OPTIMISING THE TRADE-OFF BETWEEN ACCURACY AND PRIVACY IN DATA STREAM MINING ENVIRONMENTS

A THESIS SUBMITTED TO AUCKLAND UNIVERSITY OF TECHNOLOGY
IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Supervisors

Associate Prof. Roopak Sinha

Prof. Edmund Lai

Dr. M. Asif Naeem

18th July 2022

By

U. H. W. A. Hewage

Engineering, Computer and Mathematical Sciences

Abstract

Data streams differ from static datasets due to numerous characteristics such as being incremental, high speed, high volume, subject to concept drift, and dynamically adapting. This unique nature of data streams makes Privacy-Preserving Data Stream Mining (PPDSM) rather challenging. The trade-off between data privacy and data mining accuracy is one of the significant concerns in PPDSM. Optimising this trade-off is a complicated task due to the nature of data streams. Though privacy-preserving methods are proposed to optimise this trade-off in PPDSM, there is still room for improvement in this area. Moreover, there is a lack of well-structured frameworks to perform the accuracy-privacy optimisation. This research aims to implement an appropriate perturbation method providing optimal trade-off between data privacy and data mining accuracy in PPDSM, which ultimately leads to a well-structured framework.

We proposed seven variations of noise addition methods to achieve high privacy while maintaining high accuracy. These novel methods combine cumulative noise addition, noise resetting, and cycle-wise noise addition, inspired by the well-known Logistic Function. The best-performing noise addition method from the proposed variations was used to build the Accuracy Privacy optimising Framework (APOF). The foundation of APOF is that the accuracy and privacy level depends entirely on the user, and achieving 100% accuracy and privacy is not possible. Consequently, APOF was designed to optimise the accuracy-privacy trade-off by considering the user's privacy requirements. The optimisation is achieved through a data fitting module. Finally, we

extended APOF to Enhanced-APOF to operate in data streaming environments.

The logistic cumulative noise addition outperformed other proposed noise addition methods considering accuracy and privacy. The optimised accuracy-privacy trade-off could be achieved from the cycle-wise noise addition, and the cycles were designed based on the Logistic function. We could use all these benefits by using logistic cumulative noise addition as the privacy-preserving technique in APOF. Through the data fitting module APOF, we predicted the respective accuracy level for a user-defined privacy threshold retaining a small error. APOF allows the user to fine-tune requirements if needed and further optimise the accuracy-privacy trade-off according to his/her requirements. Experimental evidence shows the Enhanced-APOF is a well-structured framework for accuracy-privacy trade-off optimisation for a data streaming environment as it was designed considering the nature of data streams. The logistic cumulative noise addition for privacy preservation, Hoeffding Adaptive Tree for classification, and data fitting for optimisation have proven to be a prominent combination to achieve accuracy-privacy trade-off optimisation.

Contents

List of Tables

List of Figures

Attestation of Authorship

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the qualification of any other degree or diploma of a university or other institution of higher learning.

Signature of candidate

Publications

Table 1: List of Published/Submitted Research Articles

Manuscript	Publication/Submission	Contribution
1	Hewage, U. H. W. A., Sinha, R. & Naeem, M. A. (2022). Privacy Preserving Data (Stream) Mining Techniques and Their Impact on Data Mining Accuracy - A Systematic Literature Review. Artificial Intelligence Review. (Manuscript No - AIRE-D-21-01064, Submitted after revisions)	Waruni Hewage: 90%, Roopak Sinha: 6%, M. Asif Naeem: 4%
2	Hewage, U.H.W.A., Pears, R. & Naeem, M. A. (2022). Optimizing the Trade-off Between Classification Accuracy and Data Privacy in the Area of Data Stream Mining. International Journal of Artificial Intelligence, 1 (1), 147-167. (Published)	Waruni Hewage: 85%, Russel Pears: 10%, M. Asif Naeem: 5%
3	Hewage, U.H.W.A, Sinha, R. & Naeem, M. A. (2022) An Accuracy-Privacy Optimization Framework Considering User's Privacy Requirements for Data Stream Mining. ACM Transactions on Knowledge Discovery from Data. (Manuscript No - TKDD-2021-11-0391, Under review)	Waruni Hewage: 85%, Roopak Sinha: 10%, M. Asif Naeem: 5%
4	Hewage, U.H. W. A, R. Sinha, and M Asif Naeem. (2022). An Efficient and Enhanced Privacy-Preserving Framework to Achieve Optimal Accuracy-Privacy Trade-off for Evolving Data Streams. ACM Transactions on Knowledge Discovery from Data. (Manuscript No - TKDD-2022-07-0261, Under review)	Waruni Hewage: 85%, M. Asif Naeem: 10%, Roopak Sinha: 5%
5	Hewage, U.H.W.A., R. Sinha, and Russel Pears. (2022). Utilizing Noise as an Attack Independent Measure for Representing Privacy in Logistic Cumulative Noise Addition. (Accepted in IEEE Women in Engineering Conference, 2022)	Waruni Hewage: 85%, Russel Pears: 8%, Roopak Sinha: 7%

We, the undersigned, hereby agree to the percentages of participation to the chapters identified above.

Table 2: Signatures of the Contributors

Name	**Signature**
U.H.W.A. Hewage	
Roopak Sinha	
M. Asif Naeem	
Russel Pears	Signed by primary supervisor Roopak Sinha, on behalf of Russel Pears as he cannot be contacted.

Acknowledgements

First and foremost, I would like to thank Prof. Russel Pears, who supported me in many ways to start my PhD journey. All the guidance and encouragement you gave as the primary supervisor in the first half of the research is invaluable.
I sincerely thank Associate Prof. Roopak Sinha for stepping up and guiding me as the primary supervisor in the second half of the PhD. I am grateful to you for your valuable time, feedback and encouragement. Thank you.
I thank my secondary supervisors, Prof. M. Asif Naeem and Prof. Edmund Lai, for all their input and feedback in making this research successful. It helped a lot to improve my work.
I am grateful to my colleagues at AUT. Thank you all for being there; our short discussions and coffees helped me relax. Also, I would like to thank my colleagues at the University of Ruhuna, Sri Lanka, for supporting me whenever needed.
I would like to express heartfelt gratitude to my parents for giving their best to make me the woman I am today. Your love, care, and prayers supported me to hold into my goals when things got hard. Amma and Thaththa, thank you so much.
Last but not least, I thank my incredible husband (Pathum Priyasanka) for being with me in every way possible, understanding me, believing in me, and encouraging me to pursue my dreams. Without you, I will not be standing here today. Thank you.

Chapter 1

Introduction

1.1 Accuracy and Privacy in Data Mining

Data mining plays a vital role in many organisations and businesses [1, 2]. It can be assumed that it will become more and more prominent and engaging in the future, considering enormous data is being used and produced [3]. Organisations use their past, and present data in their decision-making process [4, 5]. Raw data does not hold any meaning, but information retrieved from data using data mining can be invaluable [6]. Mining information from raw data can be done using different machine learning techniques such as classification, clustering, and regression analysis [4, 7]. These techniques identify significant trends and relationships in data and make valuable predictions to aid decision-making [8]. As an example, consider the product shelving criteria of supermarkets. Supermarkets mine customer data to identify their buying patterns and use it to decide how to shelve the goods. Frequently brought items can be shelved closer to each other, or items that bring more profits can be placed on the front shelves. Moreover, information retrieved from data mining can be used to optimise other operations, such as discounts on items and stock keeping. Similarly, other organisations, such as health care and financial businesses, also use data mining to support their

decision-making process to make changes and conduct necessary improvements [9].

The accuracy of data mining results is the primary measure of the correctness of the decisions made [10, 11, 12]. If data mining has low accuracy, decisions made using those data mining results are not trustworthy [13]. While it is challenging to achieve 100% accuracy, some data mining techniques have shown higher accuracy, like 95% or 98% [14, 12]. Generally, organisations expect a minimum accuracy from data mining before using it for business-critical decisions. For instance, consider the decision-making process for granting a bank loan. Banks use customer data that includes features or variables such as Name, Age, Gender, Address, Marital status, Occupation, and Monthly income to decide whether a customer is loyal [15]. A data mining technique such as classification can identify the relationships and patterns between the above-discussed variables and the customer's loyalty. It builds a model based on the relationships between variables. Then this model is used to classify a new customer as loyal or not.

Besides the data mining technique, the data quality also affects the data mining results [16]. Accurate and complete data enhances the accuracy of data mining results [16, 17]. Accuracy drops when data contains missing or incorrect values or outliers [18]. Data pre-processing is conducted to refine data before using it in data mining [19]. Pre-processing fills missing values with the most suitable values statistically and removes outliers [19, 20]. However, pre-processing can only be largely effective if the data contains many correct values.

While data mining aims to make accurate predictions, ensuring the privacy of the data used for data mining is equally important [21, 22]. Defining privacy is a complex task, and it changes with the context. One of the most generic definitions of privacy is "the degree of uncertainty according to which original private data can be inferred" [13]. It can be rephrased as the degree of protection provided to private data. People do not want to reveal their identity to the world in the process of data mining [9, 23].

Hence, unauthorised persons should not be able to identify individuals in the dataset, and their personal data should be protected. Data should only be used in data mining to make predictions, and the identity of individuals should not be revealed [24]. Especially sensitive data such as health-related and income-related data should be concealed rigorously.

Privacy-Preserving Data Mining (PPDM) was introduced to conduct data mining while preserving data privacy [25, 8, 26]. Two types of privacy preservation can be identified: input privacy and output privacy [15, 27]. Input privacy distorts original data values before data mining, while output privacy distorts data mining results. This research focuses on input privacy; henceforth, the term privacy is used for input privacy unless otherwise mentioned. There are different privacy preservation methods such as Perturbation [28, 29, 9], Anonymization [30], and Secure Multi-party Computation [29, 31]. These methods consist of techniques including but not limited to noise addition and multiplication [32, 9, 33], random rotation [34], random projection [35], k-anonymity [30], l-diversity [36], condensation [37], and encryption [31]. PPDM uses one or a combination of the above techniques to change original data values to preserve privacy and then perform data mining on the modified or changed data.

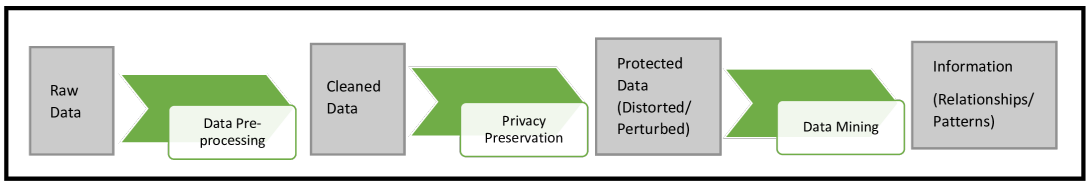

Figure 1.1: Process of Privacy-Preserving Data Mining (PPDM) using input privacy preservation techniques

1.2 The Trade-off Between Data Mining Accuracy and Data Privacy

The objective of PPDM is to change the original data values to preserve privacy while maintaining the statistical properties or relationships among data that are useful for data mining [32, 24, 26]. Statistical relationships help identify patterns in data and mine data accurately. However, changing original values degrade the quality and accuracy of the data [38, 23]. Though privacy preservation methods intend to retain statistical properties, techniques such as noise addition and multiplication highly distort data [32, 39]. It destroys vital relationships among data and prohibits achieving highly accurate data mining results.

Data privacy and accuracy are highly interlinked [38, 23]. Increasing data privacy decreases the data mining accuracy and vice versa [32]. Hence, there is a trade-off between data privacy and data mining accuracy [28, 10]. This accuracy-privacy trade-off suppresses the primary objective of PPDM, which is to achieve high privacy while maintaining high accuracy. Therefore, finding solutions to the accuracy-privacy trade-off has received increasing attention. Some research works [10, 40, 41, 42, 43, 23] discuss this issue and propose solutions, but it has yet to be fully answered. Therefore, a method to optimise the accuracy-privacy trade-off is still a question and should be investigated thoroughly.

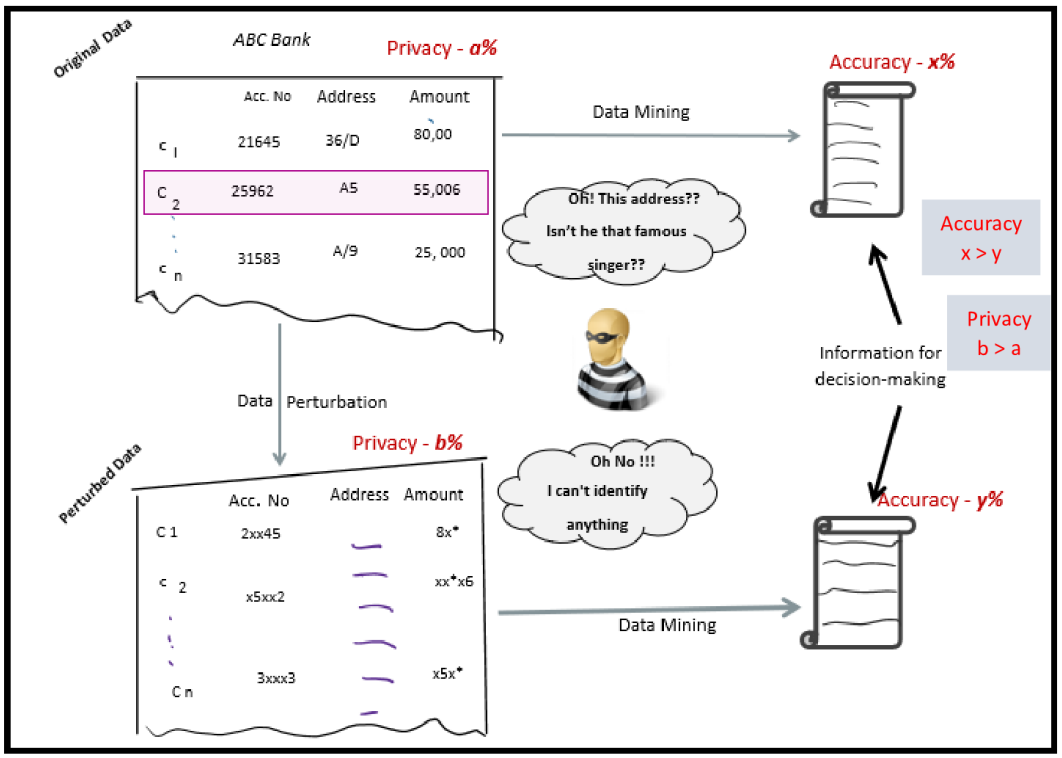

Figure 1.2: Behaviour of accuracy-privacy trade-off before and after applying privacy preservation techniques

1.3 Privacy-Preserving Data Stream Mining (PPDSM)

A data stream can be defined as a continuous flow of data generated from different sources [44]. Many application environments such as telecommunication, sensor networks, Internet of Things (IoT) and network monitoring produce data streams [18, 32]. Data stream mining is a research area to study methods and algorithms for extracting knowledge from streaming data [45]. Applying privacy-preserving techniques in data stream mining can be identified as Privacy-Preserving Data Stream Mining (PPDSM) [46, 32]. Due to the high use of technological devices and real-time data processing, the world is shifting rapidly to deal with data streams [45, 31]. Instead of conventional data mining using static databases containing historical data, it is necessary to mine real-time data from data streams. However, mining data streams is

more challenging than static databases due to its dynamic nature [47, 46]. Therefore, privacy preservation and data mining techniques should also be changed to deal with data streams.

1.3.1 Challenging Nature of Data Streams

Volume, velocity, and volatility are the three principal challenges in mining data streams [45, 31]. Data streams contain millions of records (volume), the records of a data stream can arrive at high speed (velocity), and data can vanish or change soon after being produced (volatility). Data stream mining involves other challenges, such as the need for quick data pre-processing, handling concept drift, dealing with delayed data, and real-time execution [45, 48]. Data stream mining is a continuous process as data streams are continuous, transient, and unbounded [49, 50, 51]. In static databases, data can be trained and tested repeatedly because the entire dataset is available for mining. However, data stream mining cannot be repeated as we cannot access the entire stream at once [52].

Privacy-preserving mechanisms should be improved to deal with the challenges in data stream mining. The privacy preservation technique should be able to cope with a large amount of data without any interruption to deal with the volume of the data stream [32]. As data arrives at high speed, privacy preservation should be applied quickly [47, 46], and data should be released with minimal delay. Another vital aspect of data stream mining is handling concept drift, where the underlying data distribution changes with time [53, 48, 44]. Data mining models must adapt to such changes to maintain a steady performance over time. So the privacy preservation mechanism should not make any additional impact on the data, which can cause unexpected accuracy drops.

1.3.2 Accuracy-Privacy Trade-off in PPDSM

Data stream mining techniques should be able to handle issues such as concept drift handling and the need to process data quickly. In contrast, privacy preservation techniques should deal with numerous concerns such as volume, fast execution, and consistent privacy throughout the data stream [38, 32, 36]. Failing to address these issues can degrade accuracy and privacy in higher amounts. Therefore, addressing the accuracy-privacy trade-off in data stream mining is rather complex than static databases.

All the things discussed in Section 1.3 should be considered when optimising the accuracy-privacy trade-off in data stream mining. The number of research works on the accuracy-privacy trade-off in PPDSM is comparatively lower than in PPDM. Research works such as [54, 52, 55, 32] propose techniques to address the accuracy-privacy optimisation issue in data stream mining. However, these works are either partially successful or have more potential for improvement.

1.4 Problem Definition and Motivation

Optimising the trade-off between data mining accuracy and data privacy is more challenging for PPDSM due to data streams' unique and dynamic behaviour [32, 47]. Though privacy-preserving techniques are relatively successful in working with static databases, privacy-preserving techniques for data streams need more improvements. As it is harder to increase both privacy and accuracy simultaneously, attention should be paid to inventing a novel approach for optimisation. Let us look at the following motivation scenario that explains the problem definition using a real-world example.

Assume that Ann is taking care of her grandfather. Ann uses a personal health care monitor to keep track of his health details such as heart rate and blood pressure. This device monitors Ann's grandfather's health and alerts her mobile phone if it detects

any unusual health behaviour. This health device works perfectly fine without any errors. Ann recently decided to use a privacy preservation component for the health care monitor to protect her grandfather's health data, which can be considered private and sensitive. The ongoing discussions on the importance of protecting personal data and different privacy breaches led Ann to take this decision. Due to the privacy preservation component, the original data is now protected. One day, Ann receives an alert indicating that the health care monitor has detected an unusual health behaviour. She stops all the work, calls an ambulance and goes home to check on her grandfather. However, after returning home, Ann sees that everything is normal and that the grandfather is doing well. That means the healthcare monitor has sent Ann a false alarm. That is not a big problem, as the situation did not cause severe harm. However, think of the possibility of the opposite happening. Ann's grandfather has a health emergency which needs immediate attention, but the health monitor fails to identify it. In that scenario, she does not get an alert and assumes everything is going well.

The health care monitor, which used to work absolutely fine earlier, has started to act strange after adding the privacy preservation component. What is the reason for this complication? Privacy preservation techniques distort original data to project data values from unauthorised access. However, this process reduces the accuracy of the data mining results. Due to distorted/perturbed data, data mining models fail to identify actual patterns and behaviours of the original data. This leads to getting inaccurate data mining results. Therefore, it is essential to maintain the accuracy of data mining results while protecting data privacy.

Rapidly increasing technology and usage of personal data in data mining makes privacy a necessity. Further, the world is moving into work with data streams, and PPDSM lacks privacy-preserving techniques that consider the accuracy-privacy trade-off in data stream mining. The motivations for starting and proceeding with this research arise here. Therefore, a good starting point is to implement an enhanced privacy

preservation technique for PPDM and PPDSM that performs better than the existing techniques. Then improving it to a well-structured framework with a mechanism to optimise the accuracy-privacy trade-off should be a great solution to this current issue.

1.5 Research Objectives and Contributions

This section explains the research objectives in terms of research questions and the contributions made from the research. Our research objectives are based on the issues highlighted in the literature (See Chapter 3) related to PPDSM and our motivation. The ultimate research objective is **implementing a framework to optimise accuracy-privacy trade-off in data stream mining**. To achieve this, the research process was divided into three main parts, starting by implementing an advanced perturbation method.

1.5.1 Research Questions (RQs)

Three research questions (RQs) were formulated to achieve the research objective. RQs are listed as follows.

- RQ1 - What is the most appropriate perturbation method that expresses the optimal trade-off between data privacy and data mining accuracy?
- RQ2 - What framework should be constructed to model the trade-off between privacy and accuracy for a specific user in a data mining environment?
- RQ3 - How do we extend the framework proposed in RQ2 above to a dynamically adapting or evolving data streaming environment?

1.5.2 Research Methodology

The research was carried out with the objective of finding solutions to the above-listed research questions. We followed a sequential process when addressing these research questions, as the output from RQ1 is the primary input to RQ2, and the output from RQ2 is the primary input to RQ3.

Our research approach starts with an extensive systematic literature review, and Figure1.3 represents the overall process involved.

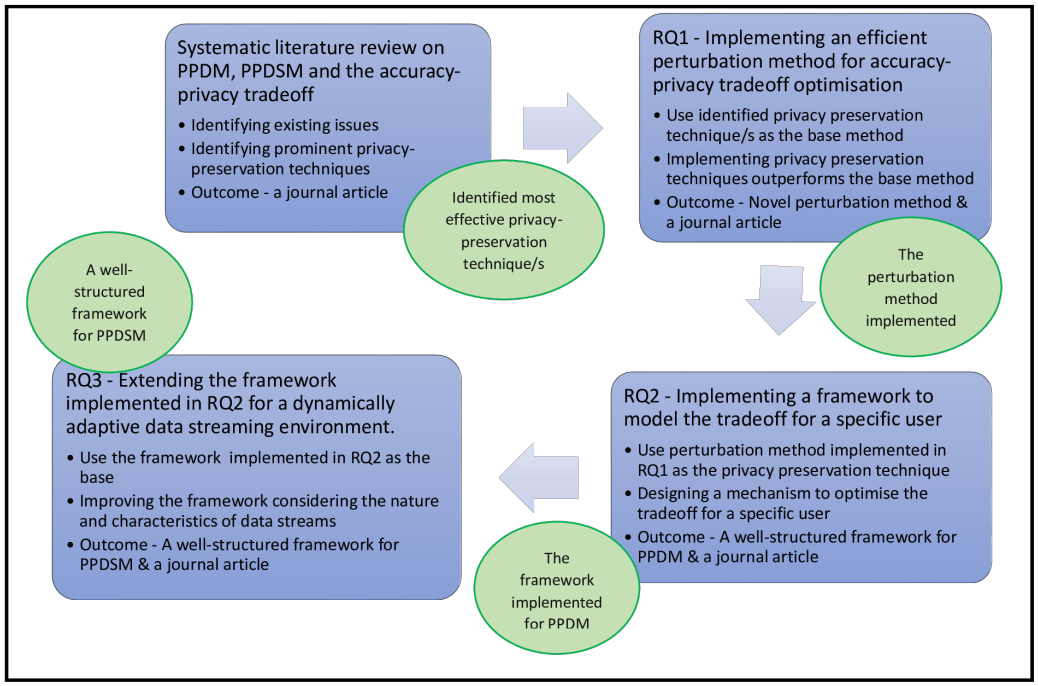

Figure 1.3: Research process

1.5.3 Research Contributions and Findings

Several contributions were made to the areas of PPDM and PPDSM in the process of finding solutions to the research questions mentioned in Section 1.5.1. This section presents all the contributions and findings made as to the outputs of each research question separately.

1. **The primary contribution from RQ1 is a noise addition-based perturbation method that helps to achieve high privacy with a good accuracy level. This method was named Logistic Cumulative Noise Addition (SRW)**. This work is presented in Chapter 5.

    - Seven variations of perturbation methods that combine Random Projection, translation, and cumulative noise addition were introduced. Random Projection-based Cumulative Noise (RPCN) [32] was used as our base method.
    - The logistic function was used to control the maximum noise added and the noise addition rate, considering the behaviour of the logistic curve. This is a novel approach to perturbation. We experimentally proved its success in noise addition as it minimizes the accuracy decrements in long-term running.
    - Different techniques such as noise resetting, cycle-wise noise addition, and the use of absolute noise values were incorporated to control the cumulative noise level further. These techniques were proven to have high performance than the base method.

2. **The Accuracy-Privacy Optimisation Framework (APOF) is the main contribution made from RQ2.** The APOF is designed to work with static databases. However, it has features that support data streams and can be used in PPDSM in some scenarios, such as when there is no concept drift in data. This work is presented in Chapter 7.

    - The APOF consists of three main modules, namely, accuracy, privacy, and data fitting.
    - Accuracy-privacy optimisation was done according to a user-given privacy threshold. It is argued that not every user needs 100% accuracy or privacy. Some may give their priority to accuracy, while others to privacy. Therefore, we can consider the user's accuracy-privacy requirements and provide them with an opportunity to fine-tune their requirements if needed.
    - The optimisation was carried out using the data fitting module. The perturbation was conducted for different noise addition rates and calculated accuracy-privacy for each noise addition rate. We used kernel methods to fit those accuracy-privacy values and predicted the expected accuracy level for a user-given privacy threshold. So users can know what the expected accuracy level for their privacy requirement is and can make changes if needed.
    - Furthermore, the data fitting module predicts the noise addition rate that needs to be used in reproducing the user's privacy level and expected accuracy level.

3. **The primary contribution from RQ3 is the Enhanced Accuracy-Privacy Optimisation Framework (Enhanced-APOF) that can be used to preserve privacy in evolving data streams.** This work is presented in Chapter 9.

- The adequacy of APOF in PPDSM was analysed to find its strengths and weaknesses. Then the following changes were made to APOF considering evolving data streams
- The perturbation method SRW was improved by adding a method to select random cycle sizes to maintain a guaranteed privacy level. Additionally, a noise resetting method was combined to minimize the effect on the accuracy in lengthy or infinite data streams.
- A comparative analysis of Hoeffding Tree (HT) and Hoeffding Adaptive Tree (HAT) in the context of PPDSM was carried out. Though HT and HAT were compared in terms of their accuracy in classification, performance comparison in PPDSM is novel.
- An efficient classifier switching method that is useful in maintaining a stable accuracy level while reducing the possible risk of overfitting was introduced.
- A window-based accuracy-privacy monitoring method was introduced to measure up-to-date accuracy-privacy values.
- Experimentally showed that a sufficiently large size window from the data stream represents the same accuracy-privacy values of the entire data stream. Hence, data fitting can be conducted using a sample from the data stream.

4. Additionally, we **experimentally evaluated noise variance as a measure of privacy using SRW**. Area Under the Curve (AUC) of the Logistic curve was used to measure the total noise variance added. It is proven that noise variance only catches the effect of noise on privacy. Nevertheless, there are other factors to consider, such as changes in the data itself, and we miss these vital factors when using noise variance as the measure of privacy. However, experiments have shown that we still can use noise variance to identify increasing or decreasing trends of privacy in SRW.

1.6 Organisation of the Thesis

This thesis presents in the "Thesis by manuscript" format, and the structure of the thesis is as follows. The chapters 2, 4, 6, and 8 provide introductions to manuscripts which present literature review, RQ1, RQ2 and RQ3 respectively. The literature review presented in Chapter 3 discusses all the related backgrounds and identifies the existing gaps. Chapters 5, 7, and 9 present the manuscripts for RQ1, RQ2 and RQ3. Each article focuses on the relevant research questions and discusses the background, methodology, results, and conclusions in detail. Chapter 10 discusses the overall research, highlighting major findings, limitations, conclusions, and future research directions. Appendix A and Appendix B outline an additional experiment conducted while addressing the main research questions. They explain the appropriateness of noise variance as a measure of privacy. Finally, Appendix C and Appendix D present a detailed description of the datasets and a list of acronyms used throughout the thesis respectively.

Chapter 2

Prelude - Manuscript 1

This Systematic Literature Review investigates existing input privacy-preserving data mining (PPDM) methods and privacy-preserving data stream mining methods (PPDSM), including their strengths and weaknesses. A further analysis was carried out to determine to what extent existing PPDM/PPDSM methods address the trade-off between data mining accuracy and data privacy which is a significant concern in the area. The systematic literature review was conducted using data extracted from 104 primary studies from 5 reputed databases. The scope of the study was defined using three research questions and adequate inclusion and exclusion criteria.

According to the results of our study, existing PPDM methods were divided into four categories: perturbation, non-perturbation, secure multiparty computation, and combinations of PPDM methods. These methods have different strengths and weaknesses concerning the accuracy, privacy, time consumption, and more. Data stream mining must face additional challenges such as high volume, high speed, and computational complexity. The techniques proposed for PPDSM are less in number than the PPDM. We categorized PPDSM techniques into three categories (perturbation, non-perturbation, and other). Most PPDM methods can be applied to classification, followed by clustering and association rule mining. It was observed that numerous studies have identified and

discussed the accuracy-privacy trade-off. However, there is a lack of studies providing solutions to the issue, especially in PPDSM.

Chapter 3

Systematic Literature Review (Manuscript 1)

3.1 Introduction

Data Mining and machine learning involve extracting knowledge from data, which makes a significant impact on the growth of organisations. Organisations use their past and current data to make decisions to improve their performance or services, where data mining comes to play [4]. This process consists of mining helpful information from raw data and making predictions that support the decision-making [56, 6]. There are two main approaches for data mining and machine learning, namely supervised learning and unsupervised learning. These include techniques such as classification, clustering, and association rules [4, 7]. These methods identify raw data patterns and produce useful information and predictions that are beneficial for the organisations.

The success of the data mining process is measured using the accuracy of data mining results [10, 12]. Furthermore, accuracy depicts the percentage of patterns learned by the data mining process. For instance, in a classification task, the percentage of correctly classified unknown data records over the total number of records is the accuracy [11].

A high accurate rate means a higher accuracy of the information produced, which leads to highly accurate decision-making. In some studies the term "utility" has also used to represent the accuracy [57, 11, 32, 58]. Ultimately, the terms "utility" and "accuracy" were used to measure the success rate of the data mining task using similar matrices. Therefore, we hereafter use the term "accuracy" to represent utility and accuracy to avoid confusion.

One of the major challenges stakeholders have to face in data mining is to protect the individual's privacy in data while using those for data mining [21]. Datasets may contain some data that data owners do not want to reveal to the outside world [24]. This data is called sensitive data [59]. For example, patients' medical history details from a hospital database or customers' bank balance details from a banking database can be considered sensitive data. Sensitive data needs to be protected so that the privacy of the individuals can be preserved in the data mining process.

Privacy-Preserving Data Mining (PPDM) [25, 8, 26] has been introduced as a solution to privacy concerns in data mining and has become a prominent area of data mining from the past few decades. PPDM methods should protect the privacy of the data while allowing the data mining process to carry out its duty as usual [24, 26]. This means PPDM methods should not cause a considerable impact on the output of the data mining [25, 26]. Two broader categories of PPDM methods, called input PPDM and output PPDM, can be seen in the literature [15, 27]. Input PPDM modifies original data before data mining to preserve privacy, while output PPDM deals with modifying the data mining output to preserve privacy. Our work focuses only on input PPDM methods as output PPDM methods mainly involve modifying data mining techniques (classifier or clustering algorithm) and need to be discussed separately. Different input privacy-preserving methods such as perturbation, anonymisation, and encryption have been proposed and practised in the data mining community. These methods have positive and negative impacts on both privacy and data mining tasks. However, the ultimate

expectation of PPDM methods is to protect data privacy so that unauthorized parties cannot identify the individuals using data. And maintain the statistical properties of the data so that it does not degrade the performance of data mining [32, 46, 28, 60]. So that the transformed and protected dataset can be used for data mining without accessing the original dataset.

A research field that studies methods and algorithms to extract knowledge from volatile streaming data [45] can be identified as data stream mining. Combining privacy preserving techniques in data stream mining is Privacy Preserving Data Stream Mining (PPDSM) [46, 32]. Mining helpful information and making predictions from data streams have additional concerns due to its behaviour. Unlike static datasets, streaming data is continuous, transient, and unbounded, which arises the need for processing quickly [18, 47, 51]. PPDSM methods should be robust to the specific behaviour of data streams [15, 38] and therefore, privacy preservation in data streams needs to be addressed differently [53, 44].

Most of the PPDM methods proposed have succeeded in preserving the privacy of data, but negatively affect the data mining results [38, 23]. PPDM methods transform original data values into another form that makes them unrecognisable by outsiders. This process can destroy the statistical properties of the data useful in data mining. Therefore, there is a trade-off between data privacy and data mining accuracy [28]. Increasing data privacy can decrease the data mining accuracy and vice versa [10]. Current researches have identified this inherent trade-off between data privacy and data mining accuracy (use as accuracy-privacy trade-off hereafter) and have proposed different PPDM methods to address the issue [23, 61, 41, 62]. Nevertheless, no perfect method has been found to optimise the accuracy-privacy trade-off, and the issue is still open to discussion.

We can find different studies in the literature that discuss PPDM and PPDSM.

However, we could not find secondary studies that discuss accuracy-privacy trade-off of PPDM/PPDSM in detail, and opening a discussion about it might be useful for the readers. Anyhow, it is impossible to talk about accuracy-privacy trade-off without having an idea about PPDM/PPDSM methods, strengths/weakness and possible challenges.Therefore. our study starts with existing input PPDM/PPDSM methods and ends by discussing accuracy-privacy trade-off. Hence, the focus of this study has covered by three major sections. First, we investigate PPDM methods, their strengths/weaknesses and applicable data mining tasks. Secondly, we investigate on PPDSM by looking into challenges in PPDSM and privacy-preserving techniques proposed for data streams. Finally, we study on the attention given to accuracy-privacy trade-off and how well the PPDM/PPDSM methods have addressed the accuracy-privacy trade-off.

The remainder of this paper has been organised as follows. Section 3.2 describes the method we followed in carrying out this systematic literature review. We discuss the results and findings of the study in Section 3.3. Finally, we discuss and conclude the knowledge gained from this study in Section 3.4 and 3.5.

3.2 SLR Protocol

This study has been carried out as a Systematic Literature Review (SLR), and the main goal of the study is to evaluate methods and techniques proposed in the area of PPDM. PPDM is a broad area that can be evaluated in many branches. Our focus is to evaluate PPDM in its applicability to data streams and its effect on the accuracy-privacy trade-off. The rest of this section explains the steps followed in conducting the SLR [63].

3.2.1 Problem Identification

Existing literature consist with plethora of secondary studies in the area of PPDM. Most of these studies [56, 64, 6, 8] summarize and evaluate the existing PPDM techniques while other studies talk about challenges and possible improvements for enhancing PPDM methods [21, 65, 66, 25]. Studies such as [4, 67, 56] present frameworks and categorizations of existing PPDM methods to provide the overall picture of the PPDM methods.

Though there are numerous studies on PPDM, only a few are focused on the application of PPDM in data stream mining. Research work such as [44, 48, 53, 45] discuss the challenges, opportunities, and possible future directions in privacy-preserving data stream mining. However, we could locate only one study [68] that discusses existing PPDM methods specifically for data streams to the best of our knowledge. A few studies [31, 69] discuss PPDM methods that can be used in big data in general and have the potential of being used in data streams as it is a category of big data. Therefore, there is a need to evaluate PPDM methods for data streams properly.

The well-known accuracy-privacy trade-off is a concern that still needs to more attention, and a considerable number of studies [59, 7, 21, 2] have identified this issue. But very few [25, 66, 70] have discussed this in detail with respect to different PPDM methods. We find the accuracy-privacy trade-off as an aspect that needs to be discussed, considering both static datasets and data streams.

By analysing the existing secondary studies, we could confirm a lack of studies that discuss the accuracy-privacy trade-off in PPDM and the existing PPDM techniques specifically for data stream mining. The motivation for our research work arises from this gap, and we try to address these issues in this comprehensive study.

3.2.2 Research Questions

After identifying gaps in existing secondary studies related to PPDM, we started our SLR by formulating three Research Questions (RQ) to address those gaps.

- RQ1: What are the existing privacy-preserving data mining methods?
  - RQ1.1: What are the strengths and weaknesses of the investigated methods?
  - RQ1.2: What data mining tasks can these methods be used for?
- RQ2: What is the nature of privacy preservation in data stream mining?
  - RQ2.1: What challenges can be identified when applying PPDM methods for data stream mining?
  - RQ2.2: What are the PPDM methods that have been proposed for data stream mining?
- RQ3: To what extent do the privacy-preserving data mining approaches identified in answering RQ1 and RQ2 address the trade-off between data privacy and classification accuracy, and what methods have been proposed to optimise the accuracy-privacy trade-off?

Concerning RQ1, we summarize the most common PPDM methods being used in the data mining community. To provide a broader insight into existing PPDM methods, we discuss the merits and demerits of each method, along with the data mining tasks they can be used for under the sub-questions of RQ1.

Under RQ2, we discuss the applicability of PPDM methods identified in RQ1 in data streams and the different PPDM methods proposed specifically for data stream mining. Here we also try to identify the challenges in applying PPDM methods in data stream mining as data streams behave differently than static datasets.

By answering RQ3, we aim to determine whether the accuracy-privacy trade-off has grabbed the attention it deserves, as it is a severe concern in the area. We investigate whether the authors have identified or discussed the above issue and the possible steps they have proposed or implemented to reduce the trade-off.

3.2.3 Search Process

A systematic manual search was conducted to find out the potential research work. First, we identified search keywords to initiate our search. The search was conducted using three sets of keywords to reduce the complexity of the searching process.

- Set 1: (PPDM OR accuracy-privacy trade*)
- Set 2: (PPDM AND utility)
- Set 3: (data stream) AND (privacy OR PPDM)

Set 1 focuses on selecting articles on privacy-preserving data mining, which may or may not include a discussion about the accuracy-privacy trade-off. We considered the articles with the term "utility" together with PPDM in Set 2, as some researchers prefer using "utility" instead of "accuracy," and those terms are being used interchangeably in the literature. Set 3 selects articles on data stream mining that talk about PPDM or privacy.

Five major databases were selected to search through, namely Scopus, IEEE, Science Direct, Springer, and ACM. The initial search was carried out to filter out the potential research work using the search strings mentioned above and filtered out the studies using the inclusion and exclusion criteria mentioned in the next section.

3.2.4 Inclusion Criteria (IC) and Exclusion Criteria (EC)

Running the three sets of keywords through five selected databases resulted in 3735 studies. The following IC and EC were used to select the relevant studies manually to address the formulated research questions.

- Inclusion Criteria (IC)
  - IC1- Studies that propose PPDM techniques for general data mining tasks.
  - IC2- Studies that discuss the challenges of applying PPDM in data streams.
  - IC3- Studies that only focus on privacy aspect of PPDM.
  - IC4- Studies that propose PPDM techniques specifically for data stream mining.
  - IC5- Studies that mainly focus on privacy-accuracy trade-off (though it's only for a specific application)
- Exclusion Criteria (EC)
  - EC1- Studies do not have a proper evaluation.
  - EC2- Studies that lacks full details of the implementation and context of the proposed methods.
  - EC3- Studies that do not carry out with relevant experimentation.
  - EC4- Studies that propose frameworks/conceptual models using existing PPDM methods.
  - EC5- Studies that only focus on a specific application/area/ or studies that have a limited usage.
  - EC6- Survey articles/secondary studies.
  - EC7- Studies that discuss impact of PPDM on different industries.

- EC8- Studies that only discuss/propose privacy breaching methods.
- EC9- Duplicate research articles.
- EC10- Studies that have new/improved versions.

Using this process, we made sure that all the relevant studies were included and irrelevant studies were excluded to increase the effectiveness of the SLR.

3.2.5 Search Execution

Figure 3.1 illustrates the process carried out to select the primary studies for SLR. The above-mentioned IC and EC were applied in different steps to filter out the articles that were out of scope considering the defined research questions. This process was carried out manually. For example, after selecting the potential articles from the initial search as the first step, IC1 and IC3 and EC5, EC6, EC7, and EC8 were applied as the second step. After this, the remaining number of articles could be reduced to 1882. The process was repeated until the most relevant articles were remaining, which turned out to be 104. We only considered the studies published in the last 20 years(2001 -2021), as most studies related to PPDM were published after 2000.

3.2.6 Data Extraction and Analysis

Data was extracted from each selected article by thoroughly reading the abstraction and conclusion and skimming through the rest of the text. The following data was collected.

- Title of the study
- Year of publication
- PPDM technique/method proposed
- Strengths and weaknesses of the proposed method/technique

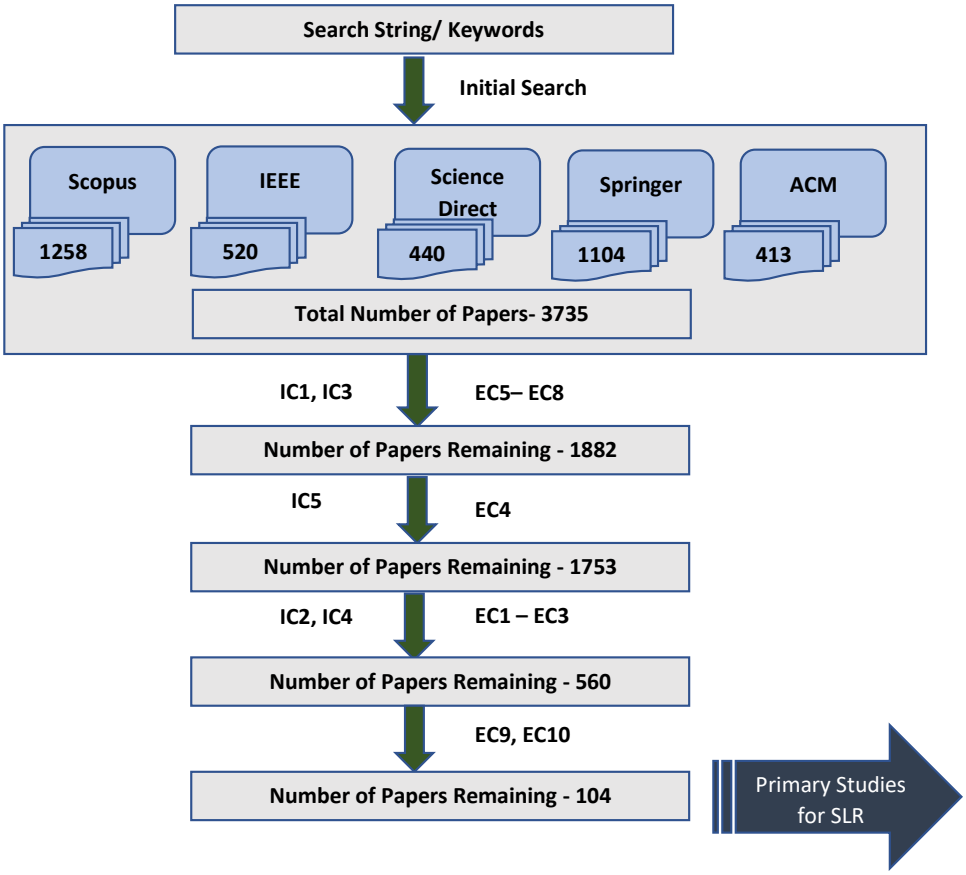

Figure 3.1: Search execution process, demonstrating all the steps followed to filter out the research articles for SLR.

- Applicable data mining tasks
- Applicability to data streams
- Challenges identified on applying PPDM to data streams
- Discussion on accuracy-privacy trade-off

Collected data were stored and analysed using MS. Excel to answer the formulated research questions. Figure 3.2 shows the distribution of the studies selected according to the year of publication. We can observe that many related studies have been published

from 2010 to 2021 than between 2000 and 2010. The number of studies published each year has been steadily rising in the last decade.

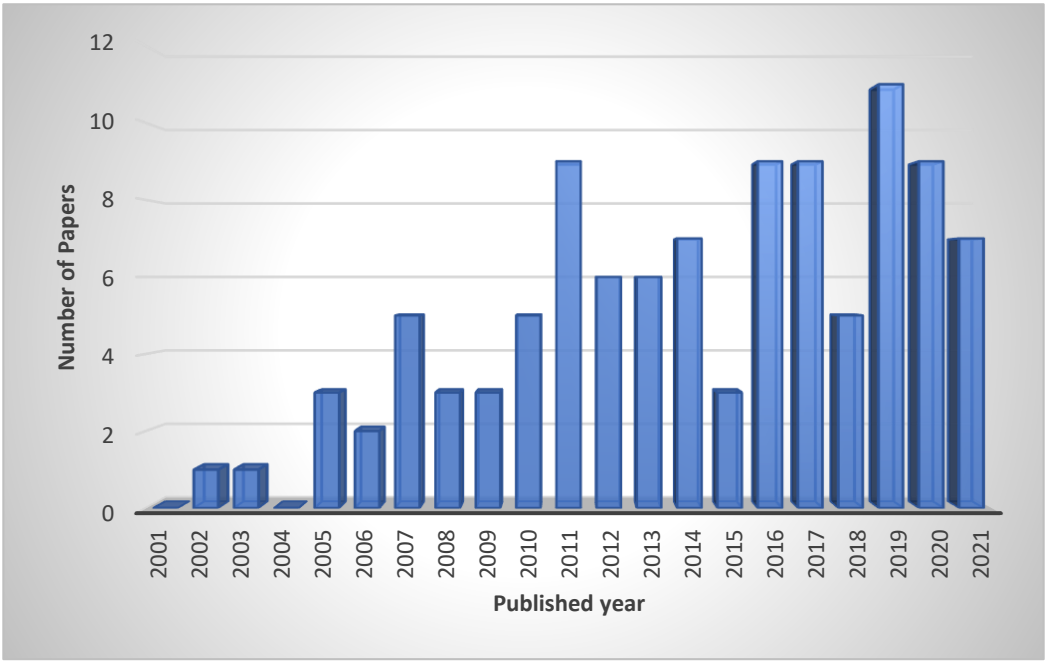

Figure 3.2: Distribution of the selected studies according to the year of publication

3.3 Results

This section summarizes the results and findings of our SLR for each research question.

3.3.1 Addressing RQ1 - Generic PPDM methods

To answer RQ1 and its sub-questions, all the different PPDM techniques and methods found in the final set of articles were studied. There are many categorizations of PPDM proposed in the literature. Authors of [71, 29] categorized PPDM techniques into two main categories, called *Secure Multi-Party Computation* and *perturbation*. In [23], PPDM has been divided into five categories namely, *Anonymisation, Perturbation, Randomization, Cryptography* and *Condensation*.

According to the analysis of extracted data, we agree with the categorization proposed in [31] as we believe it is more generic and justifiable. Therefore, we divide existing input PPDM techniques into four main categories, namely, Secure Multi-Party Computation, perturbation methods, non-perturbation methods, and combinations of the above techniques by extending the categorization provided by [31]. This section discusses all the techniques included in these four categories in detail.

Secure Multiparty Computation (SMC)

The Secure Multi-Party Computation (SMC) methods are being used for collaborative data mining and use cryptographic tools to protect data [29]. It allows different parties to jointly compute a certain functionality without revealing personal data [31]. Therefore, cryptographic methods can be used for distributed privacy and information sharing. This became popular as it provides a well-defined privacy model along with the methods for proving and quantifying [72]. However, there is a concern that the cryptographic techniques do not protect the output privacy, and instead, it stops the leakage of sensitive data in the process of computation [72]. The data mining community prefers perturbation techniques over SMC techniques because of their lower computational complexity [40, 38]. Cryptographic methods use encryption schemes that are challenging in scalability and implementation efficiency [31].

Perturbation Methods

This section discusses data perturbation methods that distort data values in specific ways to hide sensitive information while maintaining data properties that are important for data mining [28]. Data perturbation is the most commonly used privacy preserving technique in data mining because of its simplicity and computational efficiency. In [29], perturbation has been identified as altering data using statistical methodologies. However, Data perturbation methods have to pay special attention to the accuracy of data

mining, as distorting data can highly affect the data mining process. Perturbation can be divided into two categories called the value alternation approach and the probability distribution approach [9]. This section discusses the techniques that can be considered data perturbation methods.

The use of noise to distort the data is one of the earliest methods of data perturbation [32]. Additive noise and multiplicative noise are the two main usages of noise in the PPDM context [9]. Random values with zero mean and a specified variance are generated from a given distribution, such as Gaussian or Uniform distribution. Generated noise values are added to each record in additive noise environment, while each record is multiplied with the noise values in multiplicative environment [32, 9, 33]. The original data values are distorted, while the underlying data distribution can be reconstructed [14]. If the variance of added noise is high, then a high level of privacy can be expected, but it also causes a high information loss. Later, a combined version of additive and multiplicative noise is proposed in [9]. This combined approach guarantees more privacy than individual approaches.

Keke and Ling [34] first proposed a geometric transformation method named random rotation for PPDM based classification. The original dataset with m attributes is multiplied using a *(m x m)* random orthogonal matrix [32] perturbing all the attributes together [34]. A rotation-based approach that only transforms sensitive attributes is proposed in [73]. Perturbation using rotation transformation is vulnerable to rotation centre attacks [73, 34], as data closer to the origin is less perturbed than the other data records [32]. Recently, more improved versions of random rotation such as 3-D rotation transformation [74] and 4-D rotation transformation [75] has been proposed and these methods assures high data mining accuracy.

Other geometric data perturbation methods that combine random rotation, translation, and noise addition have been proposed to minimize the vulnerabilities accompanied by rotation transformation [28, 76]. These methods become robust to rotation centre

attacks by adding a translation, and robust to distance inference attacks by adding noise. However, it can still be vulnerable to background knowledge-related attacks [76].

Differential privacy [77] is a high privacy guaranteed privacy-preserving algorithm that works by adding Laplace noise to statistical databases. It ensures that an outsider cannot determine if a data item has been altered. According to the definition provided by [77] the result of the dataset is insensitive to the change of a record. Hence, makes it difficult for an attacker to gain knowledge about data. Research work such as [43, 78]) discuss research work using differential privacy as the PPDM technique.

Table 3.1 and Table 3.2 summarize different PPDM techniques in noise injection, rotation, and other geometric transformations, respectively, along with the strengths and weaknesses.

Random Projection (RP) based multiplicative data perturbation was proposed in [35]. RP projects a given dataset from a higher-dimensional space to a lower-dimensional subspace. This method is based on the Johnson-Lindenstrauss Lemma, and pair-wise distances of any two data points can be maintained within a small range [35, 32]. So, it can be considered an approximate distance preserving method. The authors of [35] have stated that RP can be more powerful when it is used with geometric transformation techniques such as scaling, rotation, and translation. Recently, a random projection-based noise addition method was proposed in [32]. This method experimentally proved high accuracy and privacy levels by combining RP, translation, and noise addition.

Condensation can also be considered as a perturbation PPDM method. It condenses data records into groups of pre-defined size k while maintaining statistical properties within the group [37]. It is not possible to distinguish one record from another within the group. Then pseudo data is generated instead of original data using the statistical information within the group. Condensation maintains inter-attribute correlations that guarantee a high accuracy level [37, 85].

Few works such as [86, 87, 88] have considered using fuzzy logic-based techniques

Table 3.1: Analysis of PPDM Methods - Perturbation (Noise Injection and Rotation)

Article	Techniques	Strengths	Weaknesses/Challenges
[9]	Additive & multiplicative noise	Robust to diversity attacks, Maximum privacy than individual approaches	—
[33]	Multiplicative noise	Minor changes to the original data	Vulnerable to attacks on additive noise
[78]	Noise addition & Differential privacy	Good accuracy and strong privacy guarantee, for both numerical & nominal attributes	Extensive experiments are needed
[79]	Perturbation	Facilitate authorized parties to reconstruct original data	Vulnerable if the reconstruction algorithm is disclosed
[34]	Random Rotation	Use weights to consider different privacy concerns	Vulnerable to rotation centre attacks
[73]	Geometric transf--ormation (rotation)	Apply rotation only on sensitive data	Vulnerable to rotation centre attacks
[74]	Three dimensional rotation	High accuracy, Data reconstruction is difficult	Attack resistance should be evaluated
[75]	Four-dimensional rotational	High accuracy	Rotation angle should be selected by human analysis
[80]	Orthogonal rotation transformation & translation	Better in both privacy & accuracy	Translation can be reversed & rotation is vulnerable to rotation centre attacks

Table 3.2: Analysis of PPDM Methods - Perturbation (Other Geometric transformations)

Article	Techniques	Strengths	Weaknesses/Challenges
[55]	Geometric transformations	Robust to reconstruction attack, Relatively faster	—
[28]	Geometric Perturbation (Multiplication, translation & distance perturbation)	Robust to rotation centre & distance inference attacks, Better privacy & accuracy	—
[76]	Random Rotation, Trans--lation & noise addition	Robust to rotation centre & distance inference attacks	Vulnerable to background knowledge related attacks
[81]	Inverse cosine based transformation	Preserves distance, high accuracy	vulnerable to attacks with background knowledge
[10]	Normalization, geometric rotation, linear regression, & multiplication	High accuracy & privacy	Accuracy may decrease when correlation coefficient of regression is not strong
[82]	Min–Max normalization & 3D shearing	High accuracy, privacy & data transformation	Attack resistance should be evaluated
[83]	Min–max normalization-based data transformation	Accuracy is well-preserved	Accuracy can decrease with multiple sensitive attributes
[84]	Random linear transformation	High privacy as both data & classifiers are perturbed	High computational overhead for large datasets

for data perturbation. A fuzzy logic-based perturbation method with less processing time has been proposed in [86]. Though the accuracy of the method is similar to the accuracy of the original dataset, privacy needs to be evaluated. A multiplication perturbation method using fuzzy logic has been implemented in [87]. This method has achieved better accuracy and privacy levels for classification and clustering. Another work that uses fuzzy models for synthetic data generation as a perturbation method can be found in [88].

Table 3.3 gives an overall idea about different techniques in random projection, condensation, fuzzy logic, and some other distortion methods in PPDM.

There are a considerable number of PPDM methods which combines different transformations and data distortion methods to achieve a better performance, considering both privacy and accuracy. Some research work [27, 91, 92, 12, 61, 93, 94, 95] use transformation techniques such as Singular Value Decomposition (SVD), Non-negative Matrix Factorization (NMF) and Discrete Wavelet Transformation (DWT) to perturb data. In [22] authors have used Principal Component Analysis (PCA) while PCA together with noise addition has been used in [42] for data perturbation. A summarization of these methods can be seen if Table 3.4.

Non-perturbation Methods

Non-perturbation methods sanitize the identifiable information to preserve privacy [31], and different anonymisation techniques are included in this category. Non-perturbation methods modify or remove only a portion of data [100] where perturbation methods distorts each data value. This process uses techniques to make a single record indistinguishable from another set of a specified number of records so that individual records cannot be identified and privacy is preserved. We discuss a set of non-perturbation-based PPDM methods in this section.

Anonymisation is the most used non-perturbation technique that involves identifying

Table 3.3: Analysis of PPDM Methods - Perturbation (Random Projection, Condensation, Fuzzy Logic and Other)

Article	Techniques	Strengths	Weaknesses/Challenges
[35]	Random projection based multiplicative noise	Better privacy than orthogonal transformation-based distance preserving perturbation	Vulnerable to attacks designed for additive noise after a logarithmic operation
[37]	Condensation	Within group statistical proper--ties gives high accuracy	Deciding the group size
[85]	Condensation (Anonymity & Pseudo data generation)	Pseudo data preserves privacy	Fixed group size increases loss in sparse regions
[14]	Condensation (Rule based)	Automatically selects the appropriate group size	—
[86]	Fuzzy logic	High accuracy & less processing time	Privacy should be evaluated
[87]	Perturbation (using Fuzzy Logic)	Better privacy and accuracy	—
[88]	Fuzzy c-regression models (FCRM)	High clusters give low information loss	—
[89]	Perturbation	Takes users' privacy consideration in to account	Maximum frequent item set length & attribute cardinality matters
[17]	Heuristic algorithms	Modifies fewer transactions, Efficient	Deciding support & confident thresholds
[90]	Randomized response	Reducing interference between perturbed & original data	Higher the distortion, lower the accuracy
[16]	Conditional probability distribution & cross sampling	Safe from linking attack, Boost accuracy when data are not sufficient	Handling multiple sensitive attributes, Order of the attributes matters
[43]	Differential privacy	Preserve statistical characteristics &	Essential parameters should be adjusted

Table 3.4: Analysis of PPDM Methods - Perturbation Using Different Transformations

Article	Techniques	Strengths	Weaknesses/Challenges
[27]	SVD, NMF, DWT	Can use when the original data is supplied by different data owners	—
[91]	SVD, PCA, NNMF	Reduces the cluster misplacement error	Carefully selecting private attributes & correlations
[92]	Sample selection & SVD	High utility and privacy than original SVD	Privacy is worst when sample rate is high
[12]	SVD, NNMF & DWT	Effective, Balanced accuracy & privacy than individual methods	Data mining task needs to be applied
[61]	NMF & SVD	Only perturb confidential attributes to maintain accuracy & privacy	Higher dimensions cause accuracy drops
[93]	SSVD	Better privacy due to double distortion	SVD computation for large datasets is expensive
[94]	SSVD & NMF	High accuracy & privacy	High execution time than individual methods
[22]	PCA	PCA lowers the complexity	—
[42]	PCA & additive perturbation	Robust to correlation-based & transform-based attacks	Slightly high execution time
[96]	Genetic Algorithms	High efficiency	Selecting fitness functions for different domains
[97]	Data Transformation	Efficiently protect boolean attributes	Accuracy, privacy should be measured using standard measures
[98]	Transformation	Only perturb confidential attributes	Transformation can be reversed
[99]	Non-linear dimen--sionality reduction	Preserves distance related properties	—
[95]			
[60]			

different parts of a data record, such as Identifiers, Quasi-identifiers, Sensitive and Non-sensitive attributes. Then it removes identifiers and modifies quasi-identifiers by performing techniques such as generalization and suppression, making a record indistinguishable from a set of other records [31]. Different anonymisation methods can be seen in the literature, such as *k*-anonymity [30], *l*-diversity [36] and *t*-closeness [101].

The basic method of anonymisation, *k*-anonymity, ensures that a single data record cannot be distinguished from at least *k-1* records [30, 58]. Identifying different parts of a data record is essential here, and then applying generalization and suppression techniques to achieve *k*-anonymized set of data. This method reduces the risk of a re-identification attack that is caused by the direct linkage of shared attributes [58]. The main weakness of the method is that it assumes that no two tuples contain data of the same person, which may not always be true [30].

Another weakness of *k*-anonymity is that it can be vulnerable to background knowledge-based attacks such as complementary release attacks and Temporal inference attacks. As a solution to this, an improved anonymisation model called *l*-diversity was introduced [36]. A table is called *l*-diverse if there are *l* well-represented values for the sensitive attribute [102]. The method provides privacy even when the data owner does not know what kind of knowledge the attacker has. However, it is difficult to implement for multiple sensitive attributes [36] and vulnerable to attacks such as similarity attacks [102].

Another anonymisation method named *t*-closeness was proposed in [101]. The requirement to achieve *t*-closeness is maintaining the distribution of a sensitive attribute in an equivalence closer to the distribution of the same attribute in the overall table. If the distance between two distributions less than the threshold *t*, it has achieved the *t*-closeness [101, 41]. This overcomes the skewness attack and similarity attack but cannot deal with identity disclosure attacks and with multiple sensitive attributes [101].

There are several more variations of anonymisation such as p-sensitive,*t*-closeness [103] have been proposed in addition to these main methods as solutions to privacy issues of the existing methods.

The main issue with all these anonymisation methods is that there is no specific computational approach to determine what data should be anonymised. This entirely depends on the expertise knowledge [103]. Different anonymisation techniques, along with their strengths and weaknesses, can be found in Table 3.5.

Methods Combining Cryptographic, Perturbation and Non-perturbation Techniques

Privacy-Preserving Data Mining methods that use different combinations of the above-discussed techniques are being used to preserve privacy. The reason for proposing these combing methods is to use the benefits of each method by reducing or eliminating the weaknesses.

In [23], authors have proposed a hybrid PPDM method by combining perturbation and anonymisation. This method uses additive noise and suppression techniques to achieve a minimum loss by avoiding the generalization involved in the anonymisation. An improved PPDM method was proposed in [107] to implement privacy separately according to the data owners' willingness. This method combines transformation and anonymisation. The hybrid approach implemented in [1] uses randomization and generalization techniques together to achieve a better accuracy level. Randomization and generalization are involved with the perturbation and non-perturbation methods, respectively. Authors of [108] propose a PPDM method by combining suppression and perturbation techniques. In this method, suppression is performed only on specific attributes, leading to a minimum information loss. A hybrid multi-group approach proposed in [109] uses randomization and SMC techniques together to achieve high accuracy and efficiency.

Table 3.5: Analysis of PPDM Methods - Non-Perturbation/ Anonymisation

Article	Techniques	Strengths	Weaknesses/Challenges
[30]	Anonymisation	Reduces re-identification by directly linking on shared attributes	No two tuples contain the same person is not real all the time
[58]	Anonymisation (k-anonymity)	Deal with large item sets	—
[36]	Anonymisation (l-diversity)	Resistant to background knowledge & homogeneity attacks	Difficult to implement when there are multiple sensitive attributes
[101]	Anonymisation (t-closeness)	Overcomes skewness & similarity attack of l-diversity	Does not deal with identity disclosure
[104]	Anonymisation ((l, d) -semantic diversity)	Reduces the risk of background knowledge based attacks	—
[41]	Anonymisation (Microaggregation)	Reduce the impact of outliers & avoid discretization of numerical data	data mining task needs to be applied
[103]	Anonymisation (p-sen--sitive, t-closeness	Robust to skewness & similarity attacks in k-anonymity & t-closeness	Generalizes Numerical attributes to categorical
[71]	Anonymisation (generalization & suppression)	Increases the performance	Privacy should be measured using standard measures
[3]	Anonymisation (generalization)	Resistant to table linkage, record linkage, attribute linkage, & probabilistic attacks	—
[105]	Anonymisation (A border based approach)	Less number of missing & artificial rules	Missing rules increases with sensitive rules
[11]	Anonymisation (clustering & s-diversity)	Resistant to similarity & probabilistic inference attack	Execution time increases linearly with the number of clusters
[106]			

Table 3.6: Analysis of PPDM Methods - Combining Cryptographic, Perturbation and Non-perturbation Techniques

Article	Techniques	Strengths	Weaknesses/Challenges
[23]	Perturbation (Additive noise) & Anonymisation (Suppression)	A minimum loss by avoiding generalization, less execution time	—
[107]	Transformation & Anonymisation	Allows implementing privacy separately for different owners	Difficult to gather the willingness of data owners
[1]	Randomization & Generalization	Minimum information loss	A data mining algorithm needs to be applied
[108]	Anonymisation (Suppre- & -ssion) Perturbation	Suppress only certain attributes, low loss	preferable for small k values
[109]	Randomization & SMC	Efficiency and accuracy by combining methods	Apply SMC only to filtered records due to high cost
[110]	Anonymisation & Noise addition	Robust to background information related attacks	When all equivalence classes can't achieve maximum entropy
[111]	Field Rotation & Bining	Preserves statistical properties such as avg, std	Possibility of not catching all the rules
[112]	Local rule sanitization (Padding and filtering)	Sharing rules not data, can be used in distributed environment	Effectiveness of the methods needs to be quantified

The existing PPDM methods consist of techniques that can be used with most of the supervised (Classification and Regression) and unsupervised (Clustering and Association Rule Mining) learning algorithms. Methods in [76, 34, 14, 113, 82] can be applied to clustering methods such as Decision Tree (DT), Naive Bayes (NB), K-Nearest Neighbour (KNN), Multi Layer Perceptron (MLP) and Support Vector Machines (SVM). PPDM methods such as [88, 98, 22, 81] can be used for clustering, while methods like [106, 90, 105] are suitable for Association Rule Mining. There are improved methods [37, 85, 95, 86] that can be used for more than one data mining task. Table 3.6 briefs the PPDM methods that combine perturbation, non-perturbation and cryptographic techniques.

The distribution of PPDM methods among data mining tasks can be seen in Figure 3.3. Most PPDM methods can be applied to classification, followed by association rule mining. While comparably fewer methods have been proposed specifically for clustering algorithms, a considerably high number of PPDM methods can be used for more than one data mining task (classification and clustering, classification and association rule mining). Moreover, some PPDM methods can be used with any data mining algorithm.

For PPDM methods we have reviewed have their strengths and weaknesses. Most of the methods are vulnerable to attacks related to background knowledge, and the researchers have identified this problem. Though we cannot provide a simplified categorization of strengths and weaknesses, we have pointed out the strengths and weaknesses of the reviewed methods in Tables 3.1, 3.2, 3.3, 3.4, 3.5 and 3.6. These tables provide a comprehensive answer to the sub-questions in RQ1 by summarizing all the different PPDM techniques we found out and the strengths, weaknesses, and challenges.

[1]These percentages are derived according to the number of studies covering generic PPDM methods from the total number of papers reviewed.

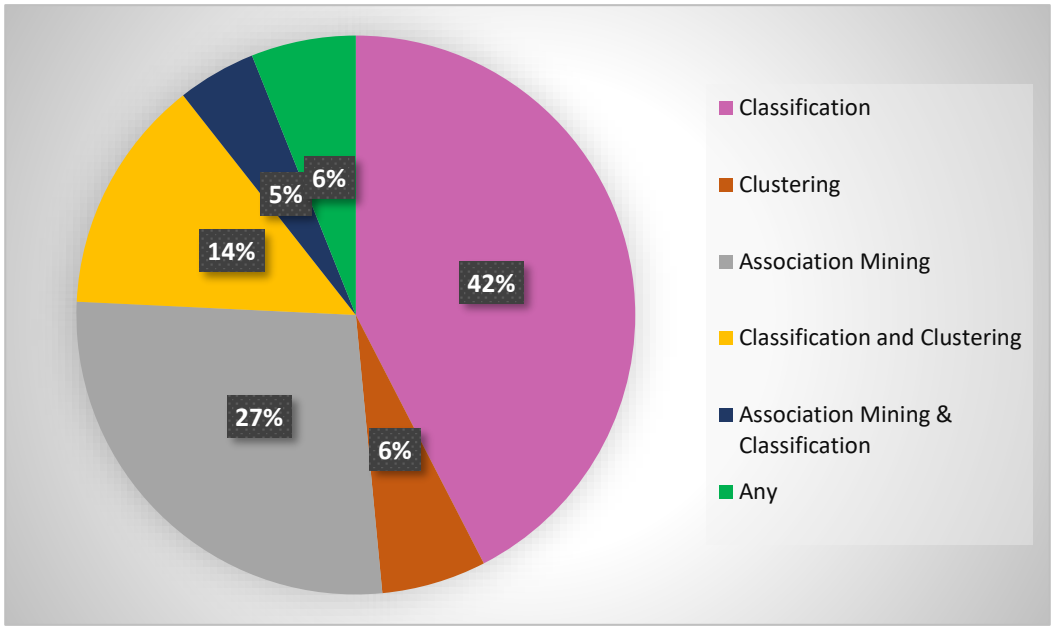

Figure 3.3: Distribution of generic PPDM methods according to the applicability of different data mining tasks[1]

3.3.2 Addressing RQ2 - PPDM for Data Streams

This section discusses the PPDM methods that can be applied to data streams and the challenges we have to overcome when successfully applying PPDM methods to data streams.

Challenges in Data Stream Mining

Most of the generic PPDM methods discussed in Section 3.3.1 cannot be directly applied to data streams due to the challenging behaviour. There are three principal challenges in mining data streams named volume, velocity, and volatility [45, 31]. Data streams have numerous challenges to consider, such as data pre-processing, analysing complex data, dealing with delayed data, and handling concept drift [45, 48]. Privacy is only one concern that data stream mining has to focus on.

Data streams are continuous, transient, and usually unbounded [49, 50, 51] in nature.

Mining data streams is a continuous process, and it cannot be redone as done for the static datasets because it is not possible to access the full set of data at once [52]. Data may reach a high speed, and therefore fast execution is needed [47, 46]. Privacy preservation needs to be performed quickly, and incoming data should be released with a minimum delay. Due to the unbounded nature of the data streams, PPDM methods should be able to cope with a massive volume of data with a fast execution time [32]. Computer memory is too small relative to the vast data volume, and all data cannot be stored [49]. Another challenge of data stream mining is the concept-drift [53, 48, 44] and it affects the PPDM process. Underlying data distribution can change with time, and data mining models should be able to adapt to the concept drift to achieve a good accuracy level [54, 52]. Privacy preservation methods should be able to cope with the effects of the concept drift.

Considering all these facts, data stream mining and privacy preservation are two conflicting tasks [15]. The data stream mining should quickly cope with the memory restrictions, while generic privacy preservation methods require multiple scans over the data, which is time and memory-consuming.

PPDM Methods for Data Stream Mining

The possibility of applying proposed methods to data streams or difficulties of adapting the methods to data streams have not been discussed in generic PPDM methods except in a few. Authors of [111] have mentioned that the proposed field rotation and binning method cannot be applied to data streams. All data should be presented to the binning process, and it cannot be done for data streams because of their incremental behaviour. The combined noise perturbation method proposed in [9] is also difficult to adapt to data streams. According to the authors, the concept of multi-level trust that is being used in this combined perturbation method is challenging to implement for data streams. The condensation-based PPDM method proposed in [85] is the only method we could

find from the selected articles that discuss the possibility of applying the method for data streams. It is suitable for both static data and dynamic data streams. However, for infinite data streams, there is a need for a mechanism to store a fixed number of condensed groups [85].

PPDM methods implemented specifically for data streams and modified versions of generic PPDM methods for use in data streams can be seen in the literature. These PPDM methods have been designed to overcome the above-discussed common challenges of the data streams. We divided these methods into three categories called Perturbation, Non-perturbation (Anonymisation), and others, based on the main techniques they have used. Considering the distribution of these methods, most methods are anonymisation-based non-perturbation methods followed by perturbation methods. A small proportion of PPDM methods use different other distortion techniques such as differential privacy, fuzzy logic, and PCA. Though we roughly categorize these methods into these categories, we observed that there is no clear boundary to define this. Most of these methods are combinations of different categories.

Anonymisation-based non-perturbation methods are among the most used PPDM techniques for data mining. An anonymisation method called FAST was presented in [114] for the fast execution of privacy preservation in data streams with less information loss. This method uses a multithreading technique through k-anonymization and can be used for clustering. Another PPDM method for clustering using k-anonymization for data streams has been introduced in [115]. This method is scalable and can be used with less communication cost and less information loss for distributed data streams. A continuously anonymising method called "CASTLE" has been implemented using k-anonymity and l-diversity in [18]. CASTLE can manage outliers and release data with a minimum delay but can be vulnerable to inference-related attacks. Microaggregation based differential private anonymisation has been proposed in [52] for classification. This method deals with concept drift by applying Kolmogorov-Smirnov statistical test

and minimizes the information loss and possible disclosure risks. The privacy preservation method discussed in [29] uses a frequency discretization technique that group records, similar to anonymisation. Moreover, sliding window-based anonymisation methods were discussed in [50, 49, 116]. The fast anonymisation method proposed in [50] can be used for associate rule mining, and the method in [49] facilitates high-speed data processing with small memory requirements. Anonymisation based on rank swapping in a sliding window discussed in [116] can reduce information loss by swapping selected tuples from the sliding window but can be impractical for infinite data streams.

Perturbation based privacy preservation methods proposed for data stream mining are [47, 51, 117, 32, 29]. In [47], "P2RoCAl", a combination of Condensation, rotation, and random swapping has been proposed for data stream classification. P2RoCAl offers a better accuracy level than similar methods and is robust to data reconstruction attacks, but condensed group size can affect the performance. It can be used in static environments as well. Statistical Disclosure Control (SDC) with different filters such as noise addition, micro-aggregation, rank swapping, and differential privacy has been used in [51]. However, the noise addition filter has the risk of disclosure. An anonymisation method based on noise addition for privacy preservation method has been proposed in [117] for clustering. This method chooses random noise within the subspace limits of the dense and non-dense subspaces to reduce information loss and enhance cluster identification. Random projection-based cumulative noise addition implemented in [32] combines three perturbation techniques, random projection, translation, and noise addition, to achieve good accuracy and privacy level. In addition to traditional independent noise addition, authors of [32] have introduced a novel noise addition method (Cumulative noise addition) that seems to be promising in performance. A random projection-based encryption method discussed in [29] provides a low computational cost with a good privacy level.

Other privacy-preserving methods proposed for data stream mining use different

techniques such as fuzzy logic and PCA [118], differential privacy [38], sliding window [46], and hashing [119]. These methods try to overcome the challenges in data stream mining by using different techniques. We observed that there are numerous methods proposed for data stream-based PPDM that fall under the category of output PPDM [15, 54], which is out of our scope.

Figure 3.4 provides an idea of how data stream-based PPDM methods are spread over different data mining tasks. Like the generic PPDM methods, most PPDSM can be applied to classification. The second-highest applicability was achieved by clustering. A few PPDSM methods have been proposed specifically for association rule mining. Some PPDSM methods can be applied to more than one data mining algorithm, which is a good sign.

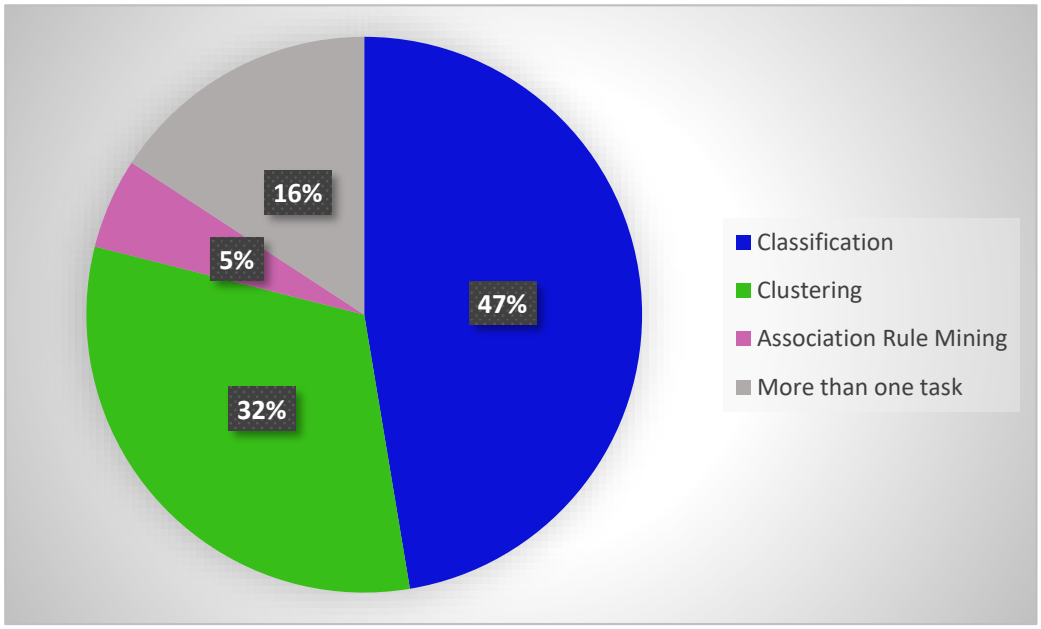

Figure 3.4: Distribution of PPDSM methods according to the applicability on different data mining tasks [2]

[2]These percentages are derived according to the number of studies covering PPDSM methods from the total number of papers reviewed

3.3.3 Addressing RQ3 - Accuracy-Privacy Trade-off

The accuracy-privacy trade-off is the most common issue in PPDM, and it should be addressed appropriately to get maximum performance. If not, the objective of PPDM methods, which is effectively protecting private data while maintaining the knowledge in original data [46, 32], can be violated. It was observed that lots of researchers have identified and discussed this trade-off, while some studies try to provide possible solutions. In this section, by answering RQ3, we discuss to what extent the accuracy-privacy trade-off has been addressed.

Accuracy-Privacy Trade-off in generic PPDM methods

Metrics of accuracy and privacy are helpful when understanding the trade-off between those two properties. Table 3.7 and 3.8 compile different evaluation metrics and measures used to calculate privacy and accuracy for generic PPDM methods. The most commonly used measure of accuracy is the error/accuracy of the data mining task. A few privacy preservation methods, such as anonymisation and rule hiding, use different techniques to measure accuracy. Differential privacy and privacy after various attacks are the most used methods. Moreover, some methods have used metrics such as VD, RP, RK, CP and CK[1] to measure privacy. However, we can see that measuring privacy and accuracy are mostly specific to data mining and privacy preservation techniques.

Regarding the accuracy-privacy trade-off discussion, We first look at the generic PPDM methods discussed in Section 3.3.1.

Research work such as [12, 98, 1, 99, 74, 55, 83, 110, 71] have identified the existing accuracy-privacy trade-off while [3, 27, 91, 9, 75, 16, 103, 90, 113, 85, 81] discuss the matter in detail. Accuracy-privacy trade-off w.r.t. rotation perturbation

[0]*VD - Value Difference, RP - Rank Position, RK - Rank Maintenance, CP - Change of Rank of Attributes, CK -Maintenance of Rank of Attributes

[0]*RMSE - Root Mean Square Error

Table 3.7: Accuracy and Privacy Evaluation metrics for generic PPDM methods

Article/s	Accuracy/Utility	Privacy
[36]	Generalization height, average size of q-blocks, discernibility metric	Bayes-Optimal Privacy based on attacks
[3]	Classification accuracy	Differential privacy based on various attacks
[58]	Utility loss using distance between large item set importance vectors of original and modified data	using the distance of importance between the original and modified data.
[78]	Classification accuracy	Differential privacy
[87]	Classification error	Considering original and perturbed data
[23]	Information loss	In the form of number of characters preserved
[27]	Distortion metrics	VD, RP, RK, CP and CK [1]
[91]	Classification accuracy, information entropy	Using sensitive information vector and a threshold
[88]	Probabilistic Information Loss, Rand Index, Jaccard Index	–
[71]	Confusion matrix	—-
[92]	Classification accuracy	VD, RP, RK,CP and CK
[12]	Classification Accuracy	VD, RP, RK,CP and CK
[93]	Classification accuracy	VD, RP, RK,CP and CK
[60]	Classification accuracy	VD, RP, RK,CP and CK
[95]	Classification accuracy	VD, RP, RK,CP and CK
[94]	Classification accuracy	VD, RP, RK,CP and CK
[61]	Classification Accuracy	VD, RP, RK,CP and CK
[41]	Information loss using SSE	–
[98]	Clustering accuracy	Using sensitive attributes
[120]	Entropy	Differential privacy
[99]	Miss-classification error	Using the probability of estimating original values
[75]	Classification accuracy	–
[108]	Classification accuracy	–
[35]	RMSE [1]	Attacks based on prior knowledge
[104]	Mean Squared Error	Using quasi-identifiers
[9]	Reconstruction Error	Normalized Estimation Error
[89]	Support error and identity error	Using prior and posterior probabilities
[17]	Hiding failure, Misses cost, Dissimilarity	–
[82]	Classification accuracy	Variance
[74]	Classification accuracy	Variance
[11]	Average Equivalence Class Size, Discernibility Metric (DM) cost, Classification accuracy	Linking attacks

Table 3.8: Accuracy and Privacy Evaluation metrics for generic PPDM methods - cont...

Article/s	Accuracy/Utility	Privacy
[55]	Classification accuracy	Attack resistance, privacy guarantee, information entropy
[40]	Classification accuracy	Attack resistance
[43]	Classification error	Differential Privacy
[16]	Predicted joint probability distribution error	Linking attacks
[62]	Classification accuracy	Percentage of completely suppressed attributes
[28]	Classification accuracy and MSE	Inference attacks
[111]	Classifier accuracy and Percentage of matching rules	–
[83]	Classification accuracy	Based on sensitive attributes
[80]	RMSE	Variance metric
[100]	Clustering accuracy	By verifying original data is modified or not
[10]	Classification accuracy, F1 score, Area Under the Curve	VD, RP, RK,CP, CK and secrecy
[105]	Dataset dissimilarity, missing rules, artificial rules	–
[121]	Classification accuracy	Encryption - trial and error, text analysis
[106]	Using number of rules	–
[14]	Classification accuracy	Confidence interval metric
[85]	Classification accuracy	Using different condensed group sizes
[73]	Correlation metric	Differential Entropy
[122]	–	Breach probability using attacks
[30]	—	Based on attacks
[34]	Classification accuracy	Multi-column privacy metric
[101]	Average group size, discernibility metric	–
[26]	Precision, Recall, F-score	Re-identification risk
[123]	Loss rule rate	Hiding failure rate

has been discussed with extensive experimental results in [34]. Authors of [122] have discussed the nature of this trade-off in distance preserving PPDM methods with examples. Accuracy-privacy behaviour of p-sensitive, t -closeness was discussed in [103]. In this method, When p decreases, utility also decreases, but good in high t values. Accuracy-privacy trade-off of additive multiplicative perturbation has been discussed in [109]. This method shows that the error and privacy increase when the trust level is increased, that denotes there is a trade-off between accuracy and privacy.

Numerous research work has proposed different solutions to optimise or reduce the accuracy-privacy trade-off. Combining suppression and perturbation to minimize the loss caused by generalization in anonymisation has been proposed in [23]. Performing the perturbation only on sensitive attributes to achieve a high accuracy level while maintaining good privacy using NMF and SVD has been discussed in [61]. The t-closeness anonymisation proposed in [101] discuss how the t parameter can be tuned to achieve a good trade-off, while [41] tries to achieve a better trade-off by applying t-closeness through micro-aggregation. Anonymisation-based clustering methods proposed in [11, 62] shows that the number of clusters formed determines the trade-off as the number of clusters increases, accuracy increases, and privacy decreases. Authors of [42] try to achieve a better accuracy-privacy trade-off by combining PCA and additive noise, but privacy increases while accuracy decreases when more noise is added. The differential privacy-based approach proposed in [43] uses an ensemble classifier to calculate the error, and this process is repeated until a pre-defined threshold is achieved. The Laplace noise added for the differential privacy is readjusted if it cannot be achieved.

A rotational transformation was combined with a translation implemented in [80] to get a good privacy level with a low accuracy loss. Authors of [109] try to balance the trade-off between accuracy and privacy using a multi-group approach. The perturbation method "NRoReM" [10] has been implemented to optimise the accuracy-privacy trade-off by combining normalization, geometric rotation, linear regression, and scalar

multiplication. In addition to the above discussed methods, research work such as [57, 94], [60, 95], [14] and [40] also implemented different techniques to optimise the accuracy-privacy trade-off.

Some research work have proposed interesting PPDM techniques but have not paid much attention to the accuracy-privacy trade-off [58, 78, 112, 96, 86, 92]; [97, 22, 104, 84, 124, 73, 125].

Some PPDM methods proposed for data streams have also made attempts to optimise the accuracy-privacy trade-off. According to the challenges identified in 3.3.2, it is clear that handling the trade-off issue in streaming environment is rather complex. However, some methods have discussed this issue [126, 46, 18] while some have tried to address it using different techniques [54, 52, 38, 55, 32].

Sequential Backward Selection (SBS) of the greedy algorithm and k-fold cross-validation to select the optimal mode in NB classification has been used in [54] to achieve a balanced accuracy-privacy trade-off. Micro-aggregation based differential private stream anonymisation has been proposed in [52], and the trade-off has been evaluated using disclosure risk and Area Under the Curve (AUC) of the classifier. The differential privacy-based PPDM method "SEAL" [38] has been proposed as a solution to the accuracy-privacy trade-off in data stream mining. To optimise the trade-off, it provides flexibility to select privacy parameters according to the domain and dataset to improve privacy and maintain the shape of the original data distribution after noise addition to improve accuracy. P2RoCAl [55] tries to achieve the same goal by combining condensation and rotation. Random projection-based cumulative noise addition [32] tries to add noise with a small variance cumulatively to minimize the effect to accuracy while maintaining good privacy. This method has experimentally proven to achieve a better trade-off. Authors of [127] have proposed a novel random projection-based noise addition method using [32] as the base technique. It uses the effect of logistic function to control the noise level but still adds it cumulatively

to achieve a high accuracy level. Meanwhile, some interesting PPDM methods in data stream mining do not include any discussion about accuracy-privacy trade-off [114, 118, 115, 51, 119, 29, 116, 117, 50, 49].

Figure 3.5 gives an overall idea about to what extent the PPDM research community has paid attention to the accuracy-privacy trade-off. It can be seen that there is a lack of attention to the accuracy-privacy trade-off in data stream-based PPDM methods relative to the generic PPDM methods. However, in both areas, some research work tries to give solutions to this issue (31.51% in generic PPDM and 26.32% in Stream-based PPDM).

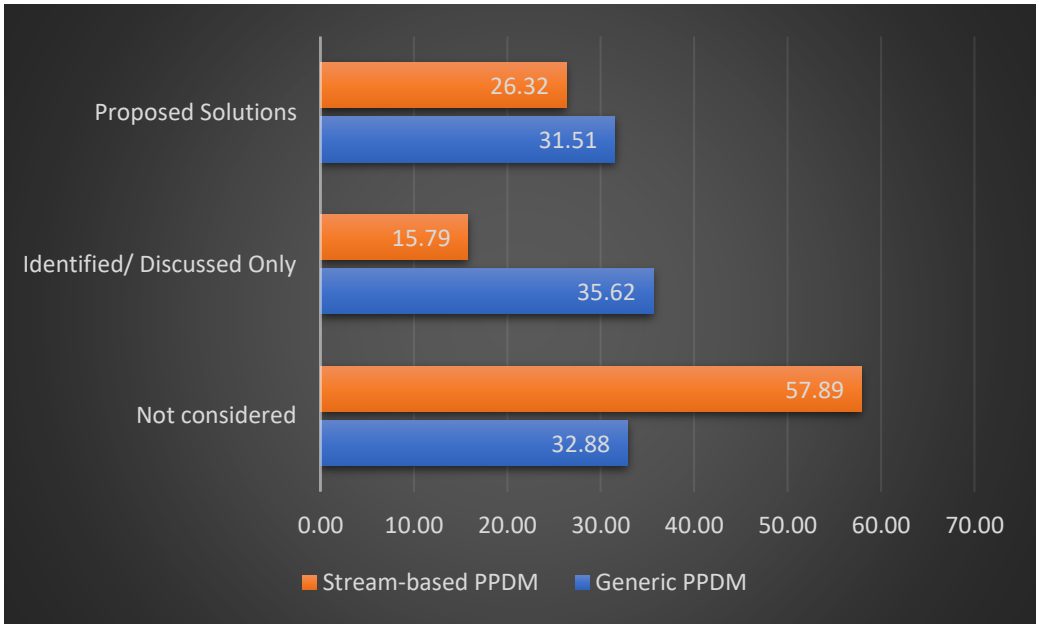

Figure 3.5: Consideration of accuracy-privacy trade-off in existing PPDM research [3]

3.4 Discussion

In this section, we discuss how we addressed the gaps in the existing secondary studies by answering the formulated research questions.

[3]These percentages are derived according to the number of studies covering accuracy-privacy trade-off in PPDM and PPDSM from the total number of papers reviewed

In RQ1 and its subsections, we tried to identify the generic PPDM methods, the strengths and weaknesses of the identified methods, and the applicable data mining tasks. There are two broad categories of PPDM methods called input and output PPDM. We considered input PPDM methods as output PPDM methods focus on changing data mining output, which is a different scenario. After reviewing selected primary studies, it was found that a plethora of techniques have been proposed for privacy preservation in data mining. We divided them into four main categories: Secure Multiparty Computation, perturbation methods, non-perturbation methods, and combinations of the above categories. These methods cover most supervised and unsupervised learning techniques (Classification, Clustering, Association Rule Mining) in data mining. The majority of the PPDM methods can be used for classification, and there are a considerable number of PPDM methods that can apply to more than one data mining algorithm (Refer Figure 3.3). Also, we observed that several studies lack standard accuracy and privacy evaluations after applying a data mining algorithm. This is a section where PPDM studies can be improved. Applying new techniques to a data mining algorithm using real or synthetic data sets provides validity and clarity and helps identify the sections that need improvements.

These generic PPDM methods have different strengths and weaknesses considering privacy, accuracy, time consumption, and more. The main reason for this is the nature of the techniques used to preserve privacy. Different techniques affect different characteristics differently. Because of this variety of advantages and weaknesses, when selecting a PPDM method, several factors such as the size of the dataset, domain, and contained sensitive data should be considered. There are no pre-defined criteria to decide on the appropriate PPDM methods for a specific dataset. Properties of the dataset and the characteristics of the privacy preservation technique should be considered when making this decision. All the facts found from the review have been summarized in Table 3.1.

A categorization model of all the input PPDM methods was created by analysing the

extracted data and considering the existing categorizations. This model helps to grasp the overall picture of existing PPDM methods. Figure 3.6 illustrates the categorization model created, summarizing all the generic PPDM techniques.

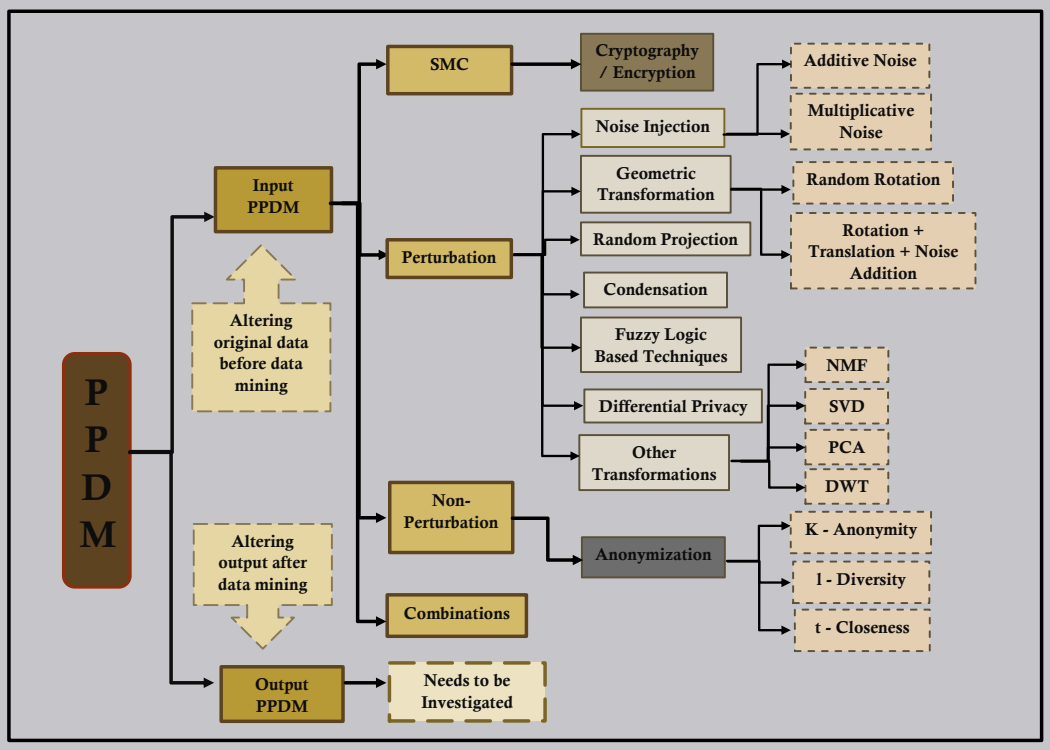

Figure 3.6: Categorization model of generic PPDM Methods

For RQ2, we investigated the applicability of PPDM techniques for data stream mining. It was observed that most of the generic PPDM methods could not be used directly for data stream mining because of the challenging nature of data streams. This includes incremental nature, high speed, vast or infinite, and possible concept drifts. Therefore, generic PPDM methods need improvements and amendments to be successfully used in data stream mining. It was observed that most of the generic PPDM methods do not discuss its applicability to data streams, which we suggest is something to be considered. If there is such discussion, it would be helpful for future development. Numerous PPDM methods proposed for data stream mining improve existing generic

PPDM methods or combinations of different techniques. Most of these methods use anonymisation techniques to preserve privacy in data streams. Perturbation methods such as noise addition are also applicable because noise can be added independently to a single record at a time. While these methods were able to overcome most of the challenges in data stream mining, they still have some concerns, such as concept drift handling and time and computational complexity, that need to be improved. We also looked into the applicability of stream-based PPDM methods in different data mining techniques. It was found that the majority of proposed stream-based PPDM methods can be used for classification, followed by clustering. Interestingly, several PPDSM methods are available for more than one data mining task, which shows the generalizability of PPSM methods in data stream mining. (Refer Figure 3.4).

The well-known trade-off between data mining accuracy and data privacy was investigated in answering RQ3 to determine how it has been addressed. A majority of research work has identified and discussed the accuracy-privacy trade-off, but a little effort has been made in proposing techniques to address the issue. The conflicting nature of accuracy and privacy is the main reason for this. Though some methods have been proposed with the pure intention of optimising the accuracy-privacy trade-off, it is impossible to achieve ideal values for both measures simultaneously. Some methods have succeeded in achieving this to some extent, but there is still considerable room for improvement. The techniques proposed to optimise the trade-off in data stream mining are less in number compared to generic PPDM methods (Refer Figure 3.5). The proposed methods to optimise the trade-off include different techniques and methods such as trying to preserve more statistical information, making changes only to sensitive attributes, parameter optimisation, and considering users' privacy requirements. We believe the optimisation of the accuracy-privacy trade-off has not received the attention it deserves, especially in data stream mining.

3.5 Conclusion and Future Directions

We can conclude that a high number of studies have been carried out to preserve privacy in generic data mining but less work in stream-based data mining. A positive remark is that these proposed Privacy-Preserving Data Mining (PPDM) methods can be used in different data mining algorithms in both supervised and unsupervised learning. However, all these methods have some strengths and weaknesses due to the techniques used to preserve privacy. Though the PPDM research community has identified the trade-off between data mining accuracy and data privacy, there is a lack of research that tries to implement techniques with extensive experimentation to optimise this trade-off. Especially, data stream mining has a great need of techniques to optimise the accuracy-privacy trade-off in data stream mining. All our findings from this study can be listed as follows;

1. There are a plethora of studies that propose different privacy-preserving techniques for PPDM.
    - The existing generic PPDM methods can be divided into four categories, namely *SMC, perturbation, non-perturbation*, and *combinations of the above*.
    - These PPDM methods can be used for different data mining algorithms, including classification, clustering, and association rule mining. Numerous methods work well on more than one data mining algorithm.
    - The existing PPDM methods have different strengths and weaknesses in several areas, including accuracy, privacy, and time complexity. These are caused by the techniques used to preserve privacy.
2. Different studies have been implemented to preserve privacy in data stream mining

- Data streams behave differently than static datasets due to characteristics such as high volume, high speed, and concept drift. Therefore, privacy preservation in data stream mining is rather challenging.
- Most of the generic PPDM methods cannot be used for data streams and need improvements to adapt to the behaviour of data streams.

3. The trade-off between data mining accuracy and data privacy is one of the main issues in PPDM that needs more attention.
    - Many studies have identified and discussed the accuracy-privacy trade-off in PPDM.
    - Numerous studies have proposed improved and advanced PPDM techniques to optimise this trade-off in generic PPDM.
    - There are only a few studies that focus on optimising the accuracy-privacy trade-off in data stream mining.

We only considered input PPDM methods in this study. Therefore, as a future direction, we would like to suggest an investigation on output PPDM methods and how they can optimise the accuracy-privacy trade-off as it seems to be used quite often in data stream mining.

Chapter 4

Prelude - Manuscript 2

Developing a perturbation method to preserve privacy without sacrificing accuracy is not an easy task. Noise addition is one of the primary perturbation methods. However, accuracy suffers significantly when noise is added with a high noise variance. Cumulative noise addition [32] proposes a novel way of adding noise while maintaining high accuracy and privacy level. Hence, it achieves a good trade-off between privacy and accuracy. However, it fails in maintaining this stable accuracy-privacy trade-off when there are many records. This can be vital for data streams, as they can contain infinite records. Therefore, cumulative noise addition should be improved for data streams with optimal accuracy and privacy trade-off.

Manuscript 2 [127] focuses on developing an enhanced noise addition-based perturbation method that provides the desired balance between data privacy and data mining accuracy. We implemented seven variations of noise addition methods based on the cumulative noise addition proposed in [32]. Reasons behind selecting [32] as the base method are its superior performance over traditional noise addition and the feasibility of applying it to data streams. We introduced several novel techniques to cumulative noise and controlled the possible excessive noise level when there are many records. These techniques include logistic function [128] based cycle-wise noise addition, use

of absolute noise values, and noise resetting. Experiments prove that these techniques effectively control the noise level of cumulative noise addition.

We selected the best-performing perturbation method out of seven variations as the winning method. This method provides the best trade-off, considering privacy and accuracy. Ultimately, manuscript 2 addresses the first research question, **"What is the most appropriate perturbation method that expresses the optimal trade-off between data privacy and data mining accuracy?"**.

Chapter 5

Optimising the Trade-off Between Classification Accuracy and Data Privacy in the Area of Data Stream Mining (Manuscript 2)

5.1 Introduction

People release their private data in various situations such as medical check-ups, requesting a bank loan, and applying for employment in their day-to-day lives. The organisations use these data to increase performance by making predictions. Assume that Sam has done a medical check-up a few months back, revealing that he has AIDS and wants to keep it a secret from society. However, suddenly his neighbourhood gets to know about this situation because a person who knows Sam personally has participated in the data analysis process of that health organisation has identified Sam from his personal details and revealed his situation. Our data is not protected anymore, and people can access this data to attack our personal lives. Therefore, organisations need a

method to protect their customers' data when they use those to make predictions and analyse performance.

Privacy-Preserving Data Mining (PPDM) helps protect data privacy when it is being used for data mining purposes [25, 8]. Data perturbation is one technique that falls under PPDM [28], which is suitable for data streams. It alters the original data values that make it difficult to recover by unauthorised people using recovery techniques but still manages to maintain the relevant properties of the dataset that are useful for data mining purposes [32]. It converts data to another form, so anyone cannot identify individuals by looking at their personal data.

One of the critical success factors of Data Mining is the availability of high-quality data to support the generation of accurate models [16]. On the other hand, sensitive data cannot be published in its original form, and thus different types of data perturbation methods have been proposed to maintain privacy. However, data perturbation can negatively affect the accuracy of prediction models. When data perturbation techniques are applied to increase data privacy, it decreases the accuracy of the classification models as perturbation distorts the original data values [38, 23]. This trade-off between data privacy and classification accuracy is an inherent problem that needs to be investigated in this area.

Increasing privacy decreases the accuracy and vice versa, which is one of the most common issues in PPDM [32, 23]. We tried to find a suitable perturbation method that minimizes the difference between privacy and accuracy values. This ultimately optimises the accuracy-privacy trade-off. Our research aims to propose a method to optimise the trade-off between accuracy and privacy to enhance the performance of data mining tasks. This article proposes seven variations of cumulative noise addition methods. These methods combine novel techniques such as logistic function [128], cycle-wise noise addition, noise resetting, and absolute noise values into an existing perturbation method called "Random Projection-based Cumulative Noise Addition"

proposed in [32].

Random projection-based cumulative noise addition [32] is the most recent work done in data perturbation using noise addition. It adds noise cumulatively to the data stream, while traditional noise addition adds the noise independently to each record. This method achieved high privacy and accuracy values, optimising the accuracy-privacy trade-off (More details will be discussed in Section 5.2.1). However, this method has practical issues in the long-term run, as it has been designed to apply to data streams. The main concern is that we cannot continue adding the noise cumulatively for a long time since noise can overpower the actual data values after some time. If this happens, it can drastically decrease the accuracy because data values become highly distorted in the long-term run. The classifier tends to learn the noise instead of the data. Our work is motivated by this issue, and we conducted this research to find possible ways to improve the work done by [32] by controlling the maximum noise added to the stream.

We mainly focused on developing an advanced perturbation method that can be used with any classification algorithm suitable for data stream mining. We used Adaptive Random Forest (ARF) as the experiments' classifier and measured the accuracy and privacy for cumulative noise addition in cooperated with different techniques to control the total noise. The main contribution of this research is an improved noise addition-based perturbation method that can be used to optimise the accuracy of privacy trade-off in a data streaming environment. In summary, the paper makes the following contributions to the field:

- Introducing seven different variations of cumulative noise addition methods which can be used as noise addition-based data perturbation techniques.
- Developing algorithms for Linear Cumulative and Logistic Cumulative Methods.
- An effective cumulative noise addition-based perturbation which can be used to optimise the trade-off between data privacy and classification accuracy.

- An evaluation of the performance of seven different cumulative noise addition methods concerning the state-of-the-art, using relative error and breach probability for different cycle sizes and growth rate values.
- A vulnerability analysis of the best performing perturbation method from the experimented seven variations of methods.

The remainder of this paper is organised as follows; Section 5.2 reviews the existing data perturbation methods, highlighting the strengths and limitations of each method. Section 5.3 consists of the proposed methodology that describes two main types of cumulative noise addition methods. The experimental setup, results and an extensive analysis of the results are presented in Section 5.4. Section 5.5 and 5.6 provide a discussion, conclusion, and future work of this work, respectively.

5.2 Related Work

Data mining is a prominent area in today's world that uses data to make predictions and improve organisations' performance by making correct decisions. This includes building learning models using supervised (ex: classification and regression) and unsupervised learning (ex: clustering and association rule mining) approaches. Modelling using different learning algorithms has been discussed in research work such as [129] which uses Multi-Relational Classifier Based on Canonical Correlation Analysis, [130] which applies the signatures to expert systems modelling using rules, and [131] that investigates on using different machine learning approaches to classify pedestrians' events based on IMU and GPS. These works prove that the data mining models can be successfully used in various application areas. We only focus on classification and, more specifically, classification algorithms that can be used to learn from data streams. Massive Online Analysis (MOA) [132] has developed a set of algorithms such

as Hoeffding Tree, Hoeffding Adaptive Tree, and Adaptive Random Forest [133] which can be used in data streaming environments.

Different data perturbation methods have been proposed in the literature to enhance data privacy in the past few decades. Data perturbation methods such as random rotation [34] and random projection [35] and Geometric perturbation [76] maintains the pair-wise distances of the records, which are helpful for data mining tasks, while methods such as additive noise [134] and condensation [37] maintain the dataset properties but sacrifice the record-wise properties.

In additive noise or randomisation, a high privacy level can be witnessed when the noise variance increases since original values are highly distorted and result in low accuracy. Though this method distorts the original data values massively, it is still vulnerable to different types of privacy breaching attacks. In [135] authors have discussed five attack techniques that can be used against randomisation. This includes Spectral Filtering [136], Singular Value Decomposition [137] and Principal Component Analysis (PCA) [138] which are based on the Eigen analysis to recover the original data from perturbed data. In addition to these, Maximum A Posteriori attack (MAP) [138] and Distribution attack [135] were also discussed as possible attack types to filter out original values from additive noise.

Random rotation or rotation perturbation was proposed to maintain the record-wise properties in a dataset to obtain a high accuracy or utility level while preserving the privacy of the data. In [39] rotation perturbation was defined as a matrix multiplication that multiplies the original data matrix by a rotation matrix which results in a perturbation matrix with the same number of records and features as the original data matrix. This transformation is orthogonal; hence, the distance between perturbed records is similar to the distance between original records, which implies that the perturbed dataset gives similar classification accuracy to the original classification results. The accuracy of the classification results is high, but several privacy-breach attacks have worked well with

this method. Distance inference attacks [76], Independent Component Analysis (ICA) [76] and [139], Known Input/Output attack [139], [122] and PCA [139] can be used to reconstruct the original data from rotation perturbation. It has been proved that the random rotation can be perfectly reversed using a few known Input/Output pairs [139].

Modified and combined versions of rotation perturbation have been proposed to overcome the privacy issue of rotation perturbation. A combination of random rotation and randomization that addresses the distance inference attack known as the general linear transformation was proposed by [140]. Authors [141] investigated privacy vulnerabilities and found out the proposed method is vulnerable to attacks in case of available background information. Random rotation, followed by a translation that addresses attacks to the rotation centre, is proposed by [76]. Attacks based on background knowledge have been developed for this method by [122].

Random projection uses the technique of matrix multiplication. It has been proposed by [35] to address the privacy issues that arose from random rotation while maintaining the distance between records to achieve a high accuracy level of the classification results. "Random projection refers to the technique of projecting a set of data points from a high-dimensional space to a randomly chosen lower-dimensional subspace" [35] by multiplying with a random matrix. The main idea of random projection is motivated by the Johnson-Linden Strauss Lemma [35]. This Lemma allows decreasing the dimensionality while maintaining the pair-wise distance of two points within an arbitrary small factor [35]. Random projection is not vulnerable to an Independent Components Analysis (ICA) based attack because the perturbed dataset's reduced dimensionality results in an under-determined system of linear equations. However, it is still vulnerable in cases where the attacker is equipped with prior knowledge about the original dataset.

Attacks to the random projection method were discussed in surveys such as [135] and [142], and these attacks are based on some prior knowledge of the original data.

Known Input/Output attack and known projection matrix attack are two attack types that can be used against random projection. The general idea of the known Input/Output attack is that the attacker has prior knowledge of a few original records and their respective perturbed records. The rest of the original data records can be recovered using those known record pairs. Authors of [143] have discussed the known Input/Output-based MAP attack. It can be used to attack random rotation perturbation even when a collection of known Input/Output pairs is less than the number of features of the original dataset known. In [144] authors have proposed to shuffle records before publishing to rectify this problem, but unfortunately, the method cannot easily be adapted to a data streaming environment.

A novel research "Random-projection based cumulative noise addition," which combines random projection, noise addition, and translation, was proposed by [32] to enhance the privacy of the perturbed data using random projection. Instead of traditional additive noise, authors have proposed a novel method called cumulative noise addition. It has been proved in this research work that it is possible to achieve a considerable level of privacy and accuracy by combining these perturbation techniques. Nevertheless, the system must face a massive noise that may overtake the original data values when the stream unfolds. That can reduce the accuracy is the main issue of this scenario.

Though the perturbation methods in the literature have proposed different techniques to increase the privacy level of data while maintaining the accuracy of the classification results, it does not seem to be entirely achieved by the existing related work. The techniques that provide better privacy levels lack the expected accuracy of the data mining results. The methods that provide significant accuracy are vulnerable to different privacy attacks. Therefore, it still requires a method that can optimise the trade-off between classification accuracy and data privacy for the betterment of the field.

5.2.1 Existing State-of-the-Art Work

After deeply analysing the experimental results of the method, random projection-based cumulative noise addition (referred to as LRW) proposed by [32] which is the most advanced noise addition technique developed, was used as the base perturbation method for this study.

In the process of random projection, the original data matrix $X(m \times n)$ is multiplied by a random matrix $R(k \times m)$ to generate a perturbation matrix $Y(k \times n)$. Each element of R is independent and identically distributed (i.i.d.) and is generated from a Gaussian distribution with mean 0 and variance σ_r^2 [139]. The projection process is represented as, $Y = \frac{1}{\sqrt{k\sigma r}}RX$ and the multiplication by the factor $\frac{1}{\sqrt{k\sigma r}}$ ensures that the column-wise inner product is preserved (records are represented as columns). According to the Johnson-Linden Strauss Lemma, random projection can be considered an approximately distance preserving perturbation. The problem with the distance preserving perturbation is that the records closer to the origin are less perturbed than records far away from the origin [76]. This allows uncovering some original records easily, even without a complex attack. To avoid this vulnerability issue, [76] has proposed a random translation method, which applies the same translation to each record. This extends the perturbation method to $Y = \frac{1}{\sqrt{k\sigma r}}RX + \Gamma$. "Applying a constant translation to all records does not affect many data mining tasks, but an attacker must sacrifice one known Input/Output pair to account for it" [32].

To add an additional degree of uncertainty to any recovery attempt that attempts to reverse random projection, the authors of [32] have introduced two types of noise addition: independent noise (randomisation) and cumulative noise. Our study focuses on the cumulative noise addition since it has successfully reduced the trade-off between privacy and accuracy to some extent. In the process of cumulative noise addition, i.i.d. Gaussian noise values are added to each record, but additionally, each of these random

values is also added to every subsequent record in the stream and can be represented using $Y = \frac{1}{\sqrt{k\sigma r}} RX + \Gamma + \Psi$. The cumulative noise addition is useful for resisting known Input/Output attacks since the attacker has to face increasing noise levels when the data stream unfolds. This allows creating the same impact with a small variance of cumulative noise instead of a large variance setting with independent noise. "Cumulative noise is designed to achieve a similar privacy benefit as independent noise, but with less impact on the accuracy of data stream mining algorithms" [32].

By analysing the experimental results, cumulative noise addition has provided low classification error with a marginally higher breach probability when compared to independent noise addition. A low classification error results from cumulative noise addition, which adds a small noise variance; hence, the distortion of original values is also minimal. However, the privacy level of cumulative noise addition is expected to increase with time as the data stream unfolds. When considering the privacy accuracy trade-off, cumulative noise addition outperformed the independent noise addition method.

Considering the data streaming environment, the stream has to face a massive amount of noise in a cumulative noise addition environment when the data stream unfolds. It is a good approach when considering the privacy aspect. However, it can negatively affect the accuracy as the classifier tries to learn the noise values rather than the original values because of the high distortion of the original data values. Therefore, a method that can control the maximum amount of noise added to the stream must be combined with the cumulative noise addition. Then it will help achieve a high accuracy level by controlling the maximum noise level while still adding it cumulatively to maintain the high privacy level.

5.3 Proposed Approach

Facing an enormous amount of noise with the time when the data stream unfolds is an inherent issue that can be experienced in any cumulative noise additive environment. Therefore, a technique that controls the maximum noise level added to the stream needs to be incorporated to enhance the accuracy without having a considerable effect on privacy.

Two main categories of cumulative noise addition, Linear Cumulative Noise Addition and Logistic Cumulative Noise Addition were introduced to achieve this objective. Under these two main approaches, different techniques such as cycle-wise noise addition, use of absolute noise values, and noise resetting method were performed to investigate the behaviour of privacy and accuracy. The ultimate objective of using these different techniques is to control the maximum noise level while still adding it cumulatively. Doing so it is expected to increase accuracy while maintaining the high privacy level provided by the cumulative noise addition and optimising the trade-off between privacy and accuracy. We used the state-of-the-art method [32] as the base of our research and expanded and improved it by cooperating with possible techniques that help to minimise the accuracy-privacy trade-off. Table 5.1 summarises all the symbols we used to build up the algorithms in Sections 5.3.1 and 5.3.2.

Table 5.1: Symbol Table

Symbol	Meaning
Y	Perturbed Dataset
R	Random Projection Matrix
X	Original Dataset
Ψ	Cumulative Noise
Γ	translation Matrix
ϖ	Logistic Cumulative Noise
cs	cycle Size

5.3.1 Linear Cumulative Noise Addition Methods

Techniques such as cycle-wise noise addition, noise resetting, and using absolute noise values have cooperated with the linear cumulative noise addition method to control the noise level of the stream. Following are the detailed description of four different variations experimented with, including the base method.

- LRW (Linear Random Walk without Resetting) - Cumulative noise is added in a random walk fashion. No absolute values and no noise value resetting are used. Note - This is the base method of the research [32]

  * $Y = \frac{1}{\sqrt{k\sigma_r}}RX + \Psi + \Gamma$

- LRWR (Linear Random Walk with Resetting) - Cumulative noise is added cycle-wise in a random walk. No absolute values are used. At the end of each cycle, the noise level is reset to zero. (cs is the cycle size.)

  * $Y = \frac{1}{\sqrt{k\sigma_r}}RX + \Sigma_{i=-cs}^{cs}(\Psi) + \Gamma$;when $i = cs$, $\Psi = 0$

- LAR (Linear Absolute with Resetting) - Cumulative noise is added in cycles using absolute noise values. In the first half and second halves of the cycle, noise values were added and subtracted. At the end of each cycle, the noise level is reset to zero.

  * $Y = \frac{1}{\sqrt{k\sigma_r}}RX + (\Sigma_{i=-cs}^{0}(+|\Psi|), \Sigma_{i=0}^{cs}(-|\Psi|)) + \Gamma$;when $i = cs$, $\Psi = 0$

- LA (Linear Absolute without Resetting) - Cumulative noise is added in cycles using absolute values. As with the scheme above in LAR, but with no noise resetting at the end of the cycle.

  * $Y = \frac{1}{\sqrt{k\sigma_r}}RX + (\Sigma_{i=-cs}^{0}(+|\Psi|), \Sigma_{i=0}^{cs}(-|\Psi|)) + \Gamma$

Figure 5.1 graphically illustrates the four variations of linear cumulative noise addition methods.

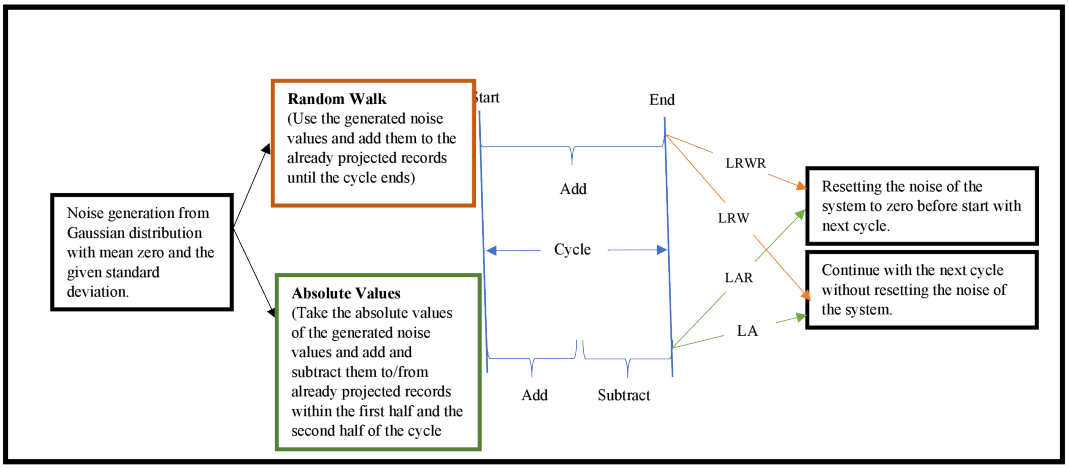

Figure 5.1: Process of Linear Cumulative Noise Addition

The entire process involved with these four variations of linear cumulative noise addition can be explained using the following pseudo-code (See Algorithm 1).

Algorithm 1 Pseudo Code - Linear Cumulative Noise Addition Methods

Input: Cycle Size (cs), Cumulative Noise Sigma (σ), Projected Dataset (X)

Process: **While** End of X
for each Record i in each cycle (defined by cs)
do
Generate noise (n) from Gaussian distribution $N(0, \sigma^2)$
Add n cumulatively to the records $(y_i = x_i + n_i)$
Use n_i **OR** absolute n_i
End Cycle
System Noise = 0 **OR** System Noise = Current Cumulative Noise
Continue Next Cycle

Output: Perturbed Dataset (Y_i)

5.3.2 Logistic Cumulative Noise Addition Methods

The well-known logistic function influences logistic cumulative noise addition [128] (illustrated in Figure 5.2). The expectation of using this concept in a cumulative noise addition environment is to control the noise in the system further while still adding it cumulatively. The mathematical representation of the logistic function is shown below. L = maximum value of the function, e = Euler's number, k = logistic growth rate and x_0 = x value of the mid-point has been used here.

$$\dot{\mathbf{f}}(x) = \frac{L}{1 + e^{-k(x-x_0)}} \tag{5.1}$$

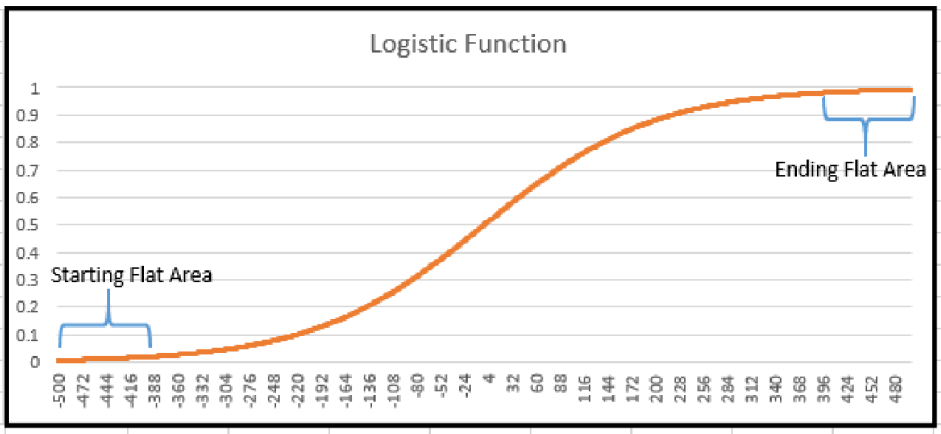

Figure 5.2: Logistic Curve
[128]

In this logistic cumulative noise addition environment, the noise was generated from a Gaussian distribution with a mean of zero and the standard deviation $f(x) \times \sigma$ (return value of the logistic function multiplied by the standard deviation of cumulative noise). From here onward, we refer to the noise generated according to the logistic process as ϖ. Since the logistic function allows controlling the growth rate and the maximum value returns from the function, it provides an effective way to control the noise addition rate and the maximum noise level, respectively. It also provides a good platform for

cycle-wise noise addition. Therefore, cooperating logistic function with cumulative noise addition appears to be a promising technique to control cumulative noise level that can positively affect the accuracy, optimising the trade-off between privacy and accuracy. Here are the experimented four variations of the logistic cumulative noise addition methods.

- SRW (Logistic Random Walk without Resetting) - Noise is added cycle-wise in a random walk fashion. No absolute values and no noise value resetting is used
  - $Y = \frac{1}{\sqrt{k\sigma_r}} RX + \Sigma_{i=-cs}^{cs}(\varpi) + \Gamma$
- SRWR (Logistic Random Walk with Resetting) - Logistic noise is added cycle-wise in a random walk fashion. No absolute values are used. At the end of each cycle, the noise level is reset to zero.
  - $Y = \frac{1}{\sqrt{k\sigma_r}} RX + \Sigma_{i=-cs}^{cs}(\varpi) + \Gamma$;when $i = cs$, $\varpi = 0$
- SAR (Logistic Absolute with Resetting) - Logistic noise is added in cycles using absolute values. As with LAR, the noise level is reset to zero at the end of each cycle.
  - $Y = \frac{1}{\sqrt{k\sigma_r}} RX + (\Sigma_{i=-cs}^{0}(+|\varpi|), \Sigma_{i=0}^{cs}(-|\varpi|)) + \Gamma$;when $i = cs$, $\varpi = 0$
- SA (Logistic Absolute without Resetting) - Logistic Noise Addition in cycles using absolute values. As with LA, no noise value resetting is used.
  - $Y = \frac{1}{\sqrt{k\sigma_r}} RX + (\Sigma_{i=-cs}^{0}(+|\varpi|), \Sigma_{i=0}^{cs}(-|\varpi|)) + \Gamma$

A graphical representation of the steps conducted in logistic cumulative noise addition methods are shown in Figure 5.3

The following pseudo-code summarizes the process followed with logistic cumulative noise addition methods (See Algorithm 2).

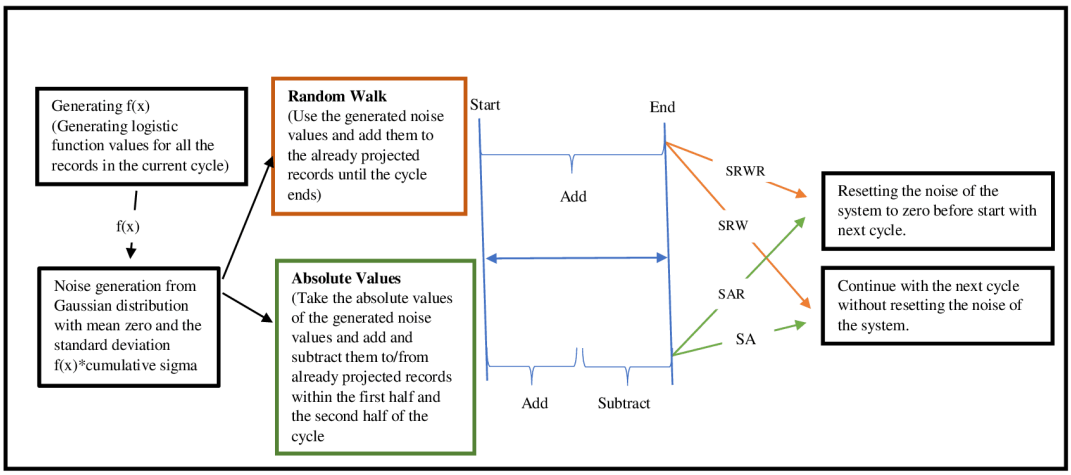

Figure 5.3: Process of Logistic Cumulative Noise Addition Methods

5.3.3 Classification and Evaluation Process

The proposed methodology uses ARF [133] as its learning algorithm. Since we use this for a data streaming or online learning environment, some specific requirements of data streams need to be considered. Processing one example at a time for at most one time, using a limited amount of time and memory, and should be ready to predict at any time are the essential requirements of a data stream [132], [133] and [145]. By considering all these requirements, training and testing of the model can be carried out in two possible ways, namely the holdout method and interleaved test-then-train or prequential method [132], [133] and [145].

The holdout method measures the performance of a single holdout set and is most useful when the division between train and test sets has been pre-defined. On the other hand, in the prequential method, each example is used to test the model before it is used to train the model, and accuracy is incrementally updated. When this is performed in the correct order, the model is constantly being tested on the samples it has not seen. The prequential method does not need a holdout set or pre-defined training and testing sets that take the maximum use of data available [132]. These properties make this method more suitable for a data streaming environment that evolves and learns incrementally.

Algorithm 2 Pseudo Code - Logistic Cumulative Noise Addition Methods

Input: Cycle Size (cs), Cumulative Noise Sigma (σ), Projected Dataset (X), Maximum Value of Logistic Function (L), Growth Rate (k)

Process: **While** End of X
for each Record i in each cycle (defined by cs)
do
Calculate the value of Logistic Function $f(x) = \frac{L}{1+e^{-k(x-x_0)}}$
Generate noise (n) from Gaussian distribution $N(0, f(x) \times \sigma)$
Add n cumulatively to the records $(y_i = x_i + n_i)$
Use n_i **OR** absolute n_i
End Cycle
System Noise = 0 **OR** System Noise = Current Cumulative Noise
Continue Next Cycle
Output: Perturbed Dataset (Y_i)

We do not have a clear idea about the amount of data or the availability rate of the data. This raises the need for accuracy to be measured over time. After considering all these factors and the usability of these methods in a practical environment, we decided to use the prequential evaluation setting implemented for our model. Therefore, this method does not produce accuracy measures for training and testing separately.

5.3.4 Accuracy-Privacy Trade-off Optimisation

The optimisation is a serious matter that needs to be handled carefully, as it depends on what kind of problem we are working with. Achieving optimisation in different scenarios has been discussed in the literature. For example, [146] discusses optimising fuzzy models for prosthetic hand myoelectric-Based control. Moreover, [147] has investigated an optimisation problem in virtual reference feedback tuning for tower crane systems, and [148] discusses predicting the minimum passing level of competency achievement, which is also an optimisation problem. However, a considerable amount of work is done in PPDM, while no discussion can be found on optimising the accuracy-privacy trade-off.

The ideal optimisation scenario for our study is where we can achieve perfect values for both privacy and accuracy. In other words, zero classification error and zero breach probability[143] which is impossible to achieve. A possible way to make this work is to minimise both classification error and breach probability. Therefore, we define the optimisation problem using classification error and breach probability according to the Privacy Accuracy Magnitude (PAM) formula *(PAM = (error)^2 + P(ϵ-privacy breach)^2)* proposed by [32].

We can say that if the PAM is less than a given error threshold (ϑ), the accuracy-privacy trade-off has been optimised.

If $PAM < \vartheta$; *Then the trade-off is optimised.*

Deciding a suitable value for ϑ is a complex and vital issue regarding the accuracy, and privacy values depend on the dataset's characteristics. Moreover, the optimal level of privacy and accuracy depends on the user's requirements, and an improved framework should be needed to handle these scenarios. Therefore, at this stage of the research, we compare the PAM values for proposed PPDM methods to identify the method that produces the minimum PAM, and how each method performs in terms of privacy and accuracy. Considering these techniques together, our proposed methodology can be briefed as in Figure 5.4. Data enters into the system sequentially, one sample at a time. These data go through the perturbation process to preserve privacy by hiding their actual values. Inside the perturbation module, we experimented with eight different cumulative noise addition methods variations to find the best method for them. Then the perturbed data is classified using Adaptive Random Forest in a prequential evaluation setting, and the accuracy of the classified data is updated for each record. Simultaneously, the privacy of the data is also measured by performing Known I/O attacks on the perturbed data. Finally, the overall error score is calculated using error and breach probability in terms of accuracy and privacy.

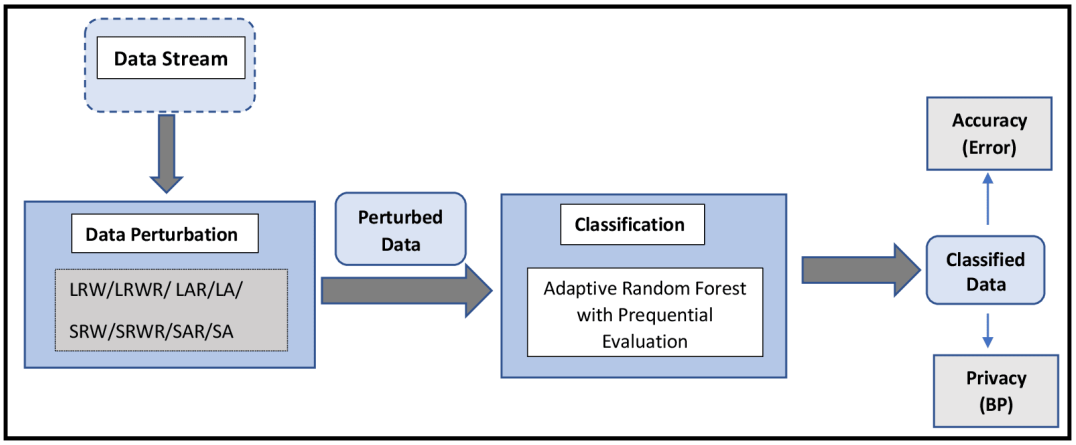

Figure 5.4: Proposed Methodology

5.4 Experiments and Results

5.4.1 Datasets

Two freely available datasets that can be considered as data streams are being used for the experiments. These can be considered streams because the records were ordered according to the time they were produced. Only the numerical features of the datasets were considered for the experiments, and both datasets contain potentially sensitive data. Data is being pre-processed accordingly. Details of the experimented datasets are represented below.

- Activity Recognition system based on Multisensory data fusion (AReM data et from UCI Machine Learning Repository) – This contains real-world data which includes 35,999 records, 6 features, and 5 classes.
- New York City Taxi Trip Duration (Taxi dataset from Kaggle, 2017) - This contains real-world data which includes 50,000 records, 7 features, and 3 classes.

5.4.2 Experimental Setup

Experiments are carried out to measure the privacy and accuracy of the datasets mentioned above. Privacy and accuracy measures used in [32] are adopted. The following Configurations are used for the experiments.

The noise was generated from a Gaussian distribution with a mean zero and a variance of 3.90×10^{-6}. Noise variance for the cumulative noise addition was selected according to the method proposed in [32] which generates a similar amount of noise when adding independent noise with a variance value of 0.0625. Cycles are virtually broken-down sets of the entire dataset, defined by cycle size. Experiments were carried out for the cycle sizes 300, 600, 1000, and 2000. Here is the configuration of other parameters used in the experiment.

General Configuration

- Classification Method – Adaptive Random Forest (ARF) with Naive Bayes leaf Prediction in a prequential evaluation setting
- Attacking Method – Known Input/Output Attack
- Cycle Sizes – 300, 600, 1000, 2000
- Variance of Cumulative Noise – 3.90 x 10^{-6}
- Number of Known Input/Output pairs – 4 per attack
- Number of Attacks – 5% of total record count
- ϵ – epsilon – 0.2

Configurations of logistic function

- Maximum Value – 1

- Growth Rate – 0.01, 0.02

5.4.3 Measuring Privacy and Accuracy

In [32] for the privacy perspective, known Input/Output MAP attacks were conducted, and the ϵ – privacy breach probability was measured, while for the accuracy perspective, the relative error was calculated. The same methods have been used in our research work also. In [32], authors have extended the known Input/Output attacking method to the random projection-based cumulative noise addition environment, which was proposed initially by [135] for multiplicative data perturbation. To measure the accuracy of the data stream, Adaptive Random Forest (ARF) proposed by [133] is used as the classifier and relative error represents the degree of success achieved by a record recovery attempt measured. It is defined as the magnitude of the difference vector between the original record and its recovered counterpart, normalized by the magnitude of the original record vector [32]. A "ϵ – privacy breach" of a record occurs if the relative error of the recovered record is less than a specified threshold (ϵ; $\epsilon > 0$) [143]. Overall error score or performance was evaluated according to the Privacy–Accuracy Magnitude (PAM) proposed in [32], and we modified the process using normalized values of relative error and breach probability to maintain fairness.

5.4.4 Results

This section explains the experimental results, including overall performance, relative error, and breach probability behaviour with different cycle sizes. The cycle sizes 300 and 2000 of the AReM dataset are represented here, and the other cycle sizes displayed a similar trend of results. The code for the experiments can be found at https://github.com/whewage/Variations-of-Cumulative-Noise-Addition.git

[1]Normalized values of relative error and breach probability were used for the

Table 5.2: Overall Performance Using PAM (AReM dataset, Cycle size 300)[1]

Method	Relative Error		Breach Probability		Overall Error Score	
	k = 0.01	k=0.02	k = 0.01	k=0.02	k = 0.01	k=0.02
LAR	0.6174	0.5502	0.1111	0.2857	0.3936	0.1349
LA	1.0000	0.9110	0.4444	0.7143	1.1975	0.1600
LRWR	0.3462	0.2945	0.4444	0.7143	0.3174	0.1199
LRW	0.6053	0.5388	0.0000	0.1429	0.3664	0.1339
SAR	0.2833	0.3482	1.0000	0.5714	1.0803	0.1228
SA	0.9540	1.0000	0.2222	0.0000	0.9595	0.1649
SRWR	0.0000	0.0000	0.8889	1.0000	0.7901	0.1038
SRW	0.2845	0.2877	0.4444	0.4286	0.2785	0.1188

Table 5.3: Overall Performance Using PAM (AReM dataset, Cycle size 2000)[1]

Method	Relative Error		Breach Probability		Overall Error Score	
	k = 0.01	k=0.02	k = 0.01	k=0.02	k = 0.01	k=0.02
LAR	0.8056	0.8127	0.1111	0.1111	0.6614	0.6728
LA	1.0000	1.0000	0.2222	0.2222	1.0494	1.0494
LRWR	0.3136	0.3385	1.0000	1.0000	1.0984	1.1146
LRW	0.2479	0.2751	0.4444	0.4444	0.2590	0.2732
SAR	0.3568	0.4606	0.8889	0.7778	0.9174	0.8171
SA	0.7606	0.7167	0.0000	0.0000	0.5785	0.5137
SRWR	0.0000	0.0000	0.5556	0.6667	0.3086	0.4444
SRW	0.0329	0.0914	0.1111	0.4444	0.0134	0.2059

convenience of understanding. According to the results of Tables 5.2 and 5.3, we can observe that for k=0.01, the lowest error score (the highest performance) was achieved by SRW for both cycle sizes. For k=0.02, with cycle sizes of 300 and 2000, the lowest error scores were given by SRWR and SRW, respectively. However, for a cycle size of 300, we can see that SRW returns the second-lowest error score, which is quite close to the error score of the SRWR. Initially, it was assumed that the noise resetting at the end of each cycle would considerably change the performance. However, from the results, we can see that it only makes a marginal improvement in accuracy `(1- relative`

`error)` but does not significantly impact the overall error score. Therefore, the perturbation method SRW (Logistic random walk without noise resetting) was the best performer as it produces the minimum overall error. Comparing the results with the state-of-the-artwork (LRW), SRW has outperformed in both cycle sizes considering both growth rate values.

Analysing the behaviour of accuracy and privacy with different cycle sizes and growth rates is important to understand the effects of those two parameters and select optimal values for these parameters. Relative error and the breach probability of the "AReM" dataset were analysed with four different cycle sizes (cs = $300, 600, 1000, 2000$) and two different growth rates (k = $0.01, 0.02$). Note that the growth rate only affects the logistic noise addition methods (highlighted in both Table 5.4 and 5.5).

Table 5.4: Behaviour of Relative Error with Different Cycle Sizes and Growth Rate Values.

Method	Cycle Size							
	300		600		1000		2000	
	k=0.01	k=0.02	k=0.01	k=0.02	k=0.01	k=0.02	k=0.01	k=0.02
LAR	0.3664		0.3907		0.404		0.4248	
LA	0.3980		0.4040		0.4224		0.4455	
LRWR	0.3440		0.3545		0.3657		0.3724	
LRW	0.3654		0.3654		0.3654		0.3654	
SAR	0.3388	0.3487	0.3596	0.3607	0.3708	0.3671	0.3770	0.3859
SA	0.3942	0.4058	0.3920	0.3973	0.4203	0.4245	0.4200	0.4142
SRWR	0.3154	0.3182	0.3308	0.3340	0.3293	0.3345	0.3390	0.3350
SRW	0.3389	0.3434	0.3403	0.3466	0.3418	0.3443	0.3425	0.3451

According to the results in Table 5.4, the relative error increases with the cycle size in both linear and logistic cumulative noise additions. A possible reason for this behaviour is that the noise in the system is high when it comes to the larger cycle sizes. On the other hand, increasing the growth rate negatively affects the accuracy since the relative error increases with the growth rate. In such cases, the classifier cannot adapt

its model fast enough to cope with the rate of increase in noise, which in turn leads to a more significant error. Nevertheless, as we can see in the table, SRWR and SRW have low relative errors in all the cases compared to the state-of-the-artwork (LRW), and LRWR and SAR have outperformed LRW in small cycle sizes.

Table 5.5: Behaviour of Breach Probability with Different Cycle Sizes and Growth Rate Values

Method	Cycle Size							
	300		600		1000		2000	
	k=0.01	k=0.02	k=0.01	k=0.02	k=0.01	k=0.02	k=0.01	k=0.02
LAR	0.025		0.025		0.005		0.005	
LA	0.040		0.033		0.015		0.010	
LRWR	0.040		0.017		0.025		0.045	
LRW	0.020		0.020		0.020		0.020	
SAR	0.065	0.035	0.017	0.033	0.030	0.020	0.040	0.035
SA	0.030	0.015	0.033	0.008	0.010	0.015	0.000	0.000
SRWR	0.060	0.050	0.025	0.042	0.055	0.045	0.025	0.030
SRW	0.040	0.03	0.017	0.058	0.015	0.045	0.005	0.020

According to Table 5.5, breach probability decreases with the cycle size and privacy increases in most cases. Unlike the classification error, we cannot see a clear movement of the behaviour of the breach probability with the growth rate. However, SA and SRW (logistic cumulative noise additions without noise resetting) show approximately similar behaviour. On the other hand, SAR and SRWR show a similar trend with different k values. However, it is insufficient to explain privacy behaviour with different growth rates. When we compare results with LRW, a clear pattern cannot be seen but SRW has produced equal/low breach probability values for k = 0.01 in $cyclesize$ = 600 and 1000 and both k values in $cyclesize$ = 2000.

Moreover, we have conducted the attacks on the logistic cycle's starting and ending flat areas instead of random attacks on the data stream. The starting flat area of the logistic cycle is the most vulnerable location for two main reasons. The first reason is

that the noise level added is deficient in that area, and the second one is that the noise addition rate is very low. On the other hand, when we consider the ending flat area, it is also vulnerable because noise is added at a constant noise addition rate, making it easier for the attackers to breach the privacy regardless of the high noise level of that area. Therefore, starting and ending flat areas have been identified as the most vulnerable to attacks. To prove our claim, we have conducted the attacks on those identified areas of the best performing perturbation method SRW.

Attacks were performed in two different manners: attacks on the flat areas of randomly selected cycles and attacks on the flat areas of each cycle. We compared those results with the breach probability of random attacks on the data stream. Table 5.6 displays the average breach probabilities of AReM and Taxi datasets after conducting 50 rounds of attacks. We conducted 50 rounds of attacks to reduce the possible bias when conducting a single round of attacks.

Table 5.6: Comparison of Breach Probabilities After Performing Attacks to Different Locations of the Data Stream

Datasets	**Starting flat period**		**Ending flat period**		**Random**
	Random cycles	Each cycle	Random cycles	Each cycle	
AReM	0.0310	0.0312	0.0296	0.0319	0.0270
Taxi	0.0273	0.0341	0.0250	0.0366	0.0200

The results prove that the selected perturbation method SRW is relatively more vulnerable when the attacks are performed in the flat areas of the logistic cycle. Both datasets have shown that it is easy to breach privacy when attacking the starting flat period of randomly selected cycles with breach probabilities of 0.002 and 0.0023 than the breach probabilities of ending flat areas of AReM and Taxi datasets, respectively. When the attacks were performed on the flat areas of each cycle, the starting flat area seemed less/equally vulnerable to the ending flat area. However, even with those

changes, SRW has maintained more than 97% of privacy in all the scenarios, proving the method's reliability.

5.5 Discussion

By analysing the results of the experiments, the perturbation method SRW can be identified as the best performer considering both privacy and accuracy perspectives. Combining the logistic function with the cumulative noise addition certainly positively impacts optimising the trade-off between privacy and accuracy. The logistic function's nature helps control the maximum noise level of the cumulative noise addition, which prevents noise from dominating the data. This increases the accuracy level, hence providing an opportunity to minimise the trade-off.

Noise addition in cycles seems effective since relative error decreases. Hence accuracy increases in smaller cycle sizes (cs). On the other hand, beach probability decreases; hence privacy increases when cycle size increases. The logistic function's growth rate(k) also impacts since relative error increases with it. Moreover, the impact of growth rate on breach probability should be investigated further since the results do not show a clear pattern. However, selecting the appropriate cycle size and growth rate values is vital. Noise resetting does not give the expected results as it only makes a marginal improvement of the accuracy but does not significantly change the overall score. Using absolute noise values is not an excellent technique to control the noise level because noise injection distorts the data irrespective of addition or subtraction.

Flat areas of the logistic cycle are the most vulnerable to the attacks, and if the attacker can find out the cycle size, he can succeed in attacking those most vulnerable areas. Therefore, cycle size is an important parameter that needs to be protected.

5.6 Conclusion and Future Work

In summary, we experimented with different techniques combined with cumulative noise addition to controlling the maximum noise added to the data stream. Controlling the maximum noise level is essential in the cumulative noise addition environment, as the system has to face a tremendous amount of noise with the time when the data stream unfolds, which can highly decrease the classification accuracy. Our objective was to maintain the maximum data privacy benefits from cumulative noise addition while having a minimal negative impact on the classification accuracy, optimising the trade-off between privacy and accuracy. According to our experiments, cumulative noise addition in cycles combined with the concept of logistic function turned out to be a promising method to control the maximum noise level of the stream. Therefore, cumulative noise addition combined with logistic function has proven to be a better approach to optimise the trade-off between privacy and accuracy by controlling the maximum noise level of the system.

As for future work, noise resetting should be investigated further conceptually and experimentally with different datasets, as it appears to be a promising method to control the noise level of the system. Moreover, noise can be added in randomly selected cycle sizes instead of a fixed cycle size to ensure high privacy by avoiding attacks on flat areas. This can be considered an essential improvement to this work. In addition, a well-formulated privacy-accuracy framework using the selected method can be the next step of the work. This framework should be able to answer the optimisation of the accuracy-privacy trade-off issue according to the user's accuracy and privacy requirements.

Chapter 6

Prelude - Manuscript 3

Privacy of data and accuracy of data mining results are highly interlinked, as increasing one can reduce the other. Different perturbation methods have been proposed to optimise this trade-off between data privacy and data mining accuracy. However, no method has been successful in achieving perfect accuracy and privacy simultaneously. Therefore, a method to optimise the accuracy-privacy trade-off practically should be investigated.

Furthermore, there is a lack of well-constructed frameworks for PPDM. A framework that can be utilised for PPDM and optimisation of accuracy-privacy trade-off is much needed in the area. Manuscript 3 focuses on developing a framework for this purpose. This Accuracy-Privacy Optimisation Framework (APOF) consists of three main modules, accuracy, privacy, and data fitting. The data fitting module is responsible for accuracy-privacy optimisation. As a practical way to solve this issue, the data fitting module considers the user's privacy requirements and predicts the expected accuracy level. This allows the user to fine-tune his or her requirements if needed.

Considering the user's requirements and using them in a data fitting module is a novel approach in PPDM. Our argument behind considering the user's requirements is that not every user needs high accuracy and privacy. Depending on the data they use and the predictions they want to make, priority can be given to either privacy or accuracy.

We use Logistic Cumulative Noise Addition (SRW) [127] proposed in Chapter 5 as the perturbation method. It has already been optimised, considering privacy and accuracy. Here, we further optimise it within a framework. The APOF is ideal for static datasets but has several features supporting data streams.

Ultimately, manuscript 3 [149] addresses RQ2, which is **"What framework should be constructed to model the trade-off between privacy and accuracy for a specific user in a data mining environment?"**. Experiments show that APOF can achieve accuracy-privacy optimisation using the data fitting module, retaining a small error.

Chapter 7

An Accuracy-Privacy Optimisation Framework Considering User's Privacy Requirements for Data Stream Mining (Manuscript 3)

7.1 Introduction

Data privacy has become a prominent topic in data mining, especially data stream mining, with considerable effort spent on improving privacy without affecting accuracy. This process is called Privacy-Preserving Data Mining (PPDM) [44, 26]. Organisations use data from various sources to enhance their decision-making processes [9]. However, data mining can undermine the privacy of individuals that provide sensitive data. The primary method to increase privacy is to perturb the data to increase privacy before mining [140]. Perturbation techniques change the actual values of the data, but this degrades the quality of the data, which ultimately negatively affects the accuracy of the data mining results or predictions [32]. There is an urgent need to develop novel

techniques to monitor and fine-tune this trade-off between data privacy and prediction accuracy in data mining systems.

Optimising the accuracy-privacy trade-off requires careful handling, as these two qualities are highly interlinked [38, 23]. Ideal levels of privacy and accuracy cannot be simultaneously achieved. Often, privacy and accuracy requirements for a project vary and depend on factors like the kind of data being processed, the level of privacy required per regulatory standards, and acceptable accuracy ranges. For example, consider a company ABC that manages a parcel delivery service and needs to mine its existing customer data to improve the efficiency of its service. Their database does not contain highly sensitive data, and the company decides to maintain 70% privacy (70% protects original data values from revealing to unauthorised people; in other words, there is a 30% of the possibility of recovering individual data values from perturbed data). In this case, it is unnecessary to exceed 70% privacy during perturbation. Consequently, we can strive to achieve the maximum possible accuracy during data mining while maintaining 70% privacy. If the level of accuracy is acceptable, we can go ahead and implement the solution without needing to consider stronger (or weaker) perturbation methods.

The main problem in PPDM is that it is impossible to achieve ideal levels of both privacy and accuracy because of the inherent trade-off between these two measures [10]. A possible way forward is to comply with the user's privacy accuracy requirements so that achieving the highest levels of privacy or accuracy is not needed. Therefore, we propose an "Accuracy-Privacy Optimisation Framework" (APOF) which predicts the respective accuracy level for a user-defined privacy threshold. Knowing the achievable accuracy level for the required privacy allows the user to fine-tune accuracy and privacy before implementation. APOF is an incorporation of various techniques such as Hoeffding Tree [132] for classification, Random Projection-based Logistic Cumulative Noise Addition (SRW) [127] as the privacy-preserving method, and Kernel Regression for the

data fitting. Authors of [127] have used the abbreviation "SRW" to avoid confusion with other methods introduced in the same article that starts with "L." It stands for "Sigmoid Random Walk," which explains the noise addition technique. We use the same to avoid any confusion.

SRW is a noise injection perturbation method that has been shown to achieve high levels of privacy and accuracy. It employs the logistic function [128] to control the noise level. Logistic function [128] is known to play an essential role in Mathematics for nearly two centuries, but its adaptation to PPDM is novel. Using the Logistic function in SRW provides benefits like achieving a good accuracy-privacy trade-off. We evaluated the performance of SRW with different noise addition rates to identify its behaviour before using it in APOF.

Classification is done using the Hoeffding Tree, and we introduce a new measure, Average Expected Loss (AEL), to represent accuracy. This considers the path and incorrect classification probabilities of the tree when calculating the loss. It provides a deep insight into accuracy rather than calculating the error rate, which only considers the number of incorrectly classified records. Though we use Hoeffding Tree as the classifier, APOF can be used with any incrementally learning classifier, and therefore it is more appropriate in data streaming environments.

The novelty of APOF is that it combines a data fitting module that optimises the accuracy-privacy trade-off according to a user-defined privacy level, allowing users to fine-tune accuracy and privacy levels if needed. We compared five kernel regression methods to select the best method to use in the data fitting module. Using this framework, the user can make iterative changes to optimise the balance between accuracy and privacy.

This research makes the following contributions to the field of PPDM:

- A novel Accuracy-Privacy Optimisation Framework (APOF) that predicts the

respective accuracy level for a user-defined privacy threshold.

- Re-evaluating the performance of the SRW perturbation method [127] for various noise addition rates to identify the best ranges of noise addition rates.

- A comparison of different kernel regression methods as a data fitting technique to optimise the accuracy-privacy trade-off

The remainder of this paper has been organised as follows. Section 7.2 explains the related work of data perturbation, classification algorithms, and kernel regression as a data fitting technique. Section 7.3 describes the proposed Methodology and Design and other related techniques used in detail, while Section 7.4 presents a discussion of experimental results. Finally, Section 7.5 presents the conclusion and future work.

7.2 Related Works

We reviewed existing works related to Privacy-Preserving Data Mining (PPDM), especially data perturbation techniques and Hoeffding Tree (HT), which can be used for data stream mining and data fitting/regression techniques. Subsequently, we highlight the existing gap in the literature related to the accuracy-privacy trade-off.

7.2.1 Data Perturbation as a PPDM Technique

In data mining, PPDM techniques protect sensitive information from unauthorized access [9]. Data perturbation, generalization, suppression, and anonymization are popular PPDM techniques [150, 151]. Data perturbation randomly hides the true form of data by changing the original data values while still preserving the relevant statistical properties of the dataset that are important for data mining [39, 46]. Different types of perturbation methods have been proposed over the past few decades including

random projection [35], random rotation [39, 34], geometric perturbation [76], additive noise [134], multiplicative noise [152, 33], 3-D [74] and 4-D [75] transformation and condensation [37]. This article only focuses on noise addition and multiplication-related perturbation techniques, as they are easily adapted to data streaming environments.

Additive and multiplicative noise are the two main types of noise injection techniques [134, 74, 39, 153]. A few other approaches combine these two techniques with other perturbation techniques to improve the performance by reducing privacy vulnerabilities [140, 9]. In additive noise, random noise values generated from a Gaussian or uniform distribution are added to each data record independently. A high privacy level can be achieved if the noise variance is high, increasing privacy but often sacrificing accuracy. Several works, such as [154, 138], discuss techniques to breach the privacy of the data perturbed using additive noise, showing that additive noise can easily be filtered out on many occasions [35].

Multiplicative noise perturbation involves multiplying the original record value with a randomly generated noise value. Two approaches for multiplicative noise injection were introduced in [33]: multiplying with a random number and logarithmic transformation. In [152] random multiplicative noise is used for protecting against clustering analysis. Multiplicative noise can be considered a random rotation with a vector, as multiplication causes a rotational transformation [39]. However, the rotational perturbation can also be attacked from methods such as distance inference attacks [76], Independent Component Analysis (ICA) [139, 76], and Known Input/Output attacks [139].

Research works such as [9] and [140] explore the potential of combining both methods to reduce the privacy risk caused by the individual weaknesses of the additive and multiplicative methods. These techniques are still vulnerable to attacks based on background knowledge [122]. In [32], a novel noise addition method called "Cumulative Noise Addition" is proposed to overcome the drawbacks of traditional additive noise.

This cumulative noise addition method adds Gaussian noise with a minimal noise variance to each record. Additionally, each of these random noise values is also added to every subsequent record in the stream. Cumulative noise addition has been combined with random projection and translation to improve privacy. Experiments have shown that this method achieves high levels of both privacy and accuracy when compared to others [32]. However, the data stream faces a significant amount of noise when the stream unfolds, which is the main drawback that eventually decreases the accuracy of predictions [127].

In [127] different techniques such as cycle-wise noise addition and noise resetting are used to improve the cumulative noise addition proposed in [32] by controlling the maximum noise level. Experiments carried out for seven different variations of cumulative noise addition showed that Logistic Cumulative Noise Addition (SRW) provided the best levels of privacy and accuracy. In SRW, Gaussian noise with mean zero and a small variance is added cycle-wise in a random walk fashion. The variation of noise changes in every step, ensuring that attackers cannot attack easily. The noise variance is decided using the well-known logistic function [128], and the dataset is virtually divided into cycles of a specific size. For the records in each cycle, the return value of the logistic function multiplied by a small additional noise variance value is used as the final noise variance. The main advantage of SRW is that the maximum noise level and noise addition rate can be controlled using the logistic function [127]. This is a considerable improvement over cumulative noise addition, as it restricts the overall levels of noise added [32]. SRW is more effective in data streaming environments with such control, as it does not sacrifice performance.

Modelling the privacy side of the framework also involves carefully selecting a method to measure privacy. Quantification of privacy is a complex matter which depends on the perturbation method and context [13]. Therefore, there is no exact definition for privacy, and existing definitions depend on the context. Several types of attack

methods, such as Maximum A Posteriori attack (MAP) [138], Distribution attack [154], and Known Input/Output attack [139] have been proposed to measure privacy in the noise injection environment. In these methods, attackers try to recover the original values from perturbed values using some prior knowledge. In [32] authors have used a known Input/Output MAP attack to breach the privacy, and then the privacy has been measured according to the ϵ-privacy. An "ϵ-privacy breach" of a record occurs if the relative error of the recovered record is less than a specified threshold (ϵ; $\epsilon > 0$) [143]. The attacking method proposed in [32] has been implemented for the cumulative noise additive environment, and authors in [127] have used the same method for attacking SRW. They also perform a vulnerability analysis of this method concerning SRW by deliberately attacking different areas of the logistic cycle. They have shown that the logistic cycle's starting and ending flat areas are more vulnerable to attacks. Still, SRW achieves a good privacy level.

Our analysis of prominent PPDM methods shows that noise addition methods can effectively be used in static data and data stream mining. Of the different types of noise addition methods, cumulative noise addition promises good accuracy and privacy levels [32]. SRW, an improved cumulative noise addition method [127] outperforms other options [32] for both privacy and accuracy and is applicable for data stream mining.

7.2.2 PPDM for Data Streams

Mining data streams has additional concerns due to the natural behaviour of data streams. Unlike static datasets, streaming data is continuous, transient, and unbounded, which arises the need to be processed quickly [47, 51]. The mining of data streams cannot be done repeatedly as for static datasets because it is impossible to access the complete set of data at once [52]. These facts should be considered when designing PPDM methods for data streams.

Different perturbation-based PPDM methods have been proposed for data stream mining, considering the above challenges. A combination of condensation, rotation and random swapping [47], Statistical Disclosure Control (SDC) with different filters such as noise addition, micro aggregation, and rank swapping [51] are some prominent PPDM techniques implemented for data stream mining. Furthermore, random projection-based cumulative noise addition [32], anonymisation-based noise addition [117] and differential privacy-based perturbation [38] have been implemented to protect the privacy in data streams. These methods have issues such as accuracy-privacy trade-off, inability to deal with the concept drift, and difficulty dealing with the infinite nature of data streams.

Noise addition, anonymization, and condensation are a few techniques that can be used for streaming data without any changes, as most of the other generic PPDM methods need improvements to adapt to the behaviour of data streams.

7.2.3 Solutions for Addressing Accuracy-Privacy Trade-off in PPDM

Numerous research work such as [75, 91, 90, 16] have discussed the issue between data privacy and data mining accuracy in detail. However, only a few have proposed solutions to optimise this trade-off. It should be adequately addressed to achieve maximum performance, as the objective of PPDM is to effectively protect sensitive data while maintaining the knowledge in original data [32, 46].

A combination of suppression and perturbation has been proposed in [23] to minimise the loss caused by generalisation in anonymisation. The authors of [41] tried to achieve a good trade-off by using t-closeness through micro aggregation. Anonymisation based on clustering methods was proposed in [11, 62] for the same purpose. A method called "NRoReM" that combines normalisation, geometric rotation, linear regression, and scalar multiplication has been proposed to optimise the accuracy-privacy

trade-off [10]. Other methods, such as [57, 14, 95] have implemented techniques like multivariate perturbation, condensation, and non-negative matrix factorisation to provide solutions to this issue.

Attention to optimising the accuracy-privacy trade-off in data stream mining is much less than generic PPDM methods. Differential privacy-based PPDM methods have been proposed in [52], and [38], and the method implemented in [55] combines condensation and rotation to optimise accuracy-privacy trade-off in data streams. Authors of [32] and [127] have proposed random projection-based noise addition methods to optimise accuracy-privacy trade-off in data stream mining. Though these methods propose different techniques, optimising the accuracy-privacy trade-off is still a work in progress. To the best of our knowledge, the only work that provides a framework to optimise the accuracy-privacy trade-off is "PPaaS" [40]. It allows selecting different perturbation methods according to the different applications, which consider users' requirements.

7.2.4 Hoeffding Tree as a Classification Algorithm for Data Stream Mining

Decision Trees, Naïve Bayes, Support Vector Machine, Neural Networks, and Random Forest are popular classes of classifier algorithms used in data streaming environments. Hoeffding Tree (HT) is a type of decision tree algorithm proposed and implemented as a part of Massive Online Analysis (MOA) by the University of Waikato [132]. In [155] HT is classified as a Very Fast Decision Tree (VFDT) which learns incrementally from streaming data using constant time and memory per example. It is one of the most simple, suitable, and efficient classification techniques for data streams [15]. Additionally, other works like [156, 157, 158] discuss the usage and performance of the HT in a data streaming environment. These works show that HT is a successful any-time incremental learner with guaranteed performance, making it capable of learning from

massive data streams.

7.2.5 Data Fitting/Regression

Regression analysis refers to statistical inferences for a model [159]. There are several types of regressions, such as linear, non-linear, and kernel regression. All these methods fit a curve to the given data and represent the most suitable curve using a mathematical function [160]. Both linear and non-linear regression types represent Y (dependent variable) as a function of X (independent variable) and calculate function parameters according to the best-fitted curve for the given data. Linear regression can be considered a particular case of non-linear regression, and any non-linear program can be used to fit a linear model [161]. These regression methods are parametric and calculate function coefficients by defining initial values to start the regression. However, kernel regression uses a non-parametric approach that does not require coefficient calculation.

Kernel Functions return the inner product between the images of two inputs in some feature space [162]. Those assign weights to the data points by considering the distance from the target point [163]. Kernel functions can be used for classification, novelty detection, and regression. Kernel methods have been used in different areas such as geometric modeling [164], protein function prediction in yeast [165], pattern analysis [162], spike rate estimation in neurophysiology [166], and signal processing [167].

Kernel functions are positive semi-definite, flexible, modular, and non-parametric [162, 168]. Being non-parametric is the most important property in our scenario because it is difficult to define accuracy as a function of privacy and vice versa. Non-parametric techniques allow data to model relationships among variables without an awareness of functional form specification. They can detect structure that sometimes remains undetected by traditional parametric estimation techniques [169]. This allows us to avoid the issue of selecting the most suitable function and parameters.

We emphasise five different kernels, which are positive, and semi-definite, namely, Gaussian, Laplacian, Rational-quadratic, Wave, and Matern-52 [170]. These are general purpose and can be used with no prior knowledge of data [171] and are positive definite or semi-definite [172]. The application and more details of these kernels can be found through: Gaussian [173, 174], Laplacian [174, 175, 176], Rational-quadratic [177, 178] Wave [179, 180] and Matern-52 [177, 181]. We compared the performance of selected kernels to identify the most suitable kernel that can be used in the data fitting module of APOF.

From the analysis of the existing works, it can be observed that different techniques are proposed to improve generic PPDM. However, only a few works have been done for privacy preservation in data streams. There are few works such as [38, 55], [32] and [127] that have paid attention to optimising the accuracy-privacy trade-off in data stream mining. Nevertheless, we could not find a well-formulated framework for this optimisation except [40] for data stream mining. From these observations, it can be concluded that though various methods have been proposed as PPDM techniques, there is a lack of well-formulated frameworks or methods to optimise the trade-off between accuracy and privacy, especially in data stream mining.

7.3 Proposed Methodology and Design

The APOF framework is a novel approach in PPDM that facilitates functionalities to comply with the user's accuracy and privacy requirements. The APOF provides the achievable accuracy level for a user-defined privacy threshold and the parameter value (noise addition rate), which can be used to perform the perturbation process to achieve the user-defined privacy threshold.

PPDM generally consists of two main processes. A perturbation process and a data mining process. We improved this traditional setup by adding a data fitting module

to comply with the user's requirements. Therefore, APOF has three main modules: accuracy, privacy, and data fitting. The following diagram (Figure7.1) illustrates the processes involved with the modules mentioned above.

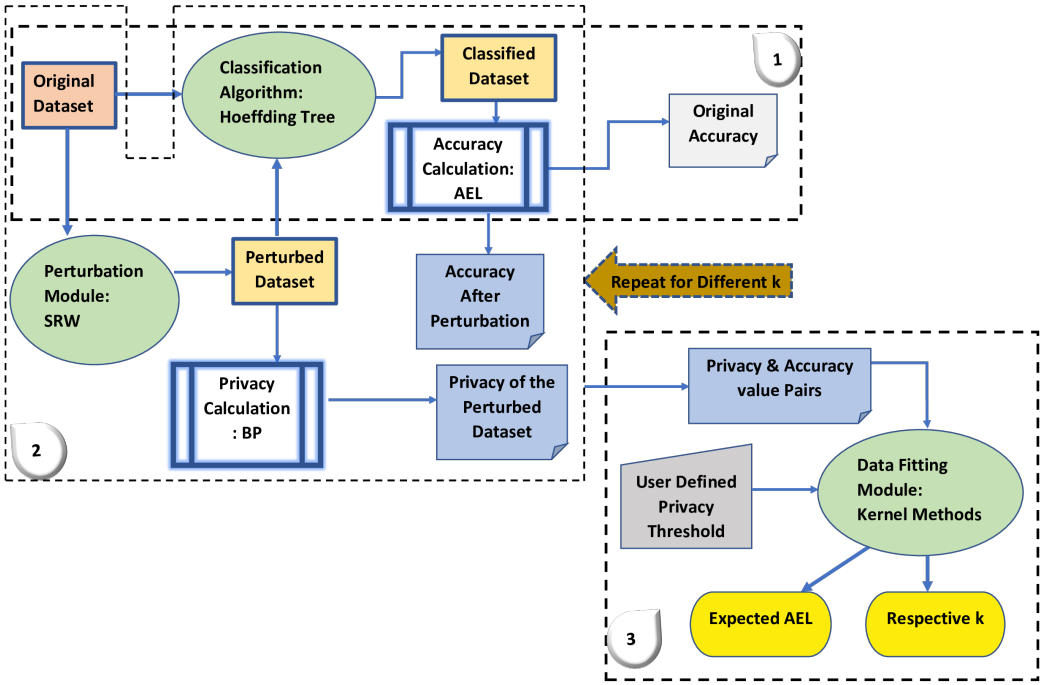

Figure 7.1: Design diagram of the APOF (1- processes involved with original dataset, 2-processes involved with the perturbed dataset, 3- data fitting module)

7.3.1 Privacy Module of the Framework

The main objective of the privacy module of the framework is to protect sensitive data and compute the expected privacy after the perturbation process. The efficiency of this module entirely depends on the perturbation method we use to perturb the original data. After carefully considering all the possibilities, we selected Random Projection-based Logistic Cumulative Noise Addition (SRW) as the perturbation method to implement in the privacy module. Noise addition is suitable for both static and data streaming environments, and SRW has been experimentally proven to achieve considerably high

accuracy and privacy values compared to other noise addition methods [127].

Deciding Possible Noise Addition Rates for SRW

SRW is a cumulative noise addition method that provides the opportunity to change the noise addition rate (k), cycle size (cs) and maximum value of the logistic function (L). It first divides the data stream into virtual cycles. Then it adds Gaussian noise with mean zero and the variance, *(f(x) ×σ)* where *f(x)* is the return value from the logistic function for each sample in the cycle, and σ is the noise variance factor. The logistic function can be defined as in [128].

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}} \tag{7.1}$$

Though the k value of the logistic function can be varied, it cannot be done for an extensive range of values. According to the authors of [127], to get the maximum benefits from the SRW, the ideal shape of the logistic curve should be preserved. Therefore, as the first step, we investigated the possible value range for *k*. (See Figure7.2)

When the *k* is less than 0.004, the curve becomes a straight line, and when the k is greater than 0.1, we can see an abrupt change from zero to the maximum value. In both these extreme situations, the natural shape of the logistic curve is distorted, and the maximum performance benefits from SRW cannot be achieved. We have identified that the shape of the logistic curve plays an important role in maintaining the balance between accuracy and privacy. This happens because the starting flat area of the curve allows the classifier to learn and adapt to the low noise variance, and the ending flat area maintains a high privacy level with high noise variance. Therefore, we decided to use the *k* values in the range of 0.005 – 0.1 with a gap of β=0.004.

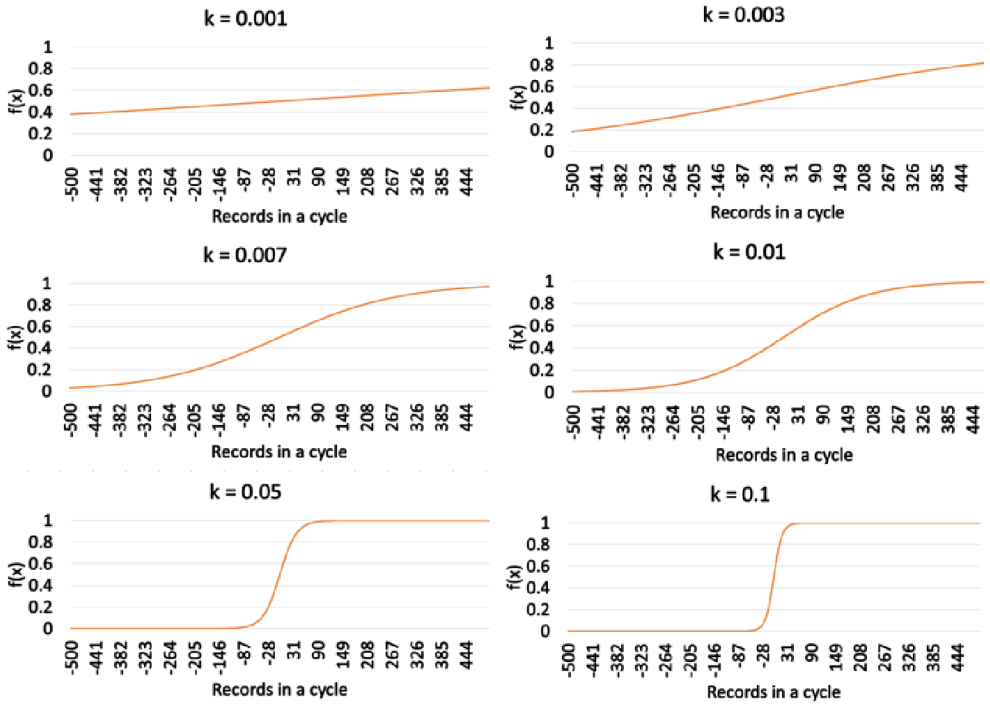

Figure 7.2: Effect of changing k to the shape of the logistic curve.

Perturbation Process

The perturbation process was carried out after deciding the possible range of noise addition rates (k). Random projection is first used on the original data [32, 127] to maintain a stronger level of privacy. We consider the original dataset as X ($m \times n$ matrix), where columns represent data records, rows represent attributes and the random projection matrix as R ($k \times m$ random matrix). Then the projected dataset Y' can be considered as a $k \times n$ matrix $(k \leqslant m)$ [35]. Random projection can be written down as follows.

$$Y' = \frac{1}{\sqrt{k\sigma r}} RX \tag{7.2}$$

The factor $\frac{1}{\sqrt{k\sigma r}}$ ensures that the column-wise inner product is preserved [127] and that is useful in maintaining the same statistical properties as the original dataset. For

experiments, the $(k = m)$ state of random projection was considered as the datasets do not have many features.

The random projection-based noise additions implemented in [32] and [127] used a translation operation (represents using Γ) to add an additional degree of privacy. According to [76] distance preserving transformations are vulnerable to rotation centre attacks as the data records closer to the origin are less perturbed than the others. Translation reduces this risk, and applying a constant translation to all records does not affect most data mining tasks [32]. The perturbed data after translation (Y'') can be represented as follows.

$$Y'' = \frac{1}{\sqrt{k\sigma r}} RX + \Gamma \tag{7.3}$$

Then, the logistic cumulative noise [127] was added to Y'' to produce the final perturbed dataset. The overall perturbation process can be defined as follows.

$$Y = \frac{1}{\sqrt{k\sigma r}} RX + \Gamma + \Sigma_{i=-cs}^{+cs}(\varpi) \tag{7.4}$$

Figure7.3 illustrates the overall perturbation process.

The dataset is virtually divided into cycles, defined by the cycle size, and the logistic function value $(f(x_i))$ is calculated for all the records in a cycle. That value is multiplied with a small noise variance factor (σ) to produce the final noise variance (υ). The noise variance factor ensures that whatever the value returns from the logistic function are scaled down to maintain a low distortion in each noise addition step. Random noise values are generated from a Gaussian distribution with zero mean and variance (υ). Generated noise values are added cumulatively to the records, meaning each value is added to the respective data record and every subsequent record. This process continues for all cycles until the end of the dataset. It is worth noting that each record in the cycle has a small but different noise variance as $f(x_i)$ is different for every record. This

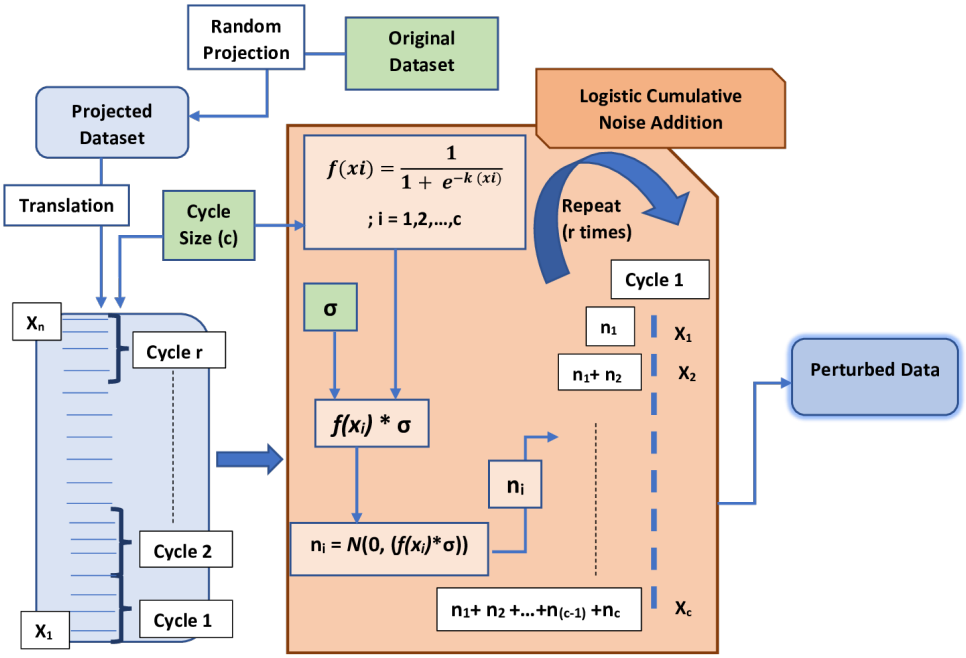

Figure 7.3: Perturbation Process - Random Projection-based Logistic Cumulative Noise Addition.

ensures more privacy against noise addition attacks, as tracking down the noise variance by an attacker is difficult.

Known I/O Attacks for Breaching Privacy

To breach privacy, we used the Maximum A Posteriori attack (MAP) which is implemented and discussed in [32] and [127], and calculated the breach probability; according to the "ϵ-privacy breach." The MAP attack was designed to recover an original record from a perturbed record using a few numbers of known Input/Output records. It assumes that the attacker knows some original and the respective perturbed records and performs attacks to recover an unknown original record using distances between them.

We used the same terminology but performed attacks differently. Instead of attacking the data stream randomly, we performed a fixed number of attacks on each cycle, making it more vulnerable. For this, we used four known Input/Output records per attack.

For the possible k values in the selected range, we conducted the perturbation process independently and recorded the Breach Probability (BP) to measure the privacy after conducting MAP attacks.

7.3.2 Accuracy Module of the Framework

The main task of the accuracy module is to calculate the accuracy of the predictions made by the data mining method. This involves comparing the accuracy of classification over the original and perturbed datasets. We used Hoeffding Tree (HT) as our data classification method specifically designed for data stream classification. HT was implemented based on the assumption that the data distribution does not change over time [132]. We used the HT with Naïve Bayes as the leaf prediction method, and other configurations were kept as defaults. We calculated the accuracy of the predicted results in terms of Average Expected Loss (AEL). AEL can be defined as follows (see Definition 1).

Definition 1. Average Expected Loss (AEL) -The average of the summation of incorrectly classified record probabilities of all leaf nodes over all the decision tree paths.

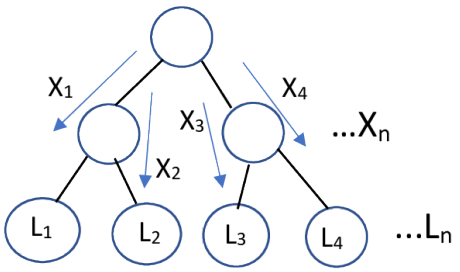

Figure 7.4: Illustration to explain the calculation of AEL using Decision Tree.

$$AEL = \Sigma_{i=0}^{n} P(X_i)(1 - P(L_i)) \tag{7.5}$$

Where;

$P(X_i)$ – Probability of classifying an instance along path X_i.

$P(L_i)$ – Correctly classifying probability of leaf node L_i.

First, we calculated the AEL of the original dataset before the perturbation and then calculated the AEL after each perturbation for all the *k* values.

Evaluation of the Results

This framework is being designed for data stream mining; therefore, the behaviour of data streams should be carefully considered when evaluating the experiments. A few of the significant requirements of data stream mining are processing one record at most one time, using a limited time and memory, and should be ready to predict at any time. [132, 48]. After considering these requirements, the evaluation process can be carried out in two ways. Those are holdout method and prequential (interleaved test-then-train) method [132, 145].

The holdout method evaluates the performance of a single holdout dataset and is suitable when the division between train and test sets can be pre-defined. The prequential method uses each record to test the model before it is used to train the model and incrementally update accuracy. When this process is performed in the correct order, the model is always being tested on the samples it has not seen. The prequential method does not require pre-defined training and testing tests and takes the maximum use of available data [132].

It was decided to use the prequential evaluation method for these experiments as the amount of data available, or the incoming data rate is not known beforehand in data streams. Also, the execution of the prequential method is more suitable for the incremental behaviour of data streams, which raises the need for accuracy to be measured over time. Therefore, this method does not involve calculating different

accuracy values for training and testing sets.

7.3.3 Optimisation Criteria

The random projection-based logistic cumulative noise addition method (SRW) has been proven to achieve high performance in [127] according to the Privacy Accuracy Magnitude (PAM) proposed in [32]. PAM has been defined as $((\texttt{error})^2 + P(\epsilon - \texttt{privacy breach})^2)$ [32] and authors of [127] have identified a criteria for optimisation. According to that criteria, accuracy-privacy trade-off is optimised if $PAM < \theta$; where θ is the error threshold. SRW has been compared with other cumulative noise addition methods and experimentally proven to have a better trade-off according to the above criteria. This means optimising the method itself using the overall performance has been successfully carried out. We try to extend this optimisation to another level by incorporating the user's privacy requirements, and we define the objective of the optimisation we carried out as follows.

To seek the minimum AEL such that

$BP(SRW) \leqslant (1 - \alpha)$ Where;

α – Privacy threshold specified by the user

$BP(SRW)$ – Breach probability of the perturbation method

Suppose the user disagrees with the AEL value that can be achieved for α specified by him. In that case, there is a possibility of making changes to the requirements before the actual perturbation starts.

7.3.4 Data Fitting Module of the Framework

The data fitting module predicts the AEL value for a user-defined privacy threshold (α) using BP values generated from the privacy module and AEL values generated from the accuracy module for all the experimented k values. Hence the previously defined optimisation is carried out using the data fitting module.

Selecting a data fitting function is a crucial decision. Data should be displayed and analysed. The data distribution should be carefully considered before selecting a model to fit the data, as the results inferred from this point depend on the selected model. Our AEL and BP results distributed over selected k values (See Section 7.4.2) showed that a first-order linear function could not be used to fit results because of the complex relationship between accuracy and privacy. A higher-order polynomial or another suitable function was needed to represent the results. If we use a higher-order polynomial function, we must decide on a function or order representing the given data and the coefficients of the selected function. The selection of a function needed careful attention as the data distribution seemed random, and a specific trend or a shape could not be found in BP and AEL values (See Figure 7.5 & Figure 7.6). Therefore, we selected kernel functions as our data fitting function as they are non-parametric, and we do not have to define the initial coefficient values.

First, we compared the curves of data fitting for different datasets by applying five selected kernels, namely, Gaussian, Laplacian, Rational Quadratic, Wave, and Matern-52, to identify the most appropriate kernel. Selecting the bandwidth or smoothing parameter is a crucial factor in applying kernel methods for regression, and different theoretical methods for selecting optimal bandwidth have been proposed in the literature. Silverman's rule-of-thumb [182], The Least square cross-validation [182, 183, 184] smoothed, and wild bootstrap [184] are mainly discussed bandwidth optimisation methods we could find in previous work. However, selecting the optimal bandwidth

is still a question, as these methods depend highly on the regression method [184]. Researchers have primarily selected the bandwidth arbitrarily because of two reasons. Firstly the theories for selecting the bandwidth have not spread through some fields of study. Secondly, the usage of inappropriate basic assumptions of the theories that only consider some specific fields [166]. However, a suitable bandwidth value helps minimise the error of the regression. In our case, we arbitrarily selected 0.01 as the bandwidth value. We kept that fixed for all the experimented kernels, as we can vary another parameter called noise or lambda (4.3 Comparison of Different Kernel Methods) to get satisfactory performance from kernel regression.

After comparing the curves of selected kernel functions for different noise values, we decided on the best kernel and the noise value applied to our scenario. Then, we used the selected kernel function to fit the AEL and BP values to predict the AEL for a user-defined privacy threshold. AEL values were assigned to the y-axis, and BP values were assigned to the x-axis. Therefore AEL is represented as a function of BP. Then we input the user-defined privacy threshold as a BP value $(1 - \alpha)$ to the fitted kernel and predicted the respective AEL value.

It is vital to re-conduct the perturbation to achieve α for the actual perturbation after deciding with the help of data fitting. To do so, we should know the relevant k value of the SRW to conduct the perturbation. Therefore, we performed another data fitting process for AEL and k values by fitting the kernel function to represent k as a function of AEL. Then we predicted the respective k for the AEL values returned for $(1 - \alpha)$.

7.3.5 Validation Experiments

After getting AEL and k values for α from the data fitting module, it is necessary to evaluate the accuracy of those values. We have two outputs from the data fitting module, AEL and k, for a user-defined privacy threshold. We conducted the validation perturbation experiments using k as the noise addition rate and recorded Average Expected Loss (AEL') and Breach Probability (BP'). Then, AEL' and BP' were compared with AEL and $(1 - \alpha)$ to see how accurate those values are.

This validation process measures the reliability of our framework. By comparing results from the validation experiment with the original experiment, we can measure the error rate of the results produced from the data fitting module. A low error rate implies that the data fitting module has successfully identified the relationship between AEL and BP and accurate predictions. If the error rate from validation experiments is high, the data fitting module has failed to recognize the relationships between AEL and BP. If this is the case, we should improve the data fitting module, and one possible way is to test with some other kernels. This allows us to see the appropriateness of using data fitting as part of APOF. The process involved with validation experiments is illustrated in Figure7.5.

7.4 Results and Discussion

This section discusses the results of accuracy, privacy, data fitting received from APOF, and the results of validation experiments that have been conducted to see the validity of APOF. It is worth highlighting that we did not compare the performance with other state-of-art perturbation methods as SRW has been experimentally proven to achieve high accuracy and privacy levels, and comparison has already been made in [32] and [127]. Here, we try to achieve a further trade-off optimisation by incorporating a data

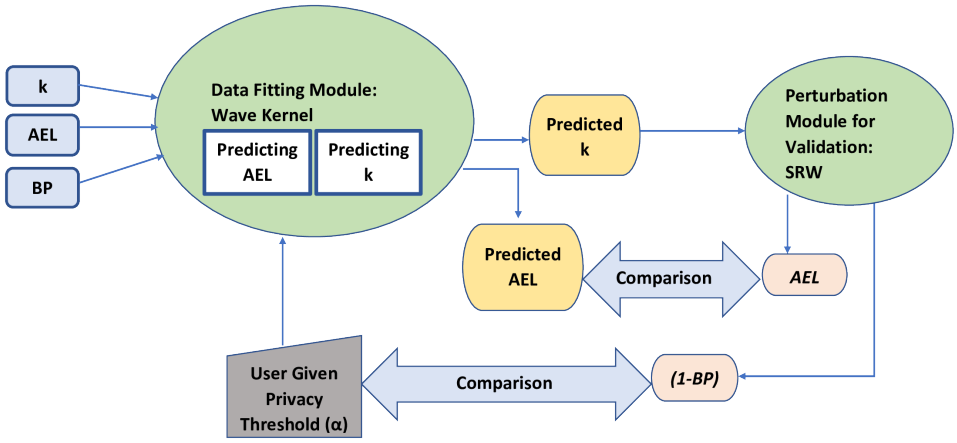

Figure 7.5: Design Diagram – Validation experiments.

fitting module to comply with the user's privacy requirements. We do not try to improve the perturbation method itself. Therefore, the data fitting module has been validated to ensure its appropriateness as a solution to the trade-off issue.

7.4.1 Datasets and Experimental Configurations

Experiments were carried out for three different datasets (AReM, Electricity, and Taxi) and for a range of noise addition rate values of the logistic function and different cycle sizes of the logistic curve. The selected datasets are considered data streams because records are time-stamped. Which means they can be ordered according to the time they were produced. Though these datasets can be treated as data streams, the availability of the concept drift is unknown. All datasets were min-max normalized in the range of 0 and 1 to maintain the fairness of the comparison. The implementation was done using "Clojure," a Java-based scripting language, and executed using Docker-based Jupyter notebooks. Table 7.1 and 7.2 summarizes the details related to the perturbation and data fitting experiments, respectively.

Table 7.1: Experimental Configuration - Perturbation

Item	Value/ Description			
Datasets	Nature	Number of Records	Number of Attributes	Target Variable
AReM from UCI (Activity Recognition system based on Multi sensor data fusion)	Real world	35,999	6 (Numeric)	Activity (Walking, Cycling, Standing, Sitting, lying)
Electricity from OpenML (Collected from Australian New South Wales electricity market)	Real world	45,312	8 (Numeric)	Change of the price (Up, Down)
Taxi from Kaggle (New York City taxi trip duration)	Real world	50,000	7 (Numeric)	Trip duration (Short, Medium, Long)
Perturbation Method	SRW			
Noise addition rate (k)	In the range of 0.005 - 0.1 with intervals of 0.004 (24 values)			
Maximum value of logistic curve (L)	1			
Cycle Sizes	1000, 2000, 4000, 8000			
Return value from logistic function	*f(x)*			
Classification Method	Hoeffding Tree			
Accuracy measurement	Average Expected Loss (AEL)			
Privacy measurement	Breach Probability (BP)			
Variance of cumulative noise	$3.90 \times 10 - 6 \times f(x)$			
MAP attack	Number of known I/O pairs – 4 per attack. Number of attacks – 5% of the records Epsilon (ϵ) – 0.2			

Table 7.2: Experimental Configuration - Data Fitting

Item	Values/Description
Datasets for data fitting (d)	k, AEL, BP results retrieved from original datasets (D).
Noise addition rates (k)	In the range of 0.005 - 0.1 with intervals of 0.004 (24 values)
Data fitting method	Kernel Regression
Experimented Kernel Types	Gaussian, Laplacian, Wave, Rational quadratic, Matern-52
Bandwidth/Smoothness	0.01
Noise/Lambda (λ)	0.1, 0.01, 0.001, 0.001
User-defined privacy thresholds tested for data fitting (α)	0.6, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 0.97
Kernel Regression implementation using	Fastmath library (generateme/fastmath "1.4.0-SNAPSHOT")
Curve fitting using	Cljplot library (cljplot "0.0.2-SNAPSHOT")

7.4.2 Accuracy-Privacy Calculation and Analysis

Privacy (BP) and Accuracy (AEL) results for all the experimented cycle sizes and k values for two datasets are being included in Table 7.3 and Table 7.4. Table 7.3 displays results for Activity Recognition system based on Multisensor data fusion (AReM) dataset and the results for the Electricity dataset is displayed in Table 7.4. Figure 7.5 and Figure 7.6 give an insight into the behaviour of AEL and BP across the *k* values. We used min-max normalized values of all three measures for a clear understanding of trends.

We cannot find any consistent relationship between the values of AEL and BP and the *k* values across all cycle sizes (See Figure 7.6). The curves have fluctuations at some points, but fluctuations have been spread over a very narrow range (See Table 7.3). AEL values vary in a range of 0.0751 for cycle size 1000, 0.0743 for cycle size 2000, 0.0694 for cycle size 4000, and 0.1076 for cycle size 8000. BP values fall in a range of 0.0156 for cycle size 1000, 0.0156 for cycle size 2000, 0.0267 for cycle size 4000, and 0.0200

Table 7.3: AEL & BP Results for Experimented k Values & Different Cycle Sizes (AREM (D))

CS	1000		2000		4000		8000	
k	AEL	BP	AEL	BP	AEL	BP	AEL	BP
0.005	0.13262	**0.03118**	0.13845	0.02004	**0.12915**	0.01782	0.09706	0.04232
0.009	0.08163	0.02672	0.13586	**0.03118**	0.10204	0.02227	0.12972	0.04232
0.013	0.08252	0.02672	0.13591	**0.03118**	0.11135	0.02227	**0.17491**	0.04009
0.017	0.07863	0.02004	0.06711	0.0245	0.08174	0.02227	0.06775	0.03786
0.021	***0.06529***	0.02227	***0.06688***	0.02895	0.08163	0.02004	0.06771	0.04009
0.025	0.10053	***0.01559***	0.07834	***0.01559***	0.08932	0.02895	0.06761	0.04009
0.029	0.11139	0.01781	0.07835	0.02895	0.07148	0.03341	0.06742	0.04009
0.033	0.10106	0.02227	0.13776	0.02895	0.07431	***0.01336***	0.06741	0.04232
0.037	0.13972	0.02895	0.13816	0.02673	0.06104	0.01782	0.06743	0.04232
0.041	**0.14034**	0.02449	0.13852	0.02895	0.07456	0.02673	0.06738	0.04009
0.045	0.14032	0.02672	0.13886	***0.01559***	0.07408	0.02004	0.06737	0.04454
0.049	0.14026	0.02672	0.13921	0.02227	0.07397	0.03341	0.06739	0.04009
0.053	0.1085	0.02004	0.1396	0.02673	0.07394	0.02895	0.06737	0.04677
0.057	0.10753	0.02895	0.13993	0.02004	0.07419	0.02895	0.06736	0.03786
0.061	0.10845	***0.01559***	0.14009	0.02673	***0.05966***	0.01559	0.06739	**0.049**
0.065	0.10849	0.02004	0.14022	0.02227	0.05999	0.02227	0.06737	0.04677
0.069	0.10842	0.02004	0.14035	0.02004	0.06006	0.0245	0.06738	0.04454
0.073	0.10846	0.02227	0.14046	0.01782	0.0601	**0.04009**	0.06737	0.03341
0.077	0.10697	0.02227	0.14061	0.02227	0.06021	0.02673	0.06735	0.04232
0.081	0.09072	0.02449	0.14074	0.0245	0.06003	0.0245	0.06738	0.04009
0.085	0.10806	0.02449	0.14089	**0.03118**	0.07409	0.02895	0.06736	0.04232
0.089	0.10805	0.01781	0.14094	0.02673	0.0741	0.02227	0.06736	***0.02895***
0.093	0.108	0.02672	0.14107	0.02673	0.08806	0.0245	0.06734	0.04454
0.097	0.107	0.01781	**0.14114**	**0.03118**	0.08804	0.03563	***0.06733***	0.04454

for cycle size 8000.

The same behaviour with the AReM dataset can be seen with the Electricity dataset for all three measures. AEL and BP have fluctuations across the *k* values and no clear trend with the cycle size. AEL values are significantly smaller, and BP values are about in the same range compared to the AReM dataset, which is good as it indicates achieving a high accuracy level while maintaining a good privacy level (See Table 7.4). The highest and lowest AEL and BP values for each cycle size are highlighted.

Experimentation results clearly show that SRW is a good perturbation approach in optimising the trade-off between accuracy and privacy. We can see that both measures try to be stable and vary within a minimal range. Though we cannot achieve 100% of both privacy and accuracy, it is possible to achieve around 97% for each measure.

The reason for achieving high accuracy and privacy is the behaviour of the logistic

Table 7.4: AEL & BP Results for Experimented k Values & Different Cycle Sizes (Electricity (D))

CS	**1000**		**2000**		**4000**		**8000**	
k	**AEL**	**BP**	**AEL**	**BP**	**AEL**	**BP**	**AEL**	**BP**
0.005	0.0177	0.02827	**0.01946**	**0.0371**	0.02196	0.0265	**0.0229**	0.0212
0.009	0.01738	0.01943	0.01713	0.0318	0.01787	0.03004	0.02246	0.0265
0.013	0.01859	0.02827	0.01717	0.0212	***0.01756***	0.02827	0.01783	0.03004
0.017	**0.02387**	0.0212	0.01687	0.02473	0.01961	0.02827	0.02041	0.02473
0.021	0.01755	0.01767	0.01899	0.0265	0.01834	0.0318	0.01874	***0.01943***
0.025	0.01785	0.0265	***0.01641***	***0.0159***	0.01953	***0.02297***	0.01913	0.0212
0.029	0.0169	0.02297	0.01872	0.03534	0.01877	0.02827	0.01812	0.03004
0.033	0.01686	0.0212	0.01764	0.02297	0.01952	**0.0371**	0.01934	0.01767
0.037	0.01626	0.01943	0.01793	0.02827	0.02035	**0.0371**	0.01856	0.0212
0.041	0.01629	0.03004	0.01792	0.0265	0.01988	0.0265	0.01906	0.03004
0.045	***0.01621***	0.02297	0.01869	0.01943	0.01943	0.03004	0.01783	0.02827
0.049	0.01754	0.0265	0.01869	0.02297	0.0195	***0.02297***	0.01906	0.0318
0.053	0.01784	***0.01413***	0.01869	0.02297	0.02153	0.03534	0.01891	0.0212
0.057	0.01767	0.01767	0.0187	0.0318	0.02108	0.0424	0.01873	0.02473
0.061	0.01785	0.0212	0.01751	0.0265	0.01925	0.0265	0.01875	0.0265
0.065	0.01782	0.0159	0.01871	0.0265	0.01939	0.03004	0.01873	0.0265
0.069	0.01782	0.03004	0.01868	0.0265	0.01803	0.04064	0.01873	0.02827
0.073	0.01766	0.0212	0.0176	0.02827	0.01793	0.0318	0.01875	**0.03887**
0.077	0.01766	0.01943	0.01762	0.02473	0.01944	0.02827	***0.01748***	0.02827
0.081	***0.01621***	0.0212	0.01761	0.02473	**0.02232**	0.03357	0.01751	0.0371
0.085	0.01622	0.0212	0.01873	0.02297	0.01972	0.03357	0.01822	0.03004
0.089	0.01638	***0.01413***	0.01875	0.0318	0.01973	**0.0371**	0.01823	0.02827
0.093	0.01637	**0.0318**	0.01875	0.0265	0.01973	0.03004	0.01809	0.0318
0.097	0.01637	0.02473	0.01855	0.02473	0.01971	0.0265	0.0177	0.0265

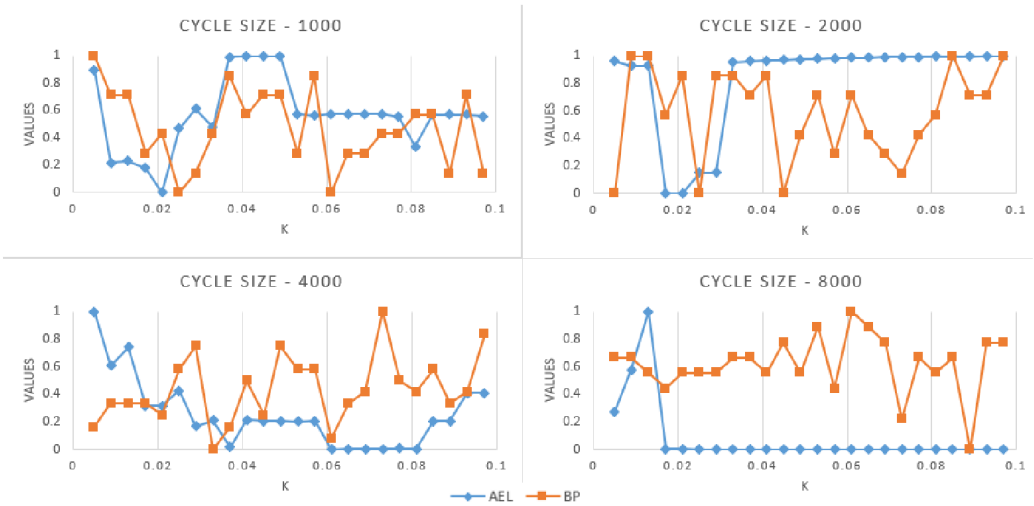

Figure 7.6: The behaviour of AEL and BP (AReM Dataset) for different cycle sizes

curve. Since we add noise with a minor variance at the beginning of the logistic curve, the classification model gets time to learn. Then in the middle part of the logistic curve, the noise variance and k increase, improving privacy. Then in the ending flat area of the curve, noise is added with a maximum variance which positively affects privacy. However, the noise addition rate becomes consistent, which helps to improve the accuracy. Therefore, both accuracy and privacy get the opportunity to be improved throughout the logistic cycle.

When considering BP and AEL, those measures do not directly relate to k. However, what has been proved is that k is a critical factor in varying BP and AEL as changing k results in achieving high or low values of those measures. Therefore, it is a valuable finding to extend this framework to its next step, deciding accuracy and privacy according to the user's requirements.

Cycle size also plays a role in obtaining various levels of privacy and accuracy, but it does not directly relate to BP and AEL. If the cycle size is too small (ex- 100), privacy can be negatively affected because the total noise variance added within a cycle can be less. On the other hand, if the cycle size is too large (ex- 20,000), then accuracy can be

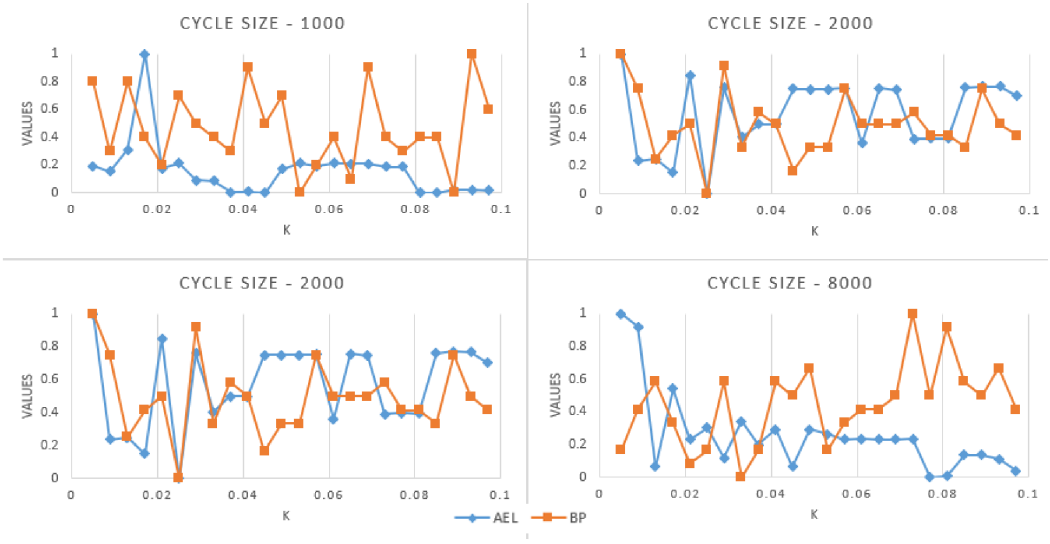

Figure 7.7: The behaviour of AEL and BP (Electricity Dataset) for different cycle sizes

negatively affected because of the high noise. If we carefully analyse the AEL values in Table 7.4, an increasing trend in AEL can be found. Therefore, a moderate value for the cycle size should be selected. Since the experiments do not provide solid grounds to decide precisely on the cycle size, we have selected 2000 as the cycle size that will be used for the next steps.

We conclude that SRW optimises the trade-off between accuracy and privacy using different noise addition rates (k). Varying k within a possible range of values helps achieve different trade-offs. This opens the possibility of combining a data fitting module into APOF.

7.4.3 Comparison of Different Kernel Methods

After conducting experiments on the perturbation process and recording AEL and BP values, we pass AEL and BP values into the data fitting module with α to predict the accuracy for α. For predicting the accuracy, it is essential to decide on the most suitable kernel function to use as the data fitting module. We conducted experiments with five

different kernel methods on three different datasets (d) to decide on the best kernel we could use regardless of the dataset.

Another critical factor we had to decide is the optimal value for lambda or noise to get the best curve to fit data, as we have fixed the bandwidth value to 0.01. As kernel regression is non-parametric, it does not produce any parameter or coefficient values as other polynomial regression methods. Therefore, we used a visualization method (curve fitting) method to select the best lambda values that capture the patterns in the datasets (d) and decide on the kernel we will select from the experimented five kernels.

For the regression, we used BP on the x-axis and AEL values on the y-axis since we want to predict the respective AEL value for a given privacy threshold. Figure7.8 and Figure7.9 show the curve fitting results (AEL against BP) of five different kernels together with different lambda values for AReM (See Figure7.8) and Electricity (See Figure7.9) datasets (d).

Figure7.8 and Figure 7.9 show that lambda significantly impacts the regression results. When lambda is high, curves are under-fitting, and lambda is small, curves tend to be over-fitting. Therefore, selecting an appropriate lambda value is crucial. After carefully examining the curves, we can see that when lambda is 0.1, the regression process cannot capture all the patterns of the dataset. In contrast, when lambda is 0.0001, the regression process tries to reach extreme dataset points. Because of this reason, we decided to use 0.001 as the lambda for our framework.

Moreover, Figure7.8 and Figure 7.9 show that Gaussian, Wave, and Matern-52 kernels perform considerably better by fitting the curves fairly for most of the data points in all the datasets. The rational-Quadratic kernel fails to capture most of the data points, and the Laplacian kernel tries to capture all the data points, hence tends to be over-fitting easily. Using this observation, we could narrow down the list of kernels into Gaussian, Wave, and Matern-52. To decide on the kernel that can be used for most datasets, we recorded the predicted AEL values of data fitting results for the

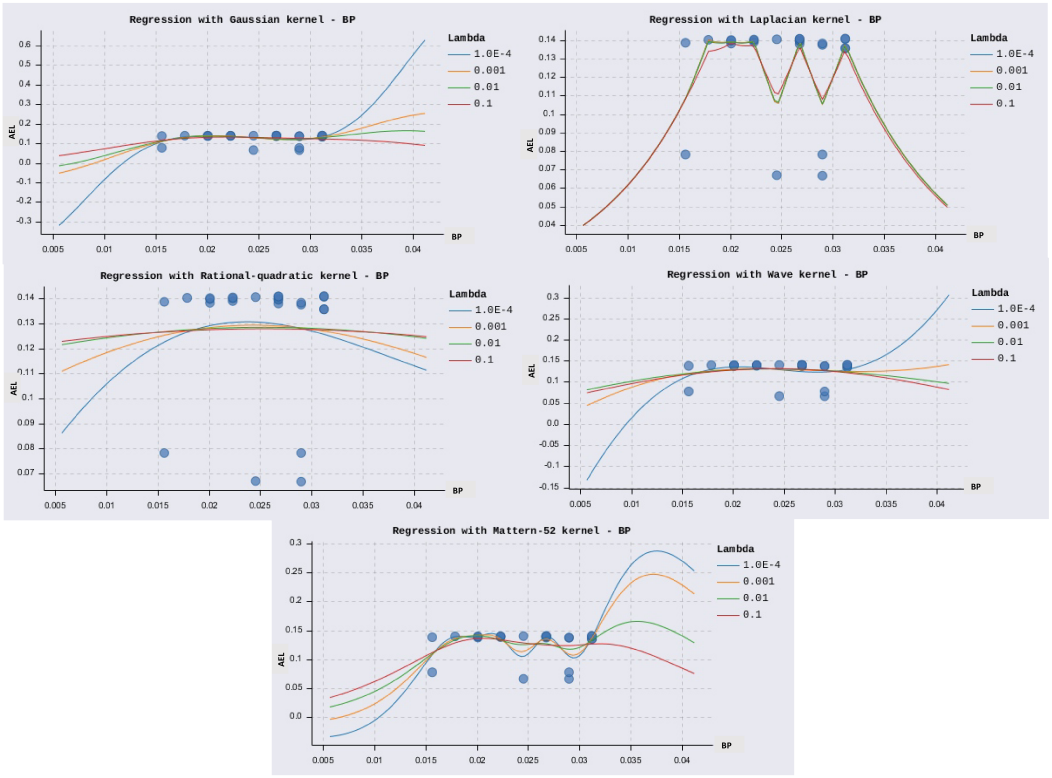

Figure 7.8: Curve fitting (AReM (d)) with different lambda values.

experimented privacy thresholds. The experimental results for the three kernels we selected are displayed in Figure7.10 for all three datasets.

By looking at the predicted AEL values for a set of user-defined privacy thresholds (See Figure 7.10), we can see that both Gaussian and Matern-52 kernels were unsuccessful in predicting AEL values for low privacy threshold values. This is because, for all the datasets, AEL values are almost zero. Though the privacy threshold is low (say 0.6 or 0.7), the dataset must sacrifice some accuracy to achieve that and AEL being zero is not practical. However, the AEL values predicted using Wave Kernel seem more sensible. Therefore, we have selected Wave Kernel as the kernel we will use for our data fitting module. When observing predicted AEL values for user-defined privacy thresholds using wave kernel, it can be seen that AEL increases when privacy increases. The increased range of the AEL for AReM, Electricity, and Taxi datasets (d) is 0.0184,

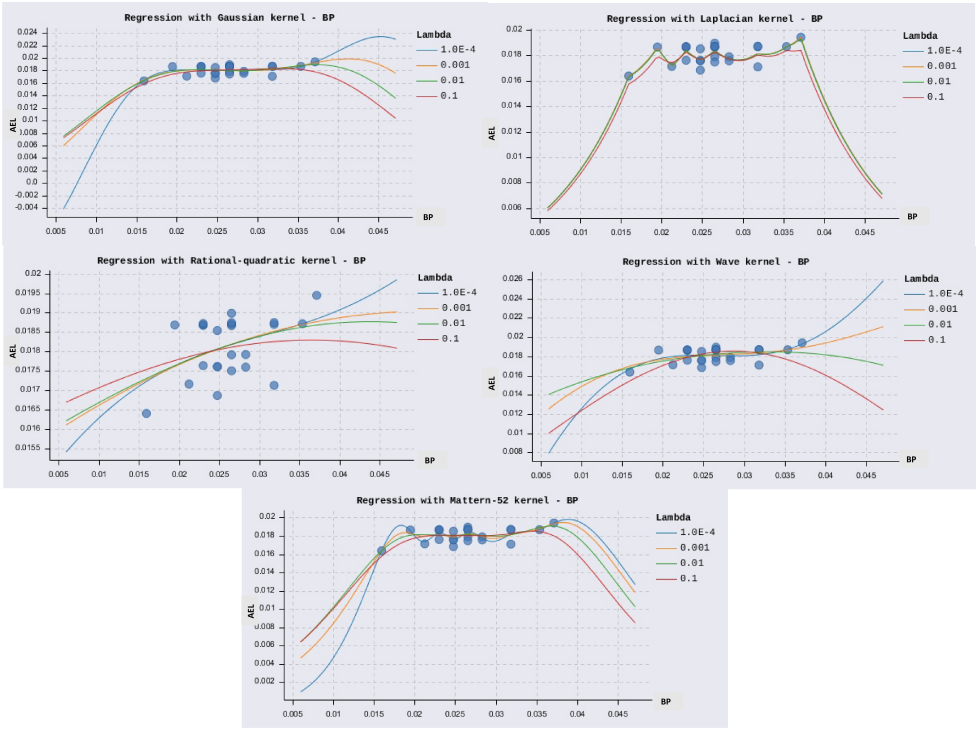

Figure 7.9: Curve fitting (Electricity (d)) with different lambda values.

0.0167 and 0.0699 respectively when privacy increases from 0.6 to 0.97. This implies that the accuracy drop according to the fitted data is also similar to that range.

7.4.4 Data Fitting and Validation Experiments

We used the Wave Kernel (with lambda = 0.001) as the data fitting function and predicted the AEL value for the privacy threshold by representing AEL as a function of privacy. Then we used the predicted AEL value to predict the respective noise addition rate (k). Therefore, we have two outputs from the data fitting module: AEL and respective k for a user-defined privacy threshold.

The results of Figure7.10 prove that the data fitting can predict the AEL value for a user-defined privacy threshold. However, it does not indicate how successful or accurate the predicted AEL values are. We carried out a validation experiment (See Table 7.5)

Privacy User threshold	Predicted AEL - AReM (d)		
	Gaussian	Wave	Mattern 52
0.6	8.341E-295	3.099E-02	2.685E-33
0.7	2.397E-156	3.847E-02	7.369E-24
0.75	2.075E-103	3.212E-02	3.515E-19
0.8	2.410E-61	4.218E-02	1.509E-14
0.85	3.482E-30	8.009E-02	5.423E-10
0.9	4.810E-10	8.657E-02	1.341E-05
0.95	1.802E-01	1.948E-01	8.567E-02
0.97	1.259E-01	1.266E-01	1.119E-01

Privacy User threshold	Predicted AEL - Electricity(d)		
	Gaussian	Wave	Mattern 52
0.6	1.752E-287	1.615E-03	3.530E-34
0.7	1.349E-151	2.188E-03	9.626E-25
0.75	6.080E-100	1.410E-03	4.567E-20
0.8	3.740E-59	2.706E-03	1.945E-15
0.85	3.052E-29	4.091E-03	6.883E-11
0.9	2.963E-10	6.378E-03	1.642E-06
0.95	1.475E-02	2.189E-02	8.612E-03
0.97	1.809E-02	1.831E-02	1.778E-02

Privacy User threshold	Predicted AEL - Taxi(d)		
	Gaussian	Wave	Mattern 52
0.6	5.248E-270	7.487E-03	1.272E-32
0.7	7.632E-139	6.795E-03	3.395E-23
0.75	1.522E-89	1.292E-02	1.581E-18
0.8	4.250E-51	3.525E-03	6.534E-14
0.85	1.582E-23	2.735E-02	2.185E-09
0.9	5.282E-07	6.248E-03	4.503E-05
0.95	7.434E-02	8.145E-02	7.473E-02
0.97	7.902E-02	7.748E-02	7.745E-02

Figure 7.10: Predicted AEL for different user-defined privacy thresholds (AReM, Electricity and Taxi datasets (d))

using predicted AEL values and respective noise addition rate values (k) to see the accuracy of the predicted results.

Table 7.5 displays the results of data fitting using wave kernel and the results of validation experiments for all three datasets.

For the privacy thresholds higher than 0.95/0.97, we can see that AEL and privacy from data fitting are approximately similar to those from validation experiments which prove that the data fitting can be used to successfully predict AEL for a user-defined privacy threshold in this scenario. If we calculate the average error for predicted AEL values for the privacy thresholds from the data fitting function, it is 0.176%, 0.036%, and 0.109% from AReM, Electricity, and Taxi datasets (d), respectively. Generally, we can conclude that this is a successful method to predict values from the same range as fitted data with a small error margin.

We experimented with predicting the AEL for lower privacy thresholds, such as 0.6

Table 7.5: Validation Results

Privacy	**Data fitting - Wave kernel**		**Validation experiments using k from data fitting.**	
Threshold	**AEL**	k	**Privacy**	**AEL**
AReM Dataset (d)				
0.97	0.126132	0.122469	0.975501	0.121488
0.98	0.136161	0.010055	0.98202	0.135987
0.985	0.108826	0.579233	0.982963	0.107357
0.99	0.155141	0.14333	0.985501	0.156683
0.995	0.156358	0.777001	0.985501	0.157356
Electricity Dataset (d)				
0.97	0.018115	0.052841	0.973498	0.01872
0.98	0.018882	0.050582	0.978198	0.018695
0.985	0.016295	0.024911	0.987731	0.015414
0.99	0.018668	0.098582	0.988431	0.018728
0.995	0.018733	0.460686	0.988431	0.018793
Taxi Dataset (d)				
0.95	0.072801	0.110657	0.9568	0.072874
0.97	0.077859	0.060662	0.9584	0.077515
0.98	0.073778	0.033562	0.9712	0.077889
0.985	0.059852	0.007068	0.96	0.06072
0.99	0.075553	0.099112	0.9568	0.076213
0.995	0.077098	0.066247	0.9712	0.07756

(60%) and 0.7 (70%). However, the data fitting module performed poorly for lower privacy thresholds for all three datasets. The most plausible reason is that the privacy values we use for data fitting are always higher than 0.95, and the data fitting function cannot predict the values out of the range. This is one of the issues related to most data fitting functions since those only work for the values in the given range. If we see the BP values from Table 7.3 and Table 7.4, those are extremely small. Hence, privacy is high because the perturbation method we used can achieve a high privacy level while maintaining a considerably high accuracy level. After fitting data with high privacy values, if we try to predict an AEL for a privacy value out of the range of fitted values, it may become incorrect. Suppose there is a need to achieve privacy values lower than about 85%. In that case, we can replace the privacy module with another perturbation method such as traditional noise addition that allows distorting data to a great extent. However, SRW has been optimised to achieve privacy of more than 85% while maintaining more than 85% of accuracy.

Another important conclusion is that the proposed perturbation method (SRW) successfully achieves high privacy while maintaining high accuracy. We cannot manipulate the *k* value to decrease privacy in higher amounts. This can be seen by comparing privacy thresholds with the privacy values retrieved from validation experiments. Though we try to predict *k* values for lower privacy levels using the data fitting function, those predicted *k* values produce higher privacy levels.

7.5 Conclusion and Future Works

The proposed Accuracy-Privacy Optimisation Framework (APOF) has been experimentally proven to optimise the accuracy-privacy trade-off in data stream mining using the SRW [127] perturbation method that achieves high accuracy and privacy values of more than 90%. A novel data fitting module helps comply with the user-defined privacy requirements.

The performance of the APOF framework depends on four key factors; the dataset, classification model, perturbation method, and data fitting method used. The APOF has accurately predicted accuracy for user-defined privacy levels with average errors of 0.176%, 0.036%, and 0.109% for AReM, Electricity, and Taxi datasets (d), respectively. APOF can be used with any classification algorithm appropriate for data stream mining and with any perturbation method if we can vary at least one of its parameters to fit into the data fitting module. Experimentally testing the accuracy behaviour of APOF with other classification algorithms is a possible future direction that helps prove the generalizability of the framework.

The data fitting module can be tested in the future to consider accuracy and find achievable privacy levels. This can be useful when the users give priority to accuracy. The exactly same regression process is involved in that, and the only difference is to use accuracy values on the x-axis and privacy values on the y-axis (See Section 7.3.4

and Section 7.4.3). Selecting a suitable kernel method can be different as this process needs to represent privacy as a function of accuracy, which is also worth investigating. Data fitting using kernel methods can accurately predict values in the range of the fitted values. The limitation of this method is that if we try to predict a value out of range, it tends to be incorrect. Fitting data values out of the range is challenging and need further investigation.

Hoeffding tree is an incremental decision tree designed for data stream mining. However, it works under the assumption that the data distribution does not change over time; that is not always true in real-time. Also, adding noise cumulatively cannot be done forever because noise values may overpower the original data values, and the classifier may start to learn noise instead of original data. This can significantly impact accuracy, especially for infinite data streams. Therefore, the cumulative noise needs to be reset to zero and restarted at some point in the data stream to avoid this issue. The immediate next step of our research is to improve APOF to detect and adapt to the concept drift in data streams and deal with infinite data streams.

Chapter 8

Prelude - Manuscript 4

Data streams differ from static datasets due to their dynamically adapting nature, including characteristics such as incremental behaviour and change in the underlying data distribution. This makes data mining and privacy preservation in data streams rather challenging. Data mining algorithms should adapt with the concept drift to maintain a steady accuracy throughout the data stream. Privacy preservation techniques should deal with infinite or lengthy data streams, and the overall process should be carried out without delay. Though Privacy-Preserving Data Stream Mining (PPDSM) operates under the same objective as PPDM, the same techniques cannot be used as it is. PPDM techniques should be changed or modified to handle the challenges that arise from data streams.

Considerable attention has been given to developing PPDSM techniques for dealing with different challenges, such as high volume and handling concept drift. However, optimising the accuracy-privacy trade-off in PPDSM has not received sufficient attention. Though research articles point out the issue, fewer of them make an effort to find solutions. Regarding the accuracy-privacy optimisation frameworks, we can only find a few in the literature. Therefore, a framework with enhanced data mining and privacy preservation techniques can answer the optimisation issue in PPDSM.

Manuscript 4 aims [185] to develop an enhanced framework for accuracy-privacy optimisation in PPDSM. It uses the APOF framework proposed in Chapter 7 as the base model. Though APOF [149] has some features suitable for data stream mining, it is not ideal for PPDSM. This is because of a few reasons, such as APOF cannot deal with concept drift and has serious accuracy and privacy concerns when working on infinite or lengthy data streams. In manuscript 4 we improve APOF to work with data streams considering its behaviour. The developed Enhanced-APOF contains features such as concept drift handling, maintaining steady accuracy and privacy for infinite or lengthy data streams, and improving the classifier's effectiveness by avoiding over-fitting.

Manuscript 4 addresses RQ3, which is "**How do we extend the framework proposed in RQ2 above to a dynamically adapting or evolving data streaming environment?**".

Chapter 9

An Efficient and Enhanced Privacy-Preserving Framework to Achieve Optimal Accuracy-Privacy Trade-off for Evolving Data Streams (Manuscript 4)

9.1 Introduction

Data mining is a branch of machine learning that plays a vital role in supporting organisations in their decision-making [4]. Organisations use different data mining techniques such as classification, clustering, and regression analysis to mine the existing data to make predictions [56, 6] about their future operations and changes. The historical data in the databases are mainly used to make predictions. These predictions from data mining and expert knowledge rule the decision-making process. However, data mining has been extended and specifically focused on "data stream mining," which uses data

from data streams instead of static databases.

We live in an era of data. Millions of data from thousands of operations are being produced in a second. Mining data in real-time is needed to get the maximum use of these data. Hence, traditional data mining that uses historical data from static databases no longer applies to this scenario. This is where the area "data stream mining" comes into play. Data streams are different in their nature and behaviour than databases. Streaming data can be continuous, transient, and unbounded, which requires them to be processed quickly [47, 51]. Therefore, data mining techniques also need to be improved and adapted to map with the nature of the data streams to get accurate results from data mining. The success of the decisions depends on the accuracy of the predictions made by the data mining in this context.

Privacy of the data being used for data mining is another aspect that needs attention in addition to the accuracy of the data mining results [21]. The identity of individuals and sensitive data related to individuals should not be revealed in data mining. Privacy-Preserving Data Mining (PPDM) [44, 26] was introduced to protect the privacy of individuals while using data to make predictions. Privacy is achieved by transforming the actual values of data to another form using different techniques such as perturbation, suppression, and anonymisation [150, 151] before using data in data mining. This process makes it difficult to identify individuals by looking at the transformed data. The objective of PPDM is not simply to protect privacy. PPDM methods should protect privacy while allowing the data mining process to carry out as usual without affecting accuracy. Therefore, PPDM methods should be able to preserve the statistical properties of data [39, 46] so that the accuracy of the data mining results will not be degraded.

However, there is an unavoidable trade-off between data privacy and data mining accuracy, as these properties are highly interlinked [38, 23]. Changing the actual values of data to preserve privacy negatively affects data mining accuracy [32]. Increasing privacy causes accuracy decrements directly and vice versa. Though it is impossible

to achieve perfect accuracy and privacy levels simultaneously [10], optimising the trade-off between privacy and accuracy is a must to achieve the objective of PPDM. Optimisation subjects to the environment it uses as the data mining method and privacy preservation method need to be modified accordingly, especially if it is a data streaming environment. PPDM methods with static databases cannot be used unchanged for data stream mining. They should be improved and modified by considering the nature and behaviour of data streams [53, 44].

Though numerous PPDM methods [47, 51, 117, 38, 32, 127] have been proposed for data stream mining, only a few of those pay attention to optimising accuracy-privacy trade-off. Moreover, there is a lack of well-formulated frameworks for this purpose in data stream mining [149]. This work aims to produce a well-formulated framework to optimise the accuracy-privacy trade-off in data stream mining using already developed techniques or frameworks. So it can be efficiently adopted and work with the nature of data streams.

The "Accuracy-Privacy Optimisation Framework" (APOF) [149] is an advanced framework that uses different techniques to optimise the trade-off between data privacy and classification accuracy. APOF was proposed for use in data stream mining but requires several improvements like handling concept drift and dealing with the lengthy or infinite data stream. APOF consists of three main modules: accuracy, privacy, and data fitting [149]. It uses the Hoeffding tree [132] as the data mining technique and random projection-based logistic cumulative noise addition (SRW) [127] as the perturbation method. APOF optimises the accuracy-privacy trade-off considering the user's privacy requirements with the help of a kernel function-based [162, 168] data fitting module [127].

However, all three modules of APOF need modifications to be used in data stream mining. Hoeffding tree is an efficient classifier for working with data streams but does not adapt to changes or concept drifts. Adapting to the concept drift is essential for

data stream mining. We cannot predict the changes in original data, and accuracy can drop drastically if the classifier fails to adapt. The perturbation method used in APOF adds noise with a small arbitrary variance cumulatively that is suitable for data streams. It helps to preserve privacy while maintaining a good accuracy level. Nevertheless, cumulative noise addition cannot be done forever for lengthy or infinite data streams, as a high noise level can negatively impact the accuracy [149, 32]. Therefore, the perturbation method should be improved to work with infinite data streams. It is impossible for the data fitting module to use the whole data stream to fit the data, as APOF does with the static databases. A section of the data stream should be acquired to fit the data. This section needs to be a well-represented sample of the whole data stream to get the optimal results according to the user's requirements.

This article produces an efficient privacy-preserving framework for data stream mining by adapting and modifying APOF as the base model. Our research makes the following contributions to privacy-preserving data stream mining.

- Analysing and adapting Accuracy-Privacy Optimisation Framework (APOF) [149] for data stream mining.
- Improving logistic cumulative noise addition [127, 149] to work with infinite data streams with guaranteed privacy.
- An efficient window-based accuracy and privacy monitoring method for data streams.
- A classifier switching method to increase the efficiency of the learning process, considering the stability of the accuracy while reducing the risk of overfitting.
- A comparative analysis of Hoeffding Tree and Hoeffding Adaptive Tree in the context of PPDM.

The remainder of this article has been organised as follows. Section 9.2 explains the existing work of privacy-preserving data stream mining. Section 9.3 describes the proposed methodology, design, and other related techniques used in detail, while Section 9.4 consists of the experimentation results and discussion. Finally, Section 9.5 presents the conclusion and future directions.

9.2 Related Literature

This section presents the existing literature on data stream mining, PPDM methods, and the accuracy-privacy trade-off in data stream mining. Primarily, we reviewed articles discussing data streams and mining data streams in general. Then we looked into PPDM methods proposed for data streams and to what extent these PPDM methods address the accuracy-privacy trade-off. The primary and vital details found through the literature review are summarized here. Subsequently, we highlight the existing gaps in the literature that directed us to continue this research.

9.2.1 Data Stream Mining

Mining data streams is rather challenging due to their unique characteristics. Volume, velocity, and volatility are the three principal challenges in data streams [45, 31]. Additionally, there are other concerns such as data pre-processing, delayed data handling, fast execution, handling concept drift, and preserving privacy [45, 48]. Data stream mining is continuous and cannot be redone for static datasets. This is due to the inability to access the whole dataset at once [52]. Execution should be carried out quickly, as data may reach a high-speed [47, 46]. Another vital issue is handling the concept drift. When the underlying data distribution changes with time, data mining techniques should adapt to the change unless accuracy has to suffer and cannot gain the maximum results from data mining [54, 52].

The above-discussed challenges should be considered when selecting a data mining model for data streams to achieve maximum performance. Support Vector Machines, Neural Networks, Decision Trees, Naïve Bayes, and Random Forest are popular classification models used for data stream mining. Hoeffding trees are decision trees specifically developed by Massive Online Analysis (MOA) by the University of Waikato for data stream mining [132]. MOA consists of Hoeffding Tree (HT), Hoeffding Adaptive Tree (HAT), Hoeffding Option tree (HOT), and many more data mining models. Hoeffding tree is a very fast decision tree [155] that learns incrementally from data streams using constant time and memory per record. It is one of the most efficient and straightforward classification techniques for data stream mining [15]. However, it operates under the assumption that the underlying data distribution does not change over time [132]. Hence, it cannot handle concept drift and can cause decrements in accuracy. On the other hand, HAT has been implemented to work with evolving data streams and is capable of adapting to the changes in the data stream itself [145]. The HAT uses the ADWIN change detector to monitor the performance of branches on the tree. If accuracy drops, those branches are replaced with more accurate branches [145, 156]. Therefore, considering all these factors, it is justifiable that the HAT is more suitable for evolving data streams.

9.2.2 PPDSM Techniques

Privacy preservation of data streams has become more complicated and challenging due to the above-discussed characteristics of data streams. Data streams can be unbounded and continuous [49, 50, 51]. Due to this, PPDM methods should be able to cope with massive volumes of data [32]. Also, privacy preservation techniques must be performed as fast as possible to release data with a minimum delay. Therefore, generic PPDM methods cannot be used unchanged for data streams and need improvements

and modifications to handle the nature of data streams. For example, the field rotation and binning method proposed in [111] cannot be used for data streams because all data should be presented to the binning process, which is complex with data streams due to the incremental behaviour.

Different PPDSM methods have been designed specifically for data streams to overcome the challenges in data stream mining. Anonymization based PPDSM methods such as *k*-anonymity [30], *l*-diversity [36] and *t*-closeness [101] are the most popular privacy preservation technique in data stream mining. Anonymization uses techniques to make a single record indistinguishable from another set of a specified number of records so that the individuals identified [30, 58]. An anonymization method called FAST uses a multithreading technique through *k*-anonymization [114]. This method provides fast execution with less information loss. A continuously anonymizing method named "CASTLE" has been implemented using *k*-anonymity and *l*-diversity [18]. It can manage outliers and release data with a minimum delay. Another anonymization method based on micro aggregation was proposed in [52]. It can deal with concept drift and provides minimal information loss. Some other anonymization-based PPDSM methods were proposed in [50, 116, 115]. The major issue of anonymization, in general, is that it assumes no two tuples contain data of the same person, which is invalid in some scenarios [30]. There is no computational approach to determine what data should be anonymized either [103].

The other category of PPDSM techniques is Perturbation. Perturbation can be defined as changing data using certain techniques to hide sensitive information while maintaining data properties important for data mining [29]. Additive and multiplicative noise [9], random rotation [34], random projection [35] and condensation [37] are some of the techniques that fall under the category of data perturbation. A PPDM technique called “P2RoCAl” that combines condensation, rotation, and random swapping was proposed in [47]. P2RoCAl provides better accuracy, but condensed group size can

affect the performance. Random projection-based cumulative noise addition implemented in [32] combines random projection, translation, and noise addition to achieve good accuracy and privacy. This method proposes a novel noise addition method called cumulative noise addition that helps to balance accuracy and privacy. Authors of [127] have improved the method proposed in [32] to control the noise level and named the noise addition method "Logistic cumulative noise addition." In addition, research works proposed in [117, 51, 29] can be considered perturbation-based PPDSM methods for data stream mining.

Additionally, other PPDSM methods that use different techniques to preserve privacy in data streams can be seen in the literature. Fuzzy logic and PCA [118], differential privacy [38], sliding window [46], and hashing [119] are a few of these techniques used in PPDM methods. All these methods try to address challenges in data stream mining, and we can see that a considerable effort has been made to improve or implement PPDM methods for data stream mining.

9.2.3 Accuracy-privacy Trade-off in Data Stream Mining

The trade-off between data privacy and data mining accuracy has grabbed more attention over the past few years. This issue should be adequately addressed, and if not, it violates the objective of PPDM, which is preserving the privacy of sensitive data while maintaining the knowledge in original data [46, 32]. Many research that proposes PPDM techniques for static data has identified and/or discussed the accuracy-privacy trade-off [12, 74, 55, 83, 9, 16, 85, 122, 75, 10].

Accuracy-privacy discussion in data stream mining is relatively low compared to static data. Due to the challenging behaviour of data streams, optimising the accuracy-privacy trade-off is much more complicated in data stream mining. However, research works such as [126, 46, 18] have discussed the accuracy-privacy trade-off, and some

have proposed solutions. Authors of [54] have used Sequential Backward Selection (SBS) of the greedy algorithm and k-fold cross-validation to select the optimal mode in NB classification to achieve a balanced accuracy-privacy trade-off. A micro-aggregation-based differential private stream anonymisation has been proposed in [52]. It evaluates the trade-off using the classifier's disclosure risk and Area Under the Curve (AUC). The differential privacy-based PPDM method "SEAL" implemented in [38] provides flexibility to select privacy parameters according to the domain and dataset to optimise the accuracy-privacy trade-off. P2RoCAl [55] tries to achieve the same by combining condensation and rotation. Random projection-based cumulative noise addition [32] adds noise with a small variance cumulatively to minimise the effect to accuracy while maintaining a good level of privacy. The improved version [127] of [32] extends the optimisation to another degree by controlling the noise addition rate and maximum noise level.

Though the above-discussed methods optimise the accuracy-privacy trade-off to a greater degree, there is still room for improvement. PPDSM still lacks sophisticated frameworks to optimise the accuracy-privacy trade-off that can be used with any data except [38] and [149]. However, these techniques do not address all the challenges related to data stream mining and still need improvements to achieve the highest results.

9.2.4 Identified Gaps in Existing Work Related to PPDSM

By analysing existing work in PPDSM, several research gaps that need to be addressed were found. The primary reason for these gaps is the challenging nature of data streams. Though different PPDM techniques have been proposed for data stream mining, few have attempted to optimise the accuracy-privacy trade-off, an essential matter in the area. Moreover, each of these proposed methods has different weaknesses, and they do not address the challenges of data streams in general. For example, some methods

cannot deal with infinite data streams, while others fail in fast execution. Therefore, there is a need for a sophisticated framework to efficiently deal with the nature of data streams in general while optimising the accuracy-privacy trade-off. Additionally, improving the efficiency of the data mining process and accuracy-privacy monitoring in data streams are some other areas that might sharpen the overall PPDM process. This research addresses these gaps to improve PPDSM.

9.3 Proposed Methodology and Design

This section explains the process of adapting and modifying APOF [149] to be used in a data streaming environment in detail. The methodology starts by analyzing APOF to identify its weaknesses and available room for improvement.

9.3.1 Analyzing the Base Model - APOF

Accuracy-Privacy Optimisation Framework (APOF) [149] was proposed to optimise the trade-off between data mining accuracy and data privacy in a data streaming environment. APOF consists of three main modules: accuracy, privacy, and data fitting. Each of these modules was designed to handle different objectives of PPDM. Figure 9.1 depicts the design diagram of APOF implemented in [149]

The accuracy module of APOF contains a classifier to predict data and calculates the accuracy in terms of Average Expected Loss (AEL) to measure the performance. It uses Hoeffding Tree (HT) as the classifier that can learn incrementally from data streams [132]. It has been categorized as a very fast decision tree, but there is a major concern related to data streams. The data distribution can change over time; hence data streams evolve. So, the classifier should detect and adapt to these changes in data. In other words, the classifier should be able to deal with concept drift to maintain a consistent accuracy level. Unfortunately, HT assumes that the data stream does not change over

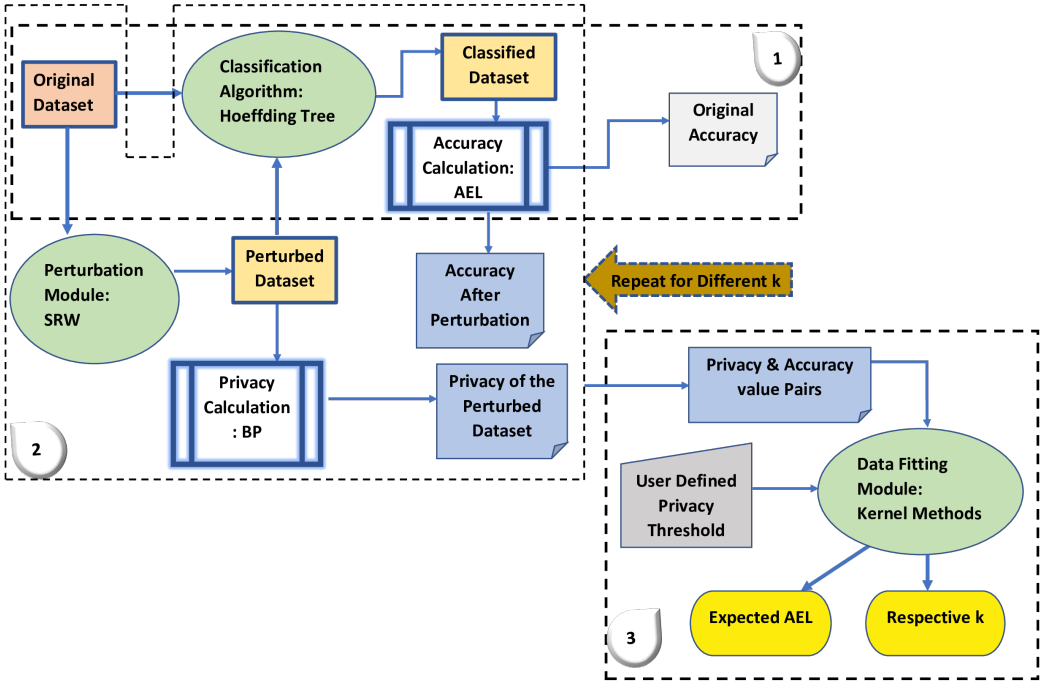

Figure 9.1: Design diagram of APOF

time [132] and cannot be used with data streams containing concept drifts. This is the major drawback of the accuracy module when it is being used in data streams. The classifier should be replaced with an appropriate alternative, considering the nature of data streams. Moreover, APOF has introduced AEL ($AEL = \Sigma_{i=0}^{n} P(X_i)(1 - P(L_i))$) [149] to measure the accuracy. It uses incorrectly classifying probability of each leaf node ($1 - P(L_i)$) and the probability of classifying a record along the path to that leaf node ($P(X_i)$). This terminology is valid when the final tree contains all the classified records.

However, the tree also changes if the data stream evolves due to the concept drift. It may have to remove or replace branches, adapting to the changes. The tree may not contain all the records it has classified by the time we access it. Therefore, AEL may not represent the overall accuracy fairly if there is a concept drift. So, it is worth paying attention to the accuracy calculation of APOF when adapting it to the data streaming

environment. The most convenient option is to use standard error instead of AEL. Additionally, overfitting the classifier is possible due to constantly learning over a long period in the case of lengthy or infinite data streams. This scenario also reduces the efficiency of the accuracy module. A method to use the classification process efficiently will increase the performance of APOF.

The privacy module of APOF consists of a perturbation method and a method to calculate privacy. APOF uses logistic cumulative noise addition (ϖ) [127] together with random projection (R) and translation (Γ) on the original dataset (X) to achieve high privacy [149]. The perturbed dataset (Y) after the overall perturbation process, can be defined as $Y = \frac{1}{\sqrt{k}\sigma_r} RX + \Gamma + \Sigma_{i=-cs}^{+cs}(\varpi)$. In logistic cumulative noise addition, noise is added in cycles, and the noise variance is decided using the logistic function [128]. So, each noise addition step uses a different noise variance that guarantees high resistance to different attacks. This perturbation method also provides a way to control the maximum noise level and the noise addition rate using the effect of the logistic function. Adding noise cumulatively ensures that accuracy does not affect to a higher degree and increases privacy at the same time [127, 149].

However, cumulative nose addition cannot be continued for a long time, especially for infinite data streams. Though it starts with a small noise variance, cumulative noise variance is high for a large number of records. This ultimately reduces the accuracy provided by APOF as the classifier begins to learn noise instead of data [149]. Therefore, a method that stops or reduces this effect should be incorporated with the perturbation method to maintain continuous performance. Using a single cycle size to define a cycle for logistic cumulative nose addition can be vulnerable to attacks if the cycle size is disclosed to unauthorized parties. The reason is that noise variance is low at the beginning of the cycle, and therefore the starting area is more vulnerable to attacks [127]. This allows an attacker to identify the cycle size by observing the perturbed data stream for a long period. Therefore, some randomness should be added (ex- random cycle

sizes) instead of a single cycle size throughout the perturbation process to provide a high privacy guarantee. Privacy is calculated in terms of breach probability [143] using Known Input/Output Attacks [139] based on the Maximum A Posteriori attack (MAP) [138]. A "ϵ-privacy breach" of a record occurs if the relative error of the recovered record is less than a specified threshold ($\epsilon;\ \epsilon > 0$) [143]. This method assumes that the attacker knows some original records and their perturbed counterparts and tries to retrieve unknown original records using those known records [32, 149]. It seems promising to calculate privacy in both static databases and data streams.

The data fitting module of APOF performs the optimisation of the accuracy-privacy trade-off according to the user-defined privacy threshold (α) [149]. It takes the original dataset as the input and performs the perturbation for a possible set of noise addition rates (k). For each k value, the accuracy and privacy are calculated. Then it conducts data fitting experiments for the calculated accuracy, and privacy values using regression with the kernel functions [162, 168]. It represents accuracy as a function of privacy using curve fitting, as it is not possible to mathematically represent the relationship between privacy and accuracy due to its unpredictable behaviour [149]. Then the data fitting module uses the fitted function to predict the achievable accuracy level for α. Additionally, this module is used to fine-tune the noise addition rate suitable for achieving accuracy for α using another data fitting function. The optimisation criterion is to seek the minimum AEL such that `breach probability (perturbation)` $\leqslant (1 - \alpha)$ [149]. The data fitting process is accurate but time-consuming for a large dataset, as it is required to repeat the experiments for different k values [149]. It is impossible to fit data for the whole data stream due to its volume and inability to access it at once. Hence, a sufficiently large sample from the data stream needs to be acquired beforehand and should perform the data fitting only to that selected sample. The success of this method highly depends on how accurately the selected sample represents the whole data stream. Furthermore, the effect of concept drift on the data fitting should

also be investigated.

Though APOF has some qualities suitable for handling privacy preservation in data stream mining, it does not cover some central issues of PPDM in data streams. The privacy module of APOF cannot deal with lengthy or infinite data streams that cause accuracy decrements and can also be vulnerable to attacks. At the same time, the accuracy module cannot deal with concept drift in the data stream, which negatively affects the accuracy. On the other hand, the data fitting module should be modified considering the nature of data streams, as it is impossible to use all data for data fitting. In addition to these primary issues, we see more potential in improving APOF to optimise the accuracy-privacy trade-off to a greater degree.

9.3.2 Adapting Accuracy Module for Data Streams

The primary improvement that has to be done with the accuracy module is changing the classifier to deal with the concept drift. After carefully searching for all the possible options, it was decided to use Hoeffding Adaptive Tree (HAT) [186, 5] as the classification algorithm for the accuracy module. HAT was designed to learn from evolving data streams without a fixed-size sliding window. This is an added advantage of HAT, as the optimal window size is a complex parameter to guess [5]. Instead of using a fixed window size, HAT uses estimators to manage statistics at each node to deal with the concept drift. HAT can be used with three different estimators; INC, EWMA, ADWIN [5, 186] we decided to use HAT with ADWIN estimator that also acts as a change detector. It monitors the performance of each branch and replaces branches with new branches when their accuracy decreases. Memory consumption of HAT is remarkably lesser, that is, $O(TAVClogW)$ where T is the number of nodes, V is the number of attributes, V is the maximum number of values for an attribute, and C is the number of classes. However, time consumption is slightly higher than others in

the same category [5]. Considering the proven performance, HT was replaced by HAT to deal with evolving data streams with guaranteed performance. Experiments were conducted to compare the performance of HT and HAT in the context of PPDM for data streams with and without concept drift.

As explained in Section 9.3.1, we replaced the accuracy calculating method of APOF (AEL) with the standard error rate to use conveniently in a data stream mining environment. This method avoids the issue of the tree removing branches when the accuracy decreases. Even if some records were already being removed with its branch, accuracy measurements still consider whether those records have been correctly classified or not. This is not possible if we use AEL as the measure of accuracy. Therefore accuracy and error are defined as follows.

$$\texttt{Accuracy} = \frac{\texttt{Number of correctly classified records}}{\texttt{Total number of records}} \tag{9.1}$$

$$\texttt{Error} = 1 - \texttt{Accuracy} \tag{9.2}$$

Efficiency is significant in data stream mining, as data streams can be lengthy or infinite. Building a decision tree throughout the entire data stream can lead to problems such as overfitting and overly complex trees that reduce the classifier's efficiency. This ultimately affects the overall performance. To avoid this, we implemented a method to switch the classifier between two modes: building and learning. The tree is building while classifying the incoming records in the building mode, and accuracy is being monitored. If the accuracy becomes stable, the tree moves into the learning mode that allows it to learn from the existing tree without further building it. Deciding when to switch between modes depends on the difference in the cumulative accuracy between

two consecutive records. If the increment of accuracy between these records is equal to or less than a small threshold Δ, we can decide that the accuracy level has reached its maximum and the tree is matured. Then the tree does not have to build anymore. At this point, the classifier switches to learning mode. If the accuracy of two consecutive records violates the above criteria, the classifier switches back to the building mode and starts building the tree again. The switching criteria can be represented as follows.

Algorithm 3 Pseudo code for classifier switching criteria

1: **if** $(\Sigma_{i=0}^{n-1} Acc(record_i) - \Sigma_{i=0}^{n-2} Acc(record_i)) \leqslant \Delta$ **then**
2: $\quad MODE \leftarrow Learning$
3: **else**
4: $\quad MODE \leftarrow Building$
5: **end if**

The above modifications make the accuracy module of APOF more suitable for efficiently and accurately operating in a data streaming environment. Figure 9.2 depicts the improved version of the accuracy module of APOF.

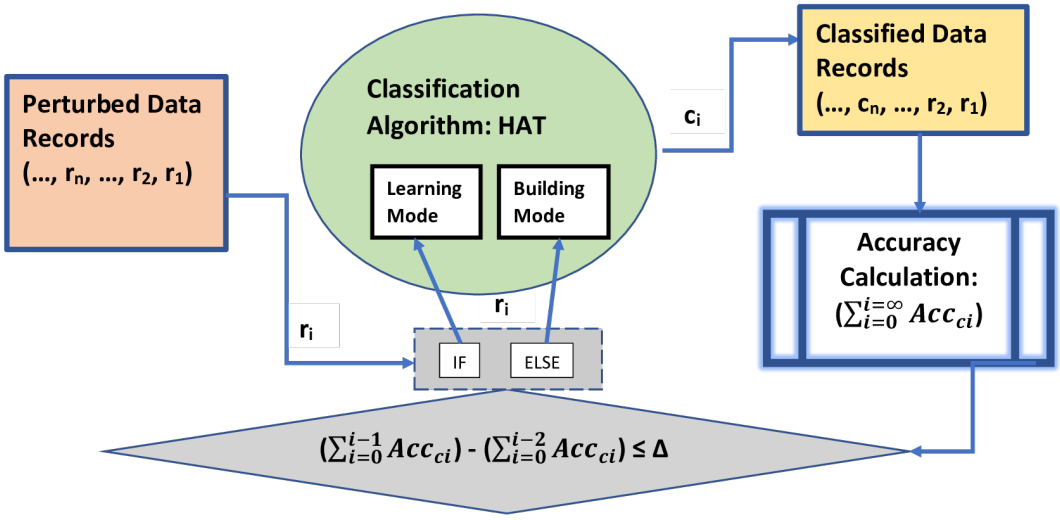

Figure 9.2: Improved accuracy module of APOF

9.3.3 Adapting Privacy Module for Data Streams

The privacy module of APOF uses logistic cumulative noise addition and random projection and translation as the perturbation method. This has been proven to achieve high privacy while maintaining a good level of accuracy [127, 149]. Nevertheless, adding noise cumulatively can be a possible negative effect on accuracy for lengthy data streams. This is because though we add noise with small variance, cumulative noise variance can be very high for a more significant number of records. The higher the noise, the more distorted the data, hence more decrement in the accuracy. As a solution to this, a noise resetting technique was implemented. After some time, the already added cumulative noise is reset to zero and starts the perturbation process from the beginning. Noise resetting helps control the noise level and prevents the classifier from learning noise instead of data in long-term running. This ultimately helps to maintain accuracy even for the infinite data streams.

Privacy is measured using performing Maximum A Posteriori attack (MAP) [138, 32, 149] to recover original records from perturbed records. This method assumes that the attacker knows some original records and their perturbed versions and uses that knowledge to recover some other unknown records. Retrieving an original record from a perturbed record requires reversing random projection, translation, and cumulative noise added to the record trying to recover. The original method in [32] assumes that the attacker knows the cumulative noise variance and uses that as prior knowledge with known I/O to recover original records. Since we use a changing noise variance $(f(x) \times \sigma)$, σ is the only constant in the noise variance. It is a factor used to scale down the noise variance, and we assume that the attacker knows the value of σ.

We implemented a method to select cycle size randomly from a given set of values to solve the possible attacks by revealing cycle size. The perturbation method can be vulnerable to background knowledge-related attacks if an attacker somehow knows the

exact cycle size. The variance of the noise added is calculated according to the logistic function. According to the logistic function, low noise is added at the beginning of the cycle. Therefore, privacy can be lower at the beginning of the cycle than in other areas. That increases the probability of recovering original data records from the perturbed records with low noise levels. Using random cycle sizes makes it difficult for attackers to guess the cycle size and attack the most vulnerable areas. The modified perturbation process is explained in Figure 9.3.

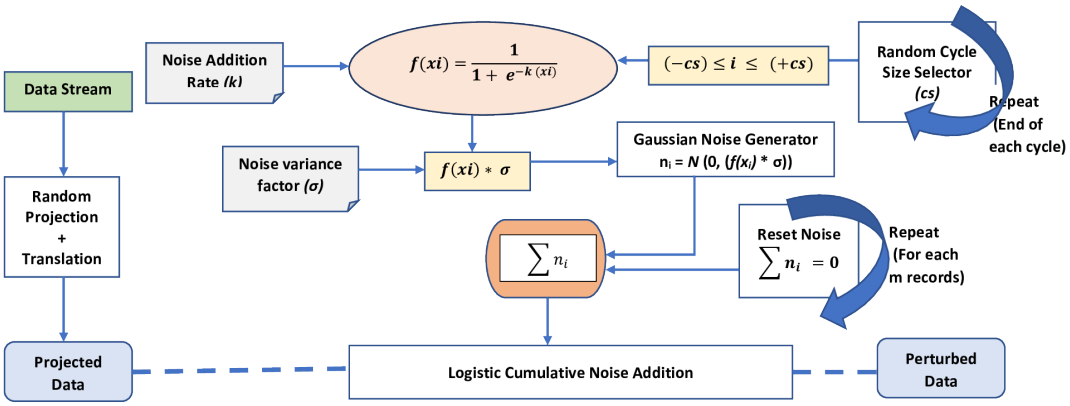

Figure 9.3: Improved privacy module of APOF

In addition to the operations of accuracy and privacy modules of APOF, a window-based accuracy-privacy monitoring method was implemented. This window-based monitoring method intends to monitor the timely accuracy and privacy values. For lengthy data streams, accuracy from the beginning to the current time may not be relevant anymore. A window of the current n number of records can be used to monitor the accuracy. On the other hand, it is practically difficult to perform attacks on the entire data stream to recover original records as data reaches high speed. Therefore, the current window of n records is used to perform attacks and calculate privacy. This method gives the accuracy and privacy values on time and reduces the memory requirements of storing many records. Finally, it increases the efficiency of APOF when used in a data streaming environment.

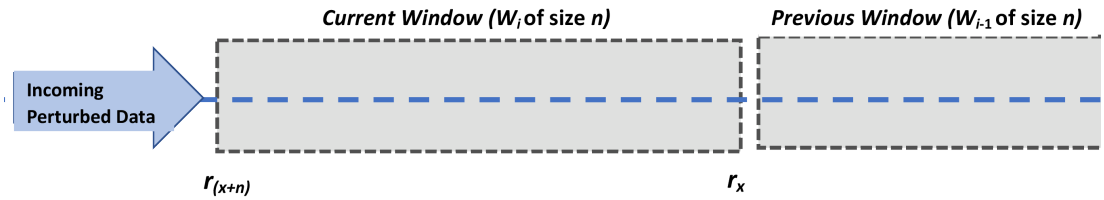

Figure 9.4: Window-based accuracy-privacy monitoring

The MAP-based known I/O attack method used in [149] was modified according to the window-based accuracy-privacy monitoring method. Earlier implementations of MAP attacks [32, 149] select both known I/O records and unknown records randomly from the dataset or data stream. However, since we monitor the accuracy in windows, we assume that known I/O pairs are spread throughout the stream, but unknown records are located within the current monitoring window. Using this terminology, we attack some randomly selected unknown records in the current window using known I/O pairs anywhere in the stream. Then we calculate the breach probability for the current window. The possible ranges of known and unknown records can be written down as follows.

Range of known I/O record (r_k) = $Index\ (r_1) \leqslant Index\ (r_k) \leqslant Index\ (r_{end})$
Range of unknown I/O record (r_u) = $Index\ (r_x) \leqslant Index\ (r_u) \leqslant Index\ (r_{x+n})$
(See Figure 9.4)

9.3.4 Adapting Data Fitting Module for Data Streams

The existing data fitting module of APOF takes accuracy-privacy value pairs for different noise addition rates returned from the whole dataset as an input and performs the data fitting [149]. However, this is not possible for the data streaming environment, as the whole data stream cannot be accessed at once. Also, if a large data set is used for data fitting, it would be time-consuming as the PPDM process should be repeatedly conducted for a set of noise addition rates. A feasible way to adapt the data fitting module to a data streaming environment is to acquire a set of data from the stream (ex:-first 10000 records) at the beginning and performs data fitting using that sample of data. This is only possible if the acquired data sample represents the whole stream. Therefore, experiments were carried out to compare the performance of the sample dataset and the whole stream. If the accuracy-privacy values returned from the sample data are approximately similar to the accuracy-privacy values from the whole stream, it can be assumed that the sample is well-representative. Furthermore, it is justifiable to use a sample set of data for the data fitting to predict the accuracy for a user-defined privacy threshold and the respective noise addition rate to achieve this accuracy and privacy levels. Therefore, the structure of the existing data fitting module of APOF will not be changed, but it should be tested with a smaller size of data samples.

9.4 Experimental Evaluation and Discussion

This section presents the results retrieved from the experiments by following the methodology explained in Section 9.3. Furthermore, we discuss the results in detail with possible reasons.

9.4.1 Datasets and Experimental Configuration

Experiments were carried out using four different datasets with and without concept drift. All the selected datasets are considered as data streams because records were ordered according to the time they were produced.

1. With concept drift
    - SEA - MOA's implementation of SEA with 3 abrupt drift points at 25000, 50000, 75000 was used. This is a synthetic dataset consists of 3 features, 2 target variables and 100000 records.
    - Radial Basis Function (RBF) - MOA's implementation of RBF with continuous fast drift of 0.001 change speed was used. This is a synthetic dataset consists of 10 features, 5 target variables and 50000 records
2. Without concept drift
    - Activity Recognition system based on Multisensor data fusion (AReM) - A real world dataset (from UCI machine learning repository) with 6 features, 5 target variables and 35999 records.
    - New York City Taxi Trip Duration (TAXI) - A real world dataset (from kaggle) with 7 features, 3 target variables and 50000 records.

TAXI dataset contains details of many individuals, and the AReM dataset contains data of a single individual retrieved from sensor monitoring. Both these datasets

Table 9.1: Experimental Configuration

Item	Values/Description
Classification Method	Hoeffding Adaptive Tree (HAT)
Accuracy measurement	Error OR (1-Accuracy)
Perturbation Method	Logistic Cumulative Noise Addition (SRW) together with Random Projection and Translation
Noise addition rate (k)	In the range of 0.005 - 0.1 with intervals of 0.004 (24 values)
Maximum value of logistic curve (L)	1
Cycle Sizes	Randomly selected from [1000, 2000, 4000, 5000, 6000]
Return value from logistic function	*f(x)*
Privacy measurement	Breach Probability (BP)
Variance of cumulative noise	$f(x) \times 3.90 \times 10 - 6$
MAP attack	Number of known I/O pairs – 2/4 per attack. Number of attacks – 5% of the window size Epsilon (ϵ) – 0.2
Window sizes for accuracy-privacy monitoring	10000, 15000, 20000
Data Fitting Function	Kernel Regression using Wave kernel
Smoothness/Bandwidth of the kernel	0.01
Kernel Regression implementation using	Fastmath library (generateme/fastmath "1.4.0-SNAPSHOT")
Curve fitting using	Cljplot library (cljplot "0.0.2-SNAPSHOT")

may contain sensitive information and are vulnerable to privacy attacks. The original feature set of the TAXI dataset was reduced to 7, considering only the numeric features, and a new target variable was created by applying equal-binning frequency. Other pre-processing steps, such as handling missing data, were also carried out appropriately.

Table 9.1 summarizes all the experimental configurations related to the accuracy module, privacy module, and data fitting.

9.4.2 Hoeffding Tree (HT) Vs. Hoeffding Adaptive Tree (HAT) in PPDSM

The performance of HT and HAT in data stream mining has been compared in research work such as [156] and [186]. However, the comparison of these two classifiers in the PPDM setting cannot be found. Therefore, it is important to compare these classifiers in the PPDSM environment to see whether they behave as expected, as there are two types of changes in data; changes made by concept drift and changes made by perturbation. Hence, we compared the performance of HT and HAT in the PPDM setting to analyze their behaviour. This experiment aims to see how both classifiers adapt to the data streams with and without concept drift along with the perturbation. Figure 9.5 & Figure 9.6 depicts the accuracy of HT and HAT for different noise addition rates (k) for AReM (no concept drift) and RBF (concept drift) data streams, respectively. Additionally, we compared HT and HAT with the current perturbation method of APOF and with the method after we made the modifications by adding random cycle sizes and noise resetting.

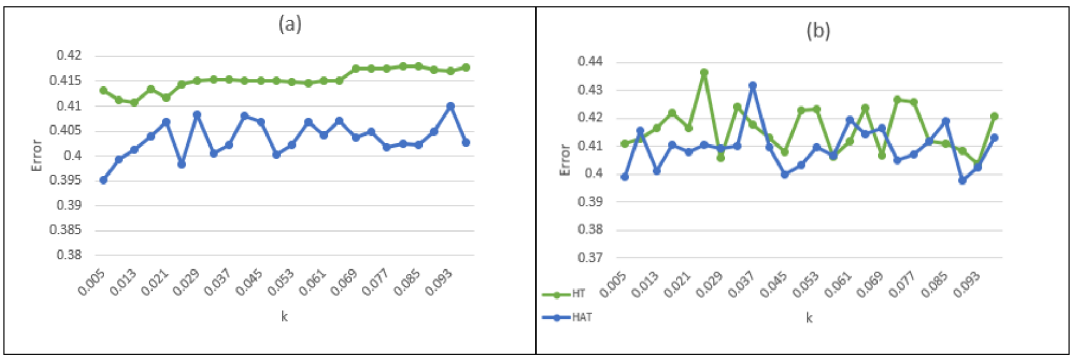

Figure 9.5: Accuracy behaviour of HT and HAT for different noise addition rates (k) - AReM Data Stream; (a)- Using existing perturbation method of APOF, (b)- After modifying the perturbation method to work with data streams

From Figure 9.5 & Figure 9.6, it can be seen that the error of HAT is less than the error of HT in all scenarios, as expected. The pattern is similar for the current

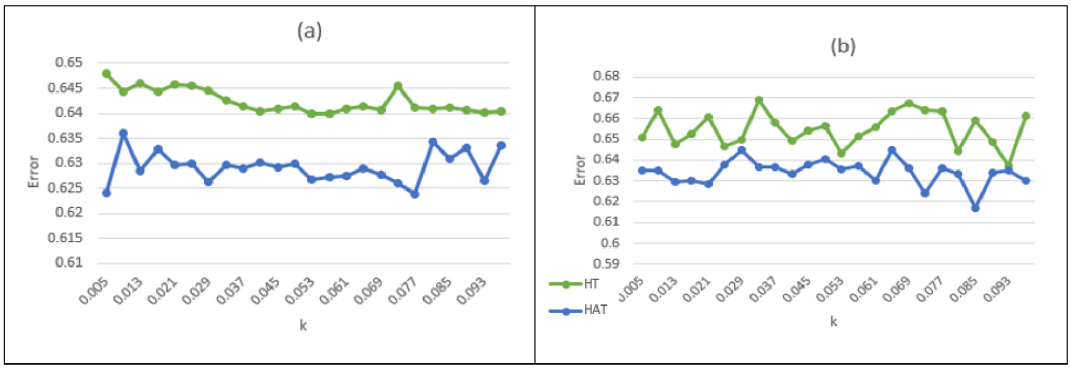

Figure 9.6: Accuracy behaviour of HT and HAT for different noise addition rates (k) - RBF Data Stream; (a)- Using existing perturbation method of APOF, (b)- After modifying the perturbation method to work with data streams

[149] and modified perturbation methods for both data streams, except the graph of the modified perturbation method for AReM data stream (Figure 9.5(b)). Though we can see some high fluctuation points and overlaps, HAT still shows a better average performance than HAT. Moreover, graphs of Figure 9.6 show that HAT works better with data streams with concept drift regardless of the perturbation method used to distort the data. Therefore, experiments hereafter use HAT as the classifier for the accuracy module.

9.4.3 Incorporating Random Cycle Sizes and Noise Resetting with Logistic Cumulative Noise Addition

Observing the perturbed data stream for a long time can provide knowledge to attackers. Detecting particular patterns or disclosing critical parameter values such as cycle size and noise addition rate may provide opportunities for an attacker to breach privacy and access sensitive original data [127, 149]. Logistic cumulative noise addition was modified considering the possible attacks to breach privacy and negative effects on the accuracy in lengthy or infinite data streams. Instead of using a single cycle size, a random cycle size selection method was introduced. A noise resetting method was

embedded to avoid possible higher data distortions in lengthy or infinite data streams. Figure 9.7 & Figure 9.8 presents the accuracy and privacy results in terms of error and breach probability for AReM and RBF data streams, respectively.

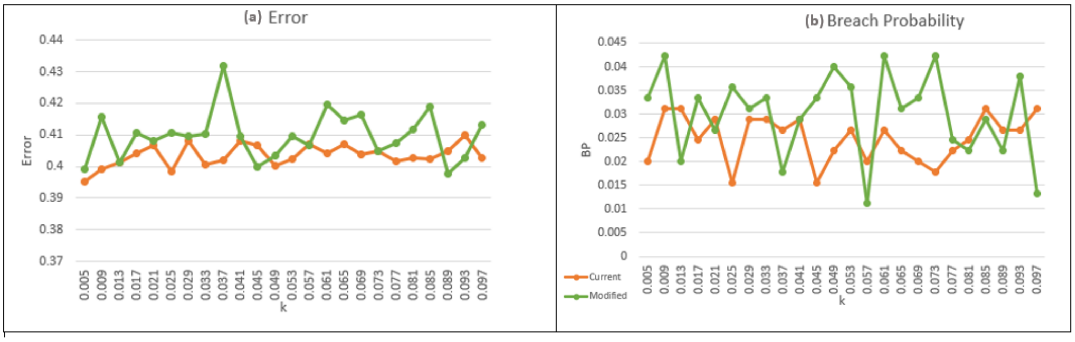

Figure 9.7: Accuracy and Privacy behaviour after perturbation for different noise addition rates (k) - AReM data stream; (a)- behaviour of Error for current and modified perturbation method of APOF, (b)- behaviour of breach probability for current and modified perturbation method of APOF.

Figure 9.8: Accuracy and Privacy behaviour after perturbation for different noise addition rates (k) - RBF data stream; (a)- behaviour of Error for current and modified perturbation method of APOF, (b)- behaviour of breach probability for current and modified perturbation method of APOF

Both error and breach probability of the modified perturbation method (with random cycle sizes and noise resetting) is slightly higher than the perturbation method proposed in [149] for both datasets. This result was expected for a data stream with limited records. Resetting the cumulative noise added to zero for every 25000 records restarts the perturbation process from the beginning. That means the distortion of data is

reduced and hence the privacy too. So the decrement in the breach probability can be explained. However, less distortion caused by noise resetting should increase the accuracy. However, adding noise in random cycle sizes degrade this effect. Selecting a cycle size randomly at the end of each cycle changes the pattern of noise variance generation and adds more randomness to the process. This randomness causes a slight decrement in accuracy but makes the data more protective.

According to the results, we can say that two techniques introduced to the perturbation (random cycle sizes and noise resetting) have opposite effects on accuracy and privacy (higher error and breach probability) in different amounts. This ultimately results in low accuracy and privacy than the perturbation method proposed in [149]. However, the purpose of introducing these two techniques should be considered. In a practical situation, if the data stream is infinite and long-term, attackers may attempt to reveal parameters such as cycle size, noise addition rate, and final noise variance by observing the behaviour of the perturbed data stream. If attackers successfully recover those parameter values, there is a possibility of recovering original data from perturbed data. On the other hand, cumulative noise addition cannot be done forever, as data can be highly distorted with a higher amount of noise. Therefore, a noise resetting technique must avoid massive accuracy drops in the long-term run. Though we must sacrifice a small amount of accuracy and privacy, these changes are necessary to adapt APOF to a data streaming environment. Otherwise, it will be challenging to handle the issues such as dealing the high volume inherent to the data stream.

9.4.4 Window-based Accuracy-Privacy-Monitoring

The window-based accuracy-privacy monitoring method discussed in Section 9.3.3 was implemented to monitor the timely accuracy and privacy values. We carried out experiments for three different Window Sizes (WS), 10000, 15000, and 20000, to see

Table 9.2: Accuracy-Privacy Monitoring Using a Window-based Method

Data Stream Name	WS = 10000		WS = 15000		WS = 20000		Entire Stream	
	Error	BP	Error	BP	Error	BP	Error	BP
AReM	0.4037	0.0260	0.4038	0.0232	0.4119	0.0300	0.4131	0.0134
TAXI	0.6075	0.0400	0.6006	0.0334	0.6047	0.0267	0.6132	0.0272
RBF	0.6345	0.0784	0.6154	0.0882	0.6430	0.0600	0.6301	0.0864
SEA	0.2945	0.0460	0.3002	0.0400	0.2931	0.0468	0.2970	0.0376

the effect of window size on accuracy and privacy. Table 9.2 summarizes the average Error and Breach Probability (BP) per window for all the experimented window sizes. Additionally, it compares the window-based method's average error and BP values with the error and BP values calculated by monitoring the entire data stream.

The results of Table 9.2 show that the average error per window is almost similar to the error of the entire data stream for all the window sizes. As this is a monitoring method, we only need to see whether the window-based method represents the actual accuracy of the data stream or not. If it does not, we can say that the window-based method misinterprets the accuracy of the entire data stream and is not appropriate as a monitoring method. The difference of the error from window-based method and entire data stream for window sizes 10000, 15000 and 20000 are [0.0094, 0.0093, 0.0012], [0.0057, 0.0126, 0.0085], [0.0044, 0.0147, 0.0129] and [0.0025, 0.0032, 0.0039] respectively for AReM, TAXI, RBF and SEA data streams. According to those results, we can ensure that the accuracy of the window-based method does not deviate much and provides a good representation of the accuracy of the entire data stream. Results show that the pattern is the same for data streams with or without concept drift. The reason is that HAT adapts to the concept drift and can maintain a stable accuracy level throughout the data stream.

BP values of the window-based method also display a similar trend to error values except for a few exceptions. According to the results from Table 9.2 difference of BP values for window sizes 10000, 15000 and 20000 and AReM, TAXI, RBF and SEA are [0.0126, 0.0098, 0.0166], [0.0128, 0.0062, 0.0005], [0.0080, 0.0018, 0.0264]

and [0.0084, 0.0024, 0.0092] respectively. For most of the window sizes for all the experimented datasets, BP values are quite similar to the breach probability of the entire stream. A few higher deviations exist, such as window size 20000 of RBF data stream. A possible reason can be that we have modified the attacking method for the window-based monitoring method. In APOF proposed in [149] attacks were performed to the entire stream, which is a large number of records. However, in the window-based method, we performed attacks on the current window, and the number of records is much lesser than the stream. Hence, there is a high possibility of finding records closer to each other, making it easier to breach privacy. As most instances have shown similar BP values, it is safer to assume that the privacy of the window-based method is similar to the privacy of the entire data stream.

The results cannot show a relationship between the window size with accuracy and privacy. For some data streams, accuracy and privacy increase with the window size, but it behaves the other way around for some. So, it may primarily depend on the dataset. From all these results, we can say that using the current window to monitor the accuracy and privacy instead of calculating accuracy and privacy over the entire stream is an efficient method. It provides timely measures by only considering the current window and reduces the memory requirements by discarding past accuracy-privacy measures.

The results from Table 9.2 prove that a window, which can be considered as a sample from the data stream, provides an accurate representation of the accuracy-privacy of the entire data stream with a marginal error. Therefore, it is appropriate to use a sufficiently large sample to fit data to decide the expected accuracy level and noise addition rate for a user-defined privacy threshold. The expected noise addition rate retrieved from data fitting $k = 0.97$ to achieve a user-given privacy threshold (for the experiments, we used it as 0.9) was used for the entire experiments. We do not present or discuss the results from data fitting, as it is repetitive to the experiments carried out in [149]. The only difference is that we used a sample from the data stream instead of using the entire

dataset for data fitting to use APOF for data streams efficiently. This reduces the time consumption and computational complexity of the data fitting module.

9.4.5 Classifier Switching Scenario

Switching the classifier to learning mode when accuracy becomes stable is a way to avoid the overfitting of the classifier caused by continuous building. This is common with most data mining algorithms, which can misinterpret actual accuracy. Figure 9.9 & Figure 9.10 show the results of accuracy with and without classifier switching method for AReM and SEA data streams, respectively. Accuracy was monitored using a window of size 10000.

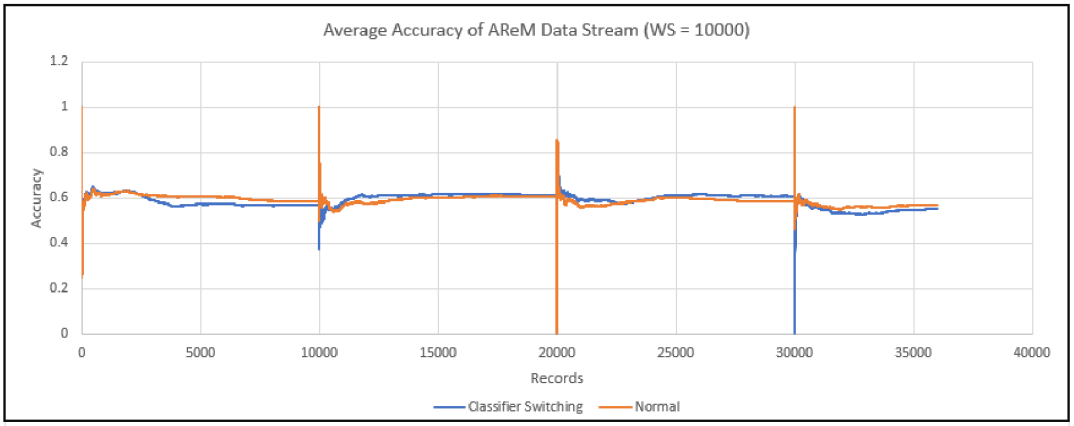

Figure 9.9: Accuracy plot of AReM data stream with and without classifier switching method.

For both data streams, the accuracy of the classifier switching method is quite similar to the accuracy of the classifier without switching. It is a little higher after the first one or two windows. Moreover, the trend of both accuracy curves is also similar. Especially for the SEA data stream, if we compare the accuracy curves of the drifting points (See black vertical lines in Figure 9.10), the classifier with the switching method has adapted more accurately to the concept drift. This means the novel classifier switching method is either similar or more accurate than a classifier without switching. It identifies

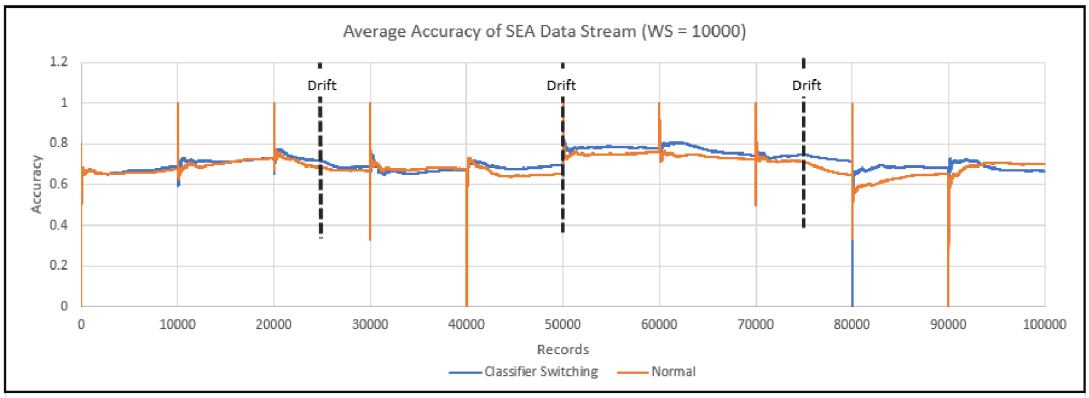

Figure 9.10: Accuracy plot of SEA data stream with and without classifier switching method.

the accuracy decrement points more accurately and quickly adapts to the change by switching mode. Therefore, the classifier switching method efficiently provides more accurate results and reduces the risk of overfitting.

9.4.6 Accuracy and Privacy Behaviour through the Windows

We compared the accuracy-privacy plots of the data streams after implementing all the modifications discussed so far. The trend of accuracy and privacy throughout the windows can be observed from these plots. The primary y-axis represents privacy in these plots, and the secondary y-axis represents accuracy with window number in the x-axis.

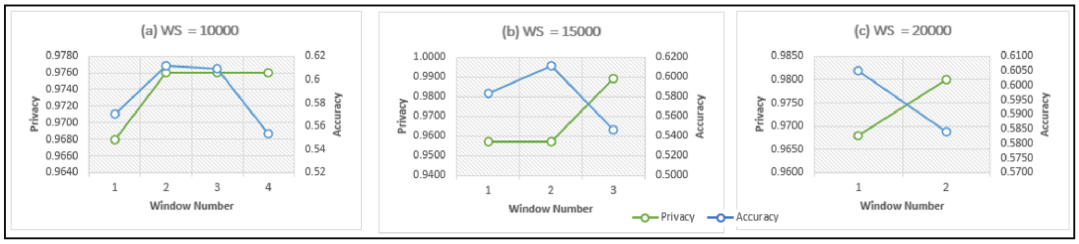

Figure 9.11: Accuracy and privacy behaviour of AReM data stream using window-based monitoring method.

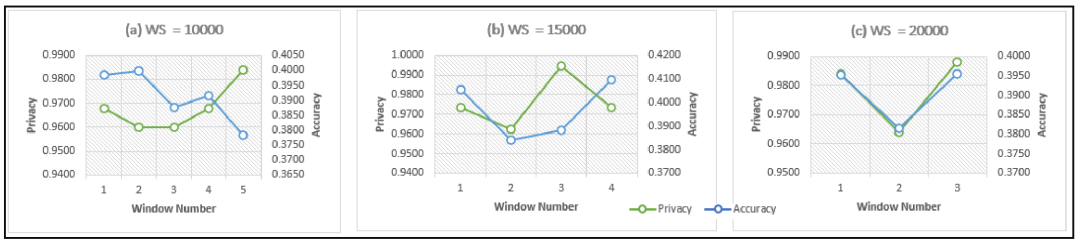

Figure 9.12: Accuracy and privacy behaviour of TAXI data stream using window-based monitoring method.

Figure 9.11 & Figure 9.12 display the accuracy-privacy plots of AReM and TAXI data streams, that do not contain any known concept drift. From the accuracy-privacy results of the AReM data stream (see Figure 9.11), it can be observed that privacy has an increasing trend. At the same time, accuracy shows a decreasing trend for all the monitoring window sizes. From first window to last window accuracy decrement values are 0.0172, 0.0338, 0.0212 and privacy increment values are 0.008, 0.0321, 0.012 for window sizes 10000, 15000 and 20000 respectively. Though there is a slight trend of decrement for accuracy and a slight trend for increment for privacy, differences are considerably low, less than 0.04.

If we analyze the accuracy-privacy trends for the TAXI data stream (see Figure 9.12), the trends are quite different from the AReM data stream. For window size 10000 (Figure 9.12(a)), we can see a clear trend of increment (increased by 0.016) for privacy and decrement (decreased by 0.0202) for accuracy. For window size 1500 (Figure 9.12(b)) and 20000 (Figure 9.12(c)), both accuracy and privacy have slightly increased if we consider the first and last window, but with fluctuations throughout the windows. Increments values are 0.0043, 0.0004 and 0.0000, 0.004 for accuracy and privacy respectively. Though slight increments and decrements with fluctuations throughout the data stream, accuracy and privacy have maintained a steady run. Plots provide solid evidence for this, as accuracy-privacy differences from the first to last are always lower than 0.04.

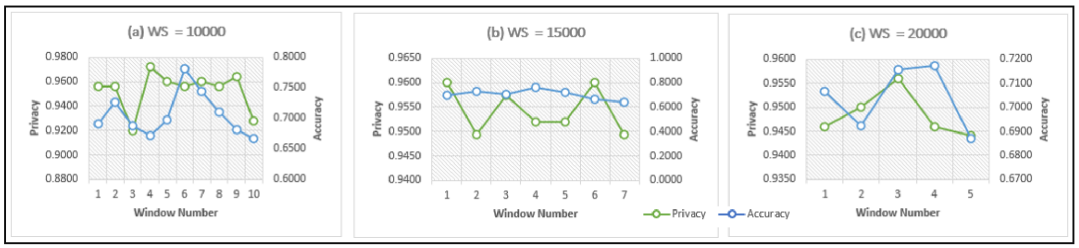

Figure 9.13: Accuracy and privacy behaviour of SEA data stream using window-based monitoring method.

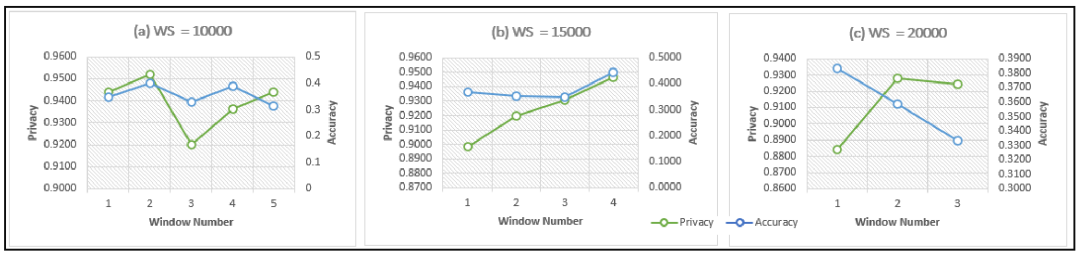

Figure 9.14: Accuracy and privacy behaviour of RBF data stream using window-based monitoring method.

Figure 9.13 & Figure 9.14 display the accuracy-privacy plots of SEA and RBF data streams. These two data streams have known concept drifts that are abrupt and continuous, respectively. According to Figure 9.13, for the SEA stream, both accuracy and privacy have slightly dropped if we see the first and last window, but increasing fluctuations can be observed in between. Accuracy has dropped by 0.0232, 0.0633, 0.0195 while privacy has lessened by 0.028, 0.01, 0.002 for window sizes 10000, 15000 and 20000. If observe the accuracy curve of, window 3, 5 and 8 of Figure 9.13(a), window 2, 4 and 6 of Figure 9.13(b) & window 2, 3 and 4 of Figure 9.13(c) behaviour of accuracy at abrupt drift points can be seen. Except few cases such as window 3 of Figure 9.13(a), window 6 of Figure 9.13(b) and window 2 of Figure 9.13(c) accuracy does not seem affected from the concept drift. Though accuracy has decreased in the windows mentioned above, those are not the minimum accuracy values recorded on the curve. Therefore, it cannot be assumed that it has happened due to the concept drift.

Table 9.3: Execution Time with Enhanced-APOF

Data Stream	Time per Record (milliseconds)
AReM	11.1777
TAXI	12.2437
RBF	11.2348
SEA	15.7352

From Figure 9.14, different behaviours of accuracy and privacy for different window sizes can be observed. Regardless of the increasing or the decreasing trend, the difference in accuracy-privacy values of the first and last windows are minor. Accuracy differences are 0.0335, 0.0765, 0.05 while privacy difference records 0.0000, 0.0481, 0.04 for window sizes 10000, 15000 and 20000. The overall accuracy of this data stream after the perturbation is low (around 0.36), but the accuracy of the original dataset before perturbation is also low (around 0.38). Therefore, we can say that the low accuracy is not caused by perturbation or continuous concept drift.

The above-discussed privacy and accuracy curves build a solid ground to state that accuracy and privacy stay stable with tiny fluctuations (always lower than 0.08) throughout the data stream. Moreover, it can be observed that the nature of accuracy-privacy behaviour, which is the opposite trend of each other, is not always true. Some curves such as Figure9.12(a) & (b), Figure 9.13(a) and Figure 9.14(b) are evidence to this as we can see increasing or decreasing trends for both accuracy and privacy.

9.4.7 Execution Time

We measured the execution time (see Table 9.3) for all four datasets to see how fast the framework can process a record. Fast execution is compulsory for data stream mining environments as data records can reach high speed. The time measured includes the time duration for the classification process with the accuracy calculation and perturbation process with privacy calculation together with all the modifications we discussed so far.

Results in Table 9.3 tell us that Enhanced-APOF can process a record in less than

a second. SEA data stream records the highest execution time per record, which is 15.7352 milliseconds. Since the measured time includes the entire PPDM process processing a record within milliseconds is quite reasonable for data streams.

9.4.8 Privacy Against Accuracy - Achieving Accuracy-Privacy Optimisation

As the optimisation criteria, we try to find the minimum error which satisfy `breach probability (perturbation)` $\leqslant (1-\alpha)$ [149] using data fitting module where α is the user-defined privacy threshold. This is the same criterion proposed for APOF, and the process and results have been explained in detail in [149]. Though we have not explained data fitting results in this article as it is repetitive to the experiments in [149], we provide experimental evidence to show how optimisation has been achieved.

Plotting privacy against accuracy or vice versa is a difficult task. Calculating the accuracy for each record in the data stream is straightforward, as we classify all the incoming records using HAT. However, there is no way to calculate privacy for each record. We calculate privacy by performing attacks on random records, and it is impractical to attack each record. Our window-based accuracy-privacy monitoring method provides a reliable solution for this. We calculate accuracy and privacy for the current window, plotting privacy against accuracy for each monitoring window. Plotting privacy against accuracy or vice versa reveals crucial things about the accuracy-privacy behaviour of the entire data stream. The primary advantage is that it allows us to investigate the relationship between privacy and accuracy.

Figure 9.15 and Figure 9.16 depict the relationship between privacy and accuracy for data streams without and with concept drift, respectively. The size of the monitoring window is 10000. The classifier HAT uses a default window of the size of 10000 records. We decided to use the same size of a window to parallel the accuracy monitoring process

to the classifier. Note that we have used normalised values of accuracy and privacy for plotting to observe their relationship. Furthermore, we have represented each graph's linear and non-linear trend lines to identify whether they have an increasing or a decreasing trend.

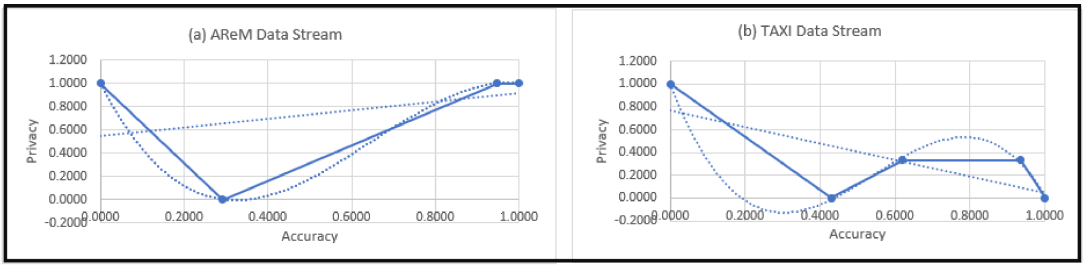

Figure 9.15: Privacy against accuracy of data streams that do not contain concept drift - (a) AReM Data Stream, (b) TAXI Data Stream

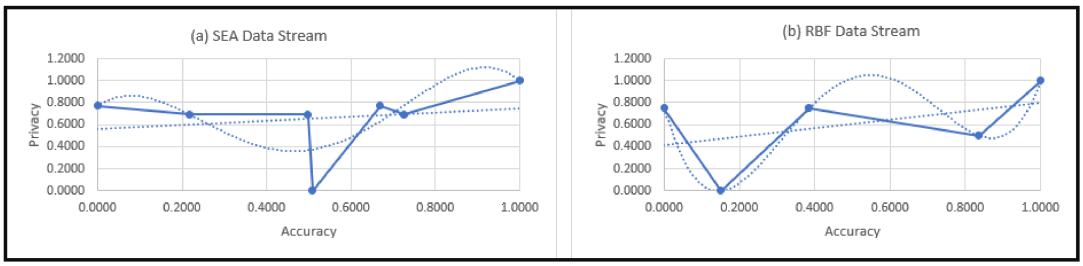

Figure 9.16: Privacy against Accuracy of data streams with concept drift - (a) SEA Data Stream, (b) RBF Data Stream

From Figure 9.15 and Figure 9.16 we can see that accuracy and privacy has a positive correlation except TAXI (Figure 9.15(b)) data stream. The correlation values for AReM, TAXI, SEA and RBF are 0.3619, -0.7180, 0.2055, 0.4337 respectively. This proves that in most cases, when accuracy increases, privacy also increases. Our framework has successfully overcome the well-known trade-off between data privacy and data mining accuracy. This proves that "When privacy increases, accuracy decreases, and vice versa" is not always valid. It depends on the dataset, and for most of the datasets, we can increase both privacy and accuracy using the proposed framework. This can happen regardless of the availability of the concept drift, given that the classifier can adapt to

the concept drift. Therefore, in addition to the optimisation we perform using the data fitting module according to the user's privacy requirement, the proposed framework has optimised the accuracy-privacy trade-off in general.

Additionally, when we observe the trend lines of each graph, we can say that non-linear or polynomial functions represent the relationship between accuracy and privacy more accurately. The linear trend lines are under-fitted and cannot precisely grasp the relationship between accuracy and privacy. Though the polynomial functions are well-suited to represent this relationship, the order of the function cannot be pre-defined. It depends on the dataset and is hard to guess without experiments.

9.5 Conclusion and Future Directions

The ultimate objective of this work was to implement an enhanced and efficient privacy-preserving framework for data stream mining. This framework should address the challenges arising from data streams due to their unique behaviour. We used the Accuracy-Privacy Optimisation Framework proposed in [149] as our base model and redesigned it to work in a data streaming environment efficiently. Experiments were carried out with four different datasets representing data streams with (RBF & SEA) and without (AReM & TAXI) concept drift. Experimental results provide solid grounds to prove that the Enhanced Accuracy-Privacy Optimisation Framework (Enhanced-APOF) is a sophisticated framework to be used in privacy-preserving data stream mining. The challenges that arise from the unique behaviour and characteristics of data streams were considered when improving APOF [149]. This work has achieved the initial objectives mentioned in Section 9.1 as follows.

As the initial step, the APOF [149] was analysed to find out the sections that need improvements when used in data stream mining (See Section 9.3.1). HT classifier of accuracy module in APOF is incapable of dealing with concept drift, as it operates under

the assumption that the underlying data distribution does not change over time. This is not true when working with real-time data streams. Therefore, we replaced the existing classifier with HAT to successfully adapt to the concept drift. The performance of HT and HAT was compared in a privacy-preserving data streaming environment, and the results show that HAT can achieve higher accuracy than HT (See Figure 9.5 & Figure 9.6). A novel classifier switching method was proposed to achieve high efficiency and avoid possible overfitting situations when working with lengthy or infinite data streams. Switching the classifier to learning mode, without building the tree anymore, when accuracy becomes stable avoids producing complex trees and increases the efficiency of the classification process. Results show that the switching method can obtain the same or higher accuracy level when compared with the classifier without switching (See Figure 9.9 & Figure 9.10).

The privacy module of APOF [149] is consisted of a perturbation method to preserve privacy. This perturbation method combines random projection, translation, and logistic cumulative noise addition and has been proven to achieve high privacy. However, our analysis of APOF revealed that adding noise cumulatively for an extended period distorts data to a high degree. This directly decreases the accuracy in long-term running, especially for lengthy or infinite data streams. Further, according to the logistic function, cycle-wise noise addition may cause privacy breaches if the cycle size is disclosed to an unauthorised person. We proposed a noise resetting technique and a random cycle selection method to avoid these issues. Results (See Figure 9.7 & Figure 9.8) show that the current accuracy and privacy drop by a small amount after modifying the privacy module. This is because data streams used for experiments have limited data records (maximum 100,000) and cannot see the real-time effects. However, considering possible high risks for privacy and accuracy in the long-term run, we believe these modifications are essential.

Moreover, a window-based accuracy-privacy monitoring method was introduced to

monitor timely accuracy-privacy measures. This avoids calculating cumulative accuracy and privacy values from the beginning of the data stream. Experimental results (See Table 9.2) show that window-based accuracy-privacy values are similar to the accuracy-privacy values of the whole stream. This proves that the window-based accuracy-privacy monitoring method is trustworthy and does not misinterpret the actual accuracy-privacy of the data stream. The window-based accuracy-privacy method opens a new avenue to adapt the data fitting module of APOF [149] to the data streaming environment. A sufficiently large sample from the data stream can be used in the data fitting module to predict the expected accuracy for a user-defined the privacy threshold and the noise addition rate, as the sample precisely represents the entire data stream.

Importantly, our experiments show that the proposed framework has successfully overcome the well-known accuracy-privacy trade-off. Though it is impossible to achieve 100% accuracy and privacy simultaneously, we can increase both accuracy and privacy using the Enhanced-APOF. This confirms that even with the challenging data streaming environment, the proposed framework works efficiently and optimises the accuracy-privacy trade-off in general. It has considered possible challenges in data streams and provides reliable solutions to those challenges. A record can be processed within milliseconds using Enhanced-APOF, which satisfies the need for fast execution for data stream mining. Therefore, Enhanced-APOF provides a sophisticated framework to optimise the accuracy-privacy trade-off in the data streaming environment.

As for future directions, the Enhanced-APOF should be tested with real-time lengthy data streams to evaluate its performance. Additionally, the effect of the modified privacy module on the accuracy can be tested with different data mining techniques appropriate for data stream mining. This will provide more insights into selecting the best data mining technique by comparing all the possibilities. If we can mathematically represent accuracy as a function of privacy and vice versa, that would significantly contribute to the field. If possible, the data fitting module can be replaced with a mathematical

representation and can predict accuracy for a user-defined a privacy threshold without any prior experiments. This reduces the computational and time complexity of the data fitting module. This is a challenging task as many variables are involved, and the Enhanced-APOF should be tested with numerous datasets to decide.

Chapter 10

Discussion

Privacy-Preserving Data Stream Mining (PPDSM) is more challenging than PPDM due to specific characteristics and behaviours of data streams [44, 48]. Volume, high speed, and concept drift are a few characteristics that make data streams evolve and dynamically adapt. Therefore, PPDM methods cannot be directly used for PPDSM. Privacy preservation techniques and data mining techniques should be changed or improved to handle the challenges of data streams [47, 32]. The trade-off between data mining accuracy and data privacy is one of the most concerning issues in both PPDM, and PPDSM [187].

10.1 Major Findings and Contributions of this Work

10.1.1 Logistic Cumulative Noise Addition - An Advanced Noise Perturbation Method

A novel privacy preservation technique named Logistic Cumulative Noise Addition (SRW) was proposed in Chapter 5 as the primary output of RQ1. It is based on cumulative noise addition (LRW) [32], which was proven to be an advanced noise

addition method. However, adding noise cumulatively cannot be continued for a long time. It gradually increases the total noise added to the data stream, and accuracy can suffer significantly. Therefore cumulative noise addition should be accompanied by a technique to control the noise level to maintain a better accuracy-privacy trade-off [32, 127].

Seven variations of cumulative noise addition methods (LAR, LA, LRWR, SAR, SA, SRWR, SRW) were implemented and compared the performance with the base method (LRW) proposed in [32]. It was clear from the experiments that the cycle-wise noise addition methods perform well than the linear noise addition methods. The natural behaviour of the logistic function [128] helps control the noise addition rate and the maximum noise added to the dataset or data stream, which ultimately helps maintain a better accuracy-privacy trade-off. The SRW method shows the minimum overall error score from cycle-wise noise addition variations followed by SRWR, SA, and SAR. The overall error scores prove that our proposed perturbation method (SRW) performs better than the base method (LRW). Hence, SRW was identified as the most appropriate perturbation method.

Experiments prove that resetting the already added cumulative noise to zero positively impacts accuracy but negatively impacts privacy. The cycle-wise noise additions with noise resetting (SRWR) show low relative error but high breach probability compared to the cycle-wise noise additions without noise resetting SRW. However, we cannot entirely ignore the idea of noise resetting, as it effectively achieves higher accuracy and can help control the noise level in lengthy or infinite data streams. Moreover, using absolute noise values instead of original noise values does not positively affect accuracy or privacy. For linear and cycle-wise noise additions, variations that use absolute noise values (LAR, LA, SAR, SA) showed the worst performance compared to other variations. We can conclude with the experimental results that either absolute values or the original noise values, adding noise would only distort the data values.

The logistic curve's starting and ending flat areas affect accuracy and privacy significantly when adding noise in cycles using the logistic function. In the starting flat area, noise is closer to zero, allowing the classifier to learn without facing too much noise. Nevertheless, less noise variance in the starting flat area provides less privacy and gives a high breach probability. In ending flat areas, noise variance is high, providing higher privacy than in the starting flat area. From the results in Table 5.6, we can see that the breach probability of the starting flat area is higher than the same of the ending flat area. Moreover, the breach probability of flat areas is higher than when attacking the data stream randomly. Consequently, intentionally attacking flat areas can significantly breach privacy.

10.1.2 Accuracy-Privacy Optimisation Framework (APOF)

Optimising the accuracy-privacy trade-off by only using data mining and privacy preservation techniques is not easy. Therefore, a component to perform the optimisation using a different technique or techniques should be embedded with the PPDM or PPDSM method. Moreover, providing accuracy and privacy levels requested by the user is essential [149]. Some users might have priority to privacy over accuracy and vice versa. Hence, optimisation of the accuracy-privacy trade-off should consider the requirements of a specific user [149]. Based on these details, we investigated what framework should be constructed to model the trade-off between privacy and accuracy for a specific user for data mining (RQ2).

The Accuracy-Privacy Optimisation Framework (APOF) (Figure 7.1) proposed in Chapter 7 considers the user's accuracy and privacy requirements when optimising the trade-off. APOF consists of three modules, accuracy, privacy, and data fitting, formulating a well-structured framework for accuracy-privacy trade-off optimisation. The novel data fitting module handles accuracy-privacy optimisation according to the

user's requirements. Experiments show that data fitting is a good technique for accuracy-privacy optimisation, but with the assumption that the dataset or data stream has no concept drift. We used 1000, 2000, 4000, and 8000 as cycle sizes and experimented with 24 noise addition rates k in the range of (0.005 - 0.1). However, AEL and BP values do not show a clear trend in deciding the best cycle size (Figure 7.6 and Figure 7.7). It could be said that if the cycle size is too small, it can negatively affect the accuracy as the total noise variance added in that cycle is low.

The Wave kernel was selected for data fitting after strategically eliminating Rational-quadratic, Gaussian, Matern-52 and Laplacian kernels. Curve fitting results showed that the Rational-quadratic kernel under-fits data while the Laplacian kernel over-fits, therefore not good choices (Figure 7.8 and Figure 7.9). Gaussian and Matern-52 failed to predict AEL for low privacy thresholds (Figure 7.10). After predicting the AEL value for α using data fitting, the user can proceed with that if it maps with overall requirements. If not, the user can fine-tune their accuracy or privacy requirements to see what difference it makes to the overall picture.

Validation experiments were conducted to evaluate the accuracy of the data fitting module. According to the validation experiments, the average errors of actual AEL and predicted AEL are 0.176%, 0.036%, and 0.109% for AReM, Electricity, and Taxi datasets. The average errors were low but differed from dataset to dataset. Results in Table 7.5 provide evidence that the accuracy-privacy optimisation done using data fitting is a successful approach. Hence, providing solutions to RQ2, APOF provides a well-structured framework for accuracy-privacy optimisation in PPDM.

10.1.3 Enhanced APOF for Data Stream Mining

As an extension to the previous section, APOF was found to be successful when used in data streaming environments. We identified the sections that need improvements on each module and developed "Enhanced Accuracy-Privacy Optimisation Framework (Enhanced-APOF)" for PPDSM in Chapter 9.

Hoeffding Adaptive Tree (HAT) [145] replaced the accuracy module's existing classifier (HT) to deal with concept drift. The comparison of errors of HT and HAT depicts that HAT is indeed a better classifier than HT to deal with concept drift (See Figure 9.6). The novel classifier switching method, cooperated with the accuracy module, allows learning from the already matured tree without building the tree continuously. This avoids the overfitting issue in the long-term run. Experimentation results in Figure 9.9 prove that the classifier switching method can be carried out without negatively affecting the accuracy. Moreover, the classifier switching method adapts well even if there is a concept drift in the data stream (Figure 9.10).

Revealing cycle size makes the perturbation method vulnerable, as attackers can attack the records with low noise variance. Therefore, selecting a cycle size randomly from a set of given sizes assures that the cycle size does not disclose easily. If noise overpowers actual data values, the classifier learns noise instead of data, decreasing accuracy. Therefore, noise resetting is needed to control the total cumulative noise in the long-term run. Anyhow, privacy is expected to decrease slightly due to this process. Accuracy and privacy graphs in Figure 9.7 and Figure 9.8 provide evidence of this assertion. However, this much sacrifice is worth considering the benefits noise resetting and random cycle size selection provide in the long run.

Monitoring accuracy and privacy using a window are effective. It helps to get up-to-date accuracy and privacy. Experiment results in Table 9.2 show that the window-based monitoring method's error and breach probability values are almost similar to the entire

data stream. That assures the window-based method does not mislead and represents the exact accuracy-privacy behaviour of the data stream. Moreover, we cannot use the entire data stream for the data fitting, as it is expensive in terms of time and computational power. Above all, we cannot access the entire data stream at once, as we do with static datasets. Therefore, we can use a sample or window of data from the data stream for data fitting. It perfectly fits data for accuracy-privacy optimisation since the accuracy and privacy values of the sample or window correctly represent the entire data stream.

Figure 9.15 and Figure 9.16 prove that the well-known accuracy-privacy trade-off is not always valid. That means it is possible to increase accuracy and privacy simultaneously using Enhanced-APOF. Though this highly depends on the data stream, achieving this objective is possible. It requires more experiments on various real-time data streams to mathematically represent this relationship. However, we still cannot precisely say how much of the accuracy increment will cause a certain amount of privacy increment or decrement and vice versa.

10.2 Linking Primary Research Outputs/Findings

The primary research output produced by answering RQ1 is an improved perturbation method (SRW) that uses logistic cumulative noise addition, random projection, and translation (Chapter 5). We used SRW as the perturbation method in the privacy module of APOF when answering RQ2. RQ2 is primarily focused on implementing a well-structured framework for accuracy-privacy optimisation in static datasets (Chapter 7). Since SRW is suitable for static datasets and data streams, it can be used without an issue. The challenges in data stream mining were not considered at this stage, but APOF could be used with data streams under certain conditions. The accuracy-privacy optimisation of APOF is carried out using the novel data fitting module. This is well-suited for static datasets, as the entire dataset is available. Still, it is expensive if the

volume of the dataset is high.

It was required to adapt APOF to the streaming environments (Chapter 9) when addressing RQ3. Each module of APOF was observed and improved to work with data streams. SRW was modified to achieve stable accuracy and privacy in long-term running. We used HAT to deal with concept drift, and a classifier switching method was introduced to avoid overfitting. The most challenging task was the adaptation of the data fitting module to PPDSM. However, we successfully used a sample or window of the data stream for data fitting. Combining all these changes into APOF, we implemented "Enhanced-APOF" as the output of RQ3.

10.3 Reproducibility

The reproducibility of research work is one of the main challenges in data stream mining research. Facilitating re-running the experiments and reproducing results are essential for the growth of the research as other researchers might want to access the work for further improvements. Therefore we have located the source code of the complete work in GitHub[1], and it is publicly available to access.

The project is an all-in-one package including source code and the datasets used in this work. Executing files in the "dataset-construction" folder creates all the datasets used in our work, which are interconnected with the source code, making it easier to run the project. Further, this includes a "Makefile" and a "README" file which mention necessary commands and other required configurations to prepare the environment for running the project. Hence, anyone interested can download the project, configure the project environment, run the project and reproduce the results.

[1] https://github.com/whewage/Enhanced-APOF.git

10.4 Computational Complexity

The computational complexity of the perturbation method is dominated by the matrix multiplication operation involved in the random projection. This matrix multiplication is $O(kn)$ where k is the number of records, and n is the number of features of the dataset/ data stream. The additional operations required in translation and logistic cumulative noise addition are insignificant compared to $O(kn)$. The computational complexity of classification using HAT is $O(n)$ as each record is classified separately. Therefore the overall computational complexity of the PPDM/PPDSM process is $O(n(k+1))$. We do not consider the computational complexity of the data fitting module as it is not part of the actual, real-time PPDSM process. Data fitting is performed prior to the actual PPDSM process to determine the appropriate accuracy privacy level and the noise addition rate.

10.5 Limitations of the Research

10.5.1 Data Fitting Module's Accuracy Decreases for Low Privacy Thresholds

The major limitation of this research work is related to the data fitting module of the APOF and Enhanced-APOF. The accuracy of the data fitting module is high for high privacy thresholds but can be low for low privacy thresholds. This means the data fitting module performs poorly in predicting the accuracy level correctly when the user-defined privacy is low. The reason behind this is familiar to most of the data fitting techniques. Data fitting produces highly accurate results if the predicted values are in the same range as those used to fit the data [149]. Accuracy drops when trying to predict values far from the range of the fitted data. In our case, Logistic Cumulative Noise Addition

(SRW) has been designed to achieve high privacy while maintaining high accuracy. Therefore, the accuracy and privacy values provided by SRW for all the tested noise addition rates are high. Since we used these high accuracy-privacy values for data fitting, it works well with the high privacy thresholds in the same range. However, it fails with the low privacy thresholds, which fall out of the fitted value range.

10.5.2 Noise Resetting at Constant Intervals Can Be a Threat

As a method to maintain high accuracy for infinite or lengthy data streams, we used a noise resetting technique. The noise was reset to zero for every 25,000 records. This can be more efficient and less vulnerable if it is possible to reset the noise randomly. The privacy of the data stream drops when we reset the noise at constant intervals. There is a possibility of this being identified by attackers in the long-term observation. Since privacy is zero at the reset point, it can be more vulnerable to attacks. However, there is a practicality issue when implementing random noise resetting. Random noise resetting points can be anywhere in the data stream. If we consider a logistic cycle, it can be in the beginning, middle, or end of the cycle. Especially when the random resetting point is in the middle of the logistic cycle, SRW might fail to retrieve all the benefits from the shape of the logistic curve. This negatively affects both accuracy and privacy, as we break down the natural shape of the logistic curve by noise resetting. Therefore, this is a complex task as we cannot control the random noise resetting point.

10.5.3 Other PPDM and PPDSM Methods

The scope of this research was limited to input PPDM and PPDSM methods. We focused only on perturbation methods in input PPDM and PPDSM due to their easy adaptability to work with data streams and the large variety of techniques included in this category. Other than input PPDM and PPDSM methods, output privacy preservation methods are suitable and widely used for data streams [15, 27]. Moreover, non-perturbation-based input privacy preservation methods such as anonymization [30] can be used for data streams. Adapting other possible input PPDM methods and output PPDM methods to APOF is considered a future avenue.

Chapter 11

Conclusions and Future Directions

11.1 Conclusions

This research was primarily focused on providing solutions to the following research questions.

- RQ1 - What is the most appropriate perturbation method that expresses the optimal trade-off between data privacy and data mining accuracy?
- RQ2 - What framework should be constructed to model the trade-off between privacy and accuracy for a specific user in a data mining environment?
- RQ3 - How do we extend the framework proposed in RQ2 above to a dynamically adapting or evolving data streaming environment?

For answering RQ1, a novel noise-addition-based perturbation method called "Logistic cumulative noise addition (SRW)" was proposed in Chapter 5. This is based on Cumulative noise addition proposed in [32] and inspired by the well-known logistic function [128]. SRW combines three perturbation techniques: Logistic cumulative noise addition, translation, and random projection. SRW outperformed the base method

achieving a good accuracy-privacy trade-off according to the PAM formula. The cycle-wise noise addition based on the logistic function is an effective method to optimise the accuracy-privacy trade-off. Noise resetting is a promising technique to control cumulative noise, especially in the long term. Finally, we can state that some parts of the perturbed dataset are more vulnerable than others. This is because of low noise variance at the beginning and high noise variance at the end of the logistic curve. Therefore, data records that fall at the beginning of the logistic curve are more vulnerable.

For answering RQ2, we implemented the Accuracy-Privacy Optimising Framework (APOF) in Chapter 7 to optimise the accuracy-privacy trade-off according to the user's privacy requirements. Achieving accuracy and privacy according to the user's requirements is essential and is the base of APOF. APOF consists of three modules: accuracy, privacy, and data fitting. The accuracy module performs the classification using the Hoeffding Tree and calculates the accuracy, while the privacy module perturbs data using SRW and calculates privacy. The data fitting module optimises accuracy-privacy according to the user's privacy requirement. Predicting the respective accuracy for a user-defined privacy threshold allows the user to get an initial idea about the expected accuracy and fine-tune his or her requirements if needed. APOF is well-suited for static datasets and can be used with data streams under specific conditions, such as when there is no concept drift. Data fitting could successfully be used for accuracy-privacy trade-off optimisation to retain a small error. However, the error of the accuracy prediction increases for low privacy thresholds.

For answering RQ3, we investigated APOF in-depth from a data stream mining perspective and proposed Enhanced-APOF in Chapter 9. The accuracy module of Enhanced-APOF uses Hoeffding Adaptive Tree as the classifier, which adapts well to the concept drift maintaining a stable accuracy. The perturbation method (SRW) was improved to deal with lengthy or infinite data streams without affecting accuracy and privacy in the long term. The data fitting module of APOF was modified to

use a sample or window from the data stream, as we cannot access or use the entire data stream to fit data. The entire PPDSM process for a record can be conducted within milliseconds. Hence the Enhanced-APOF is well-structured to deal with the dynamically adapting nature of data streams. Additionally, we can state that it is impossible to mathematically represent the relationship between accuracy and privacy from the scope of our experiments. However, we proved that the accuracy-privacy trade-off is not always valid, and Enhanced-APOF has optimised the trade-off to some extent.

Finally, it can be concluded that the Enhanced-APOF is a well-structured framework for optimising the trade-off between data mining accuracy and data privacy for a data streaming environment. The Enhanced-APOF has been tested with data streams with (SEA, RBF) and without (AReM, TAXI) concept drift and has proven to adapt successfully to the dynamic and evolving nature of data streams. It is fast and efficient. Moreover, the Logistic cumulative noise addition (SRW) as the privacy preservation technique, Hoeffding Adaptive Tree as the classifier, and data fitting as the trade-off optimisation technique for a specific user is a good combination that facilitates achieving this objective.

11.2 Future Directions

Several extensions and modifications can be considered possible future directions to the current work. The APOF was only tested to predict the accuracy level for a user-defined privacy threshold. This is suitable when the user's requirements are focused on privacy. However, for some users, their primary focus can be accuracy. In such situations, APOF should be tested to predict the privacy level for a user-defined accuracy threshold. Hence, the data fitting module of APOF should be tested and evaluated for this purpose. When this is done, APOF/Enhanced-APOF can be used regardless of the user's primary focus.

Another direction is to represent the relationship between accuracy and privacy mathematically. This is a complex task, as experiments could not prove a solid pattern of accuracy-privacy behaviour. The behaviour of accuracy and privacy depends on the dataset or data stream. The factors such as the type of data, number of attributes, and type of concept drift can cause this behaviour. Finding a way to investigate the behaviour of accuracy and privacy with these factors can be helpful for a mathematical representation. However, if we can represent accuracy as a function of privacy and vice versa, the data fitting module of APOF can be replaced with a mathematical representation. If this is possible, accuracy prediction for a user-defined privacy threshold becomes more straightforward. It can reduce the time and computational overhead in the data fitting module, increasing efficiency.

The data mining algorithms tested with APOF and Enhanced-APOF are HT and HAT. However, there are other data mining algorithms that can be used for data stream mining. We can replace the data mining algorithm of Enhanced-APOF with other possible techniques and compare the performance. It can give us a deeper understanding of how well APOF or Enhanced-APOF works with various data stream mining algorithms. Additionally, the Enhanced-APOF can be tested with input PPDMs suitable for data

stream mining other than perturbation. Input PPDMs such as anonymisation-based methods [30, 36, 101] are an excellent choice. This can be further extended to output PPDM methods as well. Additionally, a performance comparison of Enhanced-APOF with other existing data stream mining frameworks should be conducted. Only a few well-structured frameworks for PPDSM can be found in the literature. However, this type of performance comparison would help evaluate Enhanced-APOF in terms of its accuracy and processing speed.

Finally, it might be helpful to see how Enhanced-APOF works with real-time data streams. All our experiments were conducted using freely available actual datasets prepared as data streams, considering the timestamp each record was generated. Due to privacy concerns, it is not easy to access real-time data streams for testing. However, using Enhanced-APOF in real-time would provide a clear picture of its strengths and weaknesses in an actual PPDSM environment.

References

[1] S. Lohiya and L. Ragha, "Privacy preserving in data mining using hybrid approach," in *Proceedings - 4th International Conference on Computational Intelligence and Communication Networks, CICN 2012*. IEEE, 2012, pp. 743–746.

[2] P. Jain, M. Gyanchandani, and N. Khare, "Big data privacy: a technological perspective and review," *Journal of Big Data*, vol. 3, no. 1, 2016.

[3] A. N. Zaman, C. Obimbo, and R. A. Dara, "A novel differential privacy approach that enhances classification accuracy," in *ACM International Conference Proceeding Series*, vol. 20-22-July, 2016, pp. 79–84.

[4] A. Kiran and D. Vasumathi, "A Comprehensive Survey on Privacy Preservation Algorithms in Data Mining," in *2017 IEEE International Conference on Computational Intelligence and Computing Research, ICCIC 2017*. IEEE, 2018.

[5] A. Bifet and R. Gavaldà, "Adaptive Parameter-free Learning from Evolving Data Streams," *Advances in Intelligent Data Analysis VIII*, no. September, pp. 249–260, 2017.

[6] M. Dhanalakshmi and E. Siva Sankari, "Privacy Preserving Data Mining Techniques-Survey," in *International Conference on Information Communication and Embedded Systems (ICICES2014)*. IEEE, 2014, pp. 1–6.

[7] M. Narwaria and S. Arya, "Privacy preserving data mining:"A state of the art"," in *2016 International Conference on Computing for Sustainable Global Development (INDIACom)*. Bharati Vidyapeeth, New Delhi as the Organizer of INDIACom - 2016, 2016, pp. 1–15.

[8] M. Md Siraj, N. A. Rahmat, and M. M. Din, "A survey on privacy preserving data mining approaches and techniques," in *ACM International Conference Proceeding Series*, vol. Part F1479, 2019, pp. 65–69.

[9] S. Chidambaram and K. G. Srinivasagan, "A combined random noise perturbation approach for multi level privacy preservation in data mining," in *2014 International Conference on Recent Trends in Information Technology, ICRTIT 2014*. IEEE, 2014, pp. 1–6.

[10] M. K. Paul, M. R. Islam, and A. S. Sattar, "An efficient perturbation approach for multivariate data in sensitive and reliable data mining," *Journal of Information Security and Applications*, vol. 62, no. August, p. 102954, 2021. [Online]. Available: https://doi.org/10.1016/j.jisa.2021.102954

[11] J. J. V. Nayahi and V. Kavitha, "Privacy and utility preserving data clustering for data anonymization and distribution on Hadoop," *Future Generation Computer Systems*, vol. 74, pp. 393–408, 2017. [Online]. Available: http://dx.doi.org/10.1016/j.future.2016.10.022

[12] A. W. Putri and L. Hira, "Hybrid transformation in privacy-preserving data mining," in *Proceedings of 2016 International Conference on Data and Software Engineering, ICoDSE 2016*, 2017, pp. 0–5.

[13] C. C. Aggarwal and P. S. Yu, *Privacy-Preserving Data Mining -Models and Algorithms*. USA: Springer US, 2008.

[14] D. Kim, Z. Chen, and A. Gangopadhyay, "Optimizing privacy-accuracy tradeoff for privacy preserving distance-based classification," *International Journal of Information Security and Privacy*, vol. 6, no. 2, pp. 16–33, 2012.

[15] R. Kotecha and S. Garg, "Preserving output-privacy in data stream classification," *Progress in Artificial Intelligence*, vol. 6, no. 2, pp. 87–104, 2017.

[16] C. Liu and S. e. a. Chen, "A novel privacy preserving method for data publication," *Information Sciences*, vol. 501, pp. 421–435, 2019. [Online]. Available: https://doi.org/10.1016/j.ins.2019.06.022

[17] C. N. Modi, U. P. Rao, and D. R. Patel, "Maintaining privacy and data quality in privacy preserving association rule mining," in *2010 2nd International Conference on Computing, Communication and Networking Technologies, ICCCNT 2010*. IEEE, 2010, pp. 7–12.

[18] J. Cao, B. Carminati, E. Ferrari, and K. L. Tan, "CASTLE: Continuously anonymizing data streams," *IEEE Transactions on Dependable and Secure Computing*, vol. 8, no. 3, pp. 337–352, 2011.

[19] E. Acuna and E. Acuña, "Preprocessing in Data Mining," *International Encyclopedia of Statistical Science*, no. September, pp. 1083–1085, 2011.

[20] S. García, J. Luengo, and F. Herrera, *Data Preprocessing in Data Mining*, 2015, vol. 72.

[21] D. Patel and R. Kotecha, "Privacy preserving data mining: A parametric analysis," *Advances in Intelligent Systems and Computing*, vol. 516, pp. 139–149, 2017.

[22] C. Gokulnath, M. K. Priyan, E. V. Balan, K. P. Prabha, and R. Jeyanthi, "Preservation of privacy in data mining by using PCA based perturbation technique," in *2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials, ICSTM 2015 - Proceedings*, no. May. IEEE, 2015, pp. 202–206.

[23] A. Kaur, "A hybrid approach of privacy preserving data mining using suppression and perturbation techniques," in *IEEE International Conference on Innovative Mechanisms for Industry Applications, ICIMIA 2017 - Proceedings*, no. Icimia. IEEE, 2017, pp. 306–311.

[24] N. Bhandari and P. Pahwa, "Comparative Analysis of Privacy-Preserving Data Mining Techniques," in *International Conference on Innovative Computing and Communications*. Springer Singapore, 2019, pp. 535–541. [Online]. Available: http://dx.doi.org/10.1007/978-981-13-2354-6{_}54

[25] M. B. Malik, M. A. Ghazi, and R. Ali, "Privacy preserving data mining techniques: Current scenario and future prospects," in *Proceedings of the 2012 3rd International Conference on Computer and Communication Technology, ICCCT 2012*. IEEE, 2012, pp. 26–32.

[26] T. Carvalho and N. Moniz, "The Compromise of Data Privacy in Predictive Performance," in *International Symposium on Intelligent Data Analysis*, vol. 1, 2021, pp. 426–438.

[27] B. Peng, X. Geng, and J. Zhang, "Combined data distortion strategies for privacy-preserving data mining," in *ICACTE 2010 - 2010 3rd International Conference on Advanced Computer Theory and Engineering, Proceedings*, vol. 1, 2010, pp. 572–576.

[28] K. Chen and L. Liu, "Geometric data perturbation for privacy preserving outsourced data mining," *Knowledge and Information Systems*, vol. 29, no. 3, pp. 657–695, 2011.

[29] V. Rajalakshmi and G. S. Mala, "An intensified approach for privacy preservation in incremental data mining," *Advances in Intelligent Systems and Computing*, vol. 178, pp. 347–355, 2013.

[30] L. Sweeney, "k-Anonymity: A model for protecting privacy," *Ieee Security And Privacy*, vol. 10, no. 5, pp. 1–14, 2002.

[31] H.-y. Tran and J. Hu, "Privacy-preserving big data analytics - A comprehensive survey," *Journal of Parallel and Distributed Computing*, vol. 134, pp. 207–218, 2019. [Online]. Available: https://doi.org/10.1016/j.jpdc.2019.08.007

[32] B. Denham, R. Pears, and M. A. Naeem, "Enhancing random projection with independent and cumulative additive noise for privacy-preserving data stream mining," *Expert Systems with Applications*, vol. 152, 2020.

[33] J. J. Kim and W. E. Winkler, "Multiplicative Noise for Masking Continuous Data," Statistical Research Division U.S. Bureau of the Census, Washington, Tech. Rep., 2003.

[34] K. Chen and L. Liu, "A random rotation perturbation approach to privacy preserving data classification," in *International Conference on Data Mining*, 2005.

[35] K. e. a. Liu, "Random projection-based multiplicative data perturbation for privacy preserving distributed data mining," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 1, pp. 92–106, 2006.

[36] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "*l*-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data*, vol. 1, no. 1, 2007.

[37] C. C. Aggarwal and P. S. Yu, "A Condensation Approach to Privacy Preserving Data Mining," in *Advances in Database Technology - EDBT 2004*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 183–199.

[38] M. A. Chamikara, P. Bertok, D. Liu, S. Camtepe, and I. Khalil, "An efficient and scalable privacy preserving algorithm for big data and data streams," *Computers and Security*, vol. 87, p. 101570, 2019. [Online]. Available: https://doi.org/10.1016/j.cose.2019.101570

[39] K. Chen and L. Liu, "Privacy preserving data classification with rotation perturbation," *Proceedings - IEEE International Conference on Data Mining, ICDM*, pp. 589–592, 2005.

[40] M. A. Chamikara, P. Bertok, I. Khalil, D. Liu, and S. Camtepe, "PPaaS: Privacy Preservation as a Service," *Computer Communications*, vol. 173, pp. 192–205, 2021. [Online]. Available: https://doi.org/10.1016/j.comcom.2021.04.006

[41] J. Soria-Comas, J. Domingo-Ferrer, D. Sanchez, and S. Martinez, "T-closeness through microaggregation: Strict privacy with enhanced utility preservation," in *2016 IEEE 32nd International Conference on Data Engineering, ICDE 2016*, vol. 27. Institute of Electrical and Electronics Engineers Inc, 2016, pp. 1464–1465.

[42] S. Mukherjee, M. Banerjee, Z. Chen, and A. Gangopadhyay, "A privacy preserving technique for distance-based classification with worst case privacy guarantees," *Data and Knowledge Engineering*, vol. 66, no. 2, pp. 264–288, 2008.

[43] K. Mivule, C. Turner, and S. Y. Ji, "Towards a differential privacy and utility preserving machine learning classifier," *Procedia Computer Science*, vol. 12, pp. 176–181, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.procs.2012.09.050

[44] V. Tayal and R. Srivastava, *Challenges in mining big data streams*. Springer Singapore, 2019, vol. 847. [Online]. Available: http://dx.doi.org/10.1007/978-981-13-2254-9{_}15

[45] G. Krempl, I. Žliobaite, D. Brzeziński, E. Hüllermeier, M. Last, V. Lemaire, T. Noack, A. Shaker, S. Sievi, M. Spiliopoulou, and J. Stefanowski, "Open challenges for data stream mining research," *ACM SIGKDD Explorations Newsletter*, vol. 16, no. 1, pp. 1–10, 2014.

[46] C. Y. Lin, Y. H. Kao, W. B. Lee, and R. C. Chen, "An efficient reversible privacy-preserving data mining technology over data streams," *SpringerPlus*, vol. 5, no. 1, pp. 1–11, 2016.

[47] M. A. Chamikara, P. Bertok, D. Liu, S. Camtepe, and I. Khalil, "Efficient data perturbation for privacy preserving and accurate data stream mining," *Pervasive and Mobile Computing*, vol. 48, pp. 1–19, 2018. [Online]. Available: https://doi.org/10.1016/j.pmcj.2018.05.003

[48] H. M. Gomes, J. Read, A. Bifet, J. P. Barddal, and J. Gama, "Machine Learning for Streaming Data: State of the Art, Challenges, and Opportunities," *SIGKDD Explor. Newsl.*, vol. 21, no. 2, pp. 6–22, 2019. [Online]. Available: https://doi.org/10.1145/3373464.3373470

[49] W. Wang, J. Li, C. Ai, and Y. Li, "Privacy protection on sliding window of data streams," in *Proceedings of the 3rd International Conference on Collaborative Computing: Networking, Applications and Worksharing, CollaborateCom 2007*, 2007, pp. 213–221.

[50] J. Wang, C. Deng, and X. Li, "Two Privacy-Preserving Approaches for Publishing Transactional Data Streams," *IEEE Access*, vol. 6, pp. 23 648–23 658, 2018.

[51] D. Martínez Rodríguez, J. Nin, and M. Nuñez-del Prado, "Towards the adaptation of SDC methods to stream mining," *Computers and Security*, vol. 70, no. 2017, pp. 702–722, 2017. [Online]. Available: https://doi.org/10.1016/j.cose.2017.08.011

[52] M. Khavkin and M. Last, "Preserving differential privacy and utility of non-stationary data streams," in *IEEE International Conference on Data Mining Workshops, ICDMW*, vol. 2018-Novem. IEEE, 2019, pp. 29–34.

[53] A. Cuzzocrea, "Privacy-preserving big data stream mining: Opportunities, challenges, directions," in *IEEE International Conference on Data Mining Workshops, ICDMW*, vol. 2017-Novem, 2017, pp. 992–994.

[54] G. Zhang and S. Li, "Research on Differentially Private Bayesian Classification Algorithm for Data Streams," in *2019 4th IEEE International Conference on Big Data Analytics, ICBDA 2019*. IEEE, 2019, pp. 14–20.

[55] M. A. Chamikara, P. Bertok, D. Liu, S. Camtepe, and I. Khalil, "Efficient privacy preservation of big data for accurate data mining," *Information Sciences*, vol. 527, pp. 420–443, 2020.

[56] S. Dutta and A. K. Guppta, "Privacy in Data Mining - A Review," in *International Conference onComputing for Sustainable Global Development (INDIACom)*, vol. 6, 2016, pp. 556–559.

[57] O. Feyisetan, B. Balle, T. Drake, and T. Diethe, "Privacy- and utility-preserving textual analysis via calibrated perturbations," in *CEUR Workshop Proceedings*, vol. 2573, 2020, pp. 41–42.

[58] Y. C. Tsai, S. L. Wang, C. Y. Song, and I. H. Ting, "Privacy and utility effects of k-anonymity on association rule hiding," in *ACM International Conference Proceeding Series*, 2016, pp. 0–5.

[59] X. Qi and M. Zong, "An Overview of Privacy Preserving Data Mining," *Procedia Environmental Sciences*, vol. 12, no. Icese 2011, pp. 1341–1347, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.proenv.2012.01.432

[60] S. M. Kabir, A. M. Youssef, and A. K. Elhakeem, "On data distortion for privacy preserving data mining," in *Canadian Conference on Electrical and Computer Engineering*, 2007, pp. 308–311.

[61] J. Wang and J. Zhang, "Addressing accuracy issues in privacy preserving data mining through matrix factorization," *ISI 2007: 2007 IEEE Intelligence and Security Informatics*, pp. 217–220, 2007.

[62] K. S. Babu and S. K. Jena, "Balancing between utility and privacy for k-anonymity," *Communications in Computer and Information Science*, vol. 191 CCIS, no. PART 2, pp. 1–8, 2011.

[63] B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering - A systematic literature review," *Information and Software Technology*, vol. 51, no. 1, pp. 7–15, 2009. [Online]. Available: http://dx.doi.org/10.1016/j.infsof.2008.09.009

[64] S. Sharma and S. Ahuja, *Privacy Preserving Data Mining: A Review of the State of the Art BT - Harmony Search and Nature Inspired Optimization Algorithms*. Springer Singapore, 2019. [Online]. Available: http://dx.doi.org/10.1007/978-981-13-0761-4{_}1

[65] Y. Abdul, A. S. Aldeen, M. Salleh, and M. A. Razzaque, "A comprehensive review on privacy preserving data mining," *SpringerPlus*, 2015.

[66] B. Vishwakarma, H. Gupta, and M. Manoria, "A survey on privacy preserving mining implementing techniques," in *2016 Symposium on Colossal Data Analysis and Networking, CDAN 2016*. IEEE, 2016, pp. 7–11.

[67] N. Nasiri and M. Keyvanpour, "Classification and Evaluation of Privercy Preserving Data Mining Methods," in *11th International Conference on Information and Knowledge Discovery(IKT)*, 2020, pp. 17–22.

[68] A. B. Sakpere and A. V. Kayem, "A state-of-the-art review of data stream anonymization schemes," *Information Security in Diverse Computing Environments*, pp. 24–50, 2014.

[69] S. Sangeetha and G. S. Sadasivam, "Privacy of Big Data : A Review," in *Handbook of Big Data and IoT Security*. Springer Nature Switzerland AG, 2019.

[70] S. A. Shanthi and M. Karthikeyan, "A review on privacy preserving data mining," in *IEEE International Conference on Computational Intelligence and Computing Research*, vol. 4. IEEE, 2012, pp. 1–36.

[71] G. Arumugam and V. Sulekha, "IMR based anonymization for privacy preservation in data mining," in *ACM International Conference Proceeding Series*, vol. Part F1305, 2016.

[72] A. Sachan, D. Roy, and P. V. Arun, "An analysis of privacy preservation techniques in data mining," *Advances in Intelligent Systems and Computing*, vol. 178, pp. 119–128, 2013.

[73] M. Ketel and A. Homaifar, "Privacy-preserving mining by rotational data transformation," in *Proceedings of the Annual Southeast Conference*, vol. 1, 2005, pp. 1233–1236.

[74] S. Upadhyay, C. Sharma, P. Sharma, P. Bharadwaj, and K. R. Seeja, "Privacy preserving data mining with 3-D rotation transformation," *Journal of King Saud University - Computer and Information Sciences*, vol. 30, no. 4, pp. 524–530, 2018. [Online]. Available: https://doi.org/10.1016/j.jksuci.2016.11.009

[75] T. Javid and M. K. Gupta, "Privacy Preserving Classification using 4-Dimensional Rotation Transformation," in *Proceedings of the 2019 8th International Conference on System Modeling and Advancement in Research Trends, SMART 2019*, 2020, pp. 279–284.

[76] K. Chen, G. Sun, and L. Liu, "Towards Attack-Resilient Geometric Data Perturbation," in *SIAM International Conference on Data Mining*, 2007, pp. 78–89.

[77] C. Dwork, "Differential privacy: A survey of results," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 4978 LNCS, pp. 1–19, 2008.

[78] W. Tang, Y. Zhou, Z. Wu, L. Lu, and M. Li, "Naive Bayes Classification based on Differential Privacy," in *ACM International Conference Proceeding Series*, 2019.

[79] J. Wang and W. K. V. Chan, "A Design for Private Data Protection Combining with Data Perturbation and Data Reconstruction," in *ACM International Conference Proceeding Series*, 2021, pp. 545–550.

[80] K. Singh and L. Batten, "An attack-resistant hybrid data-privatization method with low information loss," *IFIP Advances in Information and Communication Technology*, vol. 401, pp. 263–271, 2013.

[81] A. K. Upadhayay and e. a. Agarwal, "Privacy Preserving Data Mining: A New Methodology for Data Transformation," in *Proceedings of the First International Conference on Intelligent Human Computer Interaction*, 2009, pp. 372–390.

[82] G. S. Kumar and K. Premalatha, "Securing private information by data perturbation using statistical transformation with three dimensional shearing," *Applied Soft Computing*, vol. 112, p. 107819, 2021. [Online]. Available: https://doi.org/10.1016/j.asoc.2021.107819

[83] A. Kiran and D. Vasumathi, *Data mining: Min–max normalization based data perturbation technique for privacy preservation*. Springer Singapore, 2020, vol. 1090. [Online]. Available: http://dx.doi.org/10.1007/978-981-15-1480-7{_}66

[84] K. P. Lin, Y. W. Chang, and M. S. Chen, "Secure support vector machines outsourcing with random linear transformation," *Knowledge and Information Systems*, vol. 44, no. 1, pp. 147–176, 2015. [Online]. Available: http://dx.doi.org/10.1007/s10115-014-0751-1

[85] C. C. Aggarwal and P. S. Yu, "On static and dynamic methods for condensation-based privacy-preserving data mining," *ACM Transactions on Database Systems*, vol. 33, no. 1, pp. 1–40, 2008.

[86] N. Meghanathan, D. Nagamalai, and S. Rajasekaran, "A Comparative Study of Data Perturbation Using Fuzzy Logic to Preserve Privacy," *Lecture Notes in Electrical Engineering*, vol. 284 LNEE, pp. 161–170, 2014.

[87] T. Jahan, G. Narsimha, and C. V. Guru Rao, "Multiplicative data perturbation using fuzzy logic in preserving privacy," in *ACM International Conference Proceeding Series*, vol. 04-05-Marc, 2016.

[88] I. Cano, S. Ladra, and V. Torra, "Evaluation of information loss for privacy preserving data mining through comparison of fuzzy partitions," in *2010 IEEE World Congress on Computational Intelligence, WCCI 2010*. IEEE, 2010.

[89] S. Agrawal and J. R. Haritsa, "A Framework for High-Accuracy Privacy-Preserving Mining," in *Proceedings of the 21st International Conference on Data Engineering*, no. Icde, 2005.

[90] L. e. a. Xiaoping, "Research on privacy preserving data mining based on randomized response," in *ACM International Conference Proceeding Series*, 2020, pp. 129–132.

[91] N. P. Nethravathi, P. G. Rao, P. D. Shenoy, K. R. Venugopal, and M. Indramma, "CBTS: Correlation based transformation strategy for privacy preserving data mining," in *2015 IEEE International WIE Conference on Electrical and Computer Engineering, WIECON-ECE 2015*. IEEE, 2016, pp. 190–194.

[92] G. Li and Y. Wang, "Privacy-preserving data mining based on sample selection and singular value decomposition," in *Proceedings - 2011 International Conference on Internet Computing and Information Services, ICICIS 2011*. IEEE, 2011, pp. 298–301.

[93] S. Xu, J. Zhang, D. Han, and J. Wang, "Singular value decomposition based data distortion strategy for privacy protection," *Knowledge and Information Systems*, vol. 10, no. 3, pp. 383–397, 2006.

[94] M. M. Hasan, S. Hossain, M. K. Paul, and A. S. Sattar, "A New Hybrid Approach For Privacy Preserving Data Mining Using Matrix Decomposition Technique," in *2019 4th International Conference on Electrical Information and Communication Technology, EICT 2019*, no. December. IEEE, 2019, pp. 20–22.

[95] G. Li and R. Xue, "A New Privacy-Preserving Data Mining Method Using Non-negative Matrix Factorization and Singular Value Decomposition," *Wireless Personal Communications*, vol. 102, no. 2, pp. 1799–1808, 2018. [Online]. Available: https://doi.org/10.1007/s11277-017-5237-5

[96] T.-p. Hong, K.-t. Yang, C.-w. Lin, and S.-l. Wang, "Evolutionary privacy-preserving data mining," in *World Automation Congress*. IEEE, 2010, pp. 2–8.

[97] R. Kaur and M. Bansal, "Transformation approach for boolean attributes in privacy preserving data mining," in *Proceedings on 2015 1st International Conference on Next Generation Computing Technologies, NGCT 2015*, no. September, 2016, pp. 644–648.

[98] S. Vijayarani and A. Tamilarasi, "An efficient masking technique for sensitive data protection," in *International Conference on Recent Trends in Information Technology, ICRTIT 2011*. IEEE, 2011, pp. 1245–1249.

[99] K. e. a. Alotaibi, "Non-linear dimensionality reduction for privacy-preserving data classification," in *Proceedings - 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust and 2012 ASE/IEEE International Conference on Social Computing, SocialCom/PASSAT 2012*. IEEE, 2012, pp. 694–701.

[100] S. Vijayarani and A. Tamilarasi, "Data transformation and data transitive techniques for protecting sensitive data in privacy preserving data mining," in *Emerging Trends in Computing, Informatics, Systems Sciences, and Engineering*, T. Sobh and K. Elleithy, Eds. New York, NY: Springer New York, 2013, pp. 345–355.

[101] N. Li, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity," in *IEEE 23rd International Conference on Data Engineering*, no. 2, 2007, pp. 106–115.

[102] J. Wang, Y. Luo, S. Jiang, and J. Le, "A survey on anonymity-based privacy preserving," in *2009 International Conference on E-Business and Information System Security, EBISS 2009*. IEEE, 2009, pp. 7–10.

[103] C. N. Sowmyarani, G. N. Srinivasan, and K. Sukanya, "A new privacy preserving measure: p-sensitive, t-closeness," *Advances in Intelligent Systems and Computing*, vol. 174 AISC, pp. 57–62, 2013.

[104] K. Oishi, "Proposal of l -Diversity Algorithm Considering Distance between Sensitive Attribute Values," in *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2017, pp. 1–8.

[105] B. Suma and G. Shobha, "Privacy preserving association rule hiding using border based approach," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 23, no. 2, pp. 1137–1145, 2021.

[106] P. Cheng, S. C. Chu, C. W. Lin, and J. F. Roddick, "Distortion-based heuristic sensitive rule hiding method - The greedy way," *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*, vol. 8481, pp. 77–86, 2014.

[107] E. Poovammal and M. Ponnavaikko, "An improved method for privacy preserving data mining," in *2009 IEEE International Advance Computing Conference, IACC 2009*, no. March, 2009, pp. 1453–1458.

[108] P. Deivanai, J. J. V. Nayahi, and V. Kavitha, "A hybrid data anonymization integrated with suppression for preserving privacy in mining multi party data," in *International Conference on Recent Trends in Information Technology, ICRTIT 2011*. IEEE, 2011, pp. 732–736.

[109] Z. Teng and W. Du, "A hybrid multi-group approach for privacy-preserving data mining," *Knowledge and Information Systems*, vol. 19, no. 2, pp. 133–157, 2009.

[110] S. G. Tsiafoulis, V. C. Zorkadis, and E. Pimenidis, "Maximum Entropy Oriented Anonymization Algorithm," *Social Informatics and Telecommunications Engineering 2012*, pp. 9–16, 2012.

[111] M. A. Kadampur and D. V. Somayajulu, "A data perturbation method by field rotation and binning by averages strategy for privacy preservation," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5326 LNCS, pp. 250–257, 2008.

[112] V. Ashok and R. Mukkamala, "Data mining without data: A novel approach to privacy-preserving collaborative distributed data mining," in *Proceedings of the ACM Conference on Computer and Communications Security*, 2011, pp. 159–164.

[113] C. Sun, H. Gao, J. Zhou, Y. Fu, and L. She, "A new hybrid approach for privacy preserving distributed data mining," *IEICE Transactions on Information and Systems*, vol. E97-D, no. 4, pp. 876–883, 2014.

[114] E. Mohammadian, M. Noferesti, and R. Jalili, "FAST: Fast anonymization of big data streams," in *ACM International Conference Proceeding Series*, vol. 04-07-Augu, 2014.

[115] M. A. Mohamed, M. H. Nagi, and S. M. Ghanem, "A clustering approach for anonymizing distributed data streams," *Proceedings of 2016 11th International Conference on Computer Engineering and Systems, ICCES 2016*, pp. 9–16, 2017.

[116] G. Navarro-Arribas and V. Torra, "Rank swapping for stream data," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8825, pp. 217–226, 2014.

[117] S. Virupaksha and V. Dondeti, "Anonymized noise addition in subspaces for privacy preserved data mining in high dimensional continuous data," *Peer-to-Peer Networking and Applications*, vol. 14, no. 3, pp. 1608–1628, 2021.

[118] P. Rajesh, G. Narisimha, and C. Rupa, "Fuzzy based privacy preserving classification of data streams," in *ACM International Conference Proceeding Series*, 2012, pp. 784–788.

[119] A. Nyati, S. K. Dargar, and S. Sharda, *Design and implementation of a new model for privacy preserving classification of data streams*. Springer Singapore, 2018, vol. 906. [Online]. Available: http://dx.doi.org/10.1007/978-981-13-1813-9{_}45

[120] P. Ah-Fat and M. Huth, "Optimal Accuracy-Privacy Trade-Off for Secure Computations," *IEEE Transactions on Information Theory*, vol. 65, no. 5, pp. 3165–3182, 2019.

[121] N. H. Tran, N. A. Le-Khac, and M. T. Kechadi, "Lightweight privacy-Preserving data classification," *Computers and Security*, vol. 97, p. 101835, 2020. [Online]. Available: https://doi.org/10.1016/j.cose.2020.101835

[122] C. R. Giannella, K. Liu, and H. Kargupta, "Breaching Euclidean distance-preserving data perturbation using few known inputs," *Data and Knowledge Engineering*, vol. 83, pp. 93–110, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.datak.2012.10.004

[123] F. Yang and X. Liao, "An optimized sanitization approach for minable data publication," *Big Data Mining and Analytics*, vol. 5, pp. 257–269, 6 2022.

[124] A. Miyaji and M. S. Rahman, "Privacy-Preserving Data Mining : A Game-Theoretic Approach," *Data and Applications Security and Privacy XXV*, pp. 186–200, 2011.

[125] T. P. Hong, C. W. Lin, K. T. Yang, and S. L. Wang, "A heuristic data-sanitization approach based on TF-IDF," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6703 LNAI, no. PART 1, pp. 156–164, 2011.

[126] J. Gitanjali, J. Indumathi, N. C. Sriman, and N. Iyengar, "A Pristine Clean Cabalistic Foruity Strategize Based Approach for Incremental Data Stream," in *IEEE 2nd International Advance Computing Conference*. IEEE, 2010, pp. 410–415.

[127] U. Hewage, R. Pears, and M. A. Naeem, "Optimizing the Trade-off Between Classification Accuracy and Data Privacy in the Area of Data Stream Mining," *International Journal of Artificial Intelligence*, vol. 1, no. 1, pp. 147–167, 2022.

[128] P. F. Verhulst, "Logistic Function," 1838. [Online]. Available: https://en.wikipedia.org/wiki/Logistic{_}function

[129] M. R. K. Raziyeh Zall, "On the Construction of Multi-Relational Classifier Based on Canonical Correlation Analysis," *International Journal of Artificial Intelligence*, vol. 17, 2019. [Online]. Available: http://www.ceserp.com/cp-jour/index.php/ijai/article/view/5274

[130] C. Pozna and R. E. Precup, "Applications of signatures to expert systems modelling," *Acta Polytechnica Hungarica*, vol. 11, no. 2, pp. 21–39, 2014.

[131] M. U. Ahmed, S. Brickman, A. Dengg, N. Fasth, M. Mihajlovic, and J. Norman, "A machine learning approach to classify pedestrians' events based on imu and gps," *International Journal of Artificial Intelligence*, vol. 17, no. 2, pp. 154–167, 2019. [Online]. Available: http://www.ceser.in/ceserp/index.php/ijai/article/view/6260/6207

[132] A. Bifet, R. Kirkby, P. Kranen, and P. Reutemann, "Massive Online Analysis (MOA) Manual," Centre for Open Software Innovation, University of Waikato, Tech. Rep. March, 2009. [Online]. Available: http://scholar.google.com/scholar?hl=en{&}btnG=Search{&}q=intitle:Massive+Online+Analysis+Manual{#}1

[133] H. M. Gomes, "Adaptive random forests for evolving data stream classification," *Machine Learning*, vol. 106, no. 9-10, pp. 1469–1495, 2017.

[134] R. Agrawal and R. Srikant, "Privacy-preserving data mining," *SIGMOD Rec.*, vol. 29, no. 2, pp. 439–450, 2000.

[135] K. Liu, C. Giannella, and H. Kargupta, *A Survey of Attack Techniques on Privacy-Preserving Data Perturbation Methods*, 06 2008, pp. 359–381.

[136] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, "On the privacy preserving properties of random data perturbation techniques," in *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2003, pp. 99–106.

[137] S. Guo, X. Wu, and Y. Li, "On the lower bound of reconstruction error for spectral filtering based privacy preserving data mining," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 4213 LNAI, pp. 520–527, 2006.

[138] Z. Huang, W. Du, and B. Chen, "Deriving private information from randomized data," *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 37–48, 2005.

[139] K. Liu, C. Giannella, and H. Kargupta, "An Attacker's View of Distance Preserving Maps for Privacy Preserving Data Mining," *In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. Vol. 4213, pp. 297–308, 2006. [Online]. Available: https://doi.org/10.1007/11871637{_}30

[140] S. Guo and X. Wu, "Deriving private information from general linear transformation perturbed data," *Proceedings of the 2006 SIAM International Conference on Data Mining*, pp. 59–69, 2006.

[141] L. Liu, J. Wang, and J. Zhang, "Privacy vulnerabilities with background information in data perturbation," in *Society for Industrial and Applied Mathematics - 9th SIAM International Conference on Data Mining 2009, Proceedings in Applied Mathematics*, vol. 3. Te?hni?al Report CMIDA-HiPSCCS 005-08, Department of Co mputer S?ien?e, University of Ken- tu?ky, KY, 2009, pp. 1268–1277.

[142] B. D. Okkalioglu, M. Okkalioglu, M. Koc, and H. Polat, "A survey: deriving private information from perturbed data," *Artificial Intelligence Review*, vol. 44, no. 4, pp. 547–569, 2015.

[143] K. Liu, "Multiplicative Data Perturbation for Privacy Preserving Data Mining," PhD thesis, University of Maryland, Baltimore County (UMBC), 2007.

[144] Y. Sang, H. Shen, and H. Tian, "Effective reconstruction of data perturbed by random projections," *IEEE Transactions on Computers*, vol. 61, no. 1, pp. 101–117, 2012.

[145] A. Bifet, "MOA Data Stream Mining: A practical approach," Centre for Open Software Innovation, University of Waikato, Tech. Rep., 2009. [Online]. Available: http://dspace.cusat.ac.in/jspui/handle/123456789/3616

[146] R. E. Precup, T. A. Teban, A. Albu, A. B. Borlea, I. A. Zamfirache, and E. M. Petriu, "Evolving Fuzzy Models for Prosthetic Hand Myoelectric-Based Control," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 7, pp. 4625–4636, 2020.

[147] R. C. Roman, R. E. Precup, C. A. Bojan-Dragos, and A. I. Szedlak-Stinean, "Combined Model-Free Adaptive Control with Fuzzy Component by Virtual Reference Feedback Tuning for Tower Crane Systems," *Procedia Computer Science*, vol. 162, no. Itqm 2019, pp. 267–274, 2019. [Online]. Available: https://doi.org/10.1016/j.procs.2019.11.284

[148] U. L. Yuhana, N. Fanani, E. M. Yuniarno, S. Rochimah, L. Kóczy, and M. H. Purnomo, "Combining fuzzy signature and rough sets approach for predicting the minimum passing level of competency achievement," *International journal of artificial intelligence*, vol. 18, pp. 237–249, 2020.

[149] U. Hewage, R. Sinha, and M. A. Naeem, "An Accuracy-Privacy Optimization Framework Considering User's Privacy Requirements for Data Stream Mining," *Transactions on Knowledge Discovery from Data*, no. 1, In review.

[150] B. C. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," *ACM Computing Surveys*, vol. 42, no. 4, 2010.

[151] T. Wang, Z. Zheng, M. H. Rehmani, S. Yao, and Z. Huo, "Privacy preservation in big data from the communication perspective—A survey," *IEEE Communications Surveys and Tutorials*, vol. 21, no. 1, pp. 753–778, 2019.

[152] S. R. M. Oliveira and O. R. Zaïane, "Privacy Preserving Clustering By Data Transformation," in *Proc. of the 18th Brazilian Symposium on Databases*, vol. 1, 2003, pp. 304—-318.

[153] H. Polat and W. Du, "Privacy-Preserving Collaborative Filtering Using Randomized Perturbation Techniques," in *Third IEEE International Conference on Data Mining, Melbourne.*, 2003, pp. 625–628.

[154] C. Giannella, H. Kargupta, and K. Liu, *A Survey of Attack Techniques on Privacy-Preserving Data Perturbation Methods*. ResearchGate, 2008, no. February 2014.

[155] C. Rosset, "A Review of Online Decision Tree Learning Algorithms," Johns Hopkins University, Tech. Rep., 2015.

[156] P. K. Srimani and M. M. Patil, "Performance analysis of Hoeffding trees in data streams by using massive online analysis framework," *International Journal of Data Mining, Modelling and Management*, vol. 7, pp. 293–313, 2015.

[157] X. C. Pham, M. T. Dang, S. V. Dinh, S. Hoang, T. T. Nguyen, and A. W. C. Liew, "Learning from Data Stream Based on Random Projection and Hoeffding Tree Classifier," in *DICTA 2017 - 2017 International Conference on Digital Image Computing: Techniques and Applications*, vol. 2017-Decem, 2017, pp. 1–8.

[158] M. Zhong, "An analysis of misclassification rates for decision trees," Ph.D. dissertation, University of Central Florida, 2007.

[159] H. H. Huang, C. K. Hsiao, and S. Y. Huang, "Nonlinear regression analysis," *International Encyclopedia of Education*, pp. 339–346, 2010.

[160] H. J. Motulsky and L. A. Ransnas, "Fitting curves to data using nonlinear regression: a practical and nonmathematical review," *The FASEB Journal*, vol. 1, pp. 365–374, 1987.

[161] H. Motulsky and A. Christopoulos, "Fitting Models to Biological Data Using Linear and Nonlinear Regression, A practical guide to curve fitting," GraphPad Software Inc., San Diego CA, Tech. Rep., 2003. [Online]. Available: www.graphpad.com

[162] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. New York: Cambridge University Press, USA, 2004.

[163] T. Hastie, R. Tibshirani, and J. Friedman, *Elements of Statistical Prediction Reasoning-Data mining, Inference and Prediction*, 2nd ed. Springer, 2009, vol. 31.

[164] M. Eigensatz, "Insights into the Geometry of the Gaussian Kernel and an Application in Geometric Modeling," Master Thesis, Swiss Federal Institute of Technology Z¨urich, 2006.

[165] G. R. Lanckriet, M. Deng, N. Cristianini, M. I. Jordan, and W. S. Noble, "Kernel-based data fusion and its application to protein function prediction in yeast." in *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 2004, pp. 300–311.

[166] H. Shimazaki and S. Shinomoto, "Kernel bandwidth optimization in spike rate estimation," *Journal of Computational Neuroscience*, vol. 29, pp. 171–182, 2010.

[167] O. Bousquet and F. Pérez-Cruz, "Kernel methods and their applications to signal processing," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 4. IEEE, 2003, pp. 860–863.

[168] T. Hofmann, B. Schölkopf, and A. J. Smola, "Kernel methods in machine learning," *Annals of Statistics*, vol. 36, pp. 1171–1220, 2008.

[169] J. Racine and Q. Li, "Nonparametric estimation of regression functions with both categorical and continuous data," *Journal of Econometrics*, vol. 119, pp. 99–130, 2004.

[170] Generateme, "Fastmath-Kernel," 2020. [Online]. Available: https://generateme.github.io/fastmath/fastmath.kernel.html{#}var-kernel

[171] Data Flair, "Kernel Functions-Introduction to SVM Kernel & Examples," 2021. [Online]. Available: https://data-flair.training/blogs/svm-kernel-functions/

[172] C. Souza, "Kernel Functions for Machine Learning Applications," pp. 1–23, 2010. [Online]. Available: http://crsouza.com/2010/03/kernel-functions-for-machine-learning-applications/

[173] C. Campbell, "An Introduction to Kernel Methods Kernel Methods," *Studies in Fuzziness and Soft Computing*, vol. 66, pp. 155–192, 2001.

[174] K. K. Delibasis, "Efficient implementation of Gaussian and laplacian kernels for feature extraction from IP fisheye cameras," *Journal of Imaging*, vol. 4, 2018.

[175] S. Melacci and M. Belkin, "Laplacian support vector machines trained in the primal," *Journal of Machine Learning Research*, vol. 12, pp. 1149–1184, 2011.

[176] R. Kondor and H. Pan, "The multiscale laplacian graph kernel," *Advances in Neural Information Processing Systems*, pp. 2990–2998, 2016.

[177] M. G. Genton, "Classes of Kernels for Machine Learning: A Statistics Perspective," *CrossRef Listing of Deleted DOIs*, vol. 1, pp. 299–312, 2000.

[178] J. Fitzsimons, "Kernel Methods: Generalisations, Scalability and Towards the Future of Machine Learning," Ph.D. dissertation, University of Oxford, 2019.

[179] D. Pernes, K. Fernandes, and J. S. Cardoso, "Directional support vector machines," *Applied Sciences (Switzerland)*, vol. 9, pp. 1–19, 2019.

[180] H. Elghawalby and E. R. Hancock, "Graph embedding using an edge-based wave kernel," Tech. Rep., 2010.

[181] F. Tronarp, T. Karvonen, and S. Sarkka, "Mixture representation of the matérn class with applications in state space approximations and Bayesian quadrature," in *IEEE International Workshop on Machine Learning for Signal Processing, MLSP*, vol. 2018-Septe, no. 2. Denmark: IEEE Computer Society, 2018.

[182] M. J. Rawa, D. W. Thomas, and M. Sumner, "Kernel density estimation and its application," in *ITM Web of Conferences, XLVIII Seminar of Applied Mathematics*, vol. 00037. EDP Sciences, 2018, pp. 102–107.

[183] E. Herrmann, T. Gasser, and A. Kneip, "Choice of bandwidth for kernel regression when residuals are correlated," *Biometrika*, vol. 79, pp. 783–795, 1992.

[184] J. Gao and I. Gijbels, "Bandwidth selection in nonparametric kernel testing," *Journal of the American Statistical Association*, vol. 103, pp. 1584–1594, 2008.

[185] U. H. W. A. Hewage, R. Sinha, and M. A. Naeem, "An Efficient and Enhanced Privacy-Preserving Framework to Achieve Optimal Accuracy-Privacy Tradeoff for Evolving Data Streams," *Transactions on Knowledge Discovery from Data*, In review.

[186] A. Bifet and R. Gavaldà, "Adaptive learning from evolving data streams," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5772 LCNS, pp. 249–260, 2009.

[187] U. H. W. A. Hewage, R. Sinha, and M. A. Naeem, "Privacy Preserving Data (Stream) Mining Techniques and Their Impact on Data Mining Accuracy - A Systematic Literature Review," *Artificial Intelligence Review*, In review.

[188] D. Agrawal and C. C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms," in *Symposium on Principles of Database Systems*. Association for Computing Machinery, 2001, pp. 247–255.

[189] A. Evfimievski and J. Gehrke, "Limiting Privacy Breaches in Privacy Preserving Data Mining," in *International Conference on Management of Data and Symposium on Principles Database and Systems*. New York, United States: Association for Computing Machinery, 2003, pp. 211–222.

[190] S. Virupaksha and V. Dondeti, "Subspace based noise addition for privacy preserved data mining on high dimensional continuous data," *Journal of Ambient Intelligence and Humanized Computing*, no. 0123456789, 2020. [Online]. Available: https://doi.org/10.1007/s12652-020-01881-8

Appendix A

Prelude - Manuscript 5

Privacy gives an idea about how much protection the data has achieved after the perturbation. On the other hand, privacy tells how robust a privacy preservation method is. Measuring the privacy of perturbed data is a vital yet complex task. Existing privacy measuring methods assume that the attacker has prior knowledge about the perturbation method or/and the dataset. In this scenario, privacy is measured after performing different types of attacks on the original dataset. However, the attacker equipped with prior knowledge may not be true all the time. Therefore, a method/measurement to represent privacy without performing attacks is helpful.

Manuscript 5 investigates the use of noise variance as an attack independent measure to represent privacy. We experimented with this method on Logistic Cumulative Noise Addition (SRW) [127] which was proposed in Manuscript 5. The total variance of the cumulative noise added throughout a cycle is measured using Area Under the Curve (AUC) of the Logistic curve. Though the AUC cannot capture the changes in the dataset, it is more effective than using only the noise variance. The main reason for this is that calculating AUC considers total noise variance and other parameters such as noise addition rate and maximum noise level. It is not easy to measure exact privacy values in this method, like in the attack-based method. Nevertheless, experiments show that the

AUC of the logistic curve can detect increasing and decreasing trends of privacy, which is similar to attack-based methods.

Appendix B

Utilizing Noise as an Attack Independent Measure for Representing Privacy in Logistic Cumulative Noise Addition (Manuscript 5)

B.1 Introduction

Privacy-Preserving Data Mining (PPDM) performs data mining tasks without directly accessing the original data values[138]. This is achieved by converting the original dataset to another form that hides the original data's actual values, providing privacy for the original data. This process is called data perturbation [127]. The objective of perturbation methods is to increase data privacy while maintaining data mining tasks' accuracy [138, 188].

Measuring privacy is the most critical task after applying a perturbation to original data. Privacy can be defined in many ways, considering the environment it has been used. A more generic definition for privacy is proposed by [13], which is "the degree of

uncertainty according to which original private data can be inferred." Most currently using privacy measuring metrics assume that some background knowledge of the original data is known to the attacker. We think that this assumption is overrated as it is not always possible for an attacker to have some knowledge of original data. An attacker can be someone entirely new to that specific set of perturbed data who does not know about the original data. In that case, it is helpful to have a method to get an idea about the privacy of the dataset without performing attacks on the perturbed dataset.

Perturbation methods that use different techniques have been discussed in the literature. This includes additive noise [134], multiplicative noise [33], random rotation [34, 39], random projection [139], and few more [76, 37]. We narrowed down our concern only to the noise addition-based methods as those can be adopted to static datasets and as well as data streams.

Additive noise generates random noise values from a Gaussian/Uniform distribution and adds to each record independently [84], while multiplicative noise multiplies each record from a randomly generated noise value [33]. The above two traditional noise addition methods improve privacy but decrease the accuracy of data mining results when the noise variance is high. To overcome this, the authors of [32] have introduced an advanced method called cumulative noise addition. Cumulative noise adds the noise values with a small variance to each record and every subsequent record. Research work [127] has improved the cumulative noise addition by combining different techniques to control the maximum noise level, making the method suitable for data streams. SRW was also proposed in [127] as an improved cumulative noise addition method.

Let us investigate attacking/data reconstruction methods that measure privacy; most of them are based on some background knowledge of the original data. These methods use some kind of attack to breach the privacy of perturbed data and attempt to recover the original data. Then measure the breach probability to represent privacy. For example, Known Input/output attack [139], distribution attack [135], Independent Component

Analysis (ICA) based attacks [76, 139], Distance inference attacks [76] and MAP attacks [138] can be considered. Reference [189], which proposes a method to set limits on privacy breaches, is the only method we could find to represent the privacy of noise addition-based methods that do not depend on any knowledge of the original data.

The impact of the perturbation method on privacy and noise addition has been discussed in many works. Privacy provided by the noise has a direct relationship with the total noise variance added to the dataset because when we increase the noise variance, the privacy of the dataset also increases. In [139], data owners specify a noise constraint S where random noise up to S should be added to ensure privacy is preserved. The signal-to-noise ratio (SNR) has been used in [190] in the data perturbation context to achieve optimal data utility while preserving privacy. Authors have defined SNR as the variance of original data over the variance of noise which is again a metric of privacy based on the noise. The authors of [138] have experimentally proved that their data reconstruction method works quite well with small noise variance. It is harder to recover data when the noise variance is high. All these works imply that the level of privacy directly correlates with the variance of the noise added.

Traditionally, the privacy provided by a perturbation method has been measured using noise variance. However, later research works such as [189] and [188] argue that noise variance alone is not an adequate indicator of privacy. Privacy also depends on the original data distribution and other parameters of the perturbation method. We agree that noise variance is not sufficient to measure privacy when the distribution of data changes. However, the fact that noise variance has a considerable impact on privacy also cannot be ignored. Suppose we assume that the data distribution is consistent, which is valid for most of the databases/traditional datasets. In that case, noise variance significantly impacts the level of privacy.

We propose an attack-independent method to represent privacy, considering the

properties of the perturbation method. To achieve this task, we use the Logistic Cumulative Noise addition (SRW) [127], the most recent development of noise additive-based perturbation methods. This work is significant due to two reasons. The first reason is that performing attacks on data based on background knowledge is not always valid, as the attacker can be someone new to the data. The second reason is that the noise variance alone cannot accurately represent privacy, as other factors affect that. Therefore, this work provides a novel approach to representing privacy by capturing other properties of the perturbation method together with the noise variance. This attack-independent approach does not require any background knowledge of the original data.

The remainder of this paper has been organized as follows. Section B.2 provides an overview of the proposed methodology to represent privacy. Experiments, Results and Discussion are explained in Section B.3, and Section B.4 outlines the conclusions and future directions.

B.2 Proposed Methodology

This work aims to propose an attack independent method which is not based on the background knowledge of original data or its distribution to represent privacy using Logistic Cumulative Noise Addition (SRW) [187]. Suppose we can capture other characteristics of the perturbation method together with the total noise variance added to the data. In that case, it is sufficient to give an idea about the expected privacy level assuming data distribution does not change over time. To achieve this, we used the concept of Area Under the Curve (AUC) in the context of SRW. It captures not only the total noise variance added but also other behaviours of the perturbation method.

B.2.1 Logistic Cumulative Noise Addition (SRW)

The SRW is a cumulative noise addition method combined with cycle-wise noise to control the maximum noise level added to data [127]. The dataset is virtually divided into cycles, defined by the cycle size, and noise is added in cycles. The variance of the noise added is decided using applying the logistic function to each cycle. From this method, we have control over the noise addition rate (k) and the maximum noise level (L). The logistic function can be defined in Equation B.1, and Fig. B.1 represents the logistic curve.

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}} \tag{B.1}$$

In SRW, noise variance changes throughout the cycle. This behaviour provides more privacy than using a constant noise variance throughout the perturbation process. We also can change k appropriately, and it decides how fast the curve reaches the maximum level. The final noise variance produced in each independent step is *(f(x) ×σ)* where σ is a small noise variance value used to control the noise level. All these behaviours and parameter values should be considered when proposing an efficient method to represent privacy.

B.2.2 Representing Privacy Using AUC

Area Under the Curve (AUC) is an interesting concept that has been used in different areas such as medicine and signal processing. It has been used to measure the total amount of drug exposure as a function of time and used to distinguish the total noise from the signal. Therefore, AUC can be used to model the total amount of some parameter as a function of another parameter. This concept can be adapted to the SRW environment since the area under the logistic curve allows to measure the total amount of noise variance added as a function of data records. Moreover, it indirectly captures

the noise addition rate, the maximum nose level, and the behaviour of the logistic curve, making the concept of AUC is ideal for representing all the aspects of the SRW perturbation method

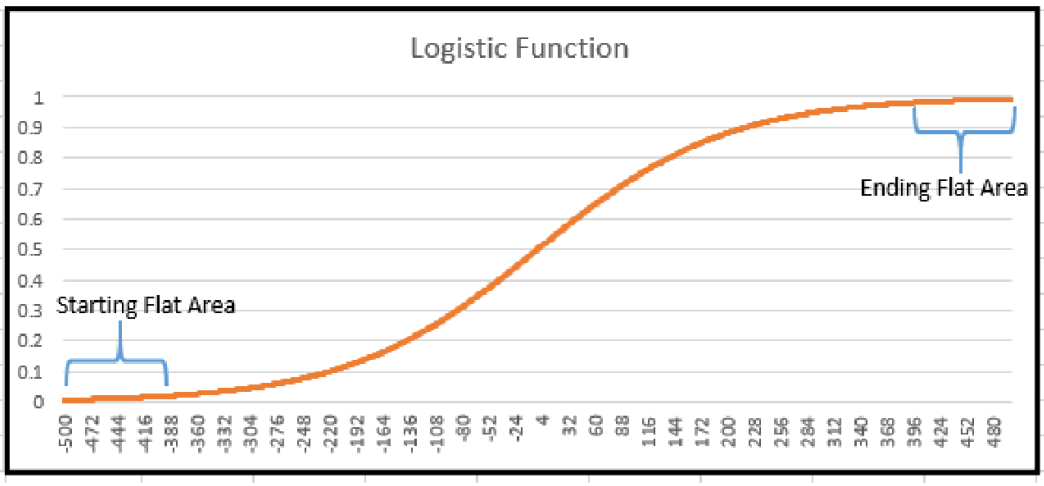

Figure B.1: Logistic Curve

Calculating the AUC of the logistic curve is straightforward. We can integrate the logistic function from lower bound (lb) of cycle size (cs) to upper bound (ub) of cycle size to calculate AUC of logistic curve (if cs = a, then lb = -a and ub = a).

$$\int_{ub}^{lb} f(x) = x + \frac{1}{k}\ln(1 + e^{-kx}) + c$$

;Where c is some constant

$$auc = f(x)_{ub} - f(x)_{lb} \tag{B.2}$$

Equation B.2 calculates the total noise variance from the logistic curve. However, we should co-operate σ to get the total overall noise variance as we generate the noise from a Gaussian distribution with mean zero and variance of ($f(x) \times \sigma$).

$$F(x) = \frac{1}{1 + e^{-kx}} \times \sigma^2$$

$$F(x) = \frac{\sigma^2}{1 + e^{-kx}} \tag{B.3}$$

Integral of $F(x)$ is:

$$\int_{ub}^{lb} F(x) = \sigma^2 x + \frac{\sigma^2}{k} \ln(1 + e^{-kx}) + c \tag{B.4}$$

And using (2), we can re-write (4) as,

$$AUC = [f(x)_{ub} - f(x)_{lb}]$$

$$AUC = auc \times \sigma^2 \tag{B.5}$$

According to the above proof, the total amount of noise variance added can be measured by multiplying the AUC of the logistic curve with additional noise variance. However, as we add noise cumulatively, AUC should be considered cumulatively. That means every data record should be added to the AUC of all the subsequent data records.

$$Total_AUC = A_1 + (A_1 + A_2) + \ldots + (A_1 + A_2 + \ldots + A_i) + (A_1 + A_2 + \ldots + A_i + \ldots + A_n) \tag{B.6}$$

; A_i - AUC up to the data record i

Equation B.6 gives the total noise variance added within one cycle of SRW and indirectly captures other properties and behaviours of the perturbation method. With the support of the above proof, we can define privacy in general for all the noise addition-based perturbation methods.

Definition (Privacy):***The percentage of protection applied to data concerning the total amount of noise variance added, given the noise addition rate (k) in a noise addition-based environment.***

B.3 Experiments

We conducted experiments for different k values of the logistic function and calculated the privacy using AUC. Finally, we normalized all the AUC values to bring them into the same range for ease of understanding. Additionally, we calculated the Breach Probability (BP) to measure privacy and compared it with AUC privacy values to see if there is any relationship. BP was calculated using MAP attacks [138]. We conducted the perturbation experiments for two datasets (AReM from UCI and Electricity from OpenML).

We recorded the total noise variance added to the dataset by calculating AUC for 24 different noise addition rates (k) and four different cycle sizes of the logistic curve. This range of k values was selected according to the details provided in [127], maintaining the ideal shape of the logistic curve to get the maximum privacy benefits. Calculated AUC values for the AReM dataset have been displayed in Table B.1. (Note that AUC values for the Electricity dataset also showed a similar trend.)

Looking at the AUC calculated for all the k values, we can see a similar trend for all four-cycle sizes. When k increases, AUC decreases, indicating that the total noise added to the data also decreases. This behaviour is expected and can be explained using the shape of the logistic curve. When k increases, starting and ending flat areas of the logistic curve also becomes lengthier (See Fig. B.1). That means the period we add the noise in its lowest variance (closer to zero) also increases, reducing the total noise variance added to the dataset. This implies that when the total noise added to data is low, privacy provided by the perturbation method is also low. Fig. B.2 displays the graphs comparing AUC and BP for cycle sizes 1000 and 4000 for AReM and cycle sizes 4000 and 8000 for electricity datasets. Min-max normalized values of both measures have been used for understandability.

Overall, we do not observe any strong relationship between AUC and BP for both

Table B.1: Behaviour of AUC for Different Cycle Sizes

	AUC for Different Cycle sizes			
k	***1000***	***2000***	***4000***	***8000***
0.005	333.683	1112.863	4084.76	15945.275
0.009	284.715	1029.27	3994.933	15855.369
0.013	266.519	1008.43	3974.033	15834.469
0.017	258.768	1000.443	3966.046	15826.482
0.021	254.915	996.565	3962.168	15822.604
0.025	252.747	994.395	3959.998	15820.434
0.029	251.411	993.059	3958.662	15819.098
0.033	250.531	992.179	3957.782	15818.218
0.037	249.92	991.568	3957.171	15817.607
0.041	249.48	991.127	3956.73	15817.166
0.045	249.151	990.799	3956.402	15816.838
0.049	248.9	990.547	3956.15	15816.586
0.053	248.703	990.351	3955.954	15816.39
0.057	248.546	990.194	3955.797	15816.233
0.061	248.419	990.067	3955.67	15816.106
0.065	248.315	989.963	3955.566	15816.002
0.069	248.228	989.876	3955.479	15815.915
0.073	248.156	989.803	3955.406	15815.842
0.077	248.094	989.742	3955.345	15815.781
0.081	248.041	989.689	3955.292	15815.728
0.085	247.995	989.643	3955.246	15815.682
0.089	247.956	989.604	3955.207	15815.643
0.093	247.921	989.569	3955.172	15815.608
0.097	247.891	989.539	3955.142	15815.578

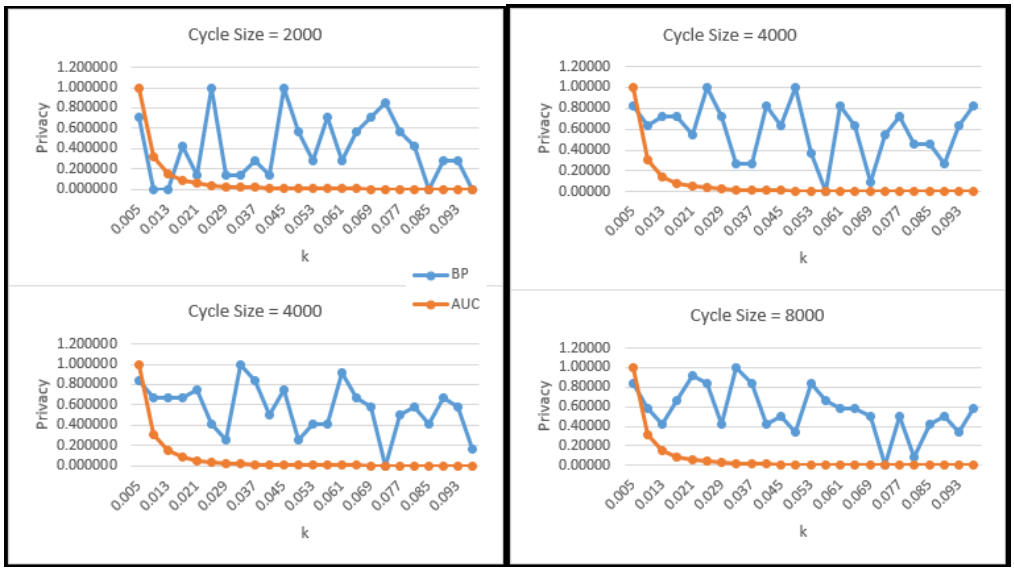

Figure B.2: AUC and BP Comparison - AReM(Left) and Electricity(Right) Dataset

datasets. BP fluctuates throughout all the k values, while AUC shows a decreasing trend. Trying to recover original data records from perturbed records using known I/O pairs and measuring the success rate of recovery when calculating BP is the primary reason for this. Hence, it considers the properties of the original data and assumes that the attacker knows some of the original data records and their perturbed counterparts. Nevertheless, AUC gives a privacy measurement depending on the perturbation method, assuming that the attacker does not know about the original data. If we carefully observe the behaviour of BP curves of both datasets, we can see a slightly decreasing trend though there are fluctuations throughout. This is a good sign indicating that the properties of the perturbation method (calculated using AUC) successfully capture the privacy trends without considering the properties of the original data. In summary, the results of the experiments show that the total noise variance calculated using the AUC of the logistic curve has a direct effect on privacy. This can be further clarified by the fact that the behaviour of AUC retrieved from the experiments can be justified by the behaviour of the perturbation method using the logistic curve. Additionally,

experiments do not display a strong relationship between BP and AUC. But a slightly similar pattern can be inferred.

B.4 Conclusions and Future Directions

In conclusion, we find that the total noise variance calculated using AUC is a justifiable measure to represent the privacy of SRW. It can be considered an attack-independent method to represent privacy. This measure also captures other properties of the perturbation method, such as noise addition rate and maximum noise variance allowed. The use of total noise variance as a measure of privacy can be extended to other noise addition-based methods, such as additive noise and multiplicative noise, effectively if it is possible to capture the properties of the perturbation method. The proposed approach can be considered valid as it shows the same trend as privacy using BP. BP assumes the availability of background knowledge, while AUC only considers the properties of the perturbation method. Incorporating the generic properties of the dataset with the proposed privacy measure is a possible future avenue, as it is essential when original data distribution changes.

Appendix C

Details of Datasets

Table C.1: Details on different datasets used in the thesis

Name & Source	Description	No.of Records	Features	Target Variable	Data Stream?
		Datasets with no known concept drift			
AReM from UCI	Activity Recognition system based on Multisensor data fusion (Real world)	35,999	6 (Numeric)	Activity (Walking, Cycling, Standing, Sitting, lying)	Yes
Electricity from OpenML	Collected from Australian New South Wales electricity market (Real world)	45,312	8 (Numeric)	Change of the price (Up, Down)	Yes
Taxi from Taxi	New York City taxi trip durations (Real world)	50,000	7 (Numeric)	Trip duration (Short, Medium, Long)	Yes
		Datasets with known concept drift			
SEA from MOA	Consists of 3 abrupt drift points at 25000, 50000, 75000 (Synthetic)	100,000	3 (Numeric)	Group A, Group B	Yes
RBF from MOA	Radial Basis Function dataset consist of continuous fast drift of 0.001 change speed (Synthetic)	50,000	10 (Numeric)	class 1, class 2, class 3, class 4, class 5	Yes

Appendix D

List of Acronyms

Table D.1: Glossary

AEL	Average Expected Loss
APOF	Accuracy Privacy Optimising Framework
ARF	Adaptive Random Forest
AUC	Area Under the Curve
BP	Breach Probability
CS	Cycle Size
DT	Decision Tree
DWT	Discrete Wavelet Transformation
HAT	Hoeffding Adaptive Tree
HT	Hoeffding Tree
I/O	Input/Output
ICA	Independent Component Analysis
KNN	K-Nearest Neighbour
LA	Linear Absolute without Resetting
LAR	Linear Absolute with Resetting
LRW	Linear Random Walk without Resetting
LRWR	Linear Random Walk with Resetting
MAP	Maximum A Posteriori attack
MOA	Massive Online Analysis
NB	Naive Bayes
NMF	Non-negative Matrix Factorization
PAM	Privacy Accuracy Magnitude
PCA	Principal Component Analysis
PPDM	Privacy Preserving Data Mining
PPDSM	Privacy Preserving Data Stream Mining
RP	Random Projection
RPCN	Random Projection-based Cumulative Noise
RQ	Research question
SA	Logistic Absolute without Resetting
SAR	Logistic Absolute with Resetting
SBS	Sequential Backward Selection
SDC	Statistical Disclosure Control
SLR	Systematic Literature Review
SMC	Secure Multiparty Computation
SRW	Logistic Random Walk without Resetting
SRWR	Logistic Random Walk with Resetting
SVD	Singular Value Decomposition
SVM	Support Vector Machines
VFDT	Very Fast Decision Tree
WS	Window Size