & Poorna, 2015; Laskar & Lachit, 2014). This algorithm was adopted to prevent information loss. The side effects were used to calculate efficiency and execution time of the algorithm's hiding failure rates. These studies implied that PPDM techniques were used to inspect side effects during the mining process of sensitive information. This indicates the need to investigate the in-depth details of the side effects instead of just calculating the number of occurrences.

### **Problem Statement**

Fournier et al. (2014) implemented an open-source data mining library named Sequential Pattern Mining Framework (SPMF). According to their website, as of the year 2021, there are around 200 data mining algorithms included in SPMF software. In addition to a user-friendly interface to run each of the algorithms, SPMF also compares the performance of algorithms.

A vast number of algorithmic techniques have been designed for Privacy Preserving Data Mining (Aggarwal & Philip, 2008). Hiding failure (HF), missing cost (MC), and artificial cost (AC) are three types of side effects traditionally used for PPDM (Lin et al., 2016; Lin et al., 2014a).

Before data sharing, failure to hide sensitive information during the sanitization process is called HF (Lin et al., 2019). MC occurs when the sanitization process hides data that is considered “not sensitive” (Brown & Kros, 2003). Failure to stop generating not useful information or artificial data is observed as AC (Lin et al., 2019).

As data sanitization plays a significant role in ensuring that the mined database maintains its confidentiality and originality; the side effects were used for evaluation and measuring the performance of PPDM algorithms during the data sanitization process (Lin et al., 2014a).

Wang et al. (2007) considered HF with three different side effects hidden rules, new rules generated, and lost rules to evaluate the characteristics of two new algorithms. The algorithms were based on the association rule technique.

Lin et al. (2019) recommended a multiobjective algorithm for PPDM by considering four side effects to measure the performance of data sanitization. The four side effects studied were HF, MC, AC, and data dissimilarity.

Moreover, Wimmer and Powell (2014) investigated the feasibility of applying the K-Anonymity PPDM algorithm with data mining and machine learning algorithms. The results were positive for testing sweeney, cancer, and income datasets. Future work was suggested to compare additional PPDM algorithms.

On the other hand, Lawrence et al. (2016) compared and analyzed various PPDM algorithms and techniques. The PPDM algorithms selected were random perturbation, k-

anonymity, horizontally partitioned distribution, vertically partitioned distribution, clustering, classification, association rule mining, secured sum computation, aggregation. PPDM techniques studied were using fuzzy logic, cryptography, and neural network learning.

However, there has been no evidence or information provided by the research studies that discuss if other non-traditional side effects which are data dissimilarity, hidden rules, new rules, and lost rules are observed in all the PPDM algorithms. Three side effects HF, MC, and AC were traditionally used (Lin et al., 2019). Why are other non-traditional side effects ignored in analyzing PPDM algorithms? Are there other unknown side effects that are yet to be studied? Lin et al. (2014a) evaluated the performance of compact prelarge Genetic Algorithm to Delete Transactions (cpGA2DT), “The side effects of artificial cost are also evaluated to show the performance of the proposed cpGA2DT” (p. 10). Analysis of the study by Lin et al. (2014a), raises a question, if there are unknown side effects of HF, MC, and AC?

The research literature lacked in providing a comparison of all PPDM algorithms based on side effects. SPMF provided a repository for all sequential pattern data mining algorithms, their implementation and performance comparisons. The Common Vulnerabilities and Exposures (CVE) website, maintains information about computer security errors. Similarly, there exists no specific database focusing on PPDM algorithm issues or side effects. Were the users finding any new side effects from existing algorithms? Where were the side effects reported? There are many PPDM algorithms developed to hide sensitive information. Hence, there is a need to maintain a common repository to keep track of algorithms developed and their side effects.

This study discovered, compared, and collected side effects among all PPDM algorithms. The first step was to find if non-traditional side effects that occurred in all PPDM algorithms. Second, it was important to find unknown side effects. Third, a common online repository was



created to report the side effects of PPDM algorithms. The present research study helped to discover side effects occurring in all PPDM algorithms and store them in a common web repository. Side effects information was gathered from previous research studies, and the PPSF tool. Finding a balance between hiding information and side effects is a crucial research area due to the severity of handling sensitive information (Chun-Wei et al., 2018). The emphasis should be given to discover different side effects to help find better PPDM solutions or algorithms to continue maintaining the quality, accuracy and confidentiality of sensitive information.

### **Dissertation Goal**

There were many kinds of research studies that have analyzed the performance of data mining algorithms, PPDM algorithms, and side effects (Celik et al., 2017; Arboleda, 2019; Hussain, 2019; Nopour et al., 2021; Abdar et al., 2015). However, extremely limited research comparing all PPDM algorithms based on their side effects has been investigated. Also, there exists no database to report or maintain information on PPDM algorithms' side effects. The initial goal of this study involved gathering information related to the side effects of PPDM algorithms. The details of the information consisted of already existing side effects, and unknown side effects. The final goal was to create an online website to report these side effects for all PPDM algorithms.

In this study, PPDM algorithms considered were based on privacy preserving techniques. The privacy preserving techniques included heuristic-based, cryptography-based, reconstruction-based, greedy-based, data hiding, knowledge hiding, and hybrid techniques (& Vaghashia & Ganatra, 2015; Bhagat & Shelke, 2015; Lin et al., 2013). Vaghashia and Ganatra 2015) studied five PPDM techniques, comparing them based on both advantages and limitations. Results reported that the randomized technique was most efficient compared to the cryptographic

technique. Cryptographic techniques ranked highest in privacy, and the other four techniques faced huge information loss. The research was from the year 2015 and covers only five PPDM techniques. Additionally, there was no information given about the greedy-based technique, and neither of the techniques specified any algorithm names nor compared their side effects.

The recommended study found the side effects occurring in every PPDM algorithm until now and explored if there are any unknown side effects to be discovered. Additionally, this research created a new online database repository to store the information of all PPDM algorithms with their side effects. A common repository was necessary to report the side effects existing across all types of PPDM algorithms.

### **Research Questions**

Research questions for this study were:

RQ1: What were the similarities and differences of the existing side effects of PPDM algorithms?

RQ2: How were the side effects related to one another?

RQ3: What were the non-traditional side effects, and do they occur in PPDM algorithms?

RQ4: What were the unknown side effects occurring in PPDM algorithms?

RQ5: Where and how were the side effects of all PPDM algorithms reported?

### **Relevance and Significance**

Exploring further on the research studies conducted in the recent years, the importance of privacy, privacy preserving, PPDM, PPDM algorithms, and PPDM side effects have been analyzed deeper in this section. The main intention of PPDM was to ensure that data privacy and quality were preserved with the evolution of various data mining techniques (Mendes & Vilela,

2017). Significant examples collected in the study gave an insight into real-time privacy breach of patient's sensitive information related to diseases and illnesses such as flu, HIV, and lung cancer in compliance with the HIPAA rules.

The privacy models developed avoided privacy breaches and ensured there was no information leak. The privacy breaches ranged from different information loss and malicious intruders' hacking over Wireless Sensor Networks (WSN). The necessity of PPDM techniques was to overcome information loss, maintain consistency in privacy levels based on complexity, metrics, and feasibility. During the process of applying these PPDM techniques or algorithms, side effects emerged where privacy breach germinated. In health care, preserving patient's sensitive data was crucial to avoid privacy breaches. This proved the importance of PPDM applications in various fields to overcome privacy breaches, and loss of sensitive information. Kamakshi and Babu (2012) innovatively discovered new PPDM techniques, which investigated the needs of big organizations and government agencies to rapidly preserve the privacy in ever-increasing data. The issues addressed were concerning the public disclosure of sensitive information gathered from banking, healthcare systems, insurance companies, and government sources. The research study portrayed the importance of preserving the data from hackers by replacing original data with realistically false ones with help of a swapping technique.

Based on the research studies from the past fifteen years, an extensive literature review on PPDM was conducted by Aldeen et al. (2015). Related to the phishing issue over the Web, the researchers explored various advantages of PPDM techniques. PPDM techniques were not as simple as they sounded, the techniques were designed and applied based on data distribution. They explained the importance of developing cost-effective, robust, and accessible PPDM techniques by discovering the major disadvantages that outperformed the advantages. The

disadvantages ranged from data disclosure, attacks through the Internet, incremental data privacy issues in cloud computing, the integrity of mining results, data utility, scalability, and performance overhead. One of the root causes for PPDM techniques in failing to hide sensitive information was the tremendous growth of Information Technologies (IT).

Research studies have established that privacy breaches occur in various fields such as health care, wireless networks, global positioning system (GPS), Internet, mobile technologies, World Wide Web, banking, cloud services, and other organizations. Privacy preservation is imperative to overcome privacy breaches such as information leaks, unauthorized access, and information misuse. Hiding sensitive information comes with a risk from the side effects by introducing redundant information or even failure to hide sensitive information (Lin et al., 2016c). These side effects costed a fortune to data providers by indirectly helping their business rivals to successfully make business decisions, by exploiting the critical sensitive information gathered from the shared database (Lin et al., 2017). The damages from the side effects were not only confined to sensitive information; but non-sensitive information also dealing with many issues (Chen et al., 2020). Issues of side effects for non-sensitive information, were increased information loss and distortion of data. An Itemset Oriented Pseudo Graph Based Sanitization (IPGBS) algorithm was implemented to minimize such occurrences of information loss or data distortion during the process of hiding non-sensitive information in both dense and sparse databases (Ergenç Bostanoğlu & Öztürk, 2020). As the demand for protecting the sensitive information of an individual or an organization was increasing, the need for innovative PPDM techniques also increased. It can be inferred that privacy preserving is significant, sequentially PPDM, and PPDM algorithms were even more important to maintain the integrity of privacy preservation. More importantly, exploring the unknown side effects during PPDM is important;

to avoid any occurrences of unknown damages which might be more severe than the currently existing ones, such as hackers gathering credit card credentials, social security numbers, medical records, stalkers/human traffickers collecting victims' personal details (phone number, address, family, pictures, and social media accounts), or even theft. Many incidents of stalking, murder, and human trafficking by spying and gathering information from social media or online databases have also occurred.

### **Barriers and Issues**

There is no online forum for the PPSF tool for reporting any issues encountered with the software. In March 2020 the PPSF tool was first downloaded to collect the details of the PPDM algorithms implemented. Since January 2021, PPSF website was temporarily unavailable to download the software. Jerry Li was one of the researchers who developed the PPSF software tool (Lin et al., 2018d). Philippe is the inventor and developer of the SPMF tool and has significantly contributed to creating the PPSF tool (Fournier-Viger et al., 2014; Lin et al., 2018d). Professors Jerry Li and Philippe informed that PPSF was currently implementing more algorithms. Hence, the website was under maintenance. Jerry Li provided the older version of the PPSF tool to continue the present research work. Any issue encountered with the tool delayed the data analysis for this research study. Response time from the PPSF project leaders was within one day. The new website that will be built might encounter delays in debugging and fixing any errors.

### **Assumptions, Limitations and Delimitations**

One limitation of this study was the selection of the datasets based on the size limit; because of the address space constraints, the 64-bit Windows operating system could handle only up to a certain dataset size. Only six PPDM algorithms were available within the PPSF tool, to

test the side effects. Delimitations of this study are PPDM algorithm performance, and the database side effects are not considered.

### **Definition of Terms**

**Data Sanitization** – “Data sanitization methods aim at making data publishable while providing protection guarantees against disclosures and at the same time maintaining the usefulness of the data.” (Sramka et al., 2010, p. 1)

**Data mining** – “Data Mining”, often also referred to as “Knowledge Discovery in Databases” (KDD), is a young sub-discipline of computer science aiming at the automatic interpretation of large datasets.” (Kriegel et al., 2007, p. 87)

**Privacy preserving data mining** – “Privacy-Preserving Data Mining (PPDM) is a data mining technique for hiding the private and critical information in a dataset.” (Wu et al., 2017, p. 10024)

**Hiding failure** – “The portion of sensitive information that is not hidden by the application of a privacy preservation technique” (Bertino et al., 2008, p. 3)

**Missing cost** – “The missing cost is the set of non-sensitive frequent itemsets appearing in the original database that cannot be discovered in the sanitized database.” (Lin et al., 2016, p. 271)

**Artificial cost** – “The artificial cost indicates that the information was not concerned as the useful knowledge from the original database but will be arisen as the rules against to the threshold value after the sanitization progress.” (Lin et al., 2019, p. 12780)

### **Acronyms Used in this Dissertation**

AC: Artificial cost

ACO: Ant Colony Optimization

ACS2DT: Ant colony system-based algorithm ant colony system-based algorithm

ADR: Action Design Research

CCO: Cybercrime classification ontology

cpGA2DT: Compact prelarge Genetic Algorithm to delete transactions

CSV: comma-separated values

CVE: Common Vulnerabilities and Exposures

DSR: Design Science Research

DSRM: Design Science Research Methodology

DSS: Database Structure Similarity

DUS: Database Utility Similarity

FPUTT: Fast Perturbation algorithm Using a Tree structure and Tables

GA: Genetic Algorithm

GPS: Global positioning system

HF: Hiding failure

HHUIF: Hiding High Utility Itemset First

HMAU: Hiding-Missing-Artificial Utility

HTML: HyperText Markup Language

HUPEumu-GRAM: High utility pattern extraction using genetic algorithms with ranked  
mutation using minimum utility threshold

IDE: Integrated Development Environment

IPGBS: Itemset Oriented Pseudo Graph Based Sanitization

IS: Information System

IT: Information Technologies

IUS: Itemsets Utility Similarity

LSH: Locality-Sensitive Hashing

MC: Missing cost

MRTDS: MapReduce Top-Down Specialization

MSCIF: Maximum Sensitive Itemsets Conflict First

MSU-MAU: Maximum Sensitive Utility-MAXimum item Utility

MSU-MIU: Maximum Sensitive Utility-MInimum item Utility (MSU-MIU)

MVC: Model View Controller a Java framework

NN: Nearest-Neighbor

NSGA II: GA based multiobjective algorithm

NSGA2DT: A newly designed multiobjective algorithm

OCR: Optical character recognition

PPSF: Privacy Preservation and Security Framework

PPUMGA+: Privacy Preserving an evolutionary sanitization algorithm using transaction  
insertion

PPUMGAT- The PPUMGAT algorithm without the pre-large concept

PPUMGAT: Privacy-Preserving Utility Mining by adopting a GA-based approach for transaction  
deletion

PPUMGAT+: The PPUMGAT algorithm with the pre-large concept

PSO: Particle Swarm Optimization

PSO2DT: Particle Swarm Optimization to Data Deletion

SLR: Systematic Literature Review

SPMF: Sequential Pattern Mining Framework

STS: Spring Source Tool

TbIAS: Text-based Intelligent Assistant system



TPD: Teacher Professional Development

WSN: Wireless Sensor Networks

### **Summary**

The introduction started with a background covering the functionalities and relationships of data sharing, information privacy, preservation of privacy, protecting confidential/sensitive information, privacy-preserving techniques, big data, knowledge-mining process, PPDM and PPDM algorithms.

The problem statement addressed the research gap to find both known and unknown side effects in PPDM algorithms. The need for and importance of comparing the side effects of all PPDM algorithms was discussed. Importantly, creating an online repository to report issues and side effects for PPDM algorithms were discussed. Five research questions were developed based on the research problem identified.

Initial and final goals were discussed. Barriers and issues were related to the PPSF tool's lack of online help/forum. The limitation was related to choosing the datasets based on size limit, as 64-bit Windows operating system should have the capacity to run the datasets of selected size. To protect confidential information with minimum side effects, PPDM algorithms play a major role in supporting privacy preservation.

## **Chapter 2**

### **Review of Literature**

#### **Introduction**

The more the sensitive information is protected, the higher the risk of possible side effects is generated (Lin et al., 2016). The purpose of this study was to report known and unknown side effects of PPDM algorithms in a newly created web repository. The two main topics identified to establish the viability for exploring the side effects of privacy preserving algorithms are PPDM algorithms and side effects. The side effects were HF, MC, and AC. One of the aims of this study was to discover the different side effects that occur when using the PPDM algorithms. Exploring the literature in connection with the research problem detected, diverse research studies which helped to hypothesize two constructs were PPDM algorithms and their side effects: hiding failure, missing cost, and artificial cost (Lin et al., 2016c). Privacy preserving focuses on protecting sensitive information (Aggarwal & Philip, 2008) during the knowledge extraction process by using data distortion, data reconstruction, and data encryption technology (Sharma et al., 2013). Types of privacy preserving techniques were based on anonymization, perturbation, randomization, condensation, heuristic approaches, reconstruction, and cryptographic approaches (Malik et al., 2012; Patel, 2016). A vast number of algorithmic techniques were designed for Privacy Preserving Data Mining (Aggarwal & Philip, 2008). These side effects were used for sanitization, evaluation, and measuring performance. They were even examined to learn how to reduce other side effects of the PPDM algorithms.

#### **Critical Review of Articles**

##### ***Privacy Preserving Data Mining Algorithms***

**Performance or Evaluation Criteria.** In recent years, the importance of privacy preserving has increased due to extensive growth of data extraction, hence Privacy Preserving Data Mining (PPDM) is incorporated to ensure there is no loss of the sensitive information during the knowledge mining process and to attain accurate results (Sharma et al., 2013). Different methods of PPDM used were association rule mining, association rule hiding, downgrading classifier effectiveness, query auditing, and inference control (Mendes & Vilela, 2017). The algorithms developed so far were unable to cope with the enormous increase in data collection, data transfer, and database size (Zakerzadeh et al., 2015). To preserve privacy in data sets, MapReduce algorithm was developed based on anonymization method. The authors experimented by applying a MapReduce algorithm in comparison with MRTDS algorithm. Integrating models such as k-anonymity and l-diversity were used to combat the large crowd effect information loss, during privacy preservation in big data. The results showed significant degradation in performance and an increase in privacy preservation iterations of the MapReduce algorithm; however, side effects were not discussed in this study.

Tamil Selvan and Veni (2015) compared PPDM based on the number of files, privacy level, throughput, and privacy preserving efficiency. Association rule mining technique was used to compare PPDM with optimal side effects. The number of files ranged from 25 to 200. Nearest-Neighbor (NN) and Locality-Sensitive Hashing (LSH) resulted in the highest privacy levels compared to HMAU and PPDM algorithms. As far as throughput is considered, HMAU scored higher than PPDM, NN, and LSH. PPDM was more efficient than the HMAU algorithm, NN, and LSH for privacy preserving efficiency parameter. Names of PPDM algorithms were not discussed.

Wu et al. (2017) introduced a new algorithm called an ant colony system-based algorithm (ACS2DT). The algorithm was developed using ant-based framework called Ant Colony Optimization (ACO). The prime intention of the ACS2DT algorithm was to increase the performance and reduce side effects of the sanitization process in contrast to evolutionary algorithms. The evolutionary algorithms were Genetic Algorithm (GA), Particle Swarm Optimization (PSO), or ACO. The authors used three real-world datasets called chess, mushroom, and food mart. The datasets were used to hide sensitive information through the transaction deletion process, parallelly minimizing the side effects. The ACS2DT algorithm was developed with the help of Java programming language on a supercomputer with a Linux operating system. The parameters considered in this experiment were the total number of transactions, number of distinct items, average transaction length, maximal length transactions, and dataset type. The overall results indicated that the ACS2DT algorithm outperformed evolutionary and greedy algorithms. The performance was based on generating a small number of three side effects: HF, AC, and MC.

A research study conducted by Lin et al. (2017) focused on minimizing the side effects HF, AC, and MC during the process of hiding High Utility Items sets (HUIs). A new algorithm called PPUMGAT was planned because the traditional approaches violated the rules to protect information and selecting transactions to minimize side effects. Genetic based method was used to design PPUMGAT algorithm. This experiment used five real-world datasets called chess, mushroom, accidents, food mart, and retail. The three criteria incorporated were runtime, side effects, and data integrity. The parameters considered were the total number of transactions, data-set type (sparse or dense), number of distinct items, average transaction length, and maximal transaction length. PPUMGA+ (a transaction insertion), PPUMGAT+ (pre-large

concept) and PPUMGAT- (without the pre-large concept) were the evolutionary algorithms used for comparison. HHUIF, a non-evolutionary algorithm was also studied. The population size was limited to 20. The results indicated that the PPUMGAT+ algorithm had a faster runtime in preserving high database integrity with 100% accuracy.

A research study conducted by Xu et al. (2015) suggested a different strategy called randomization and SMC based approaches for the privacy preservation machine learning algorithm. In addition to incorporating the dual ascent algorithm, the MapReduce framework was adopted to explore the study. The two main issues which were given importance were revelation and loss of sensitive information. A total of 101 feature attributes along with 116,289 data instances were considered. Real scenario dataset for the experiment was breast cancer, Higgs bossons and handwritten optical character recognition (OCR). Each dataset was classified based on the classification ratio, Higgs bossons were the hardest to classify. These datasets were used to analyze performance, refine unclear information, and learn the relationship between the attributes. The experiment suggested a new protocol for dispensed feature selection. Simulation results showed that, the performance of Higgs bossons was the hardest because the knowledge was not easy to read. However, there was no discussion about side effects of the algorithm.

Lin et al. (2016c) implemented new algorithms, to prevent the major issue of publicly publishing or sharing of confidential information in the data mining process. Algorithms were developed based on the concept of optimization approach. The two new algorithms are Maximum Sensitive Utility-MAXimum item Utility (MSU-MAU) and Maximum Sensitive Utility-MINimum item Utility (MSU-MIU). Another reason to introduce these algorithms is to overcome the side effects during the sanitization process in comparison to HHUIF and MSCIF algorithms. The authors assume the possible occurrences of three side effects HF, MC, and AC.

The experiment used a shopping mall dataset which includes sold food products (Lin et al., 2016c). Another dataset contained 23 species from mushroom family of *Agaricus* and *Lepiota*. Priority was given to explore the minimization of the side effects rather than execution time. FPUTT, HHUIF and MSICF generated same results for side effects testing. FPUTT algorithm was mainly considered to compare the speed of sanitization and choosing a victim item process. Whereas HHUIF and MSICF algorithms compared the performance, FPUTT and HHUIF fared same in perturbation (choose a victim) process. FPUTT performed far better in speeding the sanitization process. HF, MC, and AC were considered to evaluate the performance of algorithms in addition to Database Structure Similarity (DSS), Database Utility Similarity (DUS), and Itemsets Utility Similarity (IUS) measures. Although with scant food mart dataset MSU-MIU showed the highest results in terms of performance among other algorithms. For number of modified transactions to speed up sanitization process HHUIF and MSICF algorithms excelled. Data base structure similarity was also considered as evaluation criteria, results portray that MSU- MAU and MSU-MIU algorithms performed way better than the HHUIF and MSICF algorithms. For another evaluation criteria called IUS, MSU-MAU, and HHUIF algorithm performance were the same.

The solution executed by Li et al. (2016) without negotiating data privacy was to allow the multiple data owners to share information securely across the databases. This study was mainly performed in the vertically partitioned database. The algorithms designed were privacy preserving association rule mining and frequent itemset mining. Homomorphic encryption and secure outsource comparison schemes were used to develop the algorithms. The schemes were based on three algorithms, key generation, encryption, and decryption. Two datasets contain Belgian retail store's retail market, population, and housing census data. Java programming

language was used to implement the recommended solutions. The size of the dataset ranged from 49,046 to 88,162. Performance evaluation criteria were computational complexity, the security of the underlying homomorphic encryption scheme, and security under data owner's attacks. The results indicate that the information leak was minimal regarding privacy for the new algorithm developed. In comparison with high performance of the privacy level, efficiency (runtime) scored average compared to other algorithms.

**Limitations.** Most of the algorithms developed focused on performance and hiding sensitive information. There were few limitations observed from the above critical review of the research studies. One study successfully reported the fast execution rate of the new algorithm without considering the side effects. Side effects are a very essential part of the sanitization process to protect the leakage of sensitive information publicly. The algorithms were successfully hiding sensitive information, but the impact of non-sensitive information was not given much importance. The scope to improve the sanitization process is observed. Most of the algorithms implemented performed transaction deletion for the experiment, other than deletion data modification and noise addition should be implemented for testing the side effects' occurrences.

### *Side Effects*

HF, MC, and AC were common side effects that were primarily used to measure performance during sanitization of PPDM algorithms (Lin et al., 2016c). Side effects were caused during the process of hiding sensitive information in the database (Lin et al., 2017). The possible symptoms were hiding unrelated non sensitive information and data dissimilarities. AC referred to artificial information which should not be generated, MC was important information

that is not sensitive and should not be hidden, and HF was sensitive information which failed to hide (Lin et al., 2016b).

Few research studies applied association rule hiding algorithm such as HMAU to prevent an individual's confidential information in an organization; the side effect hiding failure rates allowed the study to calculate the efficiency and execution time of the algorithm (Shah et al., 2012; Laskar & Lachit, 2014; Gayathiri & Poorna, 2015). These studies implied that PPDM techniques were used to inspect side effects during the mining process of sensitive information. This indicates the need to investigate the in-depth details of the side effects instead of just calculating the number of occurrences.

Lin et al. (2016a) conducted the experiment using a mushroom and chess dataset. The population size was set to 20, and runtime was set for 10,000 iterations. A new algorithm PPUMGAT+ tested was compared with state-of-the-art evolutionary algorithm HUPeumGRAM. Results reported that the new algorithm performance was faster compared to existing GA-based algorithms. In contrast, the side effects such as MC and AC were not included in the experiment. The research study's one of the main intentions was to hide the sensitive high utility itemsets in privacy preservation utility mining (PPUM).

The need to test hiding failure was essential when hiding sensitive information. The results on algorithm performance raised concerns because one of the studies by Lin et al. (2014a) reported that best execution time performance was usually seen when the side effects were not considered. Thus, there might be a possibility that one of the side effects called hiding failure occurrences could be higher in such cases. Hence the need to explore the side effects was essential to measure the performance of an algorithm.



Lin et al. (2016b) developed a new algorithm called Particle Swarm Optimization to Data Deletion (PSO2DT) to give equal importance to reduce side effects and hide sensitive item sets. The algorithm was created based on PSO technique. A thorough and detailed analysis of the experiment was conducted. Compared to other algorithms, the results indicated that the PSO2DT algorithm was able to successfully hide sensitive information for all datasets except for the sparse food mart dataset. The study evaluated the side effects based on the number of occurrences, which gave an in-depth explanation of the relationship and the impact of each side effect. All the side effects were considered except for the artificial cost which rarely occurred during this experiment. PSO2DT algorithm was successfully able to minimize the side effects.

The research study by Lin et al. (2019) focused on minimizing the four side effects and maximizing hiding of sensitive information. The four side effects are HF, MC, AC, and data dissimilarity. A newly designed multiobjective algorithm (NSGA2DT) was compared with cpGA2DT and PSO2DT algorithms. NSGA2DT algorithm was designed based on NSGA II framework. Results were evaluated based on the experiments conducted on chess, mushroom, and food mart dataset. The maximum iterations and population size were set to 50. NSGA2DT execution time performance was much better compared to other algorithms. As far as side effects were concerned, NSGA2DT outdid by successfully reducing side effects even for large datasets. An important concept revealed in this study was the interlink between the four side effects even with the dense database. The higher the MC, AC, and data dissimilarity the lower the hiding failure. It was observed that there was no information obtained about any new side effects within the four existing side effects.

Lin et al. (2017) designed a new algorithm PPUMGAT to evaluate the performance by using three side effects HF, AC, and MC. Genetic based technique was used to develop the

PPUMGAT. The performance for the PPUMGAT+ algorithm with respect to HF and AC produced good results. PPUMGAT+ algorithm was successful in hiding sensitive information in HUIs. However, the MC side effect still needs deep research. As only transaction deletion was conducted, transaction insertion against all the three side effects should be tested for more accurate results.

Based on the criteria of side effects (HF, AC, and MC) excluding the execution time, the new algorithm cpGA2DT scored higher than greedy and simple GA based algorithms (Lin et al., 2014a). The study was based on genetic methodology. One interesting find of the experiment results was that the side effects of the artificial cost were mentioned. There is no explanation about the side effects within artificial cost. This throws some light on the need to explore this area of unknown side effects. Even this study considered the transaction deletion process. For execution time the greedy approach algorithm scored higher than the cpGA2DT.

Another experiment showed successful results of the algorithms PSO2DT and ACS2DT generating fewer side effects compared to GA-based and Greedy algorithms (Wu et al., 2017). The need to explore smaller number of occurrences of the three side effects was vital. Li, Lu, Choo, Datta, and Shao (2016) study considered only MC and ignored HF and AC side effects. This ignorance can cause a possibility of failing to hide sensitive information.

In exploring the research studies, the three commonly used side effects (HF, AC, and MC) were considered for the sanitization process, however, one study by Lin et al. (2019) mentions the fourth side effect called data dissimilarity. It was evident that there is a probability of more unknown side effects that occur during the process of privacy preservation. Hence the need to explore the unknown side effects was crucial for the privacy preservation process.

***Design Science Research***

Reibenspiess et al. (2020) tried to find appropriate design principles specific to a digital intrapreneurship platform to promote employees' innovative ideas. The researchers followed the DSR approach and Action Design Research (ADR) methodology. The ADR process involved four stages: problem formulation, building, intervention and evaluation, reflection and learning, and formalization of learning. The core intention for using the approach was based on three reasons:

- The research method supported information technology (IT) artifacts. The artifacts in the study were related to theoretical (researchers), technical (developers), and practical (employees). The researchers were addressing the real-world problems faced by employees in the workplace.
- The study tried to solve the real-world problems of a suggestion box system. The employees proposed their innovative ideas via the suggestion box.
- ADR's intervention blended with the study, executing digital reformation that is internal to an organization.

Donalds and Osie-Bryson's (2019) research goal was to present a cybercrime classification ontology (CCO) model for cybercrime attacks. In addition to the implemented model, a knowledge based CCO artifact was also developed. DSR methodology by Peffers et al. (2007) was used in the study. As the two main objectives of DSR are to identify a problem and then creating innovative IT artifacts to solve the problem. Similarly, the study adopted DSR to report cybercrime classification (a real-world problem) and created knowledge based CCO (innovative information system (IS) artifact). Previous research studies' models were incomplete in classifying cybercrimes and their concepts. The research gap identified was solved by

considering all the relevant information to classify cybercrimes. This was achieved by classifying and storing two real-world cybercrime attack events.

Gnewuch et al. (2017) planned to increase customer service quality by constructing DSR based cooperative and social conversational agents. DSR by Hevner et al. (2004) and Kuechler and Vaishnavi (2008), was employed. Researchers found the DSR approach to be more appropriate to address the research gap. The research gap showcased that there was insufficient design-based research literature for conversational agents. A conversational agent artifact was designed and evaluated through iteration. It involved two design cycles. Meta-requirements and design principles were suggested based on cooperative principle and social response theory.

A gamified mobile application was developed by Oppong-Tawiah et al. (2020) to promote pro-environmental employee behavior. The gamified application was developed based on design science research. The researcher's focused on clarity, flexibility, practicality, and applicability of the artifacts, hence DSR by Peffers et al. (2007) was chosen. DSR methodology involved five iterations of the design cycle. The design cycle steps are objectives for a solution, design, and development, demonstration, and evaluation. 137 students and employees of three American universities participated in the study. The study was conducted for six weeks, focusing on computer-related electricity usage. The results indicated that the application helped in reducing electricity consumption by the employees. Also, employees were motivated to be more pro-environmental.

Zschech et al. (2020) designed and developed a system called Text-based Intelligent Assistant system (TbIAS). TbIAS provided a system for inexperienced data mining professionals, that automatically selected data mining methods. The six steps of DSRM by Peffers et al. (2007) were used to build TbIAS. System design artifact instantiation was

incorporated during the design and development phase. The purpose for choosing DSR was: 1) The study involved the DSR pattern of creating socio-technical artifacts to solve an organizational problem (dependency on data mining experts for data mining method selection). 2) As part of DSR's design theorizing, the study needed to develop design principles and features. Designing and evaluating new algorithms for the creation of a new TbIAS system was very important.

Herselman and Botha (2015) primarily evaluated the Teacher Professional Development (TPD) course with help of iterative DSR process. DSRM by Peffers et al. (2007) and qualitative multiple case study methodology were used. An instantiated artifact was implemented by segregating artifacts in three phases. To solve problems DSR approach allows to gain knowledge and examine the structures and processes of existing socio-technical systems. Hence applying this approach, the researchers observed the existing system's functionalities before and after TPD module artifact implementation. Other reasons for choosing the DSR approach were: 1) Iterative evaluation suitable for the study. 2) Evaluation focused on artifacts' performance, which the study required. 3) Addressed the educational exploitation (wicked problem) of the Cofimvaba school district. 4) DSR's instantiation artifact allows innovating new solutions.

### **Summary**

There has been immense research conducted on privacy preservation. Various PPDM algorithms were created due to the high demand in protecting sensitive information. However, from prior research, it was observed that there is a lack of deep research studies specifically related to identifying the new side effects in PPDM. Most of the experiments conducted commonly use or report HF, AC, and MC as traditional side effects. Considering the traditional side effects consistently used in research studies, the question was are these the only side effects

occurring constantly? Among the studies discussed in the literature review, one reported a fourth side effect called data similarity. This presented a curiosity of any hidden side effects which are yet to be explored. An important aspect in the PPDM algorithm was giving importance to runtime execution due to the large volume of data. More importance should be given to explore the unknown side effects to prevent the highly sensitive information being stolen than runtime execution. Finally, with the various privacy preserving algorithms implemented, are there any statistics as to which have been successful in hiding sensitive information? Why were the side effects issues recurring with numerous algorithms available?

Previous research studies had little relevant information regarding any new side effects explored. Each study proposes a new algorithm for privacy preservation, yet there is no solution researched to permanently solve these side effects. The research studies examined reveal that there is a strong interlink between privacy preservation, PPDM algorithms, and side effects. The necessity to preserve sensitive data resulted in a higher number of new algorithms generated to improve the privacy preservation performance. This in turn resulted in various side effects occurrences. Hence this study helped in understanding and discovering the known side effects. An attempt was made to explore unknown side effects. These steps helped in clarifying the severity of protecting sensitive information and creating a common web repository to report the side effects of PPDM algorithms.

## **Chapter 3**

### **Methodology**

#### **Overview of Research Methodology**

A Design Science Research (DSR) study was performed for the present research work (Peffer et al., 2007). This research study was conducted in four phases: 1. Conducted a literature review of PPDM algorithms, side effects, and software tools, 2. Investigated the occurrences of non-traditional side-effects in all PPDM algorithms, 3. Discovered unknown side-effects in all PPDM algorithms, 4. Created an online repository that creates and stores side-effect information for all PPDM algorithms. The data collected from phases one to three were critical components for the fourth phase, an online web repository where the data was stored.

The present study's methodology was inspired by Zschech et al. (2020) and Herselman and Botha (2015). A review of both the research studies was in the Design Science Research section of Chapter 2. Especially, Herselman and Botha's (2015) research approach inspired this present study to adopt instantiation artifact type in phases.

#### **Research Methods**

DSR Methodology (DSRM) was used as the research design, which was based on the work from Peffer et al. (2007). As this research study was related to information systems, technology based DSR was adopted (Peffer et al., 2007). DSR focused on accomplishing the goals by implementing the functionalities and behavior of a particular object (GeertsGeerts, 2011).

Gerede and Su (2007) considered a data object to be an artifact. The changes in functionalities of these data objects uniquely explained a particular process model. Mizoguchi et al. (2016) explained artifacts as man-made physical objects based on a particular reason to create

the objects; for example, vehicles. Borgo et al. (2014) termed artifacts as technical artifacts. In context to engineering design, Borgo et al. (2014) defined technical artifact as follows:

“A physical object created by an intentionally performed production process. The process is intentionally performed by one or more agents with the goal of producing the object “a”? which is expected to realize intended behavior in some given generic technical situation” (p. 7).

On the other hand, IS or IT artifacts were technology based dynamic systems in contrast to human-created artifacts, such as bridges or paintings (Gregor & Iivari, 2007). Examples of dynamic systems include cybernetics and weather forecasts.

DSR in information systems research solved the organizational problem by creating and evaluating IT artifacts (Peppers et al., 2007). The four different approaches of DSRM were: problem-centered initiation, objective centered solution, design, and development centered initiation, and client or context centered initiation (Peppers et al., 2007; Cleven et al., 2009). DSRM, as created by Peppers et al. (2007), consisted of six activities: problem identification and motivation, defining the objective of a solution, design and development, evaluation, demonstration, and communication.

The majority of the researchers defined IS artifacts as systems or activities related to these systems (Simon, 2019; Gregor & Iivari, 2007). Hence, researchers distinguished the IS artifacts as design artifacts based on the design theory. The design theory consisted of goal (purpose), scope (aspects and criteria), structure (form), activity (function), and evolution (artifact mutability). On the other hand, Prat et al. (2014) supported IS artifacts as systems but considered that the DSR processes lead to artifacts creation. Their claim agreed with the DSRM designed by Peppers et al. (2007). Therefore, artifacts were considered as systems or objects that were created as an end product of a process or during a process.



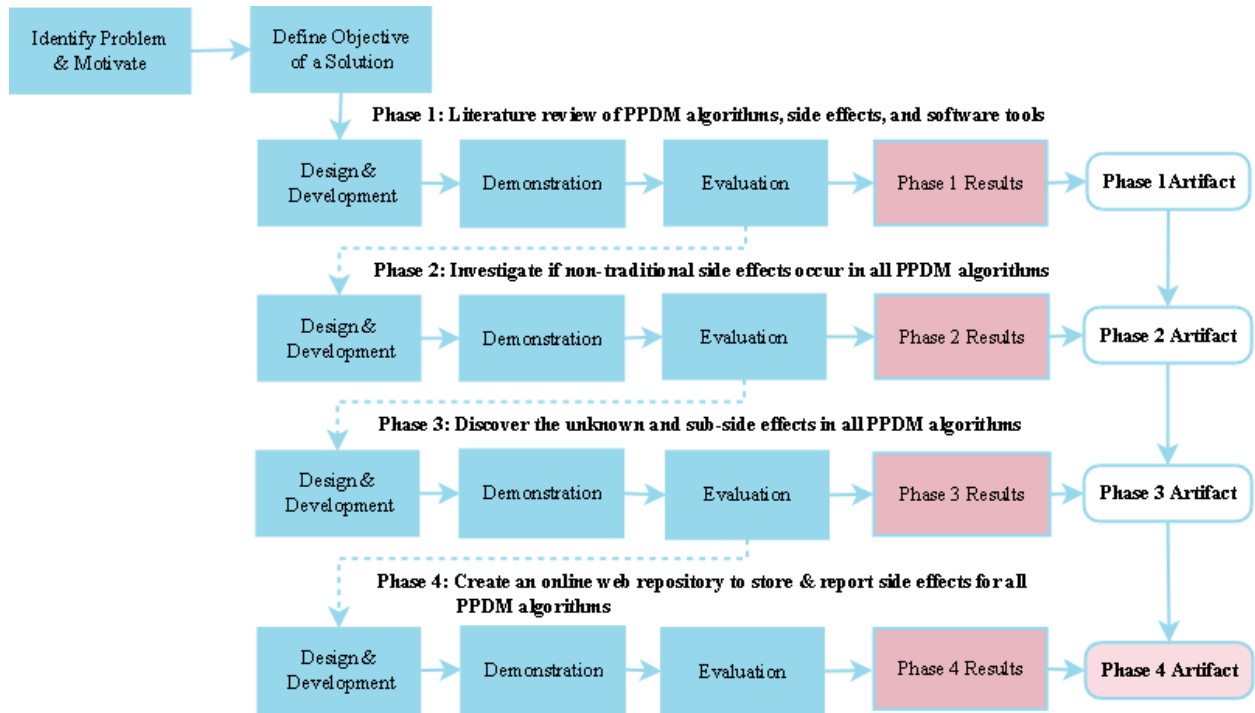
Artifacts are very important, as they provide both information and clues to continue a research process to achieve the desired goal. For example, one of the goals of this study was to collect known and unknown side effects. Design artifacts consisted of models, constructs, methods, and instantiations (Hevner et al., 2004; Lukyanenko et al., 2015; Peffers et al., 2007). Artifact types were classified based on technical, technical with social factors, socio-technical, and social (Drechsler & Dörr, 2014). Socio-technical artifacts required human intervention for a particular system that provided a desired functionality (Venable et al., 2012)

Further, artifacts were also differentiated into two types: product and process artifacts (Venable et al., 2012). The present study followed process artifacts for phases one and two; process and product artifacts for phases three and four. The reason phases three and four used both types of artifacts were because these phases used software tools to get desired results. Hence, phases one through four artifacts fell under the category of socio-technical artifacts. Herselman and Botha (2015) defined instantiation artifacts as, “Instantiations demonstrate the feasibility and effectiveness of the constructs, models or methods in an environment”

The DSRM process model (Figure 1) for this study consisted of six activities. The activities were problem identification and motivation, objectives of the solution, design and development, demonstration, evaluation, and communication. The six activities were the main focus of this research paper.

**Figure 1**

*DSRM Approach Demonstrating Different Phases of this Study*



This study implemented the instantiation artifact. This process had four phases of iteration. The initial artifact was phase one and the final artifact was phase four. Furthermore, this section is organized as below:

- Research questions and the artifacts implemented was as shown in Table 1.
- Research methodologies used for the research questions were explained.
- Phases one through three, and their corresponding research questions one through four, were discussed.
- Research frameworks used for research question five was described.
- Phase four activities related to research question five were detailed.

**Table 1***Research questions and the artifacts implementation in phases*

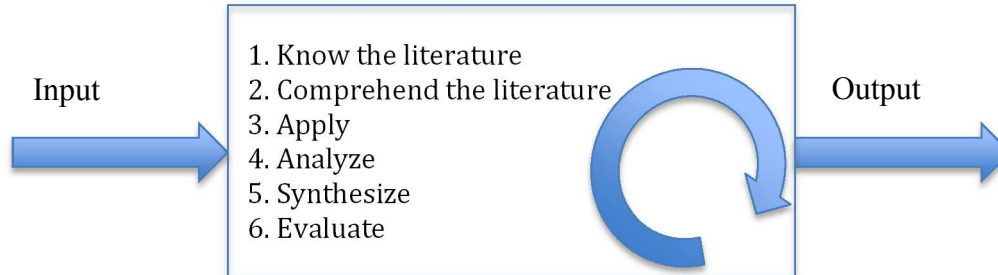
<b>Research Questions</b>	<b>Artifacts</b>
RQ1: What were the similarities and differences of the existing side effects of PPDM algorithms?	Phase 1 Artifact
RQ2: How were the side effects related to one another?	Phase 1 Artifact
RQ3: What were the non-traditional side effects, and do they occur in all PPDM algorithms?	Phase 2 Artifact
RQ4: Were there unknown side effects occurring in all PPDM algorithms?	Phase 3 Artifact
RQ5: Where and how were the side effects of all PPDM algorithms reported?	Phase 4 Artifact

Two specific literature review processes were used to answer research questions one through four. They were a literature review in Information System Research (Levy & Ellis, 2006) and Systematic Literature Review (Kitchenham, 2007; Atlam et al., 2020).

Levy and Ellis (2006) recommended the methodology for the benefit of Information System researchers at all levels. The proposed framework (Figure 2) consisted of three steps: input, process, and output. The input step involved the selection of quality journals, keyword search, backward search, forward search, and decision to finalize the search. Top 50 ranked MIS (Management Information Systems) journals and their availability in 12 literature databases were recommended. The process step included understanding, comprehending, applying, analyzing, synthesizing, and evaluating the literature. The output mainly required developing argumentation for literature writing based on the theory of argumentation. The argumentation theory was nothing but a problem that formulated a justification to motivate a research study.

**Figure 1**

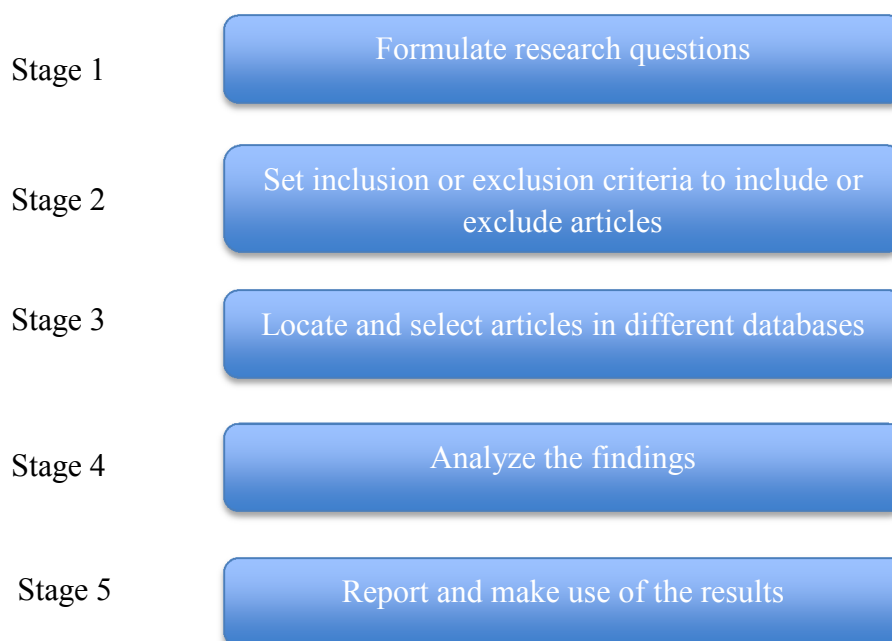
*The Three-step Literature Review Process for Information Systems Research (Levy & Ellis, 2006, p. 182)*



Kitchenham (2007) developed a Systematic Literature Review (SLR) for software engineering. SLR was also known as a Systematic review. A systematic review involved a thorough analysis of all the relevant research, to find unbiased answers for a specific research question. To examine the risk-based access control model, a systematic review was performed on finalized 44 articles (Atlam et al., 2020). Total articles searched were 1044. To study the model Atlam et al. (2020) developed five stages for systematic review. The stages were adapted from Kitchenham (2012) as shown in Figure 3.

**Figure 2**

*Five stages of Systematic Literature Review (Atlam et al., 2020, p. 7)*



Similarly, the present study used systematic literature review stages built by Atlam et al. (2020). The first stage segregated the research questions to be considered for the systematic review. The second stage included and excluded articles related to the research questions. In the third stage, the articles were searched in different databases. Few of the 12 literature databases as suggested by Levy and Ellis (2006) were considered. Search strategies recommended by Levy and Ellis (2006) were also considered. The findings were analyzed in the fourth stage. The fifth stage reported the results for each of the research questions.

Inclusion criteria was:

- Peer reviewed quality articles
- Articles published date irrespective of the year
- PPDM, PPDM algorithms, and PPDM side-effects topics will be included

- Strictly related to each research question
- Articles available in the specified 12 databases
- Articles related to information systems

Exclusion Criteria was:

- Online non-research articles
- Non-peer reviewed articles
- Articles not related to PPDM, PPDM algorithms, and PPDM side-effects
- Articles not related to information system

Data sources considered was:

- IEEE
- Elsevier ScienceDirect
- ACM Digital Library
- ProQuest
- SpringerLink
- Google Scholar
- EBSCOhost databases
- JSTOR

Search strategies was:

- Keyword search
- Backward search
- Forward search
- Decision to finalize the search

### ***Phase One***

The initial phase of this study involved in-depth learning of the side effects occurring during the PPDM process, re-confirming the datasets, further analysis of the PPDM algorithms, and exploring the PPSF software tool. As part of dissertation proposal, basic information was gathered from the literature review process. The literature review was conducted through *Google Scholar*, *MIS Quarterly*, *Springer*, *ResearchGate*, *ScienceDirect*.

- RQ1: What were the similarities and differences of the existing side effects of PPDM algorithms?
- RQ2: How were the side effects related to one another?

The first phase involved answering RQ1 and RQ2 to find the characteristics, similarities, differences, and relationship of the existing side effects of PPDM algorithms. Both the research questions used SLR approach.

### ***Phase Two***

- RQ3: What were the non-traditional side effects, and do they occur in all PPDM algorithms?

The second phase first gathered the information (names and characteristics) of non-traditional side effects. The next step investigated if these side effects have occurred in all PPDM algorithms. The data were initially gathered through the SLR process. Finally, this phase involved testing the datasets using the PPSF software tool for six PPDM algorithms. Careful observations and comparisons between the output of the datasets were analyzed. This approach was abided because non-traditional side effects should be examined, which required focused and detailed analysis. Non-traditional side effects considered for this study are data dissimilarity, hidden rules, new rules generated, and lost rules.

### **Phase Three**

- RQ4: Were there unknown side effects occurring in all PPDM algorithms?

The third phase of the experiment investigated unknown side effects in all PPDM algorithms. The initial step involved the SLR approach. Next, the PPSF tool and the datasets were used for further investigation. Six PPDM algorithms available in the PPSF tool were used to find the unknown side effects.

Moreover, the PPSF tool had only six PPDM algorithms implemented. This tool alone could not be used to find the characteristics and details of the side effects of PPDM algorithms. PPSF tool was the only tool that had open-source implementation. Other PPDM algorithms had research studies published with results, with no availability of tools or code to test the algorithms. This was the main reason to use both Literature and Systematic review processes to find more details about the PPDM algorithms and side effects.

#### ***Phase Four***

- RQ5: Where and how were the side effects of all PPDM algorithms reported?

The final phase or phase four involved finding where and how the side effects of all PPDM algorithms were reported. Based on the information collected, as there was no PPDM application already existing, an online web repository was created to report the side effects of all the PPDM algorithms.

Research question five used research methodologies based on the Web Frameworks and Web Stack (Shetty et al., 2020). The authors discussed the importance of using Web Frameworks to build web applications. Web Frameworks provided ready-to-use fundamental requirements to build web applications. Web Stack is a package consisting of different frameworks, software, web servers, databases, and operating systems used in developing web applications. Different front-end frameworks are AngularJS, ReactJS, and Vue. Certain back-end frameworks are



Spring Boot, NodeJS, and Django. The databases discussed were PostgreSQL and MongoDB. Web Stack combinations discussed were LAMP (Linux, Apache, Maven, Python), MEAN (Mongo, ExpressJS, AngularJS, NodeJS), and Spring Boot.

Soni (2017) developed a full-stack web application using different frameworks. The frameworks were based on Java Frameworks. Java is a popular open-source programming language. Full stack made use of web stack to design, implement, test, deploy, and fix errors to develop a web application. MVC (Model, View, Controller) architecture, Spring framework, Hibernate framework, and Angular JS were used to develop the UserRegistrationSystem application. The front-end framework was AngularJS. The back-end framework used was Spring Boot. H2, an embedded database, was used to store and retrieve the data. Postman, an Application Programming Interface (API) testing tool, was used to test the application. Spring Source Tool (STS), an Integrated Development Environment (IDE), was used to develop the entire application. To summarize, this study developed the web application based on MVC (Model, View, Controller) architecture, Spring framework, Hibernate framework, and Angular JS. The user interface was created using HyperText Markup Language (HTML).

### **Instrument Development and Validation**

Microsoft Excel comma-separated values (CSV) file format was used to store the data. These data were manually migrated to the PostgreSQL database.

A new web application was created to report and view the details of the side-effect. The data for the web application was retrieved from the PostgreSQL database. Postman was a powerful tool that was used to validate the API of the web application. The web application consisted of:

- A web page to view the details of all PPDM algorithms and their side-effects

- A web page to report or update the details of the side-effects for PPDM algorithms

### **Sampling**

The sample for this research study were datasets retrieved from an open-source data mining library called SPMF (Fournier-Viger et al., 2016). A total of six algorithms specific to PPDM were selected from PPSF software (Lin et al., 2018d). The names of the algorithms are shown in *Table 2*. Datasets pertaining to real-life customer transactions were considered. More details of the datasets are shown below in *Table 3*. This sampling method was used because of the importance to discover the known and unknown side effects occurring during the process of PPDM.

**Table 2***PPDM Algorithms Selected from PPSF*

<b>Algorithms</b>	<b>Full Name</b>
Greedy	Privacy Preserving Data Mining Greedy
sGA2DT	Simple Genetic Algorithm to Delete Transactions
pGA2DT	Pre-large Genetic Algorithm to Delete Transactions
cpGA2DT	Compact Prelarge Genetic Algorithm to Delete Transactions
PSO2DT	Particle Swarm Optimization to Delete Transactions
SIF-IDF	Sensitive Items Frequency-Inverse Database Frequency

**Table 3***Datasets Selected from SPMF*

<b>Dataset name</b>	<b>Description</b>
ECommerce_time_without_utility	UK based online retail data
chainstore_utility	Data from California based major grocery store
foodmart_utility	Customer transactions from retail store
accidents_utility	FIMI repository's traffic accident data
retail_utility	Belgian retail store customer transaction
mushroom_utility	Mushroom dataset from UCI repository
pumsb_utility	Population and housing census data
chess_utility	Chess dataset from UCI repository

The real-life dataset which has customer transactions and privacy preservation algorithms was used as the subject because of the sensitive information available within the data. Sensitive information was required in this study to test HF and unknown side effects. This study was conducted as a contrived study using the researcher's laptop or computer as an environment in which the subjects were normally studied.

This study incorporated the latest release version of PPSF tool. Detailed instruction for installation was obtained from the PPSF website. The latest Java version 11 was installed in Windows 10 (8u51 and above) 64-bit operating system. PPSF is a reliable instrument as this is specifically designed for PPDM with inbuilt six algorithms for testing. The validity of the instrument was promising as the algorithms were inbuilt and no alterations can be made to the original source code. The validity of the dataset was accurate as it was collected from the SPMF website in .dat or .txt format and was uploaded to PPSF software directly. The only concern

pertained to the potential for the operating system or PPSF software crashing during the process of data mining. However, the PPSF tool closed abruptly for only SIF-IDF algorithm. Hence, the results file was not generated for this algorithm. The personal computer's performance allowed proper functioning of PPSF software to deliver the results data.

### **Data Analysis**

The qualitative data analysis strategy was used to analyze the information collected from the experiment. *Table 4* shows details of the expected data for each of the phases.

**Table 4***Expected data in each phase*

<b>Phases</b>	<b>Expected data</b>
Phase 1	<ul style="list-style-type: none"> <li>• General characteristics of each of the side effects</li> <li>• Similarities of each of the side effects</li> <li>• Differences of each of the side effects</li> <li>• Relationships of each of the side effects</li> <li>• Impact of the side effects on one another</li> </ul>
Phase 2	<ul style="list-style-type: none"> <li>• Reasons: why non-traditional side effects are not commonly used?</li> <li>• Have non-traditional side effects occurred in existing PPDM algorithms?</li> <li>• Will non-traditional side effects occur for six PPDM algorithms in the PPSF tool?</li> </ul>
Phase 3	<ul style="list-style-type: none"> <li>• Find if unknown or new side effects occur for six PPDM algorithms in the PPSF tool</li> </ul>
Phase 4	<p>A web page displaying:</p> <ul style="list-style-type: none"> <li>• All the PPDM algorithms</li> <li>• PPDM techniques for each PPDM algorithm</li> <li>• All the side effects corresponding to these algorithms</li> <li>• A new webpage that will allow external users to report or update any side effects</li> </ul>

The information collected in each of the phases was in textual format. Qualitative text analysis was used to analyze the data and present the results. Kuckartz (2014) illustrates the five steps of qualitative text analysis designed by other researchers (p. 35).

The five steps were:

- Developed categories based on empirical data.
- Designed guidelines for the analysis.
- Coded the data.
- Setup tables (with crosstabs) and overviews.
- In-depth observation of individual cases.

Specifically for this study, category-based text analysis was employed to understand the data. The categories were derived from the five research questions. Profile matrix will be used to prepare and represent the data. Profile matrix table format was used to map the information (answers) gathered for each topic (associated with research questions).

Phase one data evaluation was related to topics such as characteristics, similarities, differences, relationships, and the impact of the side effects. Phase two topics included non-traditional side effects categorized into the usage and occurrences of these side effects in the past (data will be collected from literature review) and present (data will be collected by testing in PPSF tool). Phase three included topics such as unknown side effects. Phase four topics were PPDM algorithms, PPDM techniques, and side effects.

### **Formats for Presenting Results**

Results for RQ1 and RQ2 were presented in both textual and table format. RQ3 and RQ4 details were shown in table format. Finally, as part of RQ5, the data gathered from RQ1 to RQ4 were displayed on a newly created web page.

## Resources

Windows 10 (8u51 and above) 64-bit operating system, high-speed internet, Google Chrome (version 86 & 64 bit), and Microsoft Excel were used for this research. New datasets from the SPMF website were selected. The latest software from PPSF website was considered. Java programming language was used to write new code. The latest version of Java available was installed.

List of software tools with latest versions:

- Database: PostgreSQL version 13
- Web Pages (front-end): AngularJS version 21
- Back-end development: Spring version 5.3.6, Spring Boot version 2.4.5
- Communication between Spring and PostgreSQL database: Hibernate version 5.4.31

## Summary

The methodology section discussed the DSRM approach selected for the present study. Instantiation artifact type DSR with a total of four phases was considered. Each phase's relationship with the research questions was explained. Research questions one through four used two research processes. The processes were Literature Review for Information Systems and SLR. The use of these processes for each research question and their implementation in each of the phases were clarified. The PPSF software tool was used to analyze the six PPDM algorithms.

Full stack web application was developed to create the new web repository. Frameworks used to create this web application were AngularJS, Spring, and Hibernate. Postman software tool was finalized to test the API for the new web application. Data sampling was obtained from the SPMF website. Qualitative data analysis was considered to study the side effects of PPDM algorithms. Relevant resources and software applications were given in detail with versions.



## Chapter 4

### Results

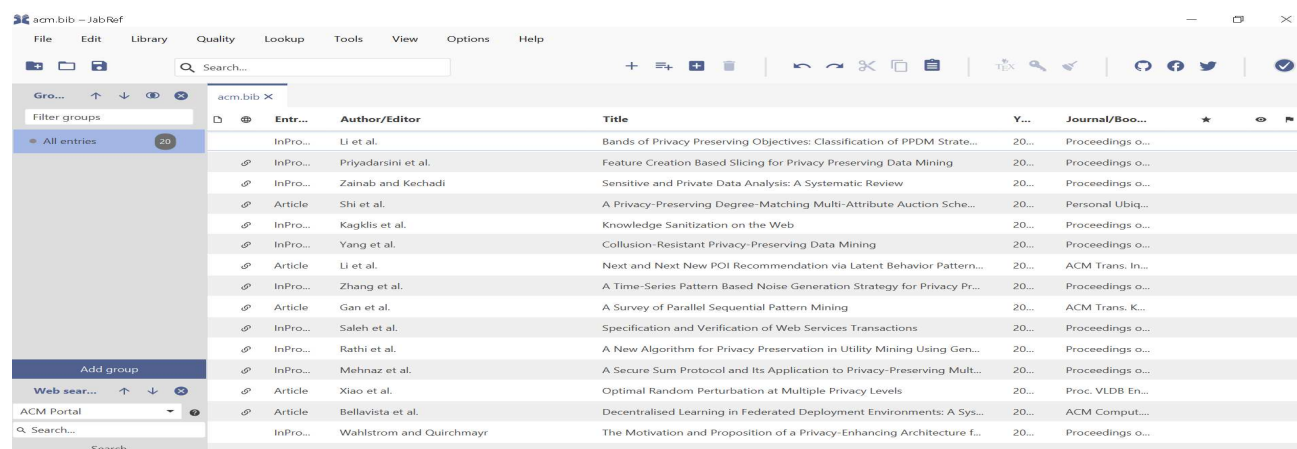
This chapter describes the results of the study. The sections of the chapter are data collection, data analysis, findings, and conclusion. Data analysis presents details of the data analyzed for each phase. The findings section shows the results of the four phases. The summary section provides the summary of the overall research based on the results of the study.

### Data Collection

This section explains the procedures used for collecting the data. The data sources (or databases) considered were IEEE, EBSCOhost, SpringerLink, and Proquest. The common keywords considered for search string were PPDM or “Privacy Preserving Data Mining Privacy Preserving Data Mining” AND cost AND failure. These keywords were specific to RQ1 and RQ2. Few databases allowed only BibTeX file export rather than CSV file format. Hence, JabRef software was used to export the search results from BibTeX to CSV format (Figure 4). JabRef is a free open-source multi-platform citation and reference manager. The official website is [www.jabref.org](http://www.jabref.org). Other databases allowed CSV file exports.

### Figure 4

#### *JabRef Tool Used to Convert BibTeX to CSV format*



EBSCOhost database used two different keyword searches:

- “PPDM” AND “failure” AND “cost”
- “Privacy Preserving Data MiningPreserving Data Mining” AND “failure” AND “cost”

For IEEE three different keyword combinations were used to get the desired search results:

- Privacy-Preserving Data MiningPreserving Data Mining AND hiding failure
- Privacy-Preserving Data Mining AND missing cost
- Privacy-Preserving Data Mining AND artificial cost

SpringerLink is another database utilized for data collection. The keywords used for each search is as follows:

- Privacy Preserving Data MiningPrivacy Preserving Data Mining AND hiding failure
- Privacy Preserving Data Mining AND missing cost
- Privacy Preserving Data Mining AND artificial cost

For Proquest, the keyword combinations used were:

- Privacy Preserving Data Mining AND failure
- Privacy Preserving Data Mining AND cost

Phase two data collection was related to RQ3. IEEE data source used four key words to obtain four different search results:

- Privacy-Preserving Data Mining AND data dissimilarity
- Privacy-Preserving Data Mining AND hidden rules
- Privacy-Preserving Data Mining AND lost rules
- Privacy-Preserving Data Mining AND new rules

EBSCOhost and ProQuest used key words such as:

- Privacy Preserving Data Mining AND data dissimilarity
- Privacy Preserving Data Mining AND hidden rules
- Privacy Preserving Data Mining AND lost rules
- Privacy Preserving Data Mining AND new rules

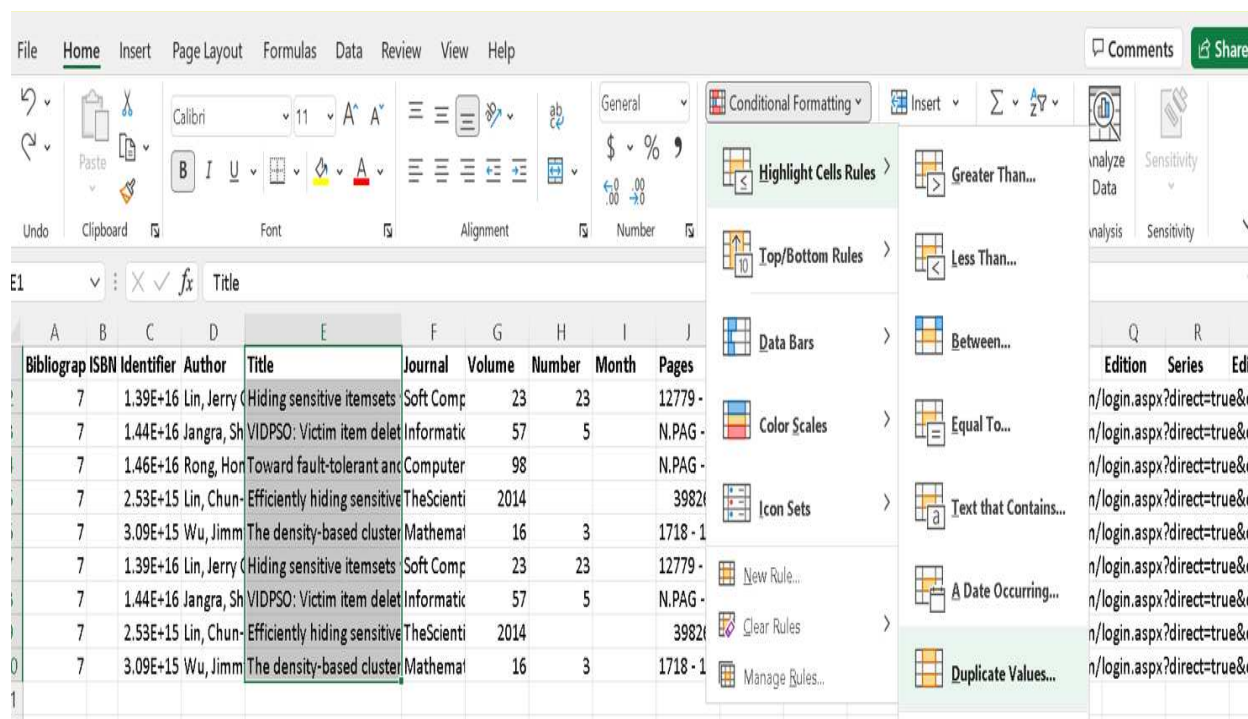
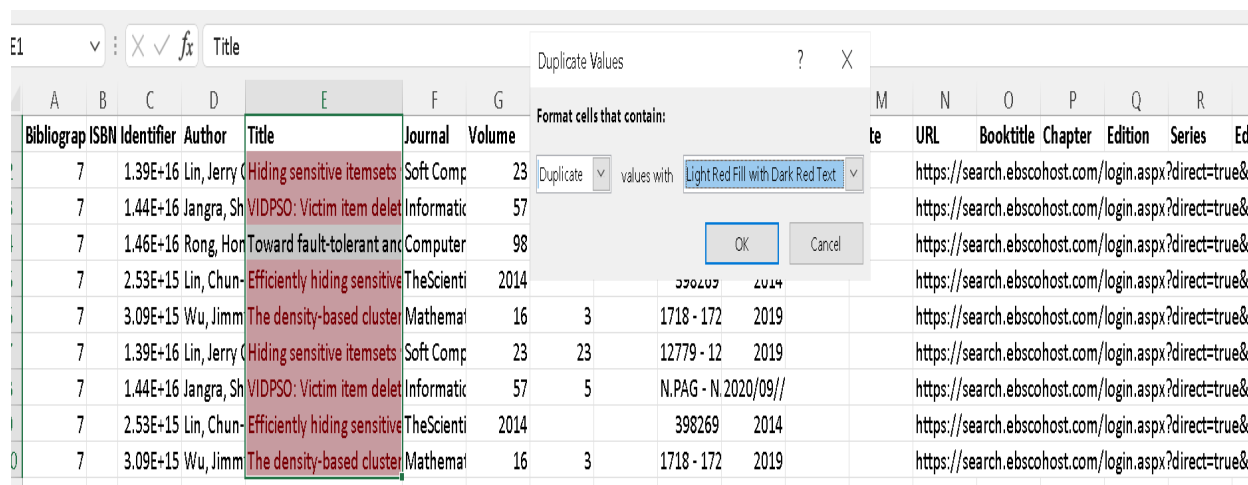
ScienceDirect used the following key words:

- Privacy Preserving Data Mining AND hidden rule AND new rule AND lost rule OR dissimilarity

ACM digital library allowed the keywords as below:

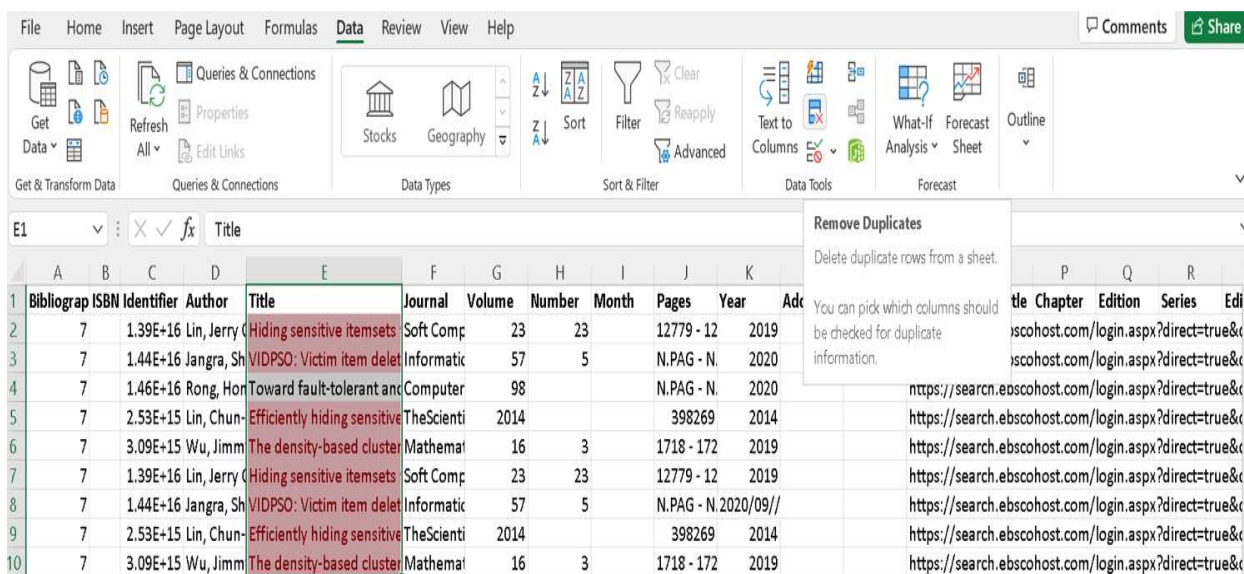
- PPDM algorithms
- Privacy Preserving Data Mining

All the data collected from each of the data sources were saved in CSV format. Microsoft Excel was used to clean the data. Duplicate records for each data source were filtered based on the title of the articles. Microsoft Excel's "Remove Duplicate" function was used to find and remove the duplicates. Figures 5 to 9 show the steps involved to remove duplicate records.

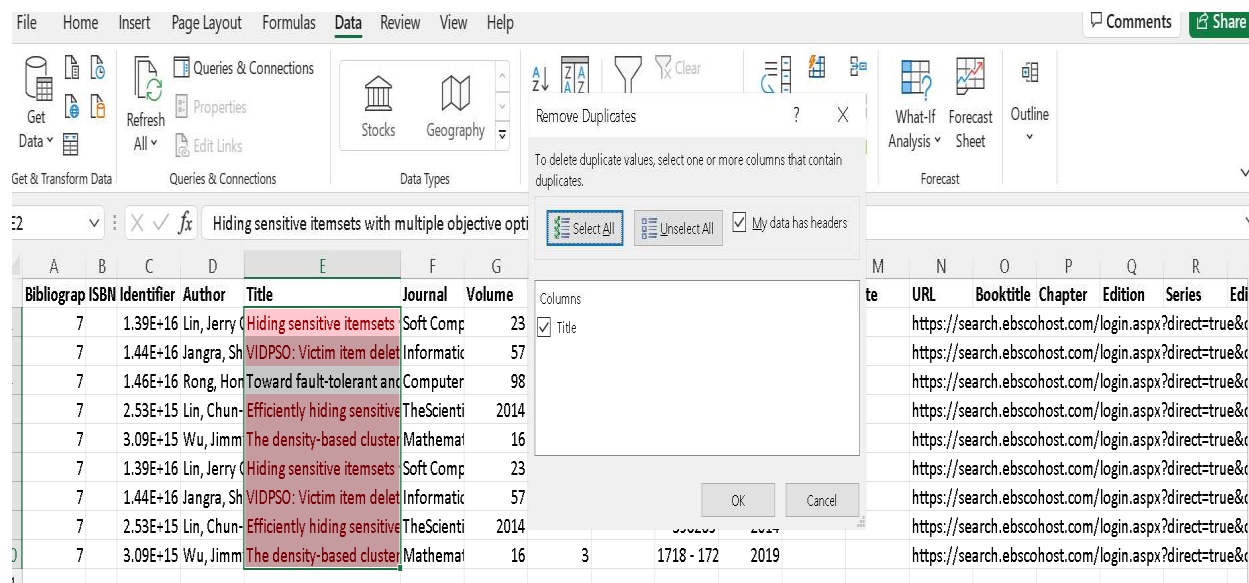
**Figure 5***Finding Duplicate Values***Figure 6***Displaying the Duplicate Values in Red*

**Figure 7**

*Selecting the “Remove Duplicate” option in “Data Tools”*

**Figure 8**

*Remove Duplicates Based on “Title”*



**Figure 9**

*Number of Duplicate Values Removed, and Unique Value Remained are Shown*

	Title	Journal	Volume	Number	Month	Pages	Year	Address	Note	URL
y C	Hiding sensitive itemsets	Soft Comp	23	23		12779 - 12	2019			https://:
Sh	VIDPSO: Victim item delet	Informati	57	5		N.PAG - N.	2020			https://:
lon	Toward fault-tolerant and	Computer	98			N.PAG - N.	2020			https://:
in-	Efficiently hiding sensitive	TheScient								https://:
im	The density-based cluster	Mathema								https://:
y C	Chun-Wei; Zhang, Yuyu; Zh	Soft Comp								https://:
Sh	Shalini; Toshniwal, Durga	Informati								https://:
in-	Wei; Zhang, Binbin; Yang,	TheScient								https://:
im	/ Ming-Tai; Lin, Jerry Chun	Mathema								https://:

Data collection was satisfactory as the exported results had more than sufficient information. For example, information about each article's title, journal, volume, number, month, abstract, keywords, year and many more were provided. During the data collection phase, the current study required only title, year, abstract, keywords, URL, and authors' information.

### Data Analysis

PRISMA is termed as Preferred Reporting Items for Systematic Reviews and Meta-Analyses (Page et al., 2021). PRISMA is effective for researchers in reporting a complete evidence-based transparent systematic review and meta-analysis information. Initially used in healthcare research, it is also applied in other research fields such as information technology, social sciences, and more. The official web location of PRISMA is at prisma-statement.org.

Data collected from all the data sources were further analyzed using a systematic review process. To conduct the review, PRISMA 2020 Version1 flow diagram was employed. PRISMA was used to screen the search results. Based on the eligibility criteria defined for this study, the instructions given in the flow diagram were thoroughly followed. Only phase one and two used PRISMA statement.

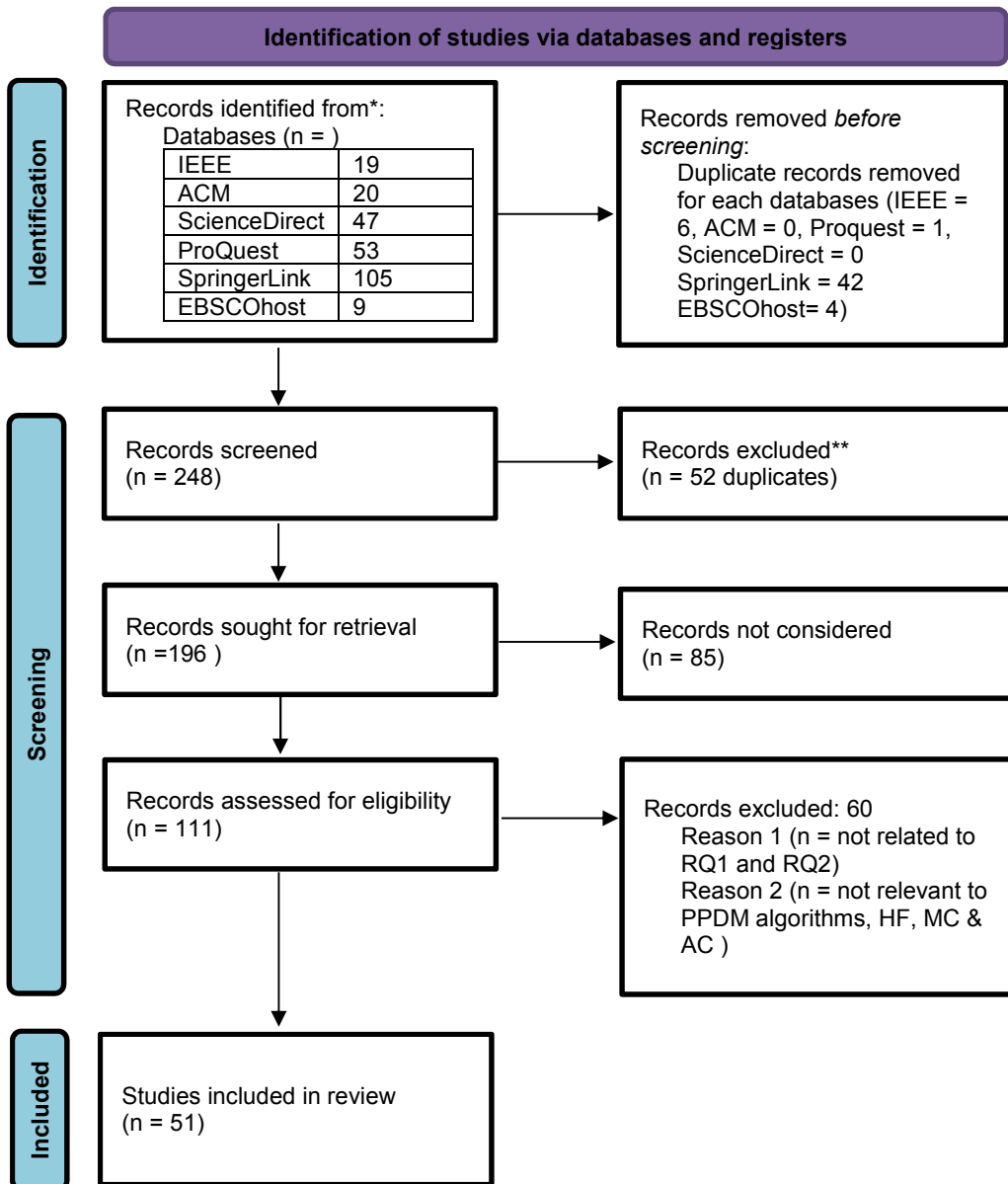
### Phase One

Phase one data analysis was related to both RQ1 and RQ2.

- RQ1: What are the similarities and differences of the existing side effects of PPDM algorithms?
- RQ2: How are the side effects related to one another?

**Figure 10**

*Flowchart for Phase One Systematic Review*



Duplicate results from IEEE database found were six from a total of nineteen . Thirteen articles were finalized. Two search results were obtained from Proquest. One duplicate result was found, and a total of fifty-two articles were selected during the identification step.

ACM digital library's advanced search allowed the use of multiple keywords in one search attempt. Hence, the results were a total of twenty without any duplicates. Similarly, ScienceDirect allowed the same search pattern as ACM and showed forty-seven search results with no duplicate articles.

In the SpringerLink database, for three different search results, there was a total of one hundred and five articles retrieved. Forty-two were duplicates, 63 duplicate and sixty-three unique articles were considered. EBSCOhost database retrieved a total of nine search results. Four duplicates were removed and five were considered for further analysis.

As part of the screening process, a total of two hundred and forty-eight articles were included after omitting the duplicates. Excluded articles after screening were fifty-two. In the records screened step, the articles were screened based on the information available in the title and abstract.

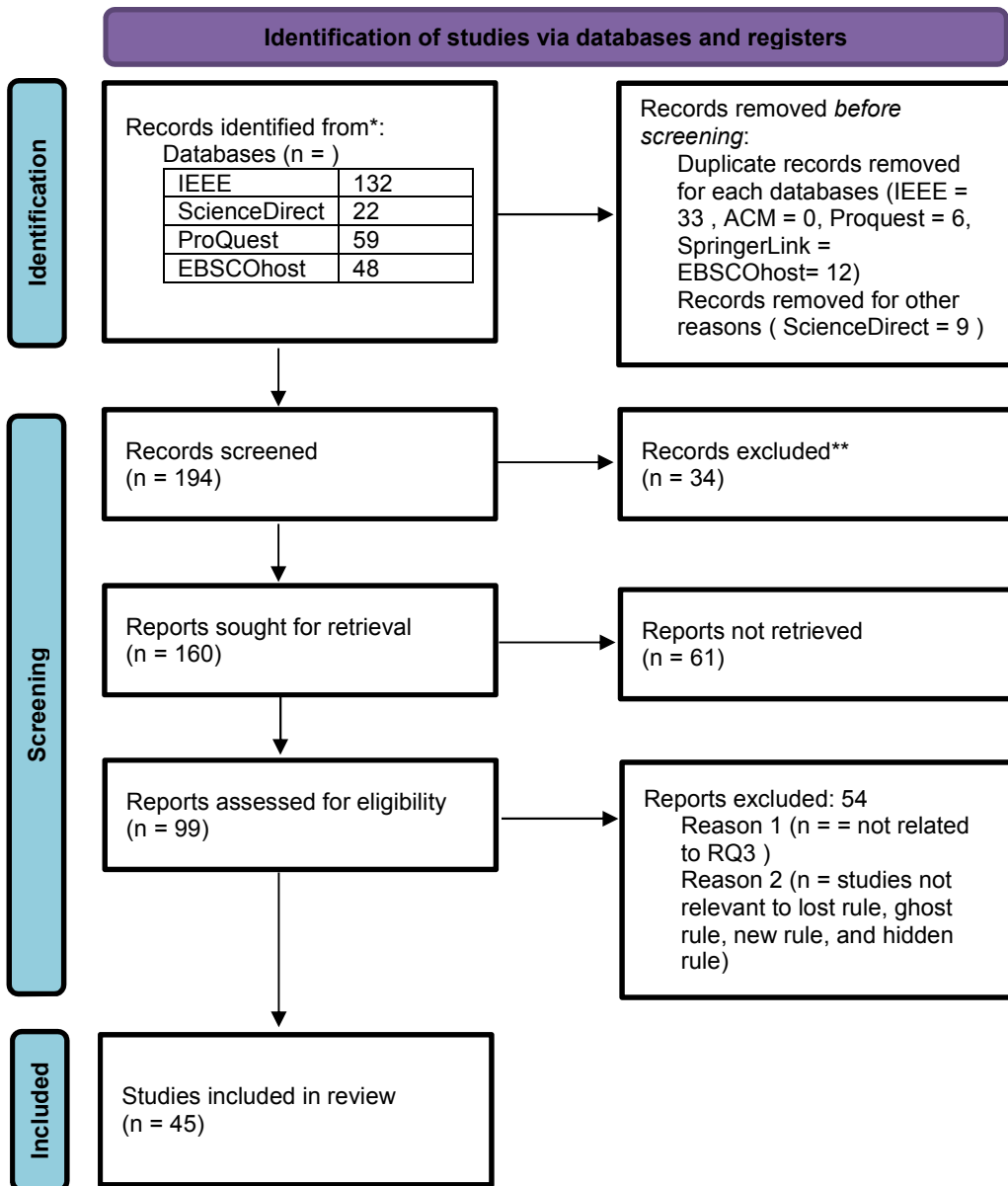
Articles sought for retrieval included full text screening of the articles minus the excluded articles from the total number of screened articles. One-hundred and ninety-six articles were sought for full text retrieval for the study. For articles not retrieved, eighty-five articles were unable to find the full text.

In the full-text screening stage, one hundred and eleven articles were assessed for eligibility. This selection was achieved by considering articles sought for retrieval minus articles not retrieved. This step also involved full-text screening to assess the eligibility to include articles for systematic review.



Articles excluded were a total of sixty. There are two reasons for excluding the articles: first, the article did not provide information for the RQ1 and RQ2 and second, although most of the excluded articles were related to privacy preserving data mining, they were not relevant to the side effects: hiding failures, missing cost, and artificial cost. Considering the eligible articles and excluded articles from the full-text screening stage, the remainder of the articles included for review were fifty-one.

### ***Phase Two***

**Figure 11***Flowchart for Phase Two Systematic Review*

Phase two data analysis was related to RQ3. The non-traditional side effects considered for the study are data dissimilarity, hidden rules, new rules, and lost rules.

- RQ3: What are the non-traditional side effects, and do they occur in PPDM algorithms?

Duplicate results from IEEE database found were thirty-three from a total of one hundred and thirty-two. Ninety-nine articles were finalized. Two search results were obtained from Proquest. Six duplicate results were found, and a total of fifty-nine articles were selected during the identification step.

ScienceDirect had a total of twenty-two search results with nine articles not considered. The nine articles had no title and missing author names. EBSCOhost database retrieved a total of 48 search results. Twelve duplicates were removed, rest were considered for further analysis. In the SpringerLink database, for three different search results, there was a total of one hundred and five articles retrieved. Forty-two were duplicates and sixty-three unique articles were considered.

ACM Digital Library's advanced search retrieved 539,746 results for each keywords pattern. Hence, results from the ACM Digital Library were considered unreliable. SpringerLink's search results were not considered due to similar reasons as the ACM Digital Library.

As part of the screening process, a total of one hundred and ninety-four articles were included after omitting the duplicates. Excluded articles after screening were thirty-four. In the records screened step, the articles were screened based on the information available in the title and abstract.

Articles sought for retrieval included full text screening of the articles minus the excluded articles from the total number of screened articles. One-hundred and sixty articles were sought for full text retrieval for the study. For articles not retrieved, sixty-one articles were unable to find the full text.

In the full-text screening stage, ninety-nine articles were assessed for eligibility. This assessment was achieved by considering articles sought for retrieval minus articles not retrieved.

This step also involved full-text screening to assess the eligibility to include articles for systematic review.

Articles excluded were a total of fifty-four. There are two reasons for excluding the articles: first, the article did not provide information for the RQ1 and RQ2 and second, although most of the excluded articles were related to privacy preserving data mining, they were not relevant to the side effects: hiding failures, missing cost, and artificial cost. Considering the eligible articles and excluded articles from the full-text screening stage. The remainder of the articles included for review were forty-five.

### ***Phase Three***

The phase three process involved testing the PPDM algorithms within PPSF tool for RQ4: What are the unknown side effects occurring in PPDM algorithms?

In the PPSF tool, the “PPDM” option was selected from “Choose an algorithm” dropdown menu. Among the six algorithms, “Greedy” algorithm was selected to upload the datasets. The retail dataset was uploaded to “Choose input database file.” The retail sensitive text file was included in “Choose input sensitive itemset file” input field. The sensitive itemset file contained the itemsets to be removed from the input file. A new greedy results text file was selected for input field “Set output file.” Input field “Minsup (%)” was set to 0.5. Next input field “Sensitive percentage (%)” was set to 0.01. The other input field W1 was set to 0.5. The last input fields W2 and W3 were set to 0.05. The “Run algorithm” button helped to run the algorithm. Depending on the file size, each algorithm took certain time to display the results. The results were shown in the bottom window of the PPSF tool, below the run algorithm button. A more detailed result was given in greedy results text file. Figure 12 gives an overview of the greedy algorithm running.

**Figure 12***Greedy Algorithm Running in PPSF Tool*

The screenshot shows the PPSF 18v1 application window. The title bar reads 'PPSF 18v1'. The main interface has a light gray background with the 'PPSF' logo in blue. Below the logo, there are several configuration options:

- Choose an algorithm:** A dropdown menu set to 'Greedy' with a '?' button next to it.
- Choose input database file:** A text box containing 'retail.txt' and a browse button '...'.
- Choose input sensitive itemets file:** A text box containing 'retailsensitive.txt' and a browse button '...'.
- Set output file:** A text box containing 'greedyresults' and a browse button '...'.
- Minsup (%):** A text box containing '0.5' with a hint '(e.g. 0.9 or 90%)'.
- Sensitive percentage (%):** A text box containing '0.01' with a hint '(e.g. 0.01 or 1%)'.
- w1 (%):** A text box containing '0.5' with a hint '(e.g. 0.9 or 90%)'.
- w2 (%):** A text box containing '0.05' with a hint '(e.g. 0.05 or 5%)'.
- w3 (%):** A text box containing '0.05' with a hint '(e.g. 0.05 or 5%)'.

Below these fields is a 'Run algorithm' button and an empty progress bar. At the bottom, a text area displays the following output:

```

Algorithm is running...
===== Greedy - STATS =====
The fitness is : 0.0
There are 176322 transactions left in the database.
Total time ~ 2367300 ms
=====

```

Similarly, the remaining five PPDM algorithms POS2DT, cpGA2DT, pGA2DT, and sGA2DT followed the above exact steps of the Greedy algorithm. The details for each dataset,

the corresponding algorithms, and the output files are shown in Table 5. SIF-DIF is the only algorithm that did not require sensitive itemset files, w1, w2, and w3 percentage to be inputted.

**Table 5**

*Algorithms and Output File for Retail dataset*

<b>Algorithms</b>	<b>Input File</b>	<b>Sensitive itemset</b>	<b>Output File</b>
<b>Greedy</b>	retail.txt	retailsensitive.txt	Greedyresults.txt
<b>sGA2DT</b>	retail.txt	retailsensitive.txt	sGA2DTresults.txt
<b>pGA2DT</b>	retail.txt	retailsensitive.txt	pGA2DTresults.txt
<b>cpGA2DT</b>	retail.txt	retailsensitive.txt	cpGA2DTresults.txt
<b>PSO2DT</b>	retail.txt	retailsensitive.txt	Pos2dtresults.txt
<b>SIF-IDF</b>	retail.txt	N/A	sifdifresults.txt

#### ***Phase Four***

After an extensive search over Google, no PPDM repository could be found. The search was based on “finding a website which reports the privacy preserving data mining algorithms’ side effects resolved”. Wikipedia has a page for PPDM definition, yet this page has no details of list of algorithms implemented to date. Decision was made to proceed with PPDM website creation due to lack of existing resources. The remainder of this section will detail about implementation of the website.

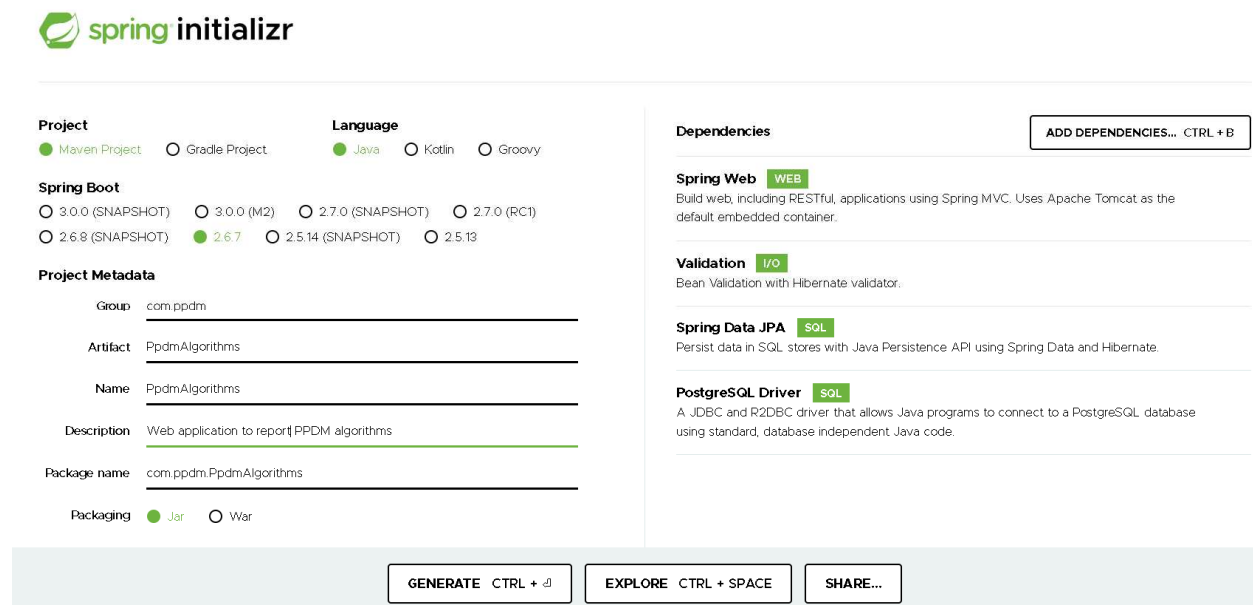
The initial step for the website involved creating Spring Boot application and testing the five REST API endpoints.

**Creating Spring Boot application.** The spring boot application was created using Spring Initializer web service. Spring Initializer was accessed via <https://start.spring.io/> as shown in Figure 13. Details were entered as given in Table 6. Spring Web, Validation, PostgreSQL driver,

Spring Data Java Persistence API (JPA) were included in dependencies section. To create the maven project file the “Generate” button was selected. Finally, PPDMAgorithms.zip file was generated to start the code development. The unzipped file was imported using Spring Tool Suite IDE as shown in Figure 14.

**Figure 13**

*Created PpdmAlgorithms Application Using Spring Initializr*



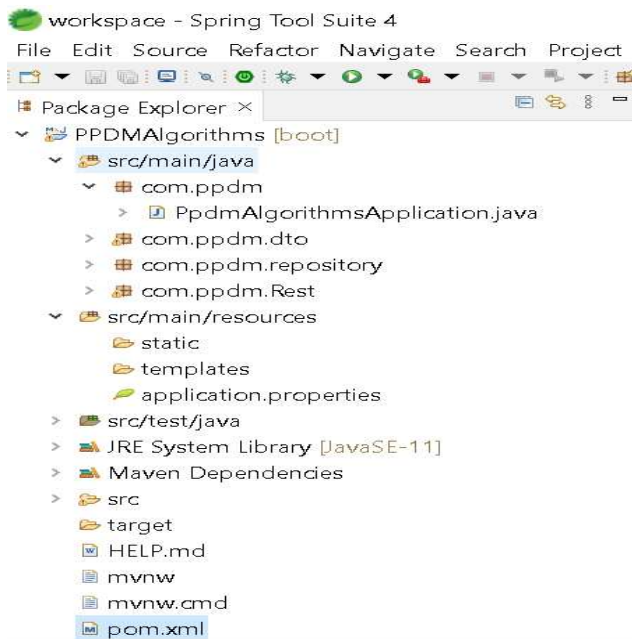
The screenshot displays the Spring Initializr web interface. The top left features the Spring Initializr logo. The main configuration area is divided into several sections:

- Project:** Includes radio buttons for **Maven Project** (selected) and **Gradle Project**.
- Language:** Includes radio buttons for **Java** (selected), **Kotlin**, and **Groovy**.
- Spring Boot:** Includes radio buttons for various versions: 3.0.0 (SNAPSHOT), 3.0.0 (M2), 2.7.0 (SNAPSHOT), 2.7.0 (RC1), 2.6.8 (SNAPSHOT), **2.6.7** (selected), 2.5.14 (SNAPSHOT), and 2.5.13.
- Project Metadata:** Includes input fields for:
  - Group:** com.ppdmi
  - Artifact:** PpdmAlgorithms
  - Name:** PpdmAlgorithms
  - Description:** Web application to report PPD algorithms
  - Package name:** com.ppdmi.PpdmAlgorithms
  - Packaging:** Includes radio buttons for **Jar** (selected) and **War**.
- Dependencies:** A list of selected dependencies with their categories:
  - Spring Web** (WEB): Build web, including RESTful, applications using Spring MVC. Uses Apache Tomcat as the default embedded container.
  - Validation** (I/O): Bean Validation with Hibernate validator.
  - Spring Data JPA** (SQL): Persist data in SQL stores with Java Persistence API using Spring Data and Hibernate.
  - PostgreSQL Driver** (SQL): A JDBC and R2DBC driver that allows Java programs to connect to a PostgreSQL database using standard, database independent Java code.

At the bottom, there are three buttons: **GENERATE** (CTRL + G), **EXPLORE** (CTRL + SPACE), and **SHARE...**

**Table 6***Maven Project Related Details*

Field	Value
Group	com.ppdM
Artifact	PpdmAlgorithms
Name	PpdmAlgorithms
Description	Web application to report PPDM algorithms
Package name	com.ppdM
Packaging	Jar
Language	Java
Java Version	17
Generated Project	Maven

**Figure 14***Spring Tool Suite Project Directory Structure*



**Setting up PostgreSQL database.** To connect to postgresQL database changes were made to pom.xml and application.properties file. First a database name called PPDMALGORITHMMS was created. In pom.xml file PostgreSQL details were included as shown in Figure 15. Similarly necessary connection details were added to application.properties file. Hibernate automatically created the database tables. The property label responsible for auto creation was “spring.jpa.hibernate.ddl-auto”. Figure 16 shows the database connection information for PostgreSQL.

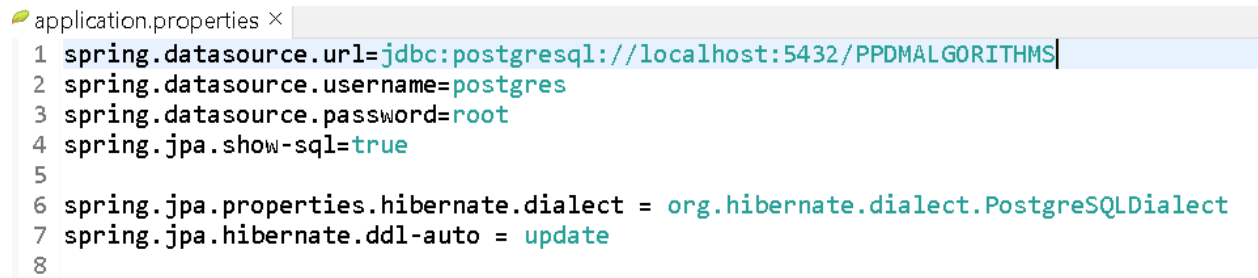
**Figure 15**

*PostgreSQL Dependency Details in pom.xml*

```
<dependency>
  <groupId>org.postgresql</groupId>
  <artifactId>postgresql</artifactId>
  <scope>runtime</scope>
</dependency>
```

**Figure 16**

*PostgreSQL Connection Details in Application Properties File*



```
application.properties ×
1 spring.datasource.url=jdbc:postgresql://localhost:5432/PPDMALGORITHMS
2 spring.datasource.username=postgres
3 spring.datasource.password=root
4 spring.jpa.show-sql=true
5
6 spring.jpa.properties.hibernate.dialect = org.hibernate.dialect.PostgreSQLDialect
7 spring.jpa.hibernate.ddl-auto = update
8
```

**Implementing Domain.** PpdmDTO is the name of the domain implementation and corresponds to the Ppdm domain Object. Implementation is located in folder src/main/java and sub package is com.ppdm.dto. PppdmDTO's complete source code is available in [https://github.com/himaait/Backup\\_Aug\\_2\\_PPDM.git](https://github.com/himaait/Backup_Aug_2_PPDM.git).

Figure 17

*PPDM Domain Implementation*

PPDM
+ppdmAlgorithmName: Long +ppdmTechniqueName: String +hidingFailure: String +missingCost: String +artificialCost: String +dataDissimilarity: String +otherSideEffects: String +email: String
+getAlgorithmId(): Long +setAlgorithmId(algorithmId:Long) +getPpdmAlgorithmName(): String +setPpdmAlgorithmName(ppdmAlgorithmName:String) +getPpdmTechniqueName(): String +setPpdmTechniqueName(ppdmTechniqueName:String) +getHidingFailure(): String +setHidingFailure(hidingFailure:String) +getMissingCost(): String +setMissingCost(missingCost:String) +getArtificialCost(): String +setArtificialCost(artificialCost:String) +getDataDissimilarity(): String +setDataDissimilarity(dataDissimilarity:String) +getOtherSideEffects(): String +setOtherSideEffects(otherSideEffects:String) +getEmail(): String +setEmail(email:String)

**Implementing Repository.** Repository interface called PpdmJpaRepository was created by extending Spring Data JPA. This implementation required JpaRepository's dependency

details to be added to Maven pom.xml, is shown in Figure 18. Repository implementation is at src/main/java/ folder and com.ppdmm.repository package. PpdmmJpaRepository's complete source code is available in [https://github.com/himaait/Backup\\_Aug\\_2\\_PPDM.git](https://github.com/himaait/Backup_Aug_2_PPDM.git).

**Figure 18**

JpaRepository Dependency in pom.xml

```
<dependency>
  <groupId>org.springframework.boot</groupId>
  <artifactId>spring-boot-starter-data-jpa</artifactId>
</dependency>
```

**Creating a RESTful API.** RESTful API was incorporated to create, update, list, and delete algorithms. A RESTful API was built by creating RESTful controller called PpdmmAlgorithmsRestController. REST endpoints implemented is as shown in Table 7. PpdmmAlgorithmsRestController's complete source code is available in Appendix A. API operations were implemented using JSON format. PpdmmAlgorithmsRestController class is located at com.ppdmm. Rest package was created within src/main/java/ folder.

**Table 7**

*REST Endpoint for PpdmmAlgorithms Application*

HTTP Methods	REST Endpoint	Description
GET	/api/ppdm/	Get all PPDM algorithms
GET	/api/ppdm/{algorithmid}	Get a PPDM algorithm by id
POST	/api/ppdm/	Create new PPDM algorithm
PUT	/api/ppdm/{algorithmid}	Update a PPDM algorithm
DELETE	/api/ppdm/	Delete an algorithm by id

The required endpoints were implemented in the controller class to retrieve and manipulate the PPDM algorithms information. The method names created for each endpoint are listAllAlgorithms, getAlgorithmById, createAlgorithm, updateAlgorithm, and deleteAlgorithm.

All the endpoints were tested using Postman app. Postman desktop agent was used to establish connection between Spring Boot application and Postman app. PPDMAlgorithms application was launched as Spring Boot application using Spring Tool Suite (Figure 19).

Postman app was accessed through <https://postman-echo.com/>.

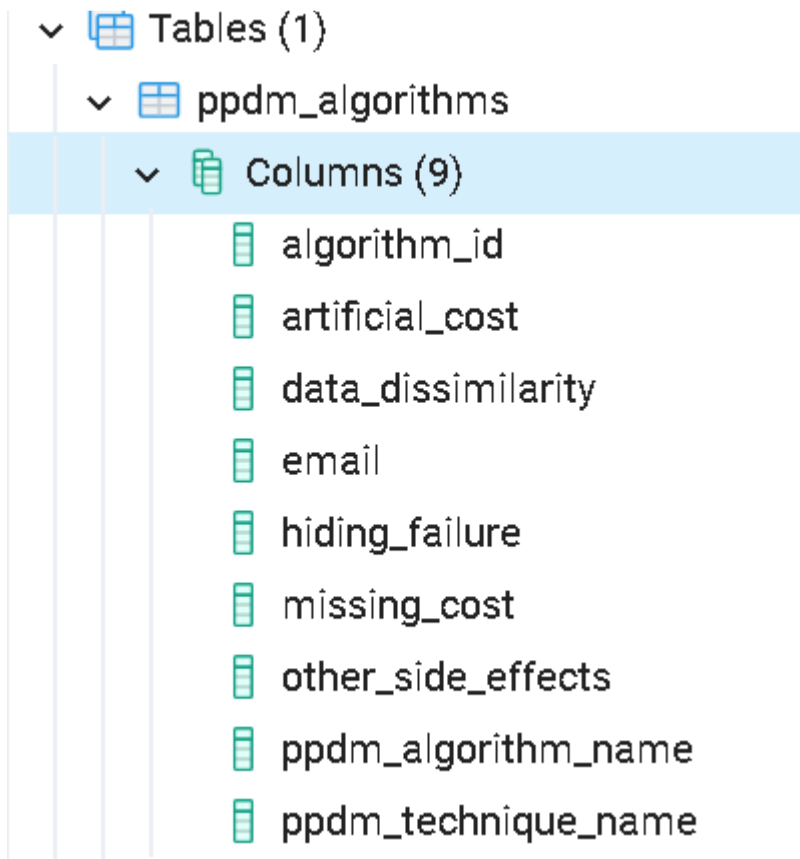
**Figure 19**

*Spring Tool Suite's Console Showing the Status as Started*

```
[
  main] com.ppdM.PpdmAlgorithmsApplication : Starting PpdmAlgorithmsApplication using Java 17.0.1 on LAPTOP-UU2K97U2 with PID 10200 (C:\Users\javau\Docu
[
  main] com.ppdM.PpdmAlgorithmsApplication : No active profile set, falling back to default profiles: default
[
  main] .s.d.r.c.RepositoryConfigurationDelegate : Bootstrapping Spring Data JPA repositories in DEFAULT mode.
[
  main] .s.d.r.c.RepositoryConfigurationDelegate : Finished Spring Data repository scanning in 115 ms. Found 1 JPA repository interfaces.
[
  main] o.s.b.w.embedded.tomcat.TomcatWebServer : Tomcat initialized with port(s): 8080 (http)
[
  main] o.apache.catalina.core.StandardService : Starting service [Tomcat]
[
  main] org.apache.catalina.core.StandardEngine : Starting Servlet engine: [Apache Tomcat/9.0.56]
[
  main] o.a.c.c.C.[Tomcat].[localhost].[/] : Initializing Spring embedded WebApplicationContext
[
  main] w.s.c.ServletWebServerApplicationContext : Root WebApplicationContext: initialization completed in 4830 ms
[
  main] com.zaxxer.hikari.HikariDataSource : HikariPool-1 - Starting...
[
  main] com.zaxxer.hikari.HikariDataSource : HikariPool-1 - Start completed.
[
  main] org.hibernate.jpa.internal.util.LogHelper : HHH000204: Processing PersistenceUnitInfo [name: default]
[
  main] org.hibernate.Version : HHH000412: Hibernate ORM core version 5.6.4.Final
[
  main] org.hibernate.annotations.common.Version : HCAANN000001: Hibernate Commons Annotations {5.1.2.Final}
[
  main] org.hibernate.dialect.Dialect : HHH000400: Using dialect: org.hibernate.dialect.H2Dialect
[
  main] o.h.e.t.j.p.i.JtaPlatformInitiator : HHH000490: Using JtaPlatform implementation: [org.hibernate.engine.transaction.jta.platform.internal.NoJtaP
[
  main] j.LocalContainerEntityManagerFactoryBean : Initialized JPA EntityManagerFactory for persistence unit 'default'
[
  main] JpaBaseConfiguration$JpaWebConfiguration : spring.jpa.open-in-view is enabled by default. Therefore, database queries may be performed during view ren
[
  main] o.s.b.a.w.s.WelcomePageHandlerMapping : Adding welcome page: class path resource [static/index.html]
[
  main] o.s.b.a.e.web.EndpointLinksResolver : Exposing 1 endpoint(s) beneath base path '/actuator'
[
  main] o.s.b.w.embedded.tomcat.TomcatWebServer : Tomcat started on port(s): 8080 (http) with context path ''
[
  main] com.ppdM.PpdmAlgorithmsApplication : Started PpdmAlgorithmsApplication in 11.607 seconds (JVM running for 14.157)
[on(3)-127.0.0.1] o.a.c.c.C.[Tomcat].[localhost].[/] : Initializing Spring DispatcherServlet 'dispatcherServlet'
[on(3)-127.0.0.1] o.s.web.servlet.DispatcherServlet : Initializing Servlet 'dispatcherServlet'
[on(3)-127.0.0.1] o.s.web.servlet.DispatcherServlet : Completed initialization in 2 ms
```

**Figure 20**

*Table Ppdm\_algorithms Auto Created by JPA's @Entity and @Table Annotations*



**Angular JS Front End Implementation.** The Angular JS platform was used to create a single page web application. AngularJS dependency information was added in Spring Boot application's pom.xml. Single page application was implemented by creating one HTML page. This method was adopted as it helps to dynamically add and remove content in a single HTML page. Single page application helps reduce wait time of loading multiple HTML pages. The index.html page was defined as single page application by including html tag <div ng-view></div>. AngularJS application was bootstrapped(started) by including ng-app in the index.html page. As part of Dependency Injection, the dependencies added for the application were ngRoute and ngResource. The file responsible for these dependencies is app.js. A ngRoute

variable helps to route the application between controller (logic) and views (web page).

However, ngResource variable helps to interact with RESTful services.

**Figure 21**

*AngularJS and BootStrap dependencies in pom.xml*

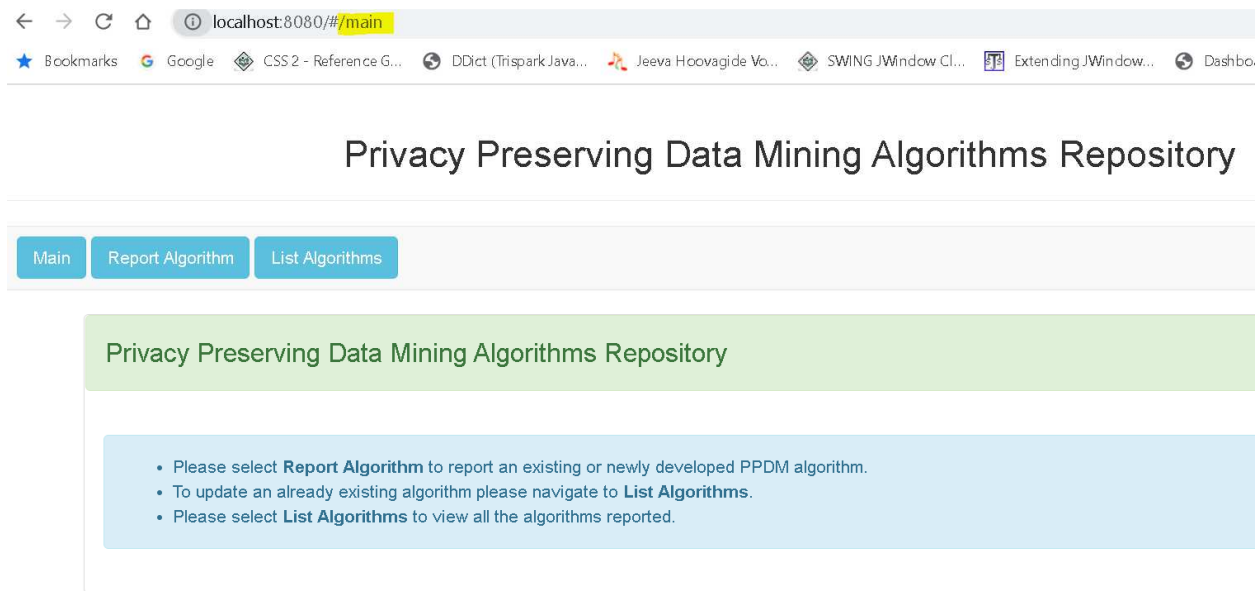
```
<dependency>
  <groupId>org.webjars</groupId>
  <artifactId>angularjs</artifactId>
  <version>1.4.9</version>
</dependency>

<dependency>
  <groupId>org.webjars</groupId>
  <artifactId>bootstrap</artifactId>
  <version> 3.3.6</version>
</dependency>
```

**Routing in AngularJS.** This is an important aspect to create a single page application. AngularJS routes for PPDMAAlgorithms application are /main, /list-all-algorithms, /add-new-algorithm, /update-algorithm, and redirect to home page. These routes help to navigate to different functionalities of the application from the URL. For example, /main route after the /# in the url lands the application's homepage. Figure 22 shows the route /main in the url.

**Figure 22**

*AngularJS Route Defined as /main for Homepage*



**Model, View, and Controller Implementation.** The view usually represents the user interface such as web pages. The main page of the application called index.html was created in src/main/resources/static. Three links called Add Algorithm, List Algorithms, and Main was added to index.html. Total of four view pages were generated for View implementation:

- Home page: src/main/resources/template/main.html
- Add Algorithm page: src/main/resources/template/addalgorithm.html
- List Algorithms: src/main/resources/template/listalgorithms.html

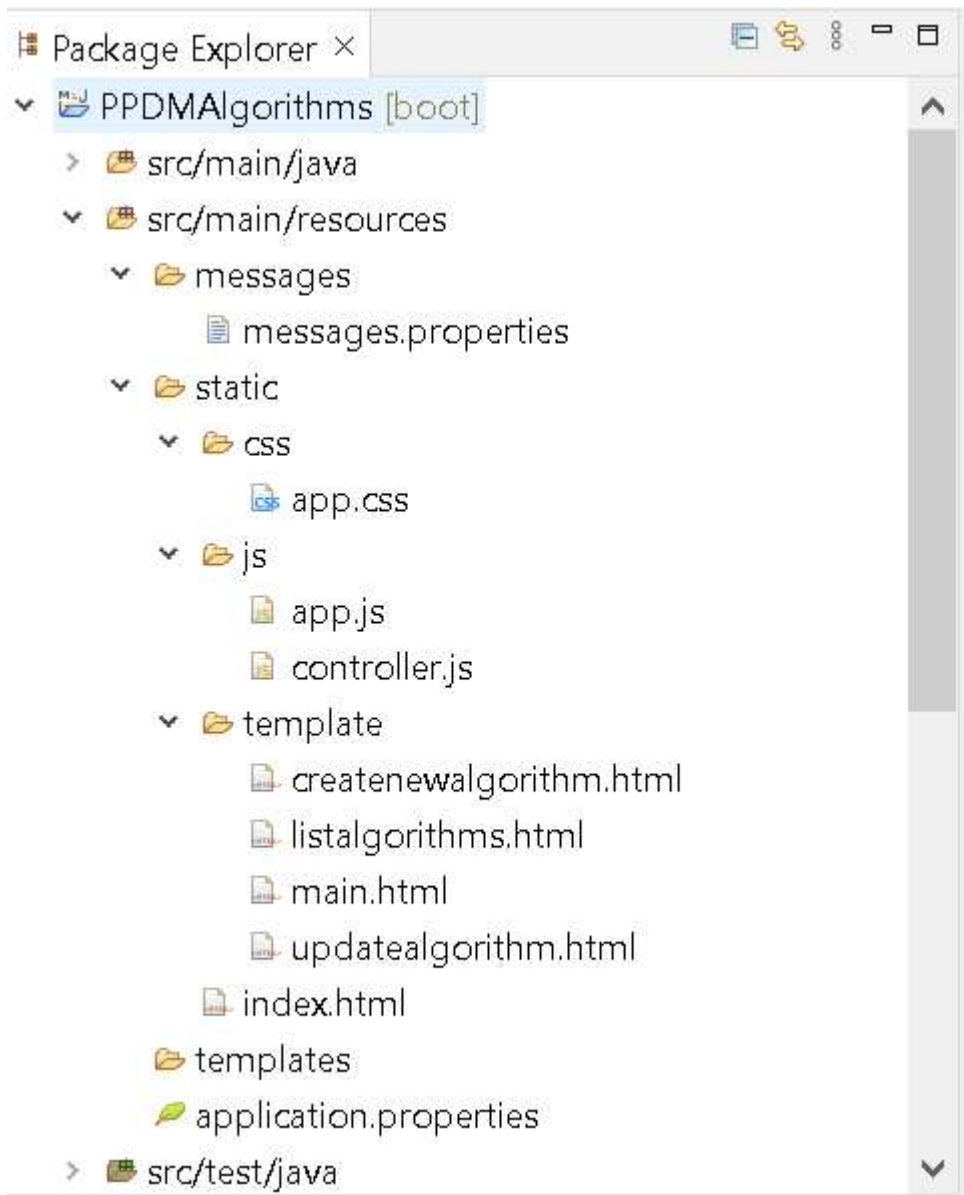
The actual implementation of accessing the view pages is executed in file app.js. This file has all the corresponding routes and application configuration. The variable angular.module was used to route and provide the required resources for the AngularJS application. The file app.js is located at src/main/resources/static/js/app.js.



The AngularJS controller was created as controller.js. The location of the controller file is shared with app.js's location. The main agenda of the controller is responsible to implement business logic for user's interaction . The model of the application represents the data for reporting a PPDM algorithm. For example, PPDMAlgorithms application's data model consists of algorithm id, name, technique, hiding failure, missing cost, artificial cost, data dissimilarity, and other side effects.

**Figure 23**

*Single Page Application Directory Structure*



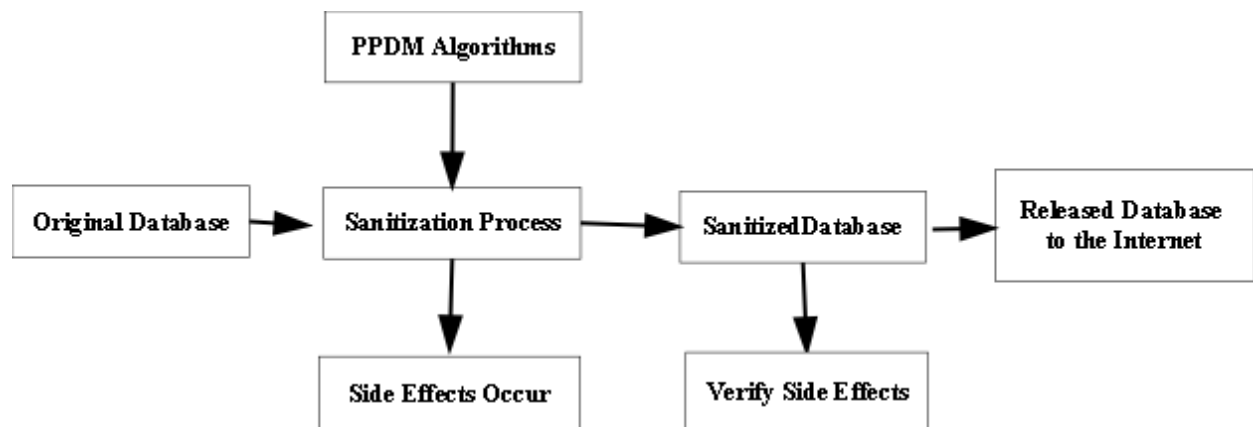
## Findings

### *Phase One*

Although the present study focused on only side effects, understanding the process of the data sanitization was necessary. This helps to know where exactly these side effects occur in data sanitization. The primary reason to understand the process is because the data sanitization process has many factors to be considered even before the side effects are observed. Figure 21 gives clarity about data sanitization and side effects occurrences (Zhang et al., 2019). It can be observed from Figure 24 that as part of sanitization process PPDM algorithms are applied. During the process of sanitization, side effects occur. The verification of the side effects generated or not can only be tested in the sanitized database. Based on the nature of the side effects, and data quality decision is made to release to the Internet.

**Figure 24**

*Data Sanitization Process*



The data mining tasks related to PPDM is association rule mining, clustering, and classification (Mendes & Vilela, 2017). Association rule mining (ARM) is a form of data sanitization which applies a rule based PPDM. Privacy preserving ARM involves finding

unspecified patterns and their association from the data within a database. The rule hiding algorithms mainly use a PPDM technique called association rule hiding (ARH) technique. The objective of ARH is to make sure to hide sensitive association rule generated from ARM (Shah et al, 2012). ARH is also known as sensitive pattern protection model (Zhang et al., 2019). The main agenda of ARH is to protect sensitive information through data modification.

A real-world example for ARM is seen applied in grocery stores. For example, customers who buy bread will mostly buy spread (peanut butter and jelly). The company collects this association of bread and spreads from the grocery's database. Then the company tries to attract customers by providing discounts or attract more customers by selling both peanut butter and jelly in one jar. This is just a simple example of ARM, several such associations can be observed in toy stores, clothing, furniture, banking, and jewelry shopping

It was observed in systematic literature review, most of the association rule related algorithms' side effects are examined during sanitization process. Side effects with respect to association rule hiding are mainly associated with hiding sensitive patterns. Association rule mining related algorithms evaluate the side effects based on both sensitive and non-sensitive patterns.

Three commonly considered side effects for evaluating PPDM algorithms are hiding failure, missing cost, and artificial cost. This information was gathered from all the research studies related to PPDM algorithms implementation. Additionally, the side effects data was also gathered from frequent itemset mining and association rule mining algorithms. As the scope of the study is related to only PPDM algorithms, hence the side effects related to only these algorithms will be discussed.

Hiding failure characteristics identified are sensitive itemset/information is still found in sanitized database. As per the hidden rule applied for an algorithm, these sensitive patterns selected from original database should not be visible in the secured/sanitized database. Hiding failure is also mentioned as “failure to hide” or “hidden failures”. As per the framework established by Bertino et al. (2005) specifically for evaluating PPDM algorithms, hiding failure helps in determining the balance between privacy and information finding.

Missing cost characteristics refers to non-sensitive itemsets/information which exists in original database but is not found in sanitized database. Missing cost is also known as missescost or missing itemsets. Non-sensitive information is considered important and useful, hence should not go missing during sanitization process. The non-sensitive information plays an important role to maintain the resemblance of the original database in sanitized database. Similarly, missing cost as observed in all studies were identified as huge amount of non-sensitive information lost. Most of the PPDM algorithms implemented were for delete transaction. The main intention of solving or identifying missing cost is sensitive information acquired through not so important information has to be protected as well.

Artificial cost characteristics were related to new or artificial itemsets/information being present in sanitized database. This artificial information is not useful and does not exist in original database, hence should not be visible in sanitized database. Artificial cost is also known as artificial itemsets or artificial patterns or artifactual patterns or new rules. In few studies the term “ghost rules” is used for new rules to identify artifactual information in sanitized database.

Few PPDM research studies discussed the relationship between the three side effects HF, MC, and AC. Researchers explained the relationship of side effects, itemsets and mined rules for data sanitization process. These side effects were used to evaluate the performance of the PPDM

algorithm techniques. Bertino et al. (2005) specifically discusses the relationship and impact of both hiding failure and misses cost have on each other. The study explains the interdependency and impact of both HF and MC have on each other. Further the researchers mention the importance of these side effects are for evaluating efficiency, performance, and quality of any PPDM algorithms developed. Similarly, both misses cost, and artificial cost was also used to evaluate the data quality in sanitized database. Hence, only AC or MC alone cannot help in determining the data quality.

**Table 8**

*Identified Common Side Effects*

<b>Side effects</b>	<b>Other names</b>
Hiding Failures	Hidden failures, failure to hide
Missing cost	Misses cost, missing itemset, data utility
Artificial cost	Artificial itemset, artificial patterns

**Table 9***Characteristics of Side Effects*

	<b>Hiding Failure</b>	<b>Missing cost</b>	<b>Artificial cost</b>
<b>Type of data/itemset</b>	Related to sensitive itemsets/information	Related to nonsensitive itemsets/information	Related to artificial or new itemsets/information
<b>Rule for sanitization</b>	Sensitive data <b>should not be revealed</b>	Nonsensitive information is considered useful and <b>should be revealed</b>	Artificial or new information is not useful or required and <b>should not be revealed</b>
<b>Original database</b>	Sensitive information exists	nonsensitive information exists	artificial or new information does not exist
<b>Sanitized database</b>	Sensitive data <b>is shown</b> in sanitized database	Nonsensitive information is <b>not found</b> in sanitized database	Unrelated new or artificial itemsets <b>is found</b> in sanitized database

To summarize the expected results defined in the methodology section, the general characteristics of each of the side effects were related to failure to hide important or sensitive data, non-sensitive information lost, and new data introduced during the sanitization process. Similarities of each of the side effects seen was that all the HF, MC, and AC were observed when PPDM algorithms was applied during sanitization process. Apart from this one similarity, there were no resemblances seen in characteristics of the side effects.

There were three differences observed for the side effects. Hiding failures were related to hidden rules. Whereas missing cost were related to lost or non-sensitive rules which were accidentally hidden. However, artificial cost mainly involved new rules or ghost rules which were accidentally created during sanitization process.

It was observed each side effect is related to a type of itemset/data. Hence the relationship of side effects is dependent on that of the itemsets. For example, missing itemsets is related to non-sensitive itemsets. When missing itemsets are identified this is referred to as missing cost or misses cost. Similarly, failing to hide itemsets is related to sensitive items. HF is the coined term for when sensitive itemsets are failed to be hidden. Artifactual itemsets or artificial itemset refers to itemsets accidentally generated. This newly generated itemset does not belong to the original database. When these artificial itemsets are identified then the AC term is used.

Hence each of the side effects is related or interdependent to determining PPDM algorithms' data quality, performance, or efficiency. Suppose MC and AC are identified, then this will determine the sanitized data's quality. The more the MC and AC the more the quality of the data is reduced and vice versa. The reason for the reduction in quality is due to data loss or the creation of unrelated new data. Increase in missing itemsets and new itemsets created do not help in replicating the original database. However, in the case of hiding failure information privacy concerns come into the limelight as sensitive data is visible publicly. In some cases, not handling the non-sensitive data/itemset also gives scope to intruders to identify sensitive information. In terms of side effects, mishandling of missing costs side effect leads to an increase in hiding failure and vice versa. For example, few research studies found that more the sensitive information is hidden (HF) more the non-sensitive information is lost (MC).



There were relationships and impacts observed between each of the side effects. In context to HF and MC, both these side effects had an impact on each other. On the other hand, AC and MC were interdependent to measure the sanitized data quality. Phase one helped in finding interesting facts about PPDM techniques, methods, and side effects from various research studies. The research studies found were between years 2005 to 2022.

**Table 9**

*Relevant information retrieved from the research studies*

Author	Year	Data Source
Lee et al.	2021	SpringerLink
Nithya et al.	2021	Science Direct
Wu et al.	2021	ProQuest
Aldeen et al.	2020	SpringerLink
Jangra et al.	2020	EBSCOhost
Liu et al.	2020	SpringerLink
Li et al.	2019	Science Direct
Lin et al.	2019	EBSCOhost
Lin et al.	2019	ProQuest
Mogtaba and Kambal	2019	SpringerLink
Wu et al.	2019	EBSCOhost
Wu et al.	2019	IEEE
Wu et al.	2019	ProQuest
Wu et al.	2019	ProQuest
Zainab et al.	2019	ACM

---

Kamakshi and Vinaya Babu	2018	SpringerLink
Murthy et al.	2018	ProQuest
Nguyen	2018	ProQuest
Telikani et al.	2018	Science Direct
Femandes and Gomes	2017	IEEE
Aghasi et al.	2016	SpringerLink
Lin et al.	2016	Science Direct
Lin et al.	2016	Science Direct
Lin et al.	2016	SpringerLink
Lin et al.	2016	SpringerLink
Priyadarsini et al.	2016	ACM
Rong et al.	2016	IEEE
Selvan and Veni	2016	ProQuest
Nanawati and Jinwala	2015	ProQuest
Sowmya et al.	2015	ProQuest
Kagklis et al.	2014	ACM
Lin et al.	2014	EBSCOhost
Lin et al.	2014	ProQuest
Mandapati et al.	2013	ProQuest
Vaidya et al.	2013	ProQuest
Li et al.	2011	ACM
Wu and Huang	2011	SpringerLink

Naeem et al.	2010	IEEE
Kuo et al.	2009	SpringerLink
Teng and Du	2009	ProQuest
Xiao et al.	2009	ACM
Bertino et al.	2008	SpringerLink
Shailaja and Rao	2008	SpringerLink
Wang and Lee	2008	Science Direct
Navale and Mali	2007	SpringerLink
Surendra and Mohan	2007	SpringerLink
Urabe et al.	2007	ProQuest
Wu et al.	2007	SpringerLink
Gurevich and Gudes	2006	IEEE
Bertino et al.	2005	SpringerLink
Navale and Mali	2005	SpringerLink

### ***Phase Two***

Results of phase two were achieved through systematic review of the literature approach and data analysis. Initial literature review phase identified data dissimilarity, hidden rules, new rules, and lost rules. There were additional four non-traditional side effects names identified during later stages of the research work. Identified non-traditional side effects is shown in the Table 10. Most of the non-traditional side effects were rules set for sanitization process. These rules are evaluated to measure the data quality, performance, or efficiency of the PPDM algorithms.

**Table 10***Non-traditional Side Effects*

<b>Non-traditional side effects</b>	<b>Other names</b>
Data dissimilarity	Dissimilarity, database dissimilarity
Hidden rules	Rule set to hide sensitive itemset
Lost rules	Missing itemset/missing cost/ lost association rule
New rules	Ghost rules
Sensitive rules	N/A
Mined rules	N/A
Artificial rules	New or Ghost rule
Non sensitive rules	N/A
Ghost rules	New rules
Spurious rules	N/A

To understand data dissimilarity, it is necessary to have an original database and a sanitized original database. Additionally, an estimated data set is also predefined to valuate in the cleaned database. In few research studies database dissimilarities was used to measure performance of an algorithm, in others it is studied as a side effect. Deleted transaction size is also used to identify data dissimilarity side effect before and after data sanitization. The simplest way to determine data dissimilarity is to find the difference in the size of the original and sanitized database. However, few studies compared the original database and sanitized database. The comparison approach was considered the best approach to identify data dissimilarity.

Hidden rules are set to hide sensitive or non-sensitive items depending on the necessity of PPDM algorithm implemented. The hidden rule's intention is to hide only sensitive or non-sensitive items. On the other hand, lost rule refers to falsely/accidentally hiding a non-sensitive rule. It was also observed that non-sensitive patterns that are falsely/accidentally hidden is also termed as "Missing Costs". However, new rule is known as erroneously generating fake/artificial rules. Falsely generating artificial (fake) itemsets/data patterns is termed as "Artificial patterns". These artificial patterns help to determine the side effect called artificial cost. Both lost rule and new rule is associated with non-sensitive rule or information.

Artificial rule is identified when rules in original database does not exist but appears after the database is sanitized (Mogtaba & Kambal, 2016). It is also referred to as artificial association rules or artifactual rules (Oliveira et al., 2004). Artifactual rules are also known as new rules or ghost rules as new patterns are generated after the original database is secured (Telikani et al., 2020). It is observed that in few algorithms' rules are set for already observed side effects. Other algorithms consider certain rules as negative or side effects of PPDM algorithm (Telikani et al., 2020).

Sensitive rule in most of the PPDM algorithm was defined to hide sensitive information for a transaction. A research study by Qi and Zong (2012) used the term "mined rule" in reference to sensitive association rules set for a PPDM algorithm. Non-sensitive rule mainly deals with non-sensitive information that should not be hidden. Ghost rule is associated with non-sensitive items/data. However, Ghalehsefidi and Dehkordi (2016) calculated the value of ghost rule based on non-sensitive items and length of left-hand side (LHS) rule (sensitive). The LHS sensitive rule is an association rule set to hide sensitive items/data.

To summarize the phase two's expected results as defined in methodology section, namely the reasons why non-traditional side effects are not commonly used. Non-traditional side effects have been used in PPDM algorithms. One important observation made was the characteristics of both the common and non-traditional side effects are similar except for the terms used. For example, relying on the explanations from previous research studies, lost rules are used to evaluate missing costs, and false rules are used for artificial patterns. If we observe the characteristic of missing cost and lost rules, both deal with non-sensitive information. Only the terms used to explain are different, lost rules use the term non-sensitive rule whereas missing cost refers to non-sensitive information.

The most commonly used non-traditional side effect was data dissimilarity. In some research studies it is considered as fourth common side effect. In another study by Mogtaba and Kambal (2016), lost rules, artificial rules, hiding failures, and dissimilarity. Researchers used side effects hiding failure, missing cost, artificial cost, and database dissimilarity to evaluate efficiency and performance of the algorithm.

To summarize, non-traditional side effects have occurred in the PPDM algorithms. PPSF tool was not reliable to test the occurrences of non-traditional side effects. Although the tool gives options to test all the six algorithms, the tool's results generated were not clear to understand HF, MC, and AC. The reason to not understand the results was due to lack of access to documentation and source code. The phase three section in the findings explains in detail how the PPSF tool was used. Phase one and two successfully collected the side effects information.

**Table 10***PPDM Algorithm Names*

<b>PPDM algorithm name</b>	<b>Abbreviation/Details</b>
ADSSRC	Advances DSSRC
RRLR	Remove and Reinsert L.H.S of rule
DSRRC	Decrease Support of R.H.S. items of Rule Cluster
HCMPSO	Based on multi-objective PSO framework
NSGA2DT	Multi-objective algorithm
FHSAR	Fast Hiding Sensitive Association Rules
RRLR	Remove and Reinsert L.H.S. of Rule
DSSRC	Decrease Support of R.H.S. items of Rule Cluster
IBABC	Improved Binary ABC
ABC4ARH	Artificial Bee Colony Association Rule Hiding

**Table 11***PPDM Algorithms and Related Side Effects Evaluated*

<b>Algorithms</b>	<b>Side effects tested</b>	<b>Data Dissimilarity</b>	<b>Comments</b>
HCMPSO	HF,MC, and AC	Yes	None
Greedy	HF,MC, and AC	Yes	None
pGA2DT	HF,MC, and AC. HF was reduced.	Dissimilarity robustness same as sGA2DT	Fewer HF compared to Greedy and sGA2DT
sGA2DT	HF,MC, and AC. HF was reduced.	Dissimilarity robustness same as pGA2DT	Had less HF compared to Greedy algorithm
NSGA2DT		Database dissimilarity helped to achieve better performance	Satisfactory results for all four side effects
FHSAR	Hiding failure	Yes	Lost rules, and artificial rules
RRLR			RRLR better than DSSRC with respect to lost rules and data modification



**Table 11 Continued**

<b>Algorithms</b>	<b>Side effects tested</b>	<b>Data Dissimilarity</b>	<b>Comments</b>
DSRRC	Failed to hide rules for multiple R.H.S items		Data modification was less
IBABC	HF and MC worse for sparse database but better for dense database		Data accuracy was best for both sparse and dense datasets. Compared with algorithms BPSO, DisABC 980 and binABC
ABC4ARH	Minimum HF, MC, and artificial patterns		artificial bee colony for association rule hiding. Data accuracy worst compared to algorithms PSO2DT, ACS2DT, sGA2DT, and COA4ARH

***Phase Three***

Phase three's expected data were defined as follows:

Find if unknown or new side effects occur for six PPDM algorithms in the PPSF tool.

Initially retail dataset was used to test the unknown or new side effects in PPSF tool. However, the results generated were not clear to understand for all the six algorithms. Hence, this researcher contacted one of the researcher of PPSF tool to confirm if the correct dataset files were used. The researcher confirmed that the dataset files were correctly used. An explanation and example for “input sensitive itemsets” file was provided. As the researcher was unable to access the original source code and the documentation, a detailed explanation for the results was not provided.

This researcher made the decision to create a small database for testing the algorithm. This step was undertaken to get a clear understanding of the results. Bertino et al. (2008) evaluated their framework by manually creating a database to test each of the algorithms. Referring to Bertino et al. (2008), phase three of this study created a small size data file to test PPDM algorithms in PPSF tool. The small input and sensitive database file manually generated is shown in Table 13 and Table 14.

Only one algorithm SIF-IDF abruptly closed the PPSF tool. Due to this behavior SIF-IDF algorithm did not run the algorithm or produce output results. The results generated for the small dataset files were still not convincing to understand or find the unknown or new side effects. This researcher made the final decision to not proceed testing the algorithms with the remaining dataset files. Figure 25 to Figure 36 gives the details of six algorithms tested using PPSF tool.

**Table 12***Small Datasets and Output Files Used for Each Algorithm*

	<b>Input File</b>	<b>Sensitive itemset</b>	<b>Output File</b>
<b>Greedy</b>	smallsize.txt	smallsensitive.txt	smallgreedyresults.txt
<b>sGA2DT</b>	smallsize.txt	smallsensitive.txt	smallsga2dtresults.txt
<b>pGA2DT</b>	smallsize.txt	smallsensitive.txt	smallpga2dtresults.txt
<b>cpGA2DT</b>	smallsize.txt	smallsensitive.txt	smallcpga2dtresults.txt
<b>PSO2DT</b>	smallsize.txt	smallsensitive.txt	smallpos2dtresults.txt
<b>SIF-IDF</b>	smallsize.txt	N/A	smallsifdifresults.txt

**Table 13***Small Dataset File's Details for Each Row*

	<b>Small dataset file details</b>
<b>Row 1</b>	1,4 2,6 3,9 4,1 5,3 6,9 7,3 8,5 9,1 10,5 11,5 12,4 13,7 14,1 15,4 16,8 17,7 18,4
<b>Row 2</b>	19,5 20,2 21,10 22,6 23,8 24,3 25,6 26,4 27,4 28,1 29,10 30,9
<b>Row 3</b>	31,2 32,3 33,2
<b>Row 4</b>	34,9 35,10 36,5
<b>Row 5</b>	37,6 38,8 39,1 40,5

**Table 14**

*Sensitive File Details Showing Sensitive Itemset Selected from Each Row in Small Dataset*

Small sensitive dataset file details	
Row 1	1 4
Row 2	20 2
Row 3	31 2
Row 4	34 9
Row 5	37 6

**Figure 25**

*PPSF Tool Showing sGA2DT Algorithm is Running*

**PPSF**

Choose an algorithm: sGA2DT ?

Choose input database file: retail.txt ...

Choose input sensitive itemsets file: retailsensitive.txt ...

Set output file: retailresults.txt ...

Minsup (%): 0.9 (e.g. 0.9 or 90%)

Sensitive percentage (%): 0.01 (e.g. 0.01 or 1%)

w1 (%): 0.9 (e.g. 0.9 or 90%)

w2 (%): 0.05 (e.g. 0.05 or 5%)

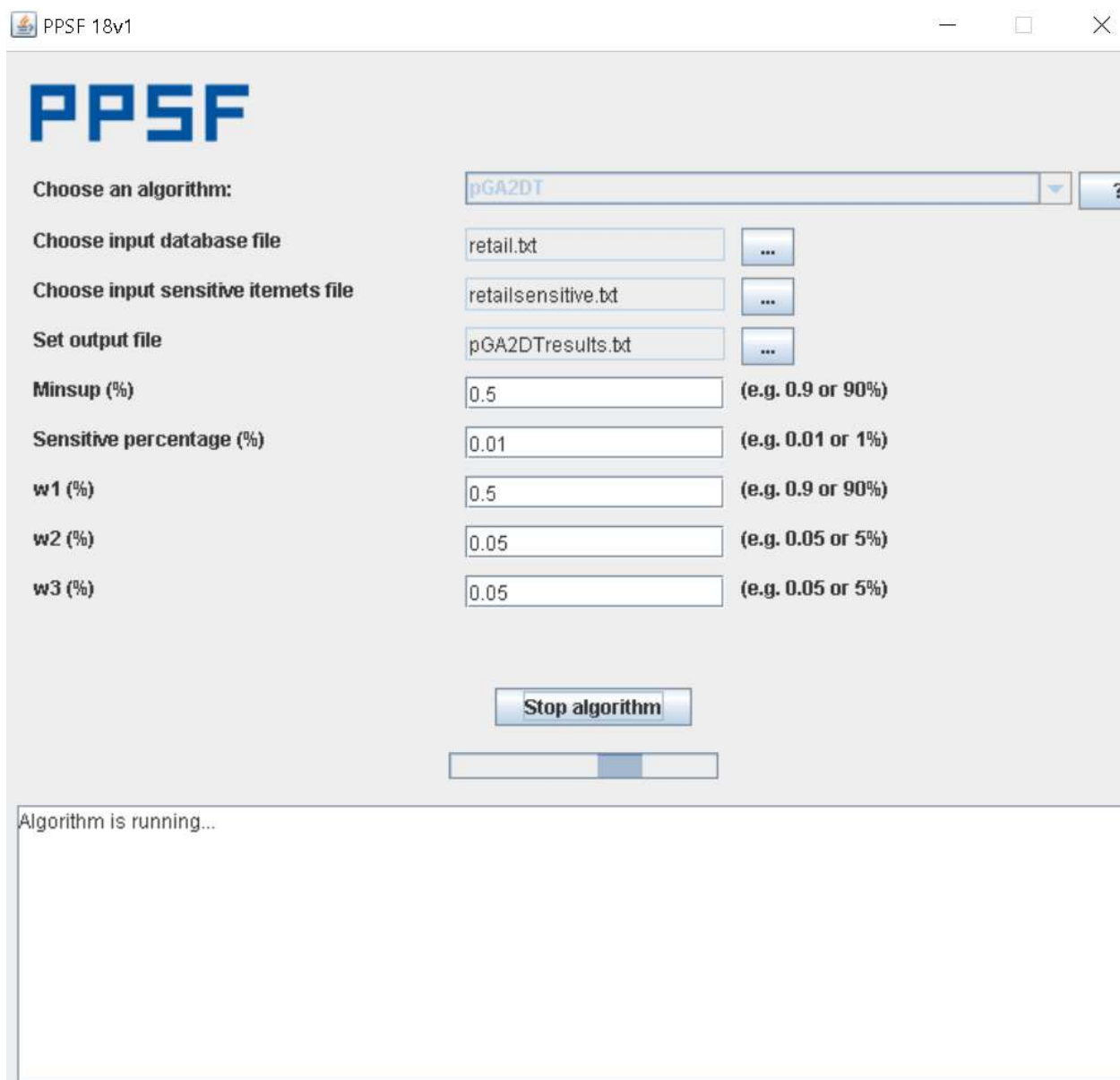
w3 (%): 0.05 (e.g. 0.05 or 5%)

Stop algorithm

Algorithm is running...

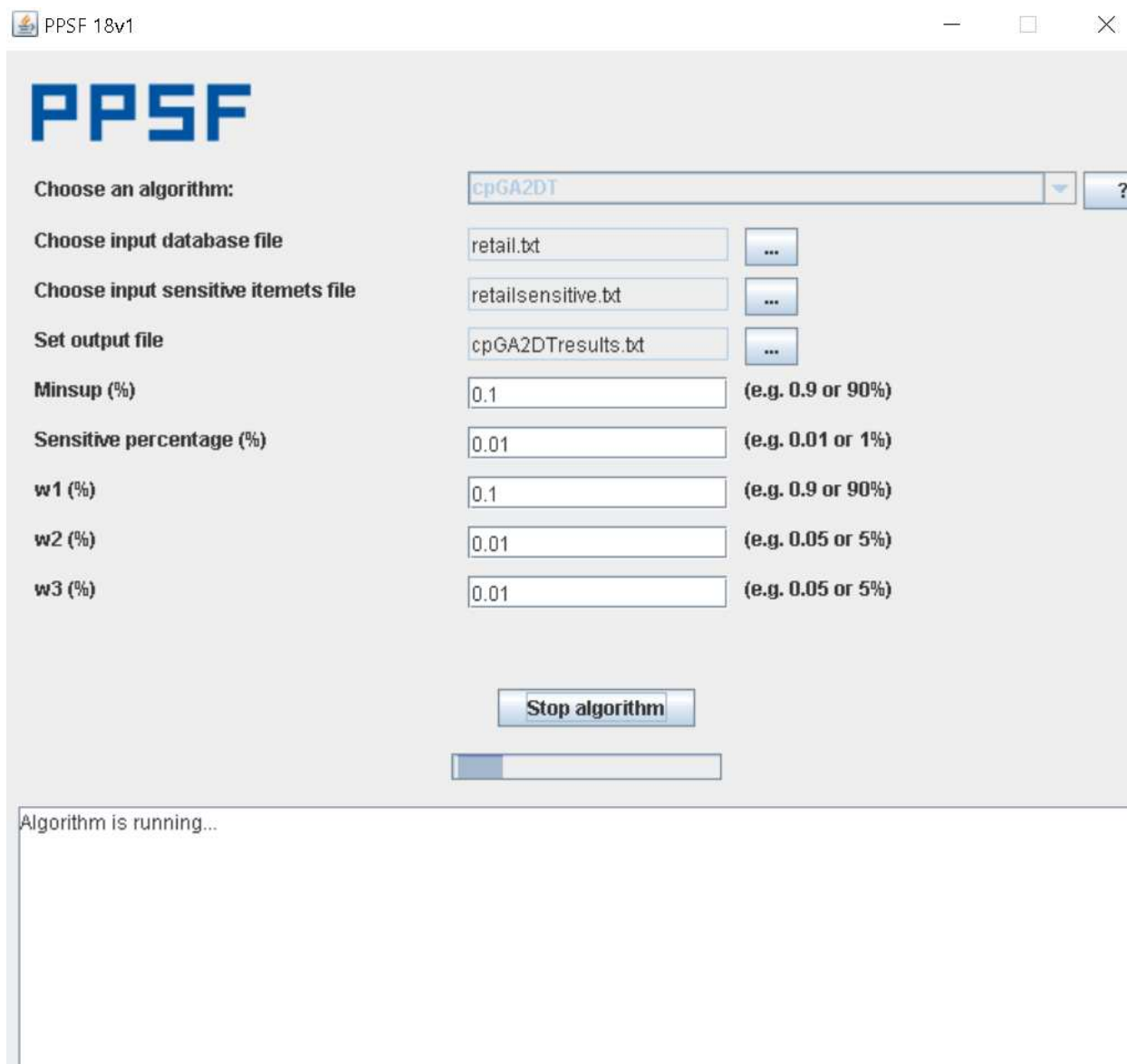
**Figure 26**

*PPSF Tool Showing pGA2DT Algorithm is Running*



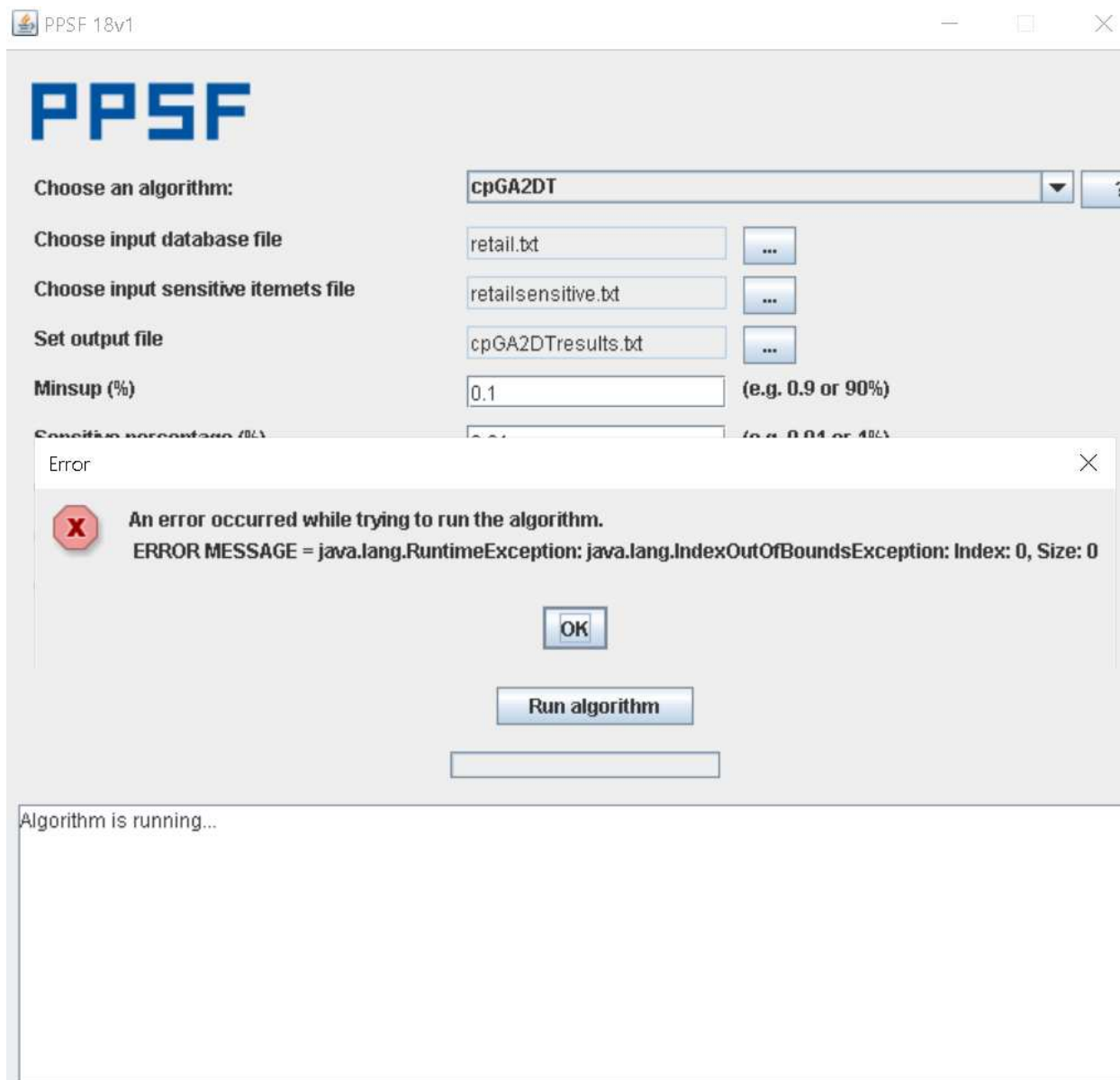
**Figure 27**

*PPSF Tool Showing cpGA2DT Algorithm is Running*



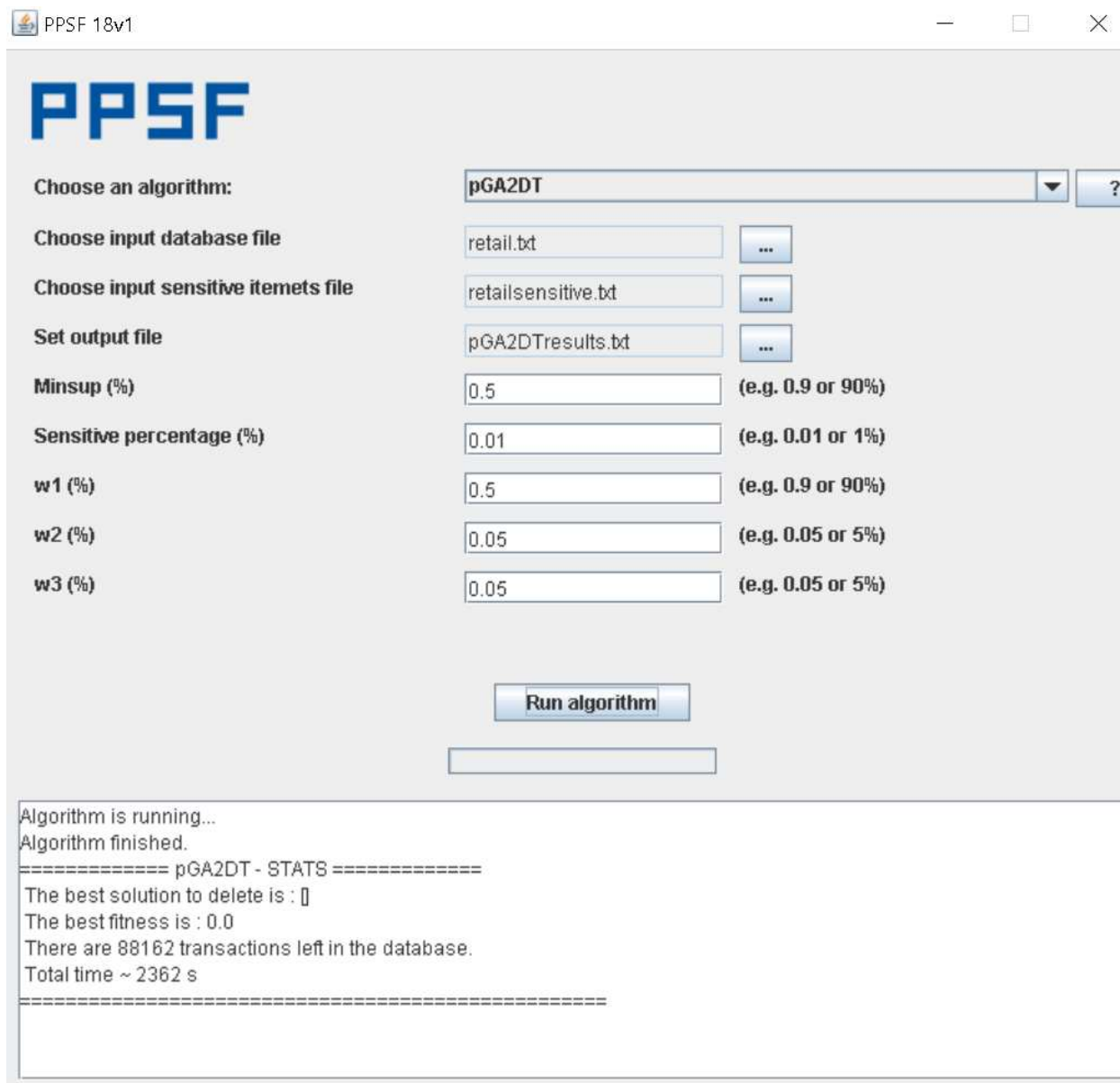
**Figure 28**

*PPSF Tool Showing cpGA2DT Algorithm Testing was Interrupted due to Index Out of Bound Error*



**Figure 29**

*PPSF Tool Showing pGA2DT Algorithm is Running and Finished*





**Figure 30**

*PPSF Tool Showing Greedy algorithm is Running and Stats After Finishing*

The screenshot shows the PPSF 18v1 application window. The title bar reads "PPSF 18v1". The main area features the "PPSF" logo in blue. Below the logo, there are several configuration options:

- Choose an algorithm:** A dropdown menu set to "Greedy" with a question mark icon to its right.
- Choose input database file:** A text box containing "retail.txt" and a browse button "...".
- Choose input sensitive itemets file:** A text box containing "retailsensitive.txt" and a browse button "...".
- Set output file:** A text box containing "greedyresults" and a browse button "...".
- Minsup (%):** A text box containing "0.5" with a hint "(e.g. 0.9 or 90%)".
- Sensitive percentage (%):** A text box containing "0.01" with a hint "(e.g. 0.01 or 1%)".
- w1 (%):** A text box containing "0.5" with a hint "(e.g. 0.9 or 90%)".
- w2 (%):** A text box containing "0.05" with a hint "(e.g. 0.05 or 5%)".
- w3 (%):** A text box containing "0.05" with a hint "(e.g. 0.05 or 5%)".

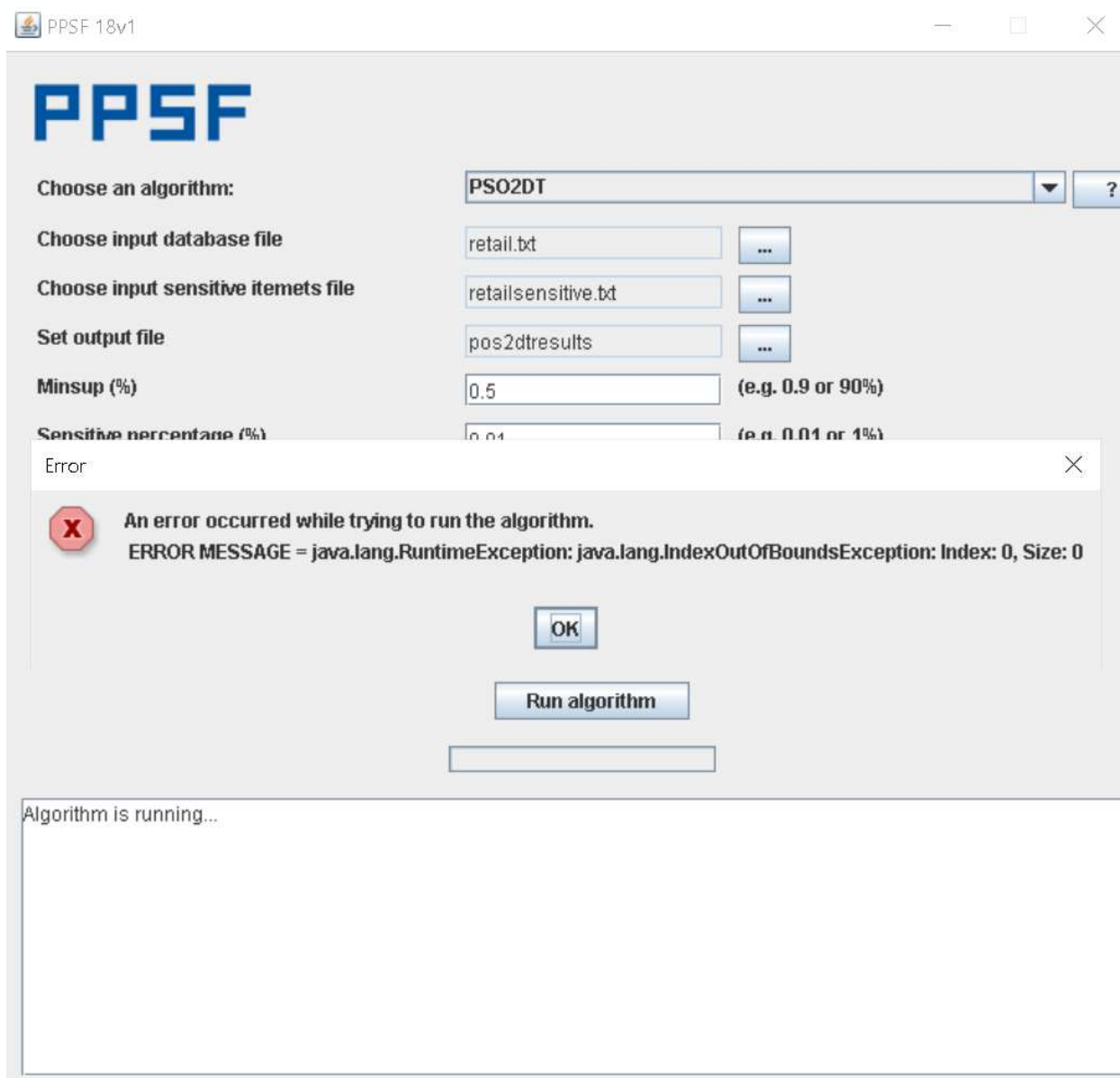
Below these settings is a "Run algorithm" button and an empty progress bar.

At the bottom, a status window displays the following text:

```
Algorithm is running...
===== Greedy - STATS =====
The fitness is : 0.0
There are 176322 transactions left in the database.
Total time ~ 2367300 ms
=====
```

**Figure 31**

*PPSF Tool Showing PSO2DT Algorithm Testing was Interrupted due to Index Out of Bound Error*



**Figure 32**

*Greedy Algorithm Testing for Customized Small Dataset*

PPSF 18v1

**PPSF**

Choose an algorithm: Greedy

Choose input database file: smallsize.txt

Choose input sensitive itemsets file: smallsensitive.txt

Set output file: smallgreedyresults

Minsup (%): 0.9 (e.g. 0.9 or 90%)

Sensitive percentage (%): 0.01 (e.g. 0.01 or 1%)

w1 (%): 0.9 (e.g. 0.9 or 90%)

w2 (%): 0.05 (e.g. 0.05 or 5%)

w3 (%): 0.05 (e.g. 0.05 or 5%)

Run algorithm

Algorithm is running...

===== Greedy - STATS =====

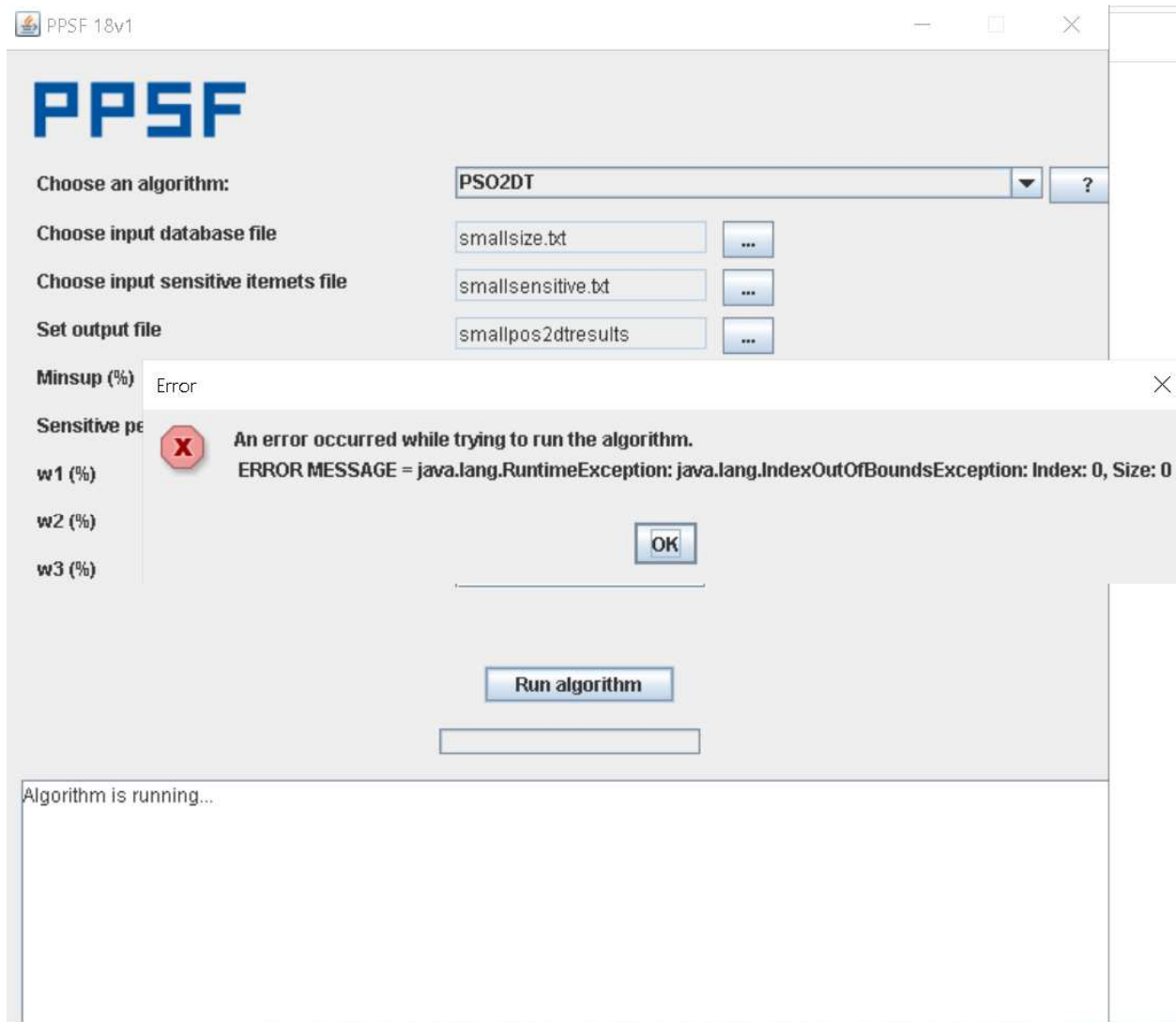
The fitness is : 0.0

There are 38 transactions left in the database.

Total time ~ 27 ms

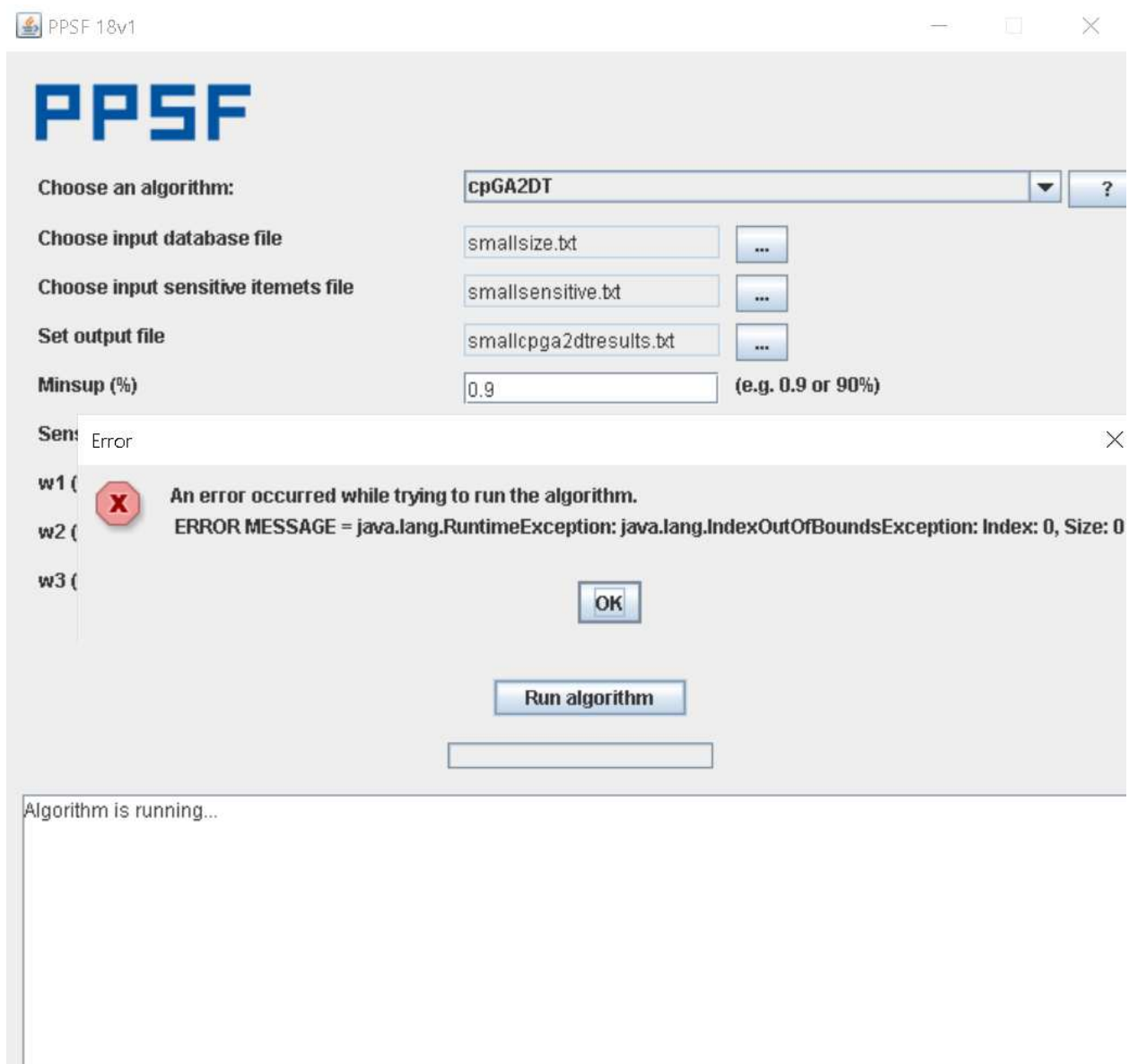
**Figure 33**

*PSO2DT Algorithm Testing for Customized Small Dataset and Index Out of Bound Error Encountered*



**Figure 34**

*cpGA2DT Algorithm Testing for Customized Small Dataset and Index Out of Bound Error Encountered*



**Figure 35**

*pGA2DT Algorithm Testing for Customized Small Dataset with After Execution Stats*

PPSF 18v1

**PPSF**

Choose an algorithm: **pGA2DT** ?

Choose input database file: smallsize.txt ...

Choose input sensitive itemsets file: smallsensitive.txt ...

Set output file: smallpga2dresults.txt ...

Minsup (%): 0.9 (e.g. 0.9 or 90%)

Sensitive percentage (%): 0.01 (e.g. 0.01 or 1%)

w1 (%): 0.9 (e.g. 0.9 or 90%)

w2 (%): 0.05 (e.g. 0.05 or 5%)

w3 (%): 0.05 (e.g. 0.05 or 5%)

**Run algorithm**

Algorithm is running...

Algorithm finished.

===== pGA2DT - STATS =====

The best solution to delete is : []

The best fitness is : 0.0

There are 4 transactions left in the database.

Total time ~ 0 s

=====

**Figure 36**

*sGA2DT Algorithm Testing for Customized Small Dataset with After Execution Stats*

PPSF 18v1

**PPSF**

Choose an algorithm: **sGA2DT** ?

Choose input database file: smallsize.bt ...

Choose input sensitive itemets file: smallsensitive.bt ...

Set output file: smallsga2dtresults ...

Minsup (%) : 0.9 (e.g. 0.9 or 90%)

Sensitive percentage (%) : 0.01 (e.g. 0.01 or 1%)

w1 (%) : 0.9 (e.g. 0.9 or 90%)

w2 (%) : 0.05 (e.g. 0.05 or 5%)

w3 (%) : 0.05 (e.g. 0.05 or 5%)

**Run algorithm**

Algorithm is running...

Algorithm finished.

===== sGA2DT - STATS =====

Final iteration time 141 ms

The best solution to delete is : []

The best fitness is : 0.0

There are 4 transactions left in the database.

Total time ~ 0 s

=====

### ***Phase Four***

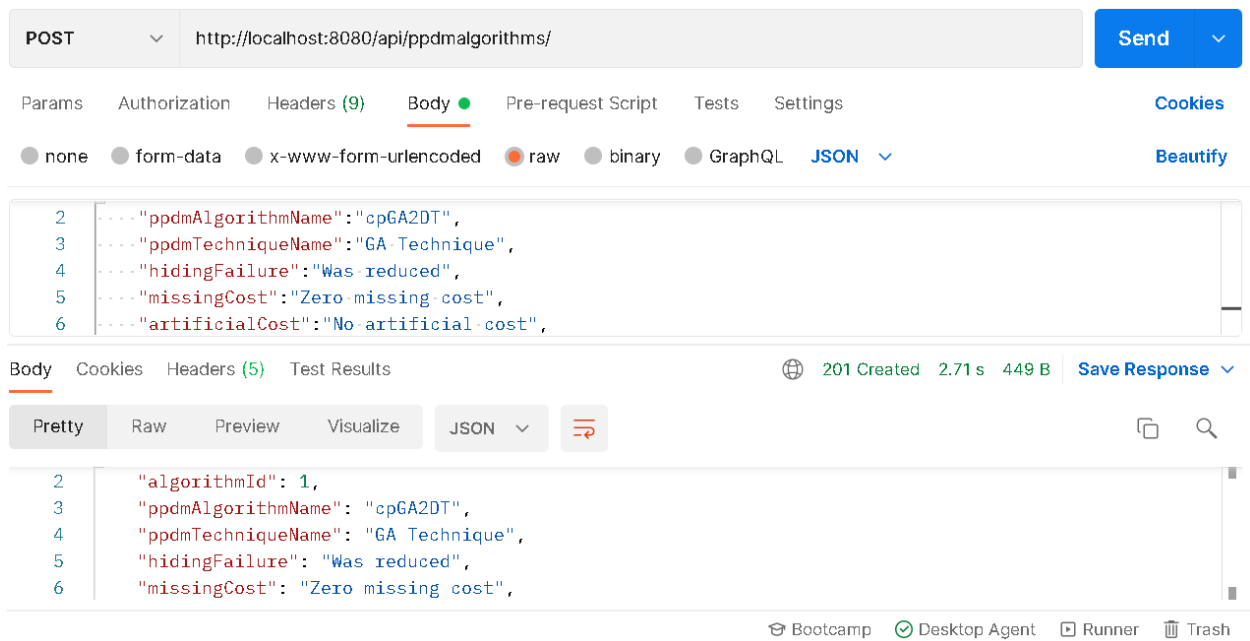
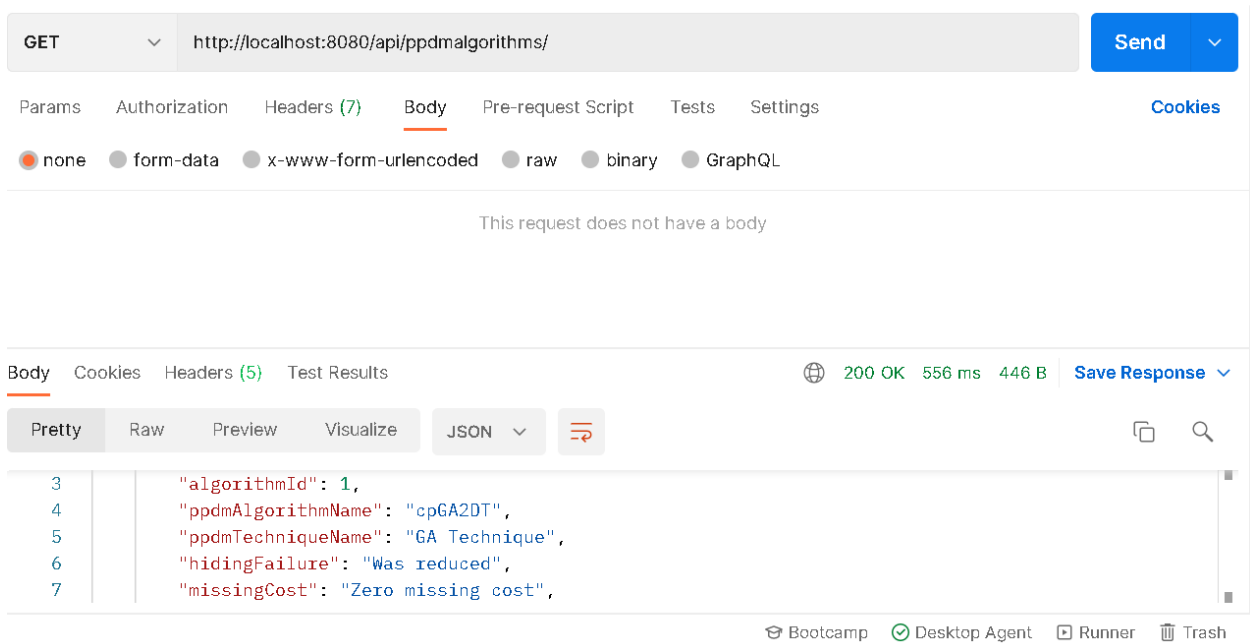
Results of REST endpoint testing are shown in Figures 34 through Figure 45. Details of the results are shown in Table 15. All the REST endpoints worked successfully. REST endpoint code is implemented before the actual user interface code is implemented. This approach was followed to make sure the backend core functionality of creating, displaying, updating, and deleting a PPDM algorithm is successful.

**Table 15**

#### *REST Endpoint Results*

<b>REST Endpoint</b>	<b>Description</b>	<b>Results</b>
/api/ppdmalgorithms/	Get all PPDM algorithms	Success
/api/ppdmalgorithms/{algorithmid}	Get a PPDM algorithm by id	Success
/api/ppdmalgorithms/	Create new PPDM algorithm	Success
/api/ppdmalgorithms/{algorithmid}	Update a PPDM algorithm	Success
/api/ppdm/	Delete an algorithm by id	Success



**Figure 37***REST Endpoint Created New PPDM algorithm Successfully***Figure 38***REST Endpoint Successfully Displaying all PPDM Algorithms*

**Figure 39***REST Endpoint Successfully Updated cpGA2DT PPDM Algorithm*

The screenshot shows a REST client interface with the following details:

- Method:** PUT
- URL:** `http://localhost:8080/api/ppdmalgorithms/cpGA2DT`
- Body:** A JSON object with the following fields:
 

```

{
  "ppdmAlgorithmName": "cpGA2DT",
  "ppdmTechniqueName": "Genetic Algorithm Technique",
  "hidingFailure": "Was reduced",
  "missingCost": "Zero missing cost",
  "artificialCost": "No artificial cost"
}
      
```
- Response:** 200 OK, 303 ms, 459 B. The response body is a JSON object:
 

```

{
  "algorithmId": 1,
  "ppdmAlgorithmName": "cpGA2DT",
  "ppdmTechniqueName": "Genetic Algorithm Technique",
  "hidingFailure": "Was reduced",
  "missingCost": "Zero missing cost"
}
      
```

**Figure 40***REST Endpoint Successfully Requesting to Delete cpGA2DT PPDM Algorithm by id "1"*

The screenshot shows a REST client interface with the following details:

- Method:** DELETE
- URL:** `http://localhost:8080/api/ppdmalgorithms/1`
- Body:** A JSON object with the following fields:
 

```

{
  "ppdmAlgorithmName": "cpGA2DT",
  "ppdmTechniqueName": "Genetic Algorithm Technique",
  "hidingFailure": "Was reduced",
  "missingCost": "Zero missing cost",
  "artificialCost": "No artificial cost"
}
      
```
- Response:** 204 No Content, 150 ms, 112 B. The response body is empty.

**Figure 41**

*REST Endpoint Successfully Deleted cpGA2DT PPDM Algorithm*

The screenshot shows a REST client interface with the following details:

- Method:** GET (dropdown menu)
- URL:** http://localhost:8080/api/ppdmalgorithms/cpGA2DT
- Buttons:** Send, Cookies
- Tabs:** Params, Authorization, Headers (7), Body, Pre-request Script, Tests, Settings
- Query Params:** A table with columns KEY, VALUE, DESCRIPTION, and Bulk Edit. The first row shows 'Key' and 'Value'.
- Body:** Pretty, Raw, Preview, Visualize, JSON (dropdown), and a copy icon.
- Response:** 404 Not Found, 14 ms, 420 B. The JSON body is:
 

```

      {
        "dataDissimilarity": null,
        "otherSideEffects": null,
        "email": null,
        "errorMessage": "Algorithm with name cpGA2DT not found"
      }
      
```
- Footer:** Bootcamp, Desktop Agent, Runner, Trash

As the REST endpoint implementation was successful, a decision was made to proceed with the view or user interface implementation. The main page of the PPDM algorithm repository is shown in Figure. Instructions are given on the main page to navigate to “Report Algorithm”, “Update Algorithm” or “List Algorithms”.

**Figure 42**

*The Main Page of PPDM Algorithm Repository*

The screenshot shows a web browser window with the URL localhost:8080/#/main. The page title is "Privacy Preserving Data Mining Algorithms Repository". The navigation bar includes buttons for "Main", "Report Algorithm", and "List Algorithms". The main content area contains the following instructions:

- Please select **Report Algorithm** to report an existing or newly developed PPDM algorithm.
- To update an already existing algorithm please navigate to **List Algorithms**.
- Please select **List Algorithms** to view all the algorithms reported.

Report Algorithm Page of PPDM Algorithm Repository

**Figure 45**

*Report Algorithm Page with Details Before Clicking Report Button*

Privacy Preserving Data Mining Algorithms Repository

Main Report Algorithm List Algorithms

**Report Algorithm**  
Please report new or already existing algorithm

Algorithm: pGA2DT

Technique: Genetic Algorithm

Hiding Failure: Reduced

Missing Cost: Not reduced

Artificial Cost: Not reduced

Data Dissimilarity: Dissimilarity robustness same as sGA2DT

Other Side Effects: None

Email: hs664@mynsu.nova.edu

Report Refresh

**Figure 46**

*List Algorithms page with Details Showing PPDM Algorithms Reported*

Privacy Preserving Data Mining Algorithms Repository

Main Report Algorithm List Algorithms

**List Algorithms**

Algorithm	Technique	Hiding Failure	Missing Cost	Artificial Cost	Data Dissimilarity	Other Side Effects	Email	Edit	Delete
pGA2DT	Genetic Algorithm	Reduced	Not reduced	Not reduced	Dissimilarity robustness same as sGA2DT	None	hs664@mynsu.nova.edu	Edit	Delete
sGA2DT	Genetic Algorithm	Reduced	Not reduced	Not reduced	Dissimilarity robustness same as pGA2DT	None	hs664@mynsu.nova.edu	Edit	Delete

**Figure 47**

*Edit in List Algorithms Page Redirects to Update Algorithms Page*

Privacy Preserving Data Mining Algorithms Repository

Main Report Algorithm List Algorithms

### Update Reported Algorithm

Algorithm name is unique

Algorithm	pGA2DT
Technique	Genetic Algorithm
Hiding Failure	Reduced
Missing Cost	Not reduced
Artificial Cost	Not reduced
Data Dissimilarity	Dissimilarity robustness same as sGA2DT
Other Side Effects	None
Email	hs664@mynsu.nova.edu

Update Refresh

**Figure 48**

*Email is Successfully Updated as Shown in List Algorithms Page*

Privacy Preserving Data Mining Algorithms Repository

Main Report Algorithm List Algorithms

### List Algorithms

Algorithm	Technique	Hiding Failure	Missing Cost	Artificial Cost	Data Dissimilarity	Other Side Effects	Email	Edit	Delete
pGA2DT	Genetic Algorithm	Reduced	Not reduced	Not reduced	Dissimilarity robustness same as sGA2DT	None	ln500@mynsu.nova.edu	Edit	Delete
sGA2DT	Genetic Algorithm	Reduced	Not reduced	Not reduced	Dissimilarity robustness same as pGA2DT	None	hs664@mynsu.nova.edu	Edit	Delete

## Summary

To summarize, the results section had data collection, data analysis, and findings for all four phases. The results contributed knowledge for full stack web development, information privacy, PPDM, PPDM algorithms, and their side effects. Different processes and tools were used for data collection. The processes were literature review and SLR. Tools used for analyzing the data were Microsoft Excel and PRISMA. Phase one had 51 articles for final review. Phase two had 45 articles for final review. Phase three used the PPSF tool to analyze the PPDM algorithms. Phase four involved java coding details for implementing the front end and backend of the web application.

Phase one's findings gave an overview of the data sanitization process. The overview was necessary to understand at which stage the PPDM algorithm was applied, when side effects were generated, and when side effects were verified. The characteristics and differences of the side effects were explained in phase one. Phase two results involved information about non-traditional side effects. Phase three results showed detailed steps for testing PPDM algorithms using the PPSF tool. This phase also included manually creating a small database to test all PPDM algorithms in the PPSF tool. The results generated for the large database were not satisfactory. Hence, the manual creation of the database was necessary. Finally, phase four showed the successfully created PPDM web repository. The web repository consisted of the main page, reporting an algorithm, and listing all algorithms.

## Chapter 5

### Conclusions, Implications, Recommendations and Summary

#### Conclusions

This study achieved the desired result by creating a common PPDM web repository for reporting the side effects. The DSR methodology helped to accomplish the goals designed. The final goal was completed by creating the web application. This was possible by both data collection and data analysis. As the information related to the known and unknown side effects had to be explored, the study was conducted in four phases. In each phase data analysis and reporting of findings were conducted independently. As part of data analysis, PRISMA was used to conduct the stage four (analyze the findings) of systematic review. Phase one and two applied systematic review, hence PRISMA was used during these phases. Phase one's data analysis finalized the articles to report and explain the results. The results achieved in phases one and two were successful. All the expected results stated in the methodology section were accomplished. The characteristics of the common side effect HF, MC, and AC were understood. Each of the side effects characteristics were related to sensitive data, non-sensitive data, and newly generated data. The relationship of side effects, itemsets, and mining rules were also determined. Additionally, it was observed that each of the side effects have alternative names. Finally, the phase one report was concluded by briefing the expected results questions for common side effects HF, MC, and AC. The results for phase one are shown in Tables 8-10.

Similarly, phase two's data analysis finalized a total of 45 research articles related to non-traditional side effects. Additional non-traditional side effect names identified were hidden rule/s, lost rule/s, new rule/s, sensitive rule/s, mined rule/s, artificial rule/s, non-sensitive rules, ghost rules, and spurious rules. Phase one, two, and four of this study accomplished the expected



results except for phase three. The PPSF tool's results in phase three did not provide satisfactory results.

The strengths of this study are clear understanding of side effects of PPDM algorithms. HF,MC,AC, data dissimilarity, and all the side effects' names ending with the term "rule/s" were explored. For example, hidden rules, ghost rules, spurious rules etc. The web repository is a valuable knowledge base for collecting the details of the PPDM algorithms' side effects. This will give an opportunity to researchers in selecting an appropriate algorithm and to start, extend, or expand a research work related to PPDM. Researchers can concentrate on improving the PPDM algorithms rather than spending time to search for the data. The web application provides a repository as a reference for all the PPDM algorithms. In addition, the web application allows to report new PPDM algorithms developed, or any issues observed in existing ones. Another strength of this study lies in the full stack implementation of the Spring web application. The latest technologies used in this study help researchers and web developers to implement similar applications. The source code developed in this study is an open source and is available in GitHub repository. The location for GitHub is <https://github.com/himaait/ppdmalgorithm>

### **Implications**

This study presents an online repository to only report the PPDM algorithms and their side effects. It also helps to understand the relationship and variations between the side effects. For example, if one PPDM algorithm reduced one side effect, the other side effects would increase or have no impact. For example, few research studies implied the impact of HF over MC and vice versa. Other studies explained the dependency of MC and AC to measure the data quality.

Although PPDM algorithms were implemented differently, their final goal was to protect sensitive information and minimize the side effects. In this study, a total of 96 articles were thoroughly reviewed to understand the PPDM algorithms, their contribution to resolving the side effects, the side effects' characteristics, and the side effects' identity with the data/itemset. The process of sanitization was explored to understand the application of PPDM algorithms. Further, various PPDM algorithms were analyzed to understand the occurrences of these side effects. The phase at which these side effects are identified was understood. Hence, this study provided a comparison for all the PPDM algorithms based on the side effects resolved. There is scope for researchers to contribute by providing more detailed instructions to test each of these algorithms with small datasets. The small datasets would be ideal to understand the side effects resolved clearly. These future implementations can be applied to remaining PPSF algorithms as well.

PPSF are developing more algorithms. Hence, this study can be used to provide better documentation for all the new algorithms developed. They can further integrate and develop this study's web repository on their official PPSF website. This study implemented a simple reporting web repository. The features included reporting a new PPDM algorithm, updating an existing algorithm, deleting, and viewing all algorithms. Web developers can develop this web repository into a discussion forum. This study's web repository can be converted to an official forum for PPDM algorithms.

### **Limitations**

This study contributed to PPDM algorithms and their corresponding side effects. The side effects information is based on the previous research studies conducted. However, there were two limitations observed while exploring these PPDM algorithms. There is no documentation instructing the process to test the side effects observed for each of the PPDM algorithms. Lack of

documentation for the PPSF tool was the main limitation for this study. Additionally, except for the PPSF tool, there is lack of open-source tools to test other PPDM algorithms. Another limitation was related to the web repository. Due to time constraints, this study did not implement a login security feature.

## **Recommendations**

The initial step is to work with the PPSF team to extend the present study to better understand the source code implementation of the six algorithms. Further, enhance or implement PPDM algorithms to find unknown side effects. In parallel, it is very important to create proper documentation for the PPSF tool. The SPMF tool is a classic example to refer to for creating a step-by-step documentation of the PPDM algorithms. SPMF documentation for 256 algorithms can be found at [www.philippe-fournier-viger.com](http://www.philippe-fournier-viger.com). Collaborating with Jerry Li's team to implement a similar concept (SPMF documentation) for the PPSF tool is highly recommended. The documentation implementation in the PPSF tool not only contributes to PPDM algorithms, but also includes many other algorithms. The documentation would add valuable knowledge for PPDM and PPSF researchers. The PPDM web repository developed in this study can be integrated as part of PPSF tool and its website as a future work. The extension of the present study is to develop a web repository for other algorithms of the PPSF tool. This will include instructions to test the algorithms and also reporting the success or failure of the algorithms. This study allows reporting a PPDM algorithm with email as mandatory. As a future enhancement, verifying if the email entered is valid can be implemented by sending "one time password" to the email. Another future enhancement would also include not displaying email publicly, and admin login.

## **Summary**

Even before identifying the problem, this study started with understanding the roots of PPDM and PPDM algorithms' origin. How the development of big data, knowledge-mining processes, and data sharing lead to protecting confidential/sensitive information, information privacy, preservation of privacy, privacy-preserving techniques, and sanitization. The privacy problems arising during these processes made PPDM and PPDM algorithms to come into existence. The initial step of the study involved identifying the problem related to PPDM, PPDM algorithms, and side effects. The three research gaps identified were: exploring the known and unknown side effects, comparing the side effects of PPDM algorithms, and an online web repository to report side effects of PPDM algorithms. Understanding the concepts of PPDM algorithms and their relationship with side effects led to recognize research questions as well. The problem statement helped to finalize five research questions. Research question one was, what were the similarities and differences of the existing side effects of PPDM algorithms? Research question two was, how were the side effects related to one another? Research question three was, what were the non-traditional side effects, and do they occur in PPDM algorithms? Research question four was, what were the unknown side effects occurring in PPDM algorithms? Research question five was, where and how were the side effects of all PPDM algorithms reported? The goal was to create an online web repository to report PPDM algorithms and the side effects. In order to achieve this goal various approaches were incorporated as phases. The literature review involved understanding the previous research studies related to PPDM algorithms and the side effects. This process gave sufficient evidence about the PPDM algorithms developed to solve the side effects. Most of the studies discussed the common side effects such as hiding failure, missing cost, and artificial cost. In certain studies data similarity was also considered as one of the side effects. Data similarity was not frequently studied as

compared to common side effects. The initial literature review process also revealed the fact that most of the PPDM algorithms were developed to measure the data quality, performance, data accuracy, and execution time. Hence, side effects were used to measure the desired performance of the developed algorithms. A deeper literature review process called systematic review process revealed more information about the side effects.

Research methodology implemented for this study was DSR. The DSR methodology with instantiation artifact type was very much required as this study was in the category of socio-technical artifact. A total of four phases were designed based on the research questions defined. Two research processes called literature review and SLR were used to get desired results for RQ1 through RQ4. RQ5 used search engines such as google search and google scholar. These search engines helped to find that PPDM algorithm reporting website does not exist. This confirmation allowed to proceed with website implementation.

Data analysis and findings for all the four phases were documented. Data collection was from selected databases such as IEEE, Elsevier ScienceDirect, ACM Digital Library, ProQuest, SpringerLink, Google Scholar, EBSCOhost databases, and JSTOR. There were few databases which were not considered due to the redundant search results. These databases not considered in phase two were ACM and SpringerLink. Microsoft excel was used to sort and clean the search results. Jabref tool was also helpful to convert BibTeX file to csv file format. PRISMA's flowchart provided an easy understanding and track the data analysis. Barriers as identified in the initial phase lack online documentation for PPSF tool. The results for phases one, two, and three were as expected. Only for phase three the expected results were not obtained.

## References

- Abdar, M., Kalhori, S. R. N., Sutikno, T., Subroto, I. M. I., & Arji, G. (2015). Comparing performance of data mining algorithms in prediction heart diseases. *International Journal of Electrical & Computer Engineering* (2088-8708), 5(6).
- Aggarwal, C. C., & Philip, S. Y. (2008). A general survey of privacy preserving data mining models and algorithms. In C. C. Aggarwal & P. S. Yu (Eds.), *Privacy preserving data mining* (pp. 11-52). Springer.
- Aghasi, M., & Oskouei, R. J. (2014, July). Privacy preserving data mining survey of classifications. In *International Workshop Soft Computing Applications* (pp. 637-649). Springer, Cham.
- Aldeen, Y. A. A. S., Salleh, M., & Razzaque, M. A. (2015). A comprehensive review on privacy preserving data mining. *SpringerPlus*, 4(1), 694.
- Aldeen, Y. A. A. S., Salleh, M., & Razzaque, M. A. (2015). A comprehensive review on privacy preserving data mining. *SpringerPlus*, 4(1), 1-36.
- Arboleda, E. R. (2019). Comparing performances of data mining algorithms for classification of green coffee beans. *Int. J. Eng. Adv. Technol*, 8(5), 1563-1567.
- Atlam, H. F., Azad, M. A., Alassafi, M. O., Alshdadi, A. A., & Alenezi, A. (2020). Risk-based access control model: A systematic literature review. *Future Internet*, 12(6), 103.
- Bélanger, F., & Crossler, R. E. (2011). Privacy in the digital age: A review of information privacy research in information systems. *MIS Quarterly*, 35(4), 1017-1041.
- Bertino, E., Fovino, I. N., & Provenza, L. P. (2005). A framework for evaluating privacy preserving data mining algorithms. *Data Mining and Knowledge Discovery*, 11(2), 121-154.

- Bertino, E., Lin, D., & Jiang, W. (2008). A survey of quantification of privacy preserving data mining algorithms. In *Privacy-Preserving Data Mining* (pp. 183-205). Springer, Boston, MA.
- Bhagat, B., & Shelke, Suchitra. (2015). Techniques for privacy preservation in data mining. *International Journal of Engineering Research & Technology*, 4(10).
- Borgo, S., Franssen, M., Garbacz, P., Kitamura, Y., Mizoguchi, R., & Vermaas, P. E. (2014). Technical artifacts: An integrated perspective. *Applied Ontology*, 9(3-4), 217-235.
- Brown, M. L., & Kros, J. F. (2003). Data mining and the impact of missing data. *Industrial Management & Data Systems*, 103(8), 611-621.
- Celik, S., Eydurán, E., Karadas, K., & Tariq, M. M. (2017). Comparison of predictive performance of data mining algorithms in predicting body weight in Mengali rams of Pakistan. *Revista Brasileira de Zootecnia*, 46, 863-872.
- Cleven, A., Gubler, P., & Hüner, K. M. (2009). Design alternatives for the evaluation of design science research artifacts. In *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology* (pp. 1-8).
- Conlon, C., Timonen, V., Elliott-O'Dare, C., O'Keeffe, S., & Foley, G. (2020). Confused about theoretical sampling? Engaging theoretical sampling in diverse grounded theory studies. *Qualitative Health Research*, 30(6), 947-959.
- Donalds, C., & Osei-Bryson, K. M. (2019). Toward a cybercrime classification ontology: A knowledge-based approach. *Computers in Human Behavior*, 92, 403-418.
- Drechsler, A., & Dörr, P. (2014). What kinds of artifacts are we designing? An analysis of artifact types and artifact relevance in IS journal publications. In *International*

- Conference on Design Science Research in Information Systems* (pp. 329-336). Springer, Cham.
- Ergenç Bostanoğlu, B., & Öztürk, A. C. (2020). Minimizing information loss in shared data: Hiding frequent patterns with multiple sensitive support thresholds. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 13(4), 309-323.
- Fernandes, M., & Gomes, J. (2017, April). Heuristic approach for association rule hiding using ECLAT. In *2017 2nd International conference on communication systems, computing and IT applications (CSCITA)* (pp. 218-223). IEEE.
- Fournier-Viger, P., Gomariz, A., Gueniche, T., Soltani, A., Wu, C. W., & Tseng, V. S. (2014). SPMF: A java open-source pattern mining library. *Journal of Machine Learning Research*, 15(1), 3389-3393.
- Fournier-Viger, P., Lin, J. C. W., Gomariz, A., Gueniche, T., Soltani, A., Deng, Z., & Lam, H. T. (2016). The SPMF open-source data mining library version 2. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 36-40).
- Gan, W., Chun-Wei, J., Chao, H. C., Wang, S. L., & Philip, S. Y. (2018). Privacy preserving utility mining: A survey. *2018 IEEE International Conference on Big Data (Big Data)* (pp. 2617-2626).
- Gayathiri, P., & Poorna, B. (2015). Association rule hiding techniques for privacy preserving data mining: A study. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 6(12).
- Geerts, G. L. (2011). A design science research methodology and its application to accounting information systems research. *International journal of accounting Information Systems*, 12(2), 142-151.



- Gerede, C. E., & Su, J. (2007). Specification and verification of artifact behaviors in business process models. In *International conference on service-oriented computing* (pp. 181-192). Springer, Berlin, Heidelberg.
- Ghalehsefidi, N. J., & Dehkordi, M. N. (2016). A hybrid algorithm based on heuristic method to preserve privacy in association rule mining. *Indian Journal of Science and Technology*, 9(27), 1-10.
- Gnewuch, U., Morana, S., & Maedche, A. (2017). Towards designing cooperative and social conversational agents for customer service. In: *Proceedings of the 38th International Conference on Information Systems (ICIS)*. 1-13
- Gregor, S., & Iivari, J. (2007). Designing for mutability in information systems artifacts. *Information systems foundations. Australian National University Press, Canberra, Australia*, 3-24.
- Gurevich, A., & Gudes, E. (2006, December). Privacy preserving data mining algorithms without the use of secure computation or perturbation. In *2006 10th International Database Engineering and Applications Symposium (IDEAS'06)* (pp. 121-128). IEEE.
- Herselman, M., & Botha, A. (2015, September). Evaluating an artifact in design science research. In *Proceedings of the 2015 Annual Research Conference on South African Institute of Computer Scientists and Information Technologists* (pp. 1-10).
- Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 75-105.
- Hussain, S., Muhammad, L. J., Ishaq, F. S., Yakubu, A., & Mohammed, I. A. (2019). Performance evaluation of various data mining algorithms on road traffic accident

- dataset. In *Information and Communication Technology for Intelligent Systems* (pp. 67-78). Springer, Singapore.
- Jangra, S., & Toshniwal, D. (2020). VIDPSO: victim item deletion based PSO inspired sensitive pattern hiding algorithm for dense datasets. *Information Processing & Management*, 57(5), 102255.
- Kagklis, V., Verykios, V. S., Tzimas, G., & Tsakalidis, A. K. (2014, June). Knowledge sanitization on the web. In *Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14)* (pp. 1-11).
- Kamakshi, P., & Babu, A. V. (2012, December). Automatic detection of sensitive attribute in PPDM. *2012 IEEE international conference on computational intelligence and computing research* (pp. 1-5). IEEE.
- Kamakshi, P., & Vinaya Babu, A. (2011, November). Framework to reduce the hiding failure due to randomized additive data modification PPDM technique. In *International Conference on Computational Intelligence and Information Technology* (pp. 65-71). Springer, Berlin, Heidelberg.
- Kitchenham, B. A. (2007). Guidelines for performing systematic literature reviews in software engineering. *Keele University and Durham University Joint Report*, 5, 1-57.
- Kitchenham, B. A. (2012, September). Systematic review in software engineering: Where we are and where we should be going. In *Proceedings of the 2nd international workshop on Evidential assessment of software technologies* (pp. 1-2).
- Kriegel, H. P., Borgwardt, K. M., Kröger, P., Pryakhin, A., Schubert, M., & Zimek, A. (2007). Future trends in data mining. *Data Mining and Knowledge Discovery*, 15(1), 87-97.

- Kuckartz, U. (2014). *Qualitative text analysis: A guide to methods, practice and using software*. Sage.
- Kuechler, B., & Vaishnavi, V. (2008). On theory development in design science research: Anatomy of a research project. *European Journal of Information Systems*, 17(5), 489-504.
- Kuo, Y. P., Lin, P. Y., & Dai, B. R. (2008, September). Hiding frequent patterns under multiple sensitive thresholds. In *International Conference on Database and Expert Systems Applications* (pp. 5-18). Springer, Berlin, Heidelberg.
- Laskar, D., & Lachit, G. (2014). A review on privacy preservation data mining (PPDM). *International Journal of Computer Applications Technology and Research*, 3(7), 403-408.
- Lee, G., Chen, Y. C., Peng, S. L., & Lin, J. H. (2011, September). Solving the sensitive itemset hiding problem whilst minimizing side effects on a sanitized database. In *International Conference on Security-Enriched Urban Computing and Smart Grid* (pp. 104-113). Springer, Berlin, Heidelberg.
- Levy, Y., & Ellis, T. J. (2006). A systems approach to conduct an effective literature review in support of information systems research. *Informing Science*, 9, 181-212.
- Li, L., Lu, R., Choo, K. K. R., Datta, A., & Shao, J. (2016). Privacy preserving outsourced association rule mining on vertically partitioned databases. *IEEE Transactions on Information Forensics and Security*, 11(8), 1847-1861.
- Li, R., de Vries, D., & Roddick, J. (2011, December). Bands of privacy preserving objectives: classification of PPDM strategies. In *Proceedings of the Ninth Australasian Data Mining Conference-Volume 121* (pp. 137-152).

- Li, R., de Vries, D., & Roddick, J. (2011, December). Bands of privacy preserving objectives: Classification of PPDM strategies. In *Proceedings of the Ninth Australasian Data Mining Conference-Volume 121* (pp. 137-152).
- Li, S., Mu, N., Le, J., & Liao, X. (2019). A novel algorithm for privacy preserving utility mining based on integer linear programming. *Engineering Applications of Artificial Intelligence*, 81, 300-312.
- Lin, C. W., Hong, T. P., & Hsu, H. C. (2014). Hiding Sensitive Itemsets with Minimal Side Effects in Privacy Preserving Data Mining. In *Intelligent Data analysis and its Applications, Volume I* (pp. 87-95). Springer, Cham.
- Lin, C. W., Hong, T. P., & Hsu, H. C. (2014). Reducing side effects of hiding sensitive itemsets in privacy preserving data mining. *The Scientific World Journal*, 1-12.
- Lin, C. W., Hong, T. P., Chang, C. C., & Wang, S. L. (2013). A greedy-based approach for hiding sensitive itemsets by transaction insertion. *Journal of Information Hiding and Multimedia Signal Processing*, 4(4), 201-227.
- Lin, C. W., Zhang, B., Yang, K. T., & Hong, T. P. (2014a). Efficiently hiding sensitive itemsets with transaction deletion based on genetic algorithms. *The Scientific World Journal*, 2014 (398269), 1-13.
- Lin, J. C. W., Fournier-Viger, P., Wu, L., Gan, W., Djenouri, Y., & Zhang, J. (2018d). PPSF: An open-source privacy preserving and security mining framework. *2018 IEEE International Conference on Data Mining Workshops (ICDMW)* (pp. 1459-1463).
- Lin, J. C. W., Gan, W., Fournier-Viger, P., Yang, L., Liu, Q., Frnda, J., Sevcik, L., & Voznak, M. (2016a). High utility-itemset mining and privacy preserving utility mining. *Perspectives in Science*, 7, 74-80.

- Lin, J. C. W., Hong, T. P., Fournier-Viger, P., Liu, Q., Wong, J. W., & Zhan, J. (2017). Efficient hiding of confidential high-utility itemsets with minimal side effects. *Journal of Experimental & Theoretical Artificial Intelligence*, 29(6), 1225-1245.
- Lin, J. C. W., Liu, Q., Fournier-Viger, P., Hong, T. P., Voznak, M., & Zhan, J. (2016b). A sanitization approach for hiding sensitive itemsets based on particle swarm optimization. *Engineering Applications of Artificial Intelligence*, 53, 1-18.
- Lin, J. C. W., Wu, J. M. T., Fournier-Viger, P., Djenouri, Y., Chen, C. H., & Zhang, Y. (2019). A sanitization approach to secure shared data in an IoT environment. *IEEE Access*, 7, 25359-25368.
- Lin, J. C. W., Wu, T. Y., Fournier-Viger, P., Lin, G., Hong, T. P., & Pan, J. S. (2015, August). A sanitization approach of privacy preserving utility mining. In *International Conference on Genetic and Evolutionary Computing* (pp. 47-57). Springer, Cham.
- Lin, J. C. W., Wu, T. Y., Fournier-Viger, P., Lin, G., Zhan, J., & Voznak, M. (2016c). Fast algorithms for hiding sensitive high-utility itemsets in privacy preserving utility mining. *Engineering Applications of Artificial Intelligence*, 55, 269-284.
- Lin, J. C. W., Wu, T. Y., Fournier-Viger, P., Lin, G., Zhan, J., & Voznak, M. (2016). Fast algorithms for hiding sensitive high-utility itemsets in privacy-preserving utility mining. *Engineering Applications of Artificial Intelligence*, 55, 269-284.
- Lin, J. C. W., Zhang, Y., Zhang, B., Fournier-Viger, P., & Djenouri, Y. (2019). Hiding sensitive itemsets with multiple objective optimization. *Soft Computing*, 23(23), 12779-12797.
- Liu, X., Chen, G., Wen, S., & Song, G. (2020). An improved sanitization algorithm in privacy preserving utility mining. *Mathematical Problems in Engineering*, 1-14.

- Liu, X., Wen, S., & Zuo, W. (2020). Effective sanitization approaches to protect sensitive knowledge in high-utility itemset mining. *Applied Intelligence*, 50(1), 169-191.
- Lukyanenko, R., Evermann, J., & Parsons, J. (2015). Guidelines for establishing instantiation validity in IT artifacts: A survey of IS research. In *International Conference on Design Science Research in Information Systems* (pp. 430-438). Springer, Cham.
- Malik, M. B., Ghazi, M. A., & Ali, R. (2012, November). Privacy preserving data mining techniques: Current scenario and future prospects. In *2012 Third International Conference on Computer and Communication Technology* (pp. 26-32). IEEE.
- Mandapati, S., Bhogopathi, R. B., & Chekka, R. B. (2013). A hybrid algorithm for privacy preserving in data mining. *International Journal of Intelligent Systems and Applications*, 5(8), 47.
- Mendes, R., & Vilela, J. P. (2017). Privacy preserving data mining: Methods, metrics, and applications. *IEEE Access*, 5, 10562-10582.
- Menzies, T., Kocaguneli, E., Turhan, B., Minku, L., & Peters, F. (2014). *Sharing data and models in software engineering*. Morgan Kaufmann.
- Mizoguchi, R., Kitamura, Y., & Borgo, S. (2016). A unifying definition for artifact and biological functions. *Applied Ontology*, 11(2), 129-154.
- Mogtaba, S., & Kambal, E. (2016, July). Association rule hiding for privacy preserving data mining. In *Industrial Conference on Data Mining* (pp. 320-333). Springer, Cham.
- Murthy, T. S., Gopalan, N. P., & Gunturu, S. (2018). A novel optimization based algorithm to hide sensitive item-sets through sanitization approach. *International Journal of Modern Education and Computer Science (IJMECS)*, 10(10), 48-55.

- Naeem, M., Asghar, S., & Fong, S. (2010, November). Hiding sensitive association rules using central tendency. In *2010 6th International Conference on Advanced Information Management and Service (IMS)* (pp. 478-484). IEEE.
- Nanavati, N. R., & Jinwala, D. C. (2015). A novel privacy-preserving scheme for collaborative frequent itemset mining across vertically partitioned data. *Security and Communication Networks*, 8(18), 4407-4420.
- Navale, G. S., & Mali, S. N. (2019). A multi-analysis on privacy preservation of association rules using hybridized approach. *Evolutionary Intelligence*, 1-15.
- Navale, G. S., & Mali, S. N. (2019). Lossless and robust privacy preservation of association rules in data sanitization. *Cluster Computing*, 22(1), 1415-1428.
- Nguyen, H. H. (2018). Privacy-preserving mechanisms for k-modes clustering. *Computers & Security*, 78, 60-75.
- Nithya, S., Sangeetha, M., Prethi, K. A., & Vellingiri, S. (2021). Impact factor based data sanitization in association rule mining. *Materials Today: Proceedings*, 45, 2653-2659.
- Nopour, R., Kazemi-Arpanahi, H., Shanbehzadeh, M., & Azizifar, A. (2021). Performance analysis of data mining algorithms for diagnosing COVID-19. *Journal of Education and Health Promotion*, 10.
- Oppong-Tawiah, D., Webster, J., Staples, S., Cameron, A. F., de Guinea, A. O., & Hung, T. Y. (2020). Developing a gamified mobile application to encourage sustainable energy use in the office. *Journal of Business Research*, 106, 388-405.
- Patel, A. K. (2016). A survey: Privacy preservation data mining techniques and geometric transformation. *International Journal of Scientific Research in Science, Engineering, and Technology*, 2(2), 106-111.

- Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of management information systems*, 24(3), 45-77.
- Prat, N., Comyn-Wattiau, I., & Akoka, J. (2014). Artifact evaluation in information systems design science research a holistic view. *PACIS*, 23, 1-16.
- Priyadarsini, R. P., Valarmathi, M. L., & Sivakumari, S. (2016, March). Feature creation based slicing for privacy preserving data mining. In *Proceedings of the 3rd IKDD Conference on Data Science, 2016* (pp. 1-8).
- Qi, X., & Zong, M. (2012). An overview of privacy preserving data mining. *Procedia Environmental Sciences*, 12, 1341-1347.
- Rajesh, N., Sujatha, K., & Lawrence, A. A. (2016). Survey on privacy preserving data mining techniques using recent algorithms. *International Journal of Computer Applications*, 133(7), 30-33.
- Reibenspiess, V., Drechsler, K., Eckhardt, A., & Wagner, H. T. (2020). Tapping into the wealth of employees' ideas: Design principles for a digital intrapreneurship platform. *Information & Management*, 103287.
- Rong, H., Wang, H., Liu, J., Wu, W., & Xian, M. (2016, August). Efficient integrity verification of secure outsourced knn computation in cloud environments. In *2016 IEEE Trustcom/BigDataSE/ISPA* (pp. 236-243). IEEE.
- Selvan, P., & Veni, S. (2016). Swarm optimizer based on the sensitive rule hiding with the constraints minimization for the data publishing. *International Journal of Advanced Research in Computer Science*, 7(4).



- Shah, K., Thakkar, A., & Ganatra, A. (2012). A study on association rule hiding approaches. *International Journal of Engineering and Advanced Technology (IJEAT)*, 1, 72-76.
- Shah, K., Thakkar, A., & Ganatra, A. (2012). Association rule hiding by heuristic approach to reduce side effects & hide multiple RHS items. *International Journal of Computer Applications*, 45(1), 1-7.
- Shailaja, G. K., & Rao, C. V. (2019). Robust and lossless data privacy preservation: optimal key based data sanitization. *Evolutionary Intelligence*, 1-12.
- Sharma, M., Chaudhary, A., Mathuria, M., & Chaudhary, S. (2013). A review study on the privacy preserving data mining techniques and approaches. *International Journal of Computer Science and Telecommunications*, 4(9), 42-46.
- Shetty, J., Dash, D., Joish, A. K., & Guruprasad, C. (2020). Review paper on web frameworks, databases and web stacks. *International Research Journal of Engineering and Technology (IRJET)*, 5734-5738.
- Simon, H. A. (2019). *The sciences of the artificial*. MIT press.
- Smith, H. J., Dinev, T., & Xu, H. (2011). Information privacy research: An interdisciplinary review. *MIS quarterly*, 35(4), 989-1015.
- Soni, R. K. (2017). *Full stack angularJS for java developers: Build a full-featured web application from scratch using angularJS with spring RESTful*. Apress.
- Sowmya, Y., Ratna, M. N., & Bindu, C. S. (2015). A review on big data mining, distributed programming frameworks and privacy preserving data mining techniques. *International journal of advanced research in computer science*, 6(1).

- Sramka, M., Safavi-Naini, R., Denzinger, J., & Askari, M. (2010, March). A practice-oriented framework for measuring privacy and utility in data sanitization systems. In *Proceedings of the 2010 EDBT/ICDT Workshops* (pp. 1-10).
- Surendra, H., & Mohan, H. S. (2019). Hiding sensitive itemsets without side effects. *Applied Intelligence*, 49(4), 1213-1227.
- Tamil Selvan, P. & Veni, S. (2015). A survival study on privacy preservation of data sharing with optimal side effects. *International Journal of Engineering Research & Technology (IJERT)*, 4(6).
- Telikani, A., & Shahbahrami, A. (2018). Data sanitization in association rule mining: An analytical review. *Expert Systems with Applications*, 96, 406-426.
- Telikani, A., Gandomi, A. H., Shahbahrami, A., & Dehkordi, M. N. (2020). Privacy-preserving in association rule mining using an improved discrete binary artificial bee colony. *Expert Systems with Applications*, 144, 113097.
- Teng, Z., & Du, W. (2009). A hybrid multi-group approach for Privacy-Preserving Data Mining. *Knowledge and information systems*, 19(2), 133-157.
- Urabe, S., Wang, J., Kodama, E., & Takata, T. (2007). A high collusion-resistant approach to distributed Privacy-Preserving Data Mining. *Information and Media Technologies*, 2(3), 821-834.
- Vaghashia, H., & Ganatra, A. (2015). A survey: Privacy preservation techniques in data mining. *International Journal of Computer Applications*, 119(4).
- Vaidya, J., Shafiq, B., Fan, W., Mehmood, D., & Lorenzi, D. (2013). A random decision tree framework for Privacy-Preserving Data Mining. *IEEE transactions on dependable and secure computing*, 11(5), 399-411.

- Venable, J., Pries-Heje, J., & Baskerville, R. (2012). A comprehensive framework for evaluation in design science research. In *International conference on design science research in information systems* (pp. 423-438). Springer, Berlin, Heidelberg.
- Vihavainen, S., Lampinen, A., Oulasvirta, A., Silfverberg, S., & Lehmuskallio, A. (2013). The clash between privacy and automation in social media. *IEEE Pervasive Computing*, 13(1), 56-63.
- Wang, E. T., & Lee, G. (2008). An efficient sanitization algorithm for balancing information privacy and knowledge discovery in association patterns mining. *Data & Knowledge Engineering*, 65(3), 463-484.
- Wang, S. L., Parikh, B., & Jafari, A. (2007). Hiding informative association rule sets. *Expert Systems with Applications*, 33(2), 316-323.
- Wimmer, H., & Powell, L. (2014). A comparison of the effects of k-anonymity on machine learning algorithms. In *Proceedings of the Conference for Information Systems Applied Research ISSN* (Vol. 2167, p. 1508).
- Wu, C. M., & Huang, Y. F. (2011). A cost-efficient and versatile sanitizing algorithm by using a greedy approach. *Soft Computing*, 15(5), 939-952.
- Wu, J. M. T., Lin, C. W., Fournier-Viger, P., Djenouri, Y., Chen, C. H., & Li, Z. (2019). The density-based clustering method for Privacy-Preserving Data Mining. *Mathematical Biosciences and Engineering*, 16(3): 1718-1728
- Wu, J. M. T., Lin, J. C. W., Djenouri, Y., Fournier-Viger, P., & Zhang, Y. (2019, June). A swarm-based data sanitization algorithm in Privacy-Preserving Data Mining. In *2019 IEEE congress on evolutionary computation (CEC)* (pp. 1461-1467). IEEE.

- Wu, J. M. T., Zhan, J., & Lin, J. C. W. (2017). Ant colony system sanitization approach to hiding sensitive itemsets. *IEEE Access*, 5, 10024-10039.
- Wu, W., Parampalli, U., Liu, J., & Xian, M. (2019). Privacy preserving k-nearest neighbor classification over encrypted database in outsourced cloud environments. *World Wide Web*, 22(1), 101-123.
- Wu, W., Xian, M., Parampalli, U., & Lu, B. (2021). Efficient privacy-preserving frequent itemset query over semantically secure encrypted cloud database. *World Wide Web*, 24(2), 607-629.
- Wu, X., Chu, C. H., Wang, Y., Liu, F., & Yue, D. (2007, May). Privacy preserving data mining research: Current status and key issues. In *International Conference on Computational Science* (pp. 762-772). Springer, Berlin, Heidelberg.
- Xiao, X., Tao, Y., & Chen, M. (2009). Optimal random perturbation at multiple privacy levels. *Proceedings of the VLDB Endowment*, 2(1), 814-825.
- Xu, K., Yue, H., Guo, L., Guo, Y., & Fang, Y. (2015, June). Privacy preserving machine learning algorithms for big data systems. In *2015 IEEE 35th international conference on distributed computing systems* (pp. 318-327).
- Zainab, S. S. E., & Kechadi, T. (2019, July). Sensitive and private data analysis: A systematic review. In *Proceedings of the 3rd International Conference on Future Networks and Distributed Systems* (pp. 1-11).
- Zakerzadeh, H., Aggarwal, C. C., & Barker, K. (2015). Privacy preserving big data publishing. In *Proceedings of the 27th international conference on scientific and statistical database management*, 26, 1-11.

Zhang, L., Wang, W., & Zhang, Y. (2019). Privacy preserving association rule mining: Taxonomy, techniques, and metrics. *IEEE Access*, 7, 45032-45047.

Zschech, P., Horn, R., Hörschele, D., Janiesch, C., & Zschech, K. (2020). Intelligent user assistance for automated data mining method selection. *Business & Information Systems Engineering*, 1-21.