PRIVACY-PRESERVING MACHINE LEARNING AND DATA SHARING IN HEALTHCARE APPLICATIONS

Doctoral Dissertation by
Amin Aminifar

Thesis submitted for
the degree of Philosophiae Doctor (PhD)
in
Computer Science:
Software Engineering, Sensor Networks and Engineering Computing

Department of Computer Science,
Electrical Engineering and Mathematical Sciences

Faculty of Engineering and Science

Western Norway University of Applied Sciences

Winter, 2022

Series of dissertation submitted to
the Faculty of Engineering and Science,
Western Norway University of Applied Sciences.

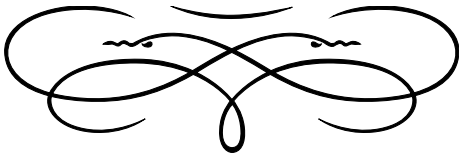
ISBN: 978-82-93677-96-3

Author: Amin Aminifar
Title: Privacy-Preserving Machine Learning and Data Sharing in Healthcare Applications

Printed production: The Communication Division /
Western Norway University of Applied Sciences

Bergen, Norway, 2022

TO MY FAMILY



PREFACE

The author of this thesis has been employed as a Ph.D. research fellow at the Department of Computer science, Electrical engineering and Mathematical sciences at Western Norway University of Applied Sciences, Norway.

The research presented in this thesis has been accomplished as part of the INTROMAT (INtroducing personalized TReatment Of Mental health problems using Adaptive Technology) project and in cooperation with the University of Bergen, Helse Bergen, and the University of Oslo, Norway.

The author of this thesis is enrolled in Computer Science: Software Engineering, Sensor Networks and Engineering Computing program.

This thesis is organized in two parts. Part I is an overview that provides an introduction to the research area of privacy-preserving machine learning and data publishing, particularly in the context of healthcare. This part also includes a discussion on the contributions of this thesis. Part II consists of a collection of published peer-reviewed research papers.

Paper A Aminifar, Amin, Yngve Lamo, Ka I Pun, and Fazle Rabbi. "A Practical Methodology for Anonymization of Structured Health Data." In Proceedings of the 17th Scandinavian Conference on Health Informatics (SHI), Linköping University Electronic Press, pp. 127-133. 2019.

Paper B Aminifar, Amin, Fazle Rabbi, Ka I Pun, and Yngve Lamo. "Diversity-Aware Anonymization for Structured Health Data." In Proceedings of the 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 2148-2154. 2021.

Paper C Aminifar, Amin, Fazle Rabbi, Ka I Pun, and Yngve Lamo. "Privacy Preserving Distributed Extremely Randomized Trees." In Proceedings of the 36th Annual ACM Symposium on Applied Computing, pp. 1102-1105. 2021.

Paper D Aminifar, Amin, Fazle Rabbi, and Yngve Lamo. "Scalable Privacy-Preserving Distributed Extremely Randomized Trees for Structured Data With Multiple Colluding Parties." In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2655-2659. 2021.

Paper E Aminifar, Amin, Fazle Rabbi, Ka I Pun, and Yngve Lamo. "Monitoring Motor Activity Data for Detecting Patients' Depression Using Data Augmentation and Privacy-Preserving Distributed Learning." In Proceedings of the 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 2163-2169. 2021.

Paper F Aminifar, Amin, Matin Shokri, Fazle Rabbi, Ka I Pun, and Yngve Lamo. "Extremely Randomized Trees with Privacy Preservation for Distributed Structured Health Data." In IEEE Access, vol. 10, pp. 6010-6027. 2022.

ACKNOWLEDGMENTS

My Ph.D. education and this thesis would not be feasible without the assistance of certain people. I would like to acknowledge the cooperation and support that I received during my academic life.

Firstly, I would like to acknowledge my supervisors, Professor Yngve Lamo, Associate Professor Violet Ka I Pun, and Associate Professor Fazle Rabbi. Thank you, Yngve, Violet, and Fazle, for all the discussions, collaboration, and the opportunity to study for a Ph.D. at the Western Norway University of Applied Sciences.

I would also like to thank my Ph.D. program coordinator and other university officials at Western Norway University of Applied Sciences. Professor Håvard Helstrup always supported me during my Ph.D. with his guidance and follow-ups in my cases. Thank you very much Håvard!

During my Ph.D., we had many meetings with members of the INTROMAT project. I would like to acknowledge the members of the INTROMAT project and the meetings that we had together. In particular, I would like to thank Dr. Enrique Garcia Ceja and Associate Professor Michael Riegler for sharing material and information about their published research articles.

I would also like to thank my laboratory members at Western Norway University of Applied Sciences for the friendly discussions and time we had together. It was nice to know all of you, my peer students at laboratory/university. Moreover, I would like to acknowledge those who provide me with practical information and guidance during my doctoral studies.

I am glad to have great friends from the years I spent in universities until my Ph.D. period. I would appreciate them for the great time that we spent together and the good memories that we share and wish them the best in their lives. In particular, I would like to thank Matin Shokri for our collaboration on the last paper in this thesis.

Last but not least, I would like to give my deepest gratitude to my family, which I would not be able to do my education without them. Education was one of the main principles in our family, and I had exceptional support during my doctoral studies. Thank you very much!

ABSTRACT

Artificial intelligence (AI) and automated decision-making have the potential to improve accuracy and efficiency in healthcare applications. In particular, AI is proved to outperform human experts in certain domains. However, the application of AI and machine learning for automated decision-making in healthcare comes with challenges, such as security and privacy preservation. Such issues are among the primary concerns that must be addressed as they may negatively affect individuals. For instance, a patient's privacy is violated if sharing his/her medical data with a third-party data recipient reveals that he/she had a medical condition. Furthermore, particular guidelines, e.g., General Data Protection Regulation (GDPR), are proposed to legally protect the privacy of patients that has to be observed while employing AI and machine learning in this domain.

In order to address such privacy concerns, in this thesis, we consider two principal directions for the analysis of data and concentrate our research on them. In one primary direction, the analysis is performed on the published/shared data. Therefore, the data holder needs to consider particular measures to protect the privacy of data subjects, for instance, by perturbing the data before publishing. In this thesis, along this direction, we propose an anonymization framework, formulated as an optimization problem, for datasets with both categorical and numerical attributes. The proposed framework is based on clustering the data samples by considering the diversity issue in anonymization to reduce the risks of identity and attribute linkage attacks. Our method achieves anonymity by formulating and solving this problem as a constrained optimization problem, by jointly considering the k-anonymity, l-diversity, and t-closeness privacy models. We evaluate our framework on popular publicly available structured healthcare data.

The other primary direction is to perform analysis without publishing the data. In such settings, we consider multiple parties, each of which holds a different part of the data. The objective is to analyze the data held on these parties without direct access to the data record values. In this thesis, along this direction, we present a scalable privacy-preserving distributed learning framework based on the Extremely Randomized Trees (ERT) algorithm and Secure Multiparty Computation (SMC) techniques. We build a machine learning model based on the entire dataset by analyzing the data locally at each party and combining the results of this analysis. We evaluate the distributed implementation of our technique based on healthcare datasets collected in the INTROMAT project and demonstrate its prediction performance.

In summary, the research in this thesis contributes to the possibility of exploiting health data in the healthcare setting for analysis and automatic decision-making without privacy violation. This has a long-term potential for better decision-making in the healthcare context, diagnosis, and treatment, at an affordable cost.

SAMMENDRAG

Kunstig intelligens (AI) og automatiserte beslutningsprosesser har potensial til å forbedre nøyaktigheten og effektiviteten i helsetjenesten. Spesielt har AI vist seg å kunne utkonkurrere menneskelige eksperter på visse områder. Imidlertid har bruken av AI og maskinlæring for automatisert beslutningstaking i helsevesenet visse utfordringer, som for eksempel sikkerhet og personvern. Slike spørsmål er blant de viktigste som må tas opp, da de kan påvirke enkeltpersoner på en negativ måte. For eksempel blir pasienters personvern krenket hvis deres medisinske data med en tredjepart viser at hadde en en spesiell medisinsk tilstand. Videre det retningslinjer, for eksempel Personvernforordningen (GDPR), for beskytte personvernet til pasienter som overvåkes ved bruk av AI og maskinlæring på dette området.

For å slike personvernhensyn tar vi i denne tesen for oss to hovedretninger for konsentrerer vår forskning om disse. I hovedretning utføres analysen på de publiserte/delte helseopplysningene. Derfor må databehandleren vurdere spesielle tiltak for å beskytte datasubjektenes personvern, for eksempel ved å forandre dataene før de publiseres. I denne tesen, foreslår vi et anonymiseringsrammeverk, formulert som et optimaliseringsproblem, for datasett med både kategoriske og numeriske attributter. Det foreslåtte rammeverket er basert på gruppering av dataprøver ved å vurdere mangfoldsproblemet i anonymisering for å redusere risikoen for identitets- og attributt-koblingsangrep. Vår metode oppnår anonymitet ved å formulere og løse dette problemet som et begrenset optimaliseringsproblem, ved vurdere personvernmodellene k-anonymity, l-diversity og t-closeness. Vi evaluerer rammeverket for populære, offentlig tilgjengelige strukturerte helsedata.

Den andre hovedretningen er å utføre dataanalyse uten å publisere helseopplysningene. I slike miljøer vurderer vi flere parter, som hver har en forskjellig deler av opplysningene. Målet er å analysere opplysningene fra disse partene uten direkte tilgang til dataregistreringsverdiene. I denne tesen, presenterer vi et skalerbart rammeverk for distribuert læring av personvern basert på teknikkene Extremely Randomized Trees (ERT) algoritmen og Secure Multiparty Computation (SMC). Vi bygger en maskinlæringsmodell basert på hele datasettet ved å analysere data lokalt og kombinere. Vi evaluerer den distribuerte implementeringen av teknikken vår og demonstrerer ytelsen til teknikken.

Oppsummert bidrar forskningen i denne tesen til å kunne utnytte helseopplysninger for dataanalyse og automatisk beslutningstaking uten brudd på personvernet. Dette vil, på lang sikt, bedre beslutningstaking innen helsesektoren, diagnostikk og behandling, til en rimelig pris.

Contents

Part I

OVERVIEW

CHAPTER 1

INTRODUCTION

1.1 Research Context

We have been witnessing a significant increase in healthcare expenditure over the past few decades. For instance, the healthcare costs in the United States of America had a rise from 27.2 billion dollars (only 5 percent of GDP) in 1960 to 4124 billion dollars (19.7 percent of GDP) in 2020 [18, 92]. In addition to the economic costs for society and government, health problems impose financial burdens on individual patients and their family members.

The costs for medication and treatment are among the direct health expenditures, yet health problems can also impose indirect costs due to loss of jobs, replacement of employees on sick leave, and reduced productivity from a medical condition [51]. Take a patient suffering from depression as an example. Aside from the treatment expenses, he/she may experience reduced productivity due to his/her condition. In addition to the socioeconomic burden, health problems also affect the patients' quality of life.

Healthcare subjects related to the brain and mental health are among the primary domains that should be addressed. This is because one in every four people develops one or more mental or behavioral disorders at some stage in his/her life [152]. Mental health disorders are the largest contributor to chronic conditions in Europe [5]. In addition, neuropsychiatric disorders are the first reason for years lived with disability in Europe, and subsequent to cardiovascular diseases and cancer, the third leading cause of disability-adjusted life years in Europe [5]. Human Brain Project [4] and INTROMAT [13] are among the European projects that address healthcare challenges related to the brain and mental health problems. In particular, INTROMAT is a research and development project in Norway that employs adaptive technology for confronting issues in mental health.

Healthcare technologies have substantial contributions to increasing the quality of life and reducing socioeconomic burdens. Early diagnosis of the disease assists medical experts in fighting and controlling it easier rather than facing the disease when it is no longer curable. For instance, a young woman whose breast cancer is diagnosed in the early stages of the disease using the new medical image processing methods has a higher chance of being cured. Moreover, the costs would be much lower since treating cancer in the advanced stages is considerably harder and more expensive, which may not always lead to the patient's full recovery.

Across approaches and technologies related to the healthcare domain, artificial

intelligence (AI) and automated decision-making have recently attracted a lot of attention. Such techniques have the potential to improve accuracy and efficiency, including in the healthcare and medical domain, and can be utilized jointly with other healthcare technologies, e.g., smart wearable devices. AI is proven to outperform human experts in several application domains. For instance, AlphaGo, based on state-of-the-art algorithms, i.e., Deep Neural Networks (DNN) and Reinforcement Learning (RL), defeated the human European champion in the game of Go [167]. AlphaGo was the first computer program that could defeat a human professional player in a full-sized game of Go.

In healthcare, for instance, the application of DNN for the classification of rhythms in electrocardiography (ECG) signals has been proposed in [91]. Hannun et al. used 91,232 single-lead ECGs, collected from 53,877 patients, to classify ECGs into 12 rhythm classes. The classification results show that DNN outperforms cardiologists considering the F1-score and sensitivity metrics. This study shows that employing an end-to-end deep learning approach to classify arrhythmia can lead to high diagnostic performance and similar to the performance of human experts (cardiologists). The utilization of such approaches can reduce the ratio for misdiagnosed ECGs and assist domain experts in diagnosis.

AI is a valuable asset in decision-making in the healthcare and mental health domain as well as in smart wearable devices for long-term and personalized monitoring of patients. However, there are several challenges involved in the adoption of AI in the healthcare domain and wearable and mobile health technologies, most importantly the privacy of personal medical data, which is the topic of this thesis.

In connection to the privacy challenges involved in the adoption of AI in the healthcare domain, in Europe, the General Data Protection Regulation (GDPR) [22] is enforced to protect individuals' privacy. By adopting GDPR, restrictions concerning the utilization of personal data are imposed. For instance, one principle is *purpose limitation*, i.e., for processing the data, a company or an organization must have a specific purpose and must inform individuals about that purpose before collecting their personal data. Another principle in GDPR is *data minimization*, which is to collect merely the necessary personal data that is required for that purpose. The other principle is *storage limitation*, meaning that the personal data can only be stored while the storage is necessary for that purpose. Moreover, according to GDPR, the data subject has the right to ask the data controller for the erasure of his/her personal data without delay.

In the United States, the administrative sections also propose privacy frameworks in order to protect individuals' privacy. For instance, in [94] which is a White House report, the authors introduce several principles to ensure consumers' privacy. The following are several examples of such principles: 1) Individual Control: it gives the right to consumers to have control over what data can be collected and how it can be used. 2) Respect for Context: it gives the right to consumers to expect that companies merely collect and use their personal data in a consistent context with what consumers provide their data in. 3) Focused Collection: it gives the right to consumers to limit the amount of data that companies can collect and retain.

These regulations posed in Europe and the United States show the importance of data privacy. The restrictions posed by administrative sections must be taken into consideration while employing technology. In the healthcare domain, particularly,

we must observe such regulations and carefully consider them due to the sensitivity of the domain and health information. This thesis considers the problem of privacy-preserving data analysis in the healthcare domain.

Examples of Privacy Violation Here, three notable examples of privacy violation after sharing the data is provided.

Example 1: In [174], the author provides an example in which the governor of Massachusetts was reidentified through linking two datasets. Thus, the private information of the subject was revealed. The following explains how the subject was reidentified.

The Group Insurance Commission (GIC), which is responsible for purchasing health insurance for state employees in Massachusetts, collected patients' data with nearly 100 attributes. The collection of patients' data was according to the recommendation from the National Association of Health Data Organizations (NAHDO). This dataset contains attributes that NAHDO recommended. The left circle in Figure 1.1 shows several attributes that this data includes.

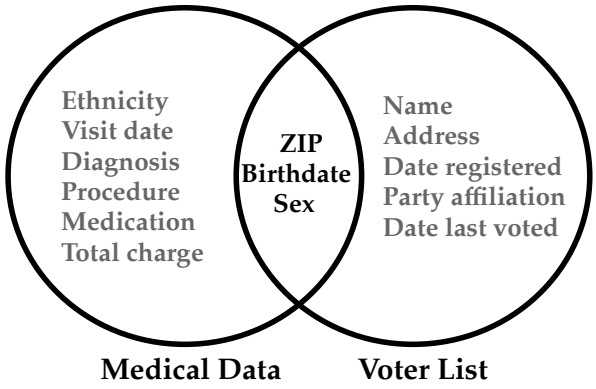


Fig. 1.1: Linking for reidentification [174]

GIC believed that the data was anonymized. Thus, they shared the collected data with researchers and sold a copy of it to the industry. On the other hand, another dataset called voter list was accessible in return for a small sum of money. The data was the voter registration list for Cambridge, Massachusetts, and included address, ZIP code, name, gender, and birthdate of data subjects. The right circle in Figure 1.1 shows several attributes that this data includes.

Having both of these datasets, the adversaries could link (match) them and reidentify particular records. This is because the dataset contained similar attributes, i.e., ZIP code, birth date, and gender. Therefore, by matching the common attributes, one could link the records from the voter list dataset to records from the medical dataset. For instance, Massachusetts' governor of the time, William Weld, was on both datasets. One could match his record information from the voter list dataset to one record from the medical dataset and infer about the attributes that only existed on medical data and not in the voter list dataset, e.g., ethnicity, diagnosis, medication.

Example 2: In another case from Australia, many individuals' privacy was shown to be prone to be violated after their data was published online [16, 21, 24, 27, 30, 63]. In August 2016, the federal Department of Health published the records of medical

bills of 10% of Australians. This was due to the policy of open government data by the Australian government. The data included all publicly reimbursed medical and pharmaceutical bills during 1984-2014. The ID information of suppliers and patients was encrypted, and because of this, one could identify which bills belong to the same patient or supplier.

First, the authors in [63], were able to decrypt the IDs of suppliers. Second, the authors showed that records for patients in this dataset could be reidentified without decryption. This is simply done by matching the known information about the patients with unencrypted parts of the records in the published dataset. The authors demonstrate that de-identification can easily fail.

Example 3: In another case in [148], the authors propose a class of statistical deanonymization attacks. High-dimensional data, such as account owners' preferences or records of transactions, are subject to such attacks. The presented technique in this study tolerates data perturbation and imprecision of the adversary's background knowledge. Furthermore, this approach is applicable even when a subset of the original data is published.

Finally, the authors in [148], practiced their proposed method for deanonymization to the Netflix Prize dataset. Netflix [19] is among the most prominent movie rental websites. This dataset includes the records for movie rating of 500,000 Netflix subscribers, which was anonymized before being published. The authors show that one with limited knowledge about a subscriber can identify his/her record in the published dataset. The adversary's background knowledge about subscribers is obtained through another online service provider, i.e., Internet Movie Database (IMDb) [12], in this study.

1.2 Problem Outline

The application of AI and its subset, machine learning, for automated decision-making in healthcare comes with challenges, such as security and privacy. Such issues are among the primary concerns that must be addressed as they can negatively affect individuals. For instance, patients' privacy is violated if sharing their data with others reveals that they or their family suffer from a medical condition. Therefore, new regulations, such as GDPR [22], are being introduced to deal with such issues.

In the case of this thesis, which is funded by the INTROducing Mental health through Adaptive Technology (INTROMAT) project, we are dealing with data that is related to the healthcare domain. The INTROMAT project's goal is to use technology to improve public mental health. Mental illness is an important problem that accounts for more than 20% of the years lived with disability worldwide [183]. Digital technology has been shown to assist in the prevention and treatment of mental health disorders [106, 176]. INTROMAT aims to facilitate effective e-mental health interventions through an interdisciplinary research team. To this end, several forms of data are available in the INTROMAT project that could be used to facilitate effective e-mental health interventions. Such data can be analyzed, particularly utilizing machine learning techniques, and the results could be used in e-mental health intervention services.

Such mental-health-related data may be collected and held in one center, e.g., a hospital, or it can be collected and held by several parties, e.g., multiple hospitals or patients' mobile phones. We are required to analyze such data to extract useful

information from it. For instance, in one case, the domain experts in INTROMAT were interested in learning a classification model for the prediction of depression in individuals based on the motor activity data collected by wearable devices. In another case in INTROMAT, the domain experts were interested in using patients' interaction data with the treatment system developed for the internet intervention of mental health issues to evaluate their progress and possibly adapt their tasks on the treatment system based on their interaction data. In such cases, and generally, for any mental-health-related data, the privacy preservation of data subjects is essential.

Thus, in the context of this research, we deal with confidential health data that is supposed to be utilized for data analysis purposes. The data may be collected and held in several centers and may not be shared with other centers due to privacy and legal concerns. In this thesis, we develop solutions for analyzing such data that address our privacy concerns.

1.3 Research Questions and Design

As discussed earlier, this research is funded by the INTROMAT project, which addresses mental-health issues with innovative technology. In this project, we require to analyze healthcare data, particularly by machine learning algorithms, to provide benefits for individuals based on the analysis results. However, analyzing such sensitive data comes with challenges.

The main focus of this research project is the study of privacy for data analysis and machine learning for extracting patterns from data in the healthcare context. We study the privacy issues in the application of machine learning for the objectives like exploiting new patterns from the data. This thesis addresses such issues and approaches the research with the questions presented in the following.

After performing the literature review and identifying the research questions, we focused on addressing them. Based on the requirements, e.g., how the data is stored, we proposed and developed our solutions. We evaluated our developed solution based on healthcare data and updated our solution according to the results. Finally, we stated our findings in scientific articles. The research design is discussed in more detail in Chapter 2.

The following are the questions for our research that are addressed in this thesis:
Research Question 1: What are the challenges for data analysis, particularly based on machine learning techniques and in the healthcare domain?

- What are the problems related to privacy of patients?
- What are the existing solutions for privacy concerns in performing machine learning for the analysis of healthcare data?
- What other criteria should be taken into consideration in such solutions?

Research Question 2: Can privacy-preserving data publishing methods serve as a solution for addressing privacy concerns in healthcare data analysis when the data is stored in one center?

- How can we use an anonymization solution for addressing privacy requirements in the analysis of health data?
- What are the shortcomings of such techniques?
- What are the alternative solutions for addressing such shortcomings?

Research Question 3: How can we address the privacy challenges of data analysis in the healthcare domain when the data is distributed among several parties?

- What are the criteria to be considered while proposing and developing such solutions?
- How can we develop a privacy-preserving machine learning solution while considering such criteria?
- What are the characteristics and limitations of such solutions?

1.4 List of Papers

The studies directly address the topic in this thesis are listed as follows:

Paper A: Aminifar, Amin, Yngve Lamo, Ka I Pun, and Fazle Rabbi. "A Practical Methodology for Anonymization of Structured Health Data." In Proceedings of the 17th Scandinavian Conference on Health Informatics (SHI), Linköping University Electronic Press, pp. 127-133. 2019.

Paper B: Aminifar, Amin, Fazle Rabbi, Ka I Pun, and Yngve Lamo. "Diversity-Aware Anonymization for Structured Health Data." In Proceedings of the 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 2148-2154. 2021.

Paper C: Aminifar, Amin, Fazle Rabbi, Ka I Pun, and Yngve Lamo. "Privacy Preserving Distributed Extremely Randomized Trees." In Proceedings of the 36th Annual ACM Symposium on Applied Computing, pp. 1102-1105. 2021.

Paper D: Aminifar, Amin, Fazle Rabbi, and Yngve Lamo. "Scalable Privacy-Preserving Distributed Extremely Randomized Trees for Structured Data With Multiple Colluding Parties." In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2655-2659. 2021.

Paper E: Aminifar, Amin, Fazle Rabbi, Ka I Pun, and Yngve Lamo. "Monitoring Motor Activity Data for Detecting Patients' Depression Using Data Augmentation and Privacy-Preserving Distributed Learning." In Proceedings of the 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 2163-2169. 2021.

Paper F: Aminifar, Amin, Matin Shokri, Fazle Rabbi, Ka I Pun, and Yngve Lamo. "Extremely Randomized Trees with Privacy Preservation for Distributed Structured Health Data." In IEEE Access, vol. 10, pp. 6010-6027. 2022.

The following publication is related to the work done during this Ph.D. project, but not directly covered in this thesis:

- Kumar Mukhiya, Suresh, Amin Aminifar, Fazle Rabbi, Ka I Pun, and Yngve Lamo. "Artificial Intelligence in Mental Health." In Frontiers in Artificial Intelligence: Models, Algorithms and Application Areas, Bentham Science Publishers, pp. 13-34. 2021.

Other publications (previous to this Ph.D. work) related to machine learning, wearable devices, and healthcare applications are available in [165, 169, 170].

1.5 Contributions

As discussed above, accessing the data, particularly healthcare data, for analysis comes with the challenge of privacy and legal concerns. However, to provide value from the available personal data, we need to confront and solve such challenges. In order to address the privacy and legal concerns, in this thesis, we consider two principal directions for the analysis of data.

One primary direction is for when the data is stored in one central location. In this direction, the analysis is performed on the published data. Therefore, the data holder must consider particular measures to protect the privacy of data subjects. For instance, the data holder perturbs the data, based on anonymization methods, before publishing in order to thwart specific attack models in the scenario in which data should be published. Therefore, in such approaches, the data is published, but in a privacy-preserving fashion. These approaches are referred to as privacy-preserving data publishing techniques [84].

The other primary direction is for when the data is distributed among multiple parties. In this direction, the analysis is performed without publishing the data. In this approach, we consider multiple parties, each of which holds a different part of the data. The objective is to analyze the data held on these parties without direct access to the data record values. This can be performed by means of distributed data analysis algorithms combined with Secure Multiparty Computation (SMC) techniques, e.g., for securely computing the result of an operation without revealing the secret values. These approaches belong to the family of privacy-preserving distributed data mining and machine learning [31].

Figure 1.2 shows the concepts related to the whole structure of the thesis and displays the relationship among the presented items. Both research directions, i.e., privacy-preserving data publishing and privacy-preserving distributed machine learning, aim for the analysis of data. The data analysis, in this thesis, usually is performed by employing machine learning algorithms, for instance, for the classification of data records. The proposed methods, in both directions, is considered for healthcare applications where we have data privacy concerns. In Chapter 4, we discuss the details related to the evaluation in terms of privacy and classification performance for both of these directions.

In the anonymization scenario for privacy-preserving data publishing, the data is stored in a centralized location. Then, based on the specified attack and privacy models,

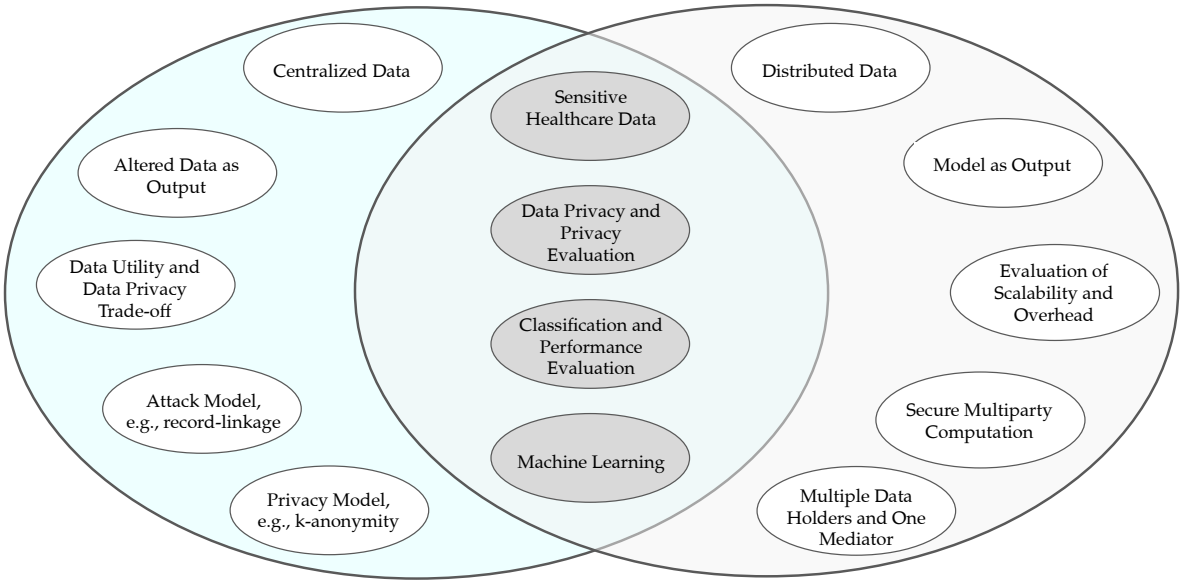


Fig. 1.2: Concepts for the two research directions of this thesis, i.e., privacy-preserving data publishing (left) and privacy-preserving distributed machine learning (right)

the data is altered by the data holder in a way that the altered data preserves utility to its recipients and preserves privacy according to the privacy model. Altering the data improves privacy but decreases data utility, which is due to the trade-off between data utility and privacy. Then, the anonymized data is shared for future data analysis, visualization, etc. Therefore, the final outcome in this process is an anonymized dataset, which can be used in a wide variety of tasks.

In the privacy-preserving distributed machine learning, the training data is stored on several data holder parties, and one central server (mediator) orchestrates the whole learning process. The number of data holder parties can be high in certain scenarios, and the scalability of the privacy-preserving distributed machine learning frameworks should be taken into consideration. The evaluation of the scalability and overhead for the privacy-preserving distributed machine learning is discussed in Chapter 4. In certain methods, The application of secure multiparty computation techniques, which are used for secure computation based on private information held on multiple parties, is necessary for protecting the privacy of patients. By using such an approach, instead of the data, a machine learning model is learned and shared with future users. Therefore, the final outcome is a machine learning model.

Figure 1.3 illustrates the process for the two directions of this thesis. The data collection and data storage phases are gone through before the alteration and analysis of the data in privacy-preserving data publishing and privacy-preserving distributed machine learning scenarios. In the anonymization scenario for privacy-preserving data publishing, based on the attack and privacy models, the data is altered by the data holder, and the anonymized data is shared for future data analysis. Therefore, the final outcome in this process is an anonymized dataset, which can be used in a wide variety of tasks, e.g., training machine learning models for prediction, visualization, etc. In the

privacy-preserving distributed machine learning direction, the data is decentralized and may not be shared due to privacy concerns. By employing a privacy-preserving distributed machine learning algorithm, a machine learning model for a specific task is learned and shared with future users. Therefore, the final outcome is a machine learning model specific to a certain task.

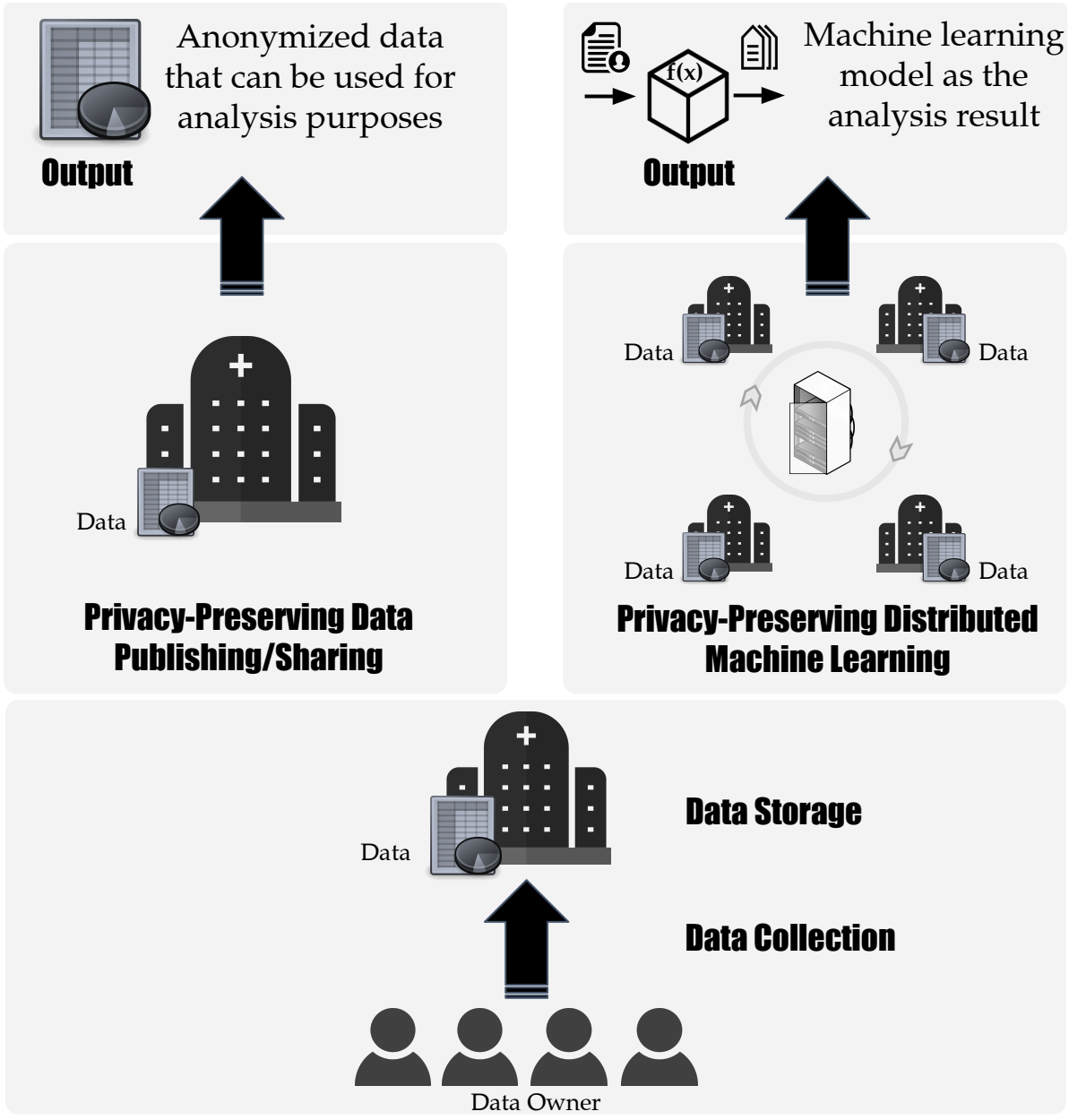


Fig. 1.3: The process for privacy-preserving data publishing (left) and privacy-preserving distributed machine learning (right)

1.5.1 *Privacy-Preserving Data Publishing*

Here, we consider the problem of privacy-preserving data publishing. In privacy-preserving data publishing, in particular anonymization, the input to our method is the raw data, which can include explicit identifiers, quasi-identifiers, and sensitive

and non-sensitive attributes. In addition, the data holder identifies attack models and decides on privacy models that address the attack models. The goal is to generate an anonymized version of data to be published with the data recipient, which is consistent with the privacy models considered. The output of the method is the altered data that is consistent with the considered privacy models. The anonymized data includes an anonymous version of quasi-identifiers, plus sensitive and non-sensitive attributes.

The following publications address the problems in privacy-preserving data publishing:

Paper A: A Practical Methodology for Anonymization of Structured Health Data

> In this paper, we investigate the possibility of adopting cryptographic algorithms in the context of anonymization of structured data. The basic idea is to map the original dataset to a dataset which is subject to less re-identification risks. Adopting cryptographic algorithms can potentially provide privacy preservation by construction. We validate the data utility preservation of the proposed approach based on the Adult dataset [28].

Paper B: Diversity-Aware Anonymization for Structured Health Data

> In this paper, we propose a method for anonymizing and sharing data that addresses the record-linkage and attribute-linkage attack models. Our method achieves anonymity by formulating and solving this problem as a constrained optimization problem, by jointly considering the k-anonymity, l-diversity, and t-closeness privacy models, for the first time to the best of our knowledge. We evaluate the proposed approach and demonstrate its relevance based on the widely-used Heart Disease dataset [66].

1.5.2 Privacy-Preserving Distributed Machine Learning

Here, we consider the problem of privacy-preserving distributed machine learning. In privacy-preserving distributed machine learning, we have the training data distributed over several parties, which is the input to our framework. Our assumption is that the data holders do not share their training data but collaborate in the learning process. The goal is to collaboratively build a machine learning model based on all available data without violation of privacy. In addition to privacy, the learned model's performance and the learning process's overhead should be taken into consideration. The output of our framework is a machine learning model that can be further used for a specific task, e.g., classification.

The following publications address the problems in privacy-preserving distributed machine learning:

Paper C: Privacy Preserving Distributed Extremely Randomized Trees

> In this paper, we consider the classification problem and show how the Extremely Randomized Trees (ERT) algorithm could be adapted for settings where (structured) data is distributed over multiple sources. We propose the Privacy-Preserving Distributed ERT (PPD-ERT) approach for privacy-preserving utilization of the ERT algorithm in a distributed setting. To the best of our knowledge, this is

the first application of the ERT algorithm in the distributed setting with privacy consideration (without sharing the raw data or intermediate training values), without any loss in classification performance.

Paper D: Scalable Privacy-Preserving Distributed Extremely Randomized Trees for Structured Data With Multiple Colluding Parties

In this paper, we extend the distributed Extremely Randomized Trees (ERT) approach with respect to privacy and scalability. First, we extend distributed ERT to be resilient with respect to the number of colluding parties in a scalable fashion. Then, we extend the distributed ERT to show its relevance in settings with limited participation of data holder parties, without any major loss in classification performance. We refer to our proposed approach as k-PPD-ERT or Privacy-Preserving Distributed Extremely Randomized Trees with k colluding parties.

Paper E: Monitoring Motor Activity Data for Detecting Patients' Depression Using Data Augmentation and Privacy-Preserving Distributed Learning

In this paper, we present an approach for extracting classification models for predicting depression based on a new augmentation technique for motor activity data in a privacy-preserving fashion. The augmentation addresses several problems with the dataset, e.g., imbalanced number of recorded days for individuals in the data, and improves the classification performance. In this study, we employ our proposed approach for privacy-preserving distributed machine learning for data collected by wearable devices/sensors to ensure the preservation of the privacy of sensitive information for the patients in the context of depression and mental health disorders.

Paper F: Extremely Randomized Trees with Privacy Preservation for Distributed Structured Health Data

In this work, we build upon our previous work [37] and propose a scalable privacy preserving framework for distributed machine learning based on the extremely randomized trees algorithm, with linear overhead in the number of parties. We use two popular publicly available healthcare datasets for performance evaluation, i.e., the Heart Disease [66] and the Breast Cancer Wisconsin (Diagnostic) [134] datasets. This data represents medical applications where missing values are present, and our algorithm is designed to handle such scenarios. Finally, we present the implementation of our technique over Amazon's AWS cloud and evaluate it in a real-world setting based on the mental health datasets associated with the Norwegian INTROducing Mental health through Adaptive Technology (INTROMAT) project.

1.6 Thesis Structure

There are two parts to this thesis. Part I consists of the problem specification, literature review, research methods, findings, and evaluations, as well as conclusions. Part

II consists of the publications for the research performed during this Ph.D. on the problems and research domains described in this thesis.

Part I of the thesis is divided into five chapters that cover the following subjects:

Chapter 1: Introduction This chapter covers the thesis research context, problem outline, and research questions in the problem domain and research methods. The list of articles, a short review of our contributions, and thesis structure are also provided in Chapter 1.

Chapter 2: Research Context and Design This chapter discusses the research context related to this thesis and describes our research focus and the relation to the presented research questions, research method, and research design.

Chapter 3: Results This chapter covers the proposed methods developed in this Ph.D. project, which are related to our discussed research questions. The technical details of our contributions are discussed in Chapter 3.

Chapter 4: Evaluation and Discussion In this chapter, we provide an overview of thesis contributions. We evaluate our contributions against the research goals and discuss our contributions against the state-of-the-art. The validity threads and reflections on the research context are also discussed in this chapter.

Chapter 5: Conclusion and Future work This chapter provides a summary of the findings and contributions of this thesis. The future trends and directions for the research performed in this thesis are also provided in Chapter 5.

CHAPTER 2

RESEARCH CONTEXT AND DESIGN

2.1 Research Focus and Research Questions

The funding project of this research focuses on the application of technology for improving public mental health. One of the related tasks in this regard, as discussed in the following sections, is the analysis of data in the mental health domain based on machine learning algorithms. In this thesis, we focused on such tasks in the project and introduced several research questions, which were provided in Chapter 1. The following are a short review of research questions in this thesis:

- **Research Question 1:** What are the challenges for data analysis, particularly based on machine learning techniques and in the healthcare domain?
- **Research Question 2:** Can privacy-preserving data publishing methods serve as a solution for addressing privacy concerns in healthcare data analysis when the data is stored in one center?
- **Research Question 3:** How can we address the privacy challenges of data analysis in the healthcare domain when the data is distributed among several parties?

The main challenge for health data analysis that is the concern of this thesis is the privacy of data owners (Research Question 1). In this thesis, we focus on finding and examining solutions for addressing this issue. In order to address such privacy concerns, as mentioned in Chapter 1, we consider two primary research directions for the analysis of health data and concentrate our research on them. The first direction, i.e., privacy-preserving data publishing, is for when the data is stored in a centralized location, and in this direction, the analysis is performed on the published/shared data. The second direction, i.e., privacy-preserving distributed machine learning, is for when the data is distributed among multiple parties, and in this direction, the analysis is performed without publishing the data. The objective is to analyze the data held on these parties without direct access to the data record values. When proposing a solution in each of these two directions, we should consider relevant criteria, e.g., privacy, overhead, and classification performance.

2.1.1 *Privacy-Preserving Data Publishing*

High-quality data is a fundamental requirement for performing analysis and mining for the extraction of knowledge and providing benefits to individuals. The data in its raw format has the highest utility. Therefore, at one extreme, to have the highest data utility, one can share the raw data intact; however, the privacy of individuals is violated if the raw data is shared. Hence, at the other extreme, one can share no data to preserve the privacy of data individuals [64], which provides no data utility. As both utility and privacy of the data are essential, we consider a trade-off between data utility and privacy for sharing the data. Thus, the one who has access to the raw data can alter the data based on this trade-off and share it afterward.

The privacy protection practice relying on controls and legislation for restricting the publishing and usage of data is not always effective. Such policies require either a high level of trust in the recipient of data or extreme distortion in data. However, the adversaries are not expected to follow the rules in policies and guidelines. Moreover, the rules and policies do not guarantee that the data will not accidentally be available to the adversary. For example, in 2011, [15, 17, 20] report that a laptop containing sensitive details of more than eight million patients went missing in a medical research organization in NHS North Central London health authority. The report in [17] mentions that these records include details about mental illness, HIV, abortion, and cancer, and it could be a disastrous tool in the hand of a blackmailer. Several more examples of such cases can be found in [187]. Therefore, developing tools and methods for publishing data in a hostile environment is a necessary requirement. The published data must be practically useful and preserve the privacy of the subjects [84].

Figure 2.1 explains the scenario for privacy-preserving data publishing. In this scenario, data owners are patients or data subjects from whom personal data is collected, and the data and its associated rights belong to them. The data holder collects and holds the raw data from the data owner. Based on the described needs and objectives, the data holder must publish the data with others. Since the data holder publishes the data, it can also be called the data publisher. The receiver of the published data is called the data recipient. The data recipient performs the data mining tasks on the published data [83]. For example, a hospital may collect patients' personal data and publish it to a research organization. In our example, the patients are data owners, the hospital is the data holder or publisher, and the research organization is the data recipient.

In this thesis, for privacy-preserving data publishing, we assume that the data holder (for example hospitals) is trusted by the data owners (for example patients). It means that the data owners are willing to share personal data with the data holder. This trust, however, does not extend to the data recipient. Therefore, the data holder must publish an altered version of data to protect the privacy of data owners.

We consider the following assumptions, which are common in privacy-preserving data publishing [83]:

- *The data holder is not an expert*. In privacy-preserving data publishing scenarios, we do not require the data holder to have knowledge about the data mining tasks. The data recipient performs the tasks for data mining. The data holder might not know the data recipient or the tasks of mining that the data recipient conducts

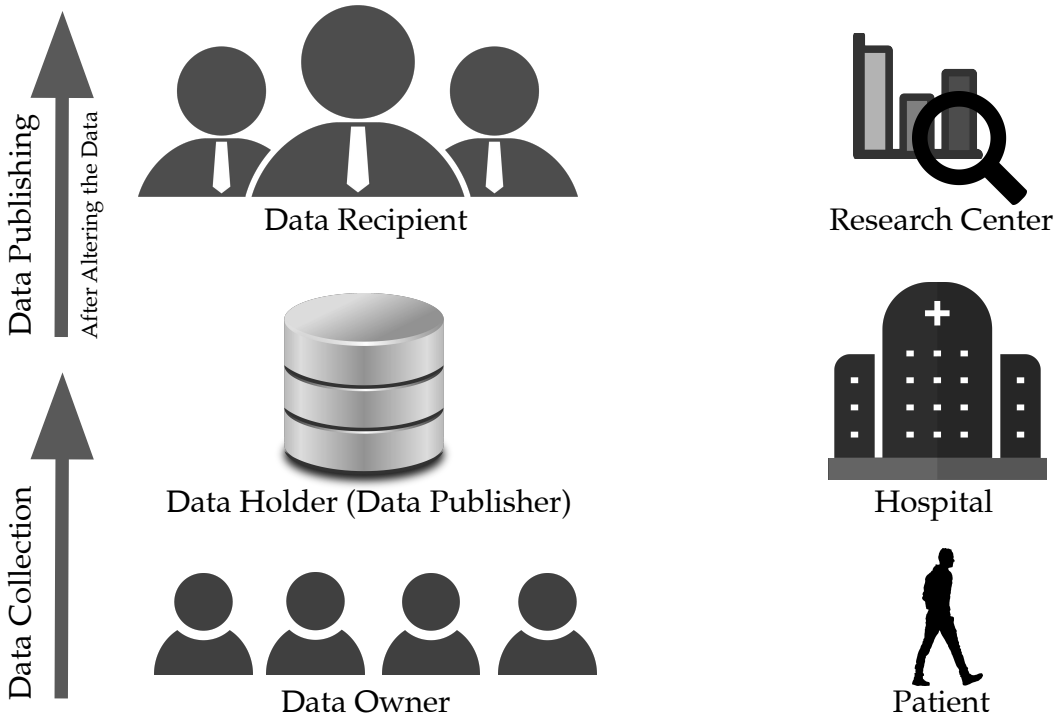


Fig. 2.1: Data collection and publishing procedure (roles on the left and examples on the right)

on data. However, due to the regulations and public advantages of associated research, the data holder should publish the data. In such scenarios, the data holder should merely be able to anonymize the data.

In particular scenarios, the data holder seeks the data mining results while not having the resources to do it. Thus, the data holder outsources the mining task to external resources. In this case, the data mining task is known to the data holder. However, it must still release an altered version of data to the data recipient for analysis.

- The data recipient can be the adversary. As mentioned above, the data owners' trust in the data holder is not transitive to the data recipient. In privacy-preserving data publishing scenarios, we assume that the adversary receives the data published to the data recipient. For instance, suppose that patients' personal data in a hospital is published to a research center for analysis. In general, a research center could be trustworthy, but it is not guaranteed that every employee at the research center who has access to the published data is trustworthy.

- *The data is required to be published, not the data mining results*. In order to obtain more accurate data mining results, the data mining experts require to have the data, not partial or statistical information about the data. The data can provide more utility to the data recipient by providing the opportunity of exploration, e.g., by visualization. Accessing the published data will also give more flexibility to the data recipient in performing the mining tasks. For example, by having the data, the data recipient can try using different algorithms and parameters to train multiple models and select the best possible options. This assumption

is consistent with the first assumption about the limited knowledge of the data holder on data mining.

- *The published records are truthful*. The correctness of the analysis results depends on the truthfulness of the records published by the data holder. If the published records do not correspond to existing data holders in real life, the results will not be practical and valuable. For example, a hospital publishes data about patients' lung cancer. A research center is interested in investigating the connection between smoking and lung cancer in individuals. If the published records to the research center do not actually correspond to a patient, the obtained results by the center are not correct or relevant.

By taking the above restrictions and assumptions into consideration, we develop our solution and investigate if privacy-preserving data publishing methods can serve as a solution for addressing privacy concerns in health data analysis where the data is stored in one center (Research Question 2). Connected to Research Question 2, our research includes considering anonymization solutions for addressing the privacy concerns in the analysis of health data, their shortcomings, and alternative solutions for addressing such shortcomings.

2.1.2 *Privacy-Preserving Distributed Machine Learning*

In many real-world scenarios, the data is inherently stored on multiple sites. For instance, as different patients go to different centers for receiving various medical services, the patients' information can be stored on the servers based at different hospitals or medical centers. In order to have high-quality data mining results, the utilization of the entire available data is required for performing data mining tasks. However, due to the legal and privacy concerns, the centers holding parts of the data cannot share them with other centers for the analysis. One of the main directions in this area is to perform the data mining tasks on the data stored on different sites without violating the privacy of data subjects.

Privacy-preserving distributed machine learning focuses on developing methods and approaches for addressing the privacy concern in this problem. Figure 2.2 explains a scenario in which the data required for learning is distributed over multiple parties. The data holder parties in this scenario are hospitals. Data holder parties cannot share their data with other data holders or a center for performing the learning task due to privacy issues. However, they can communicate with each other or a central node and collaborate in a privacy-preserving manner to perform the task of learning. The central node in the figure is a server that can be based in a center, e.g., a medical center, and in this thesis, is called the mediator. The mediator and data holder parties can communicate on a network, e.g., the Internet, to learn a model, e.g., classification model, based on the held data on all parties and in a privacy-preserving fashion.

The data can be distributed horizontally or vertically. For each situation, several different methods and approaches are proposed [108, 179]. The following explains each of these scenarios:

- Horizontally partitioned data: Different records of data are store on different sites (data holder parties). All attributes of each record are store on one site.

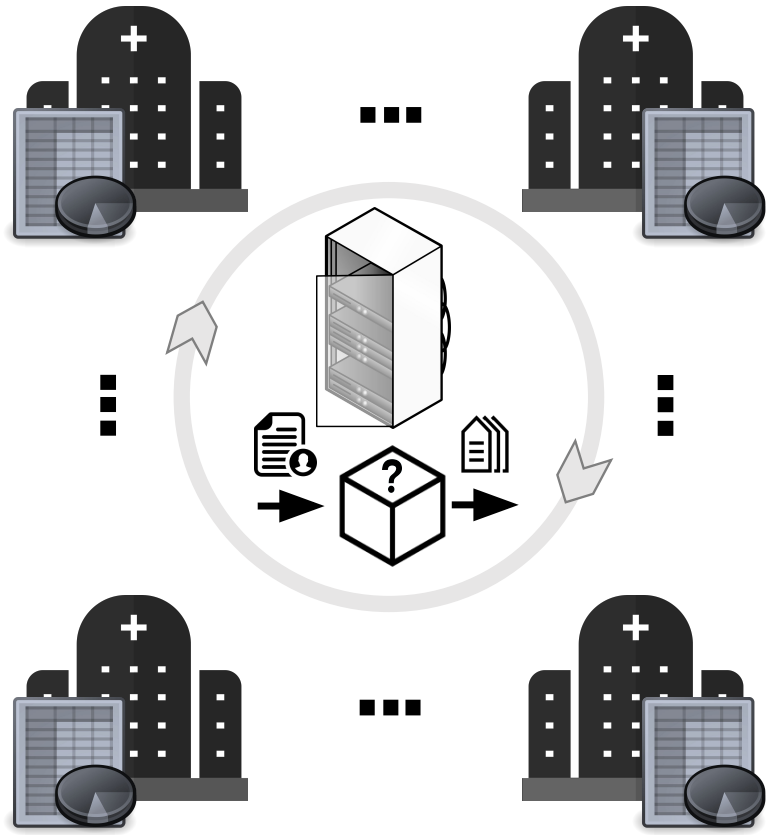


Fig. 2.2: This figure illustrates the setting for our privacy-preserving distributed machine learning [38]. Each data holder, i.e., hospital, holds a portion of the training data, and the mediator coordinates the process and communicates with hospitals to build a machine learning model based on the local data on each party.

- Vertically partitioned data: When the data is vertically partitioned, different attributes of each record may be stored on different sites.

Existing literature on data mining over distributed platforms incorporate approaches based on cryptographic and secure multiparty computing techniques [44, 60, 108, 124, 179]. In the secure computation embodied in such approaches, we are interested in the result of a computation without knowing about the basic values needed for this computation. Therefore, such techniques utilize secure computation to compute a partial result needed for the learning process without revealing the private values required for computation. However, such methods significantly increase communication and computing overhead.

Considerable communication and computing overhead make such approaches inefficient and impractical for many real-world scenarios, where we have large-scale data or limited communication and computing features, e.g., in mobile phones or resource-limited wearable devices [79, 157, 169, 170]. Several state-of-the-art solutions, such as [75, 114, 136], aim to address learning in distributed settings in terms of reducing communication and computational overheads. This is because the complexity and scalability of the approach, along with the quality of data mining results and privacy, are among the three primary metrics for evaluating privacy-preserving data mining algorithms [48].

By taking the above-mentioned requirements into consideration, in this thesis, we propose a framework for privacy-preserving distributed machine learning. In particular, we investigate if we can address the privacy challenges of data analysis in the healthcare domain, where the data is distributed among several parties employing such approaches (Research Question 3). This includes proposing privacy-preserving distributed machine learning solutions for addressing privacy requirements in the analysis of health data while considering their characteristics and limitations.

2.2 Research Methods

The research method followed in this project includes the following steps: identifying the problem domain, reviewing the literature, identifying the gap, developing a solution, and analyzing the proposed solution and updating it based on the analysis results. As discussed in Chapter 1, in this research project, *the problem* is to analyze health data that is considered sensitive, particularly using machine learning algorithms, while addressing privacy concerns. We also introduced the research questions, which are in accordance with our problem and requirements.

We *reviewed the literature* and explored the related works to address the identified problems and research questions. We found two primary research directions that address the problems specified in this research. Each of these two directions has its own advantages and disadvantages in different settings and cannot be disregarded or be considered as the ultimate solution. Then, we reviewed the literature for each area to learn about the related methods and concepts and identify the challenges and limitations in each direction.

We carried out *theoretical and experimental analyses* of our proposed solutions. Based on the reviewed approaches and backgrounds in each research direction, we proposed and developed our solutions to the problems outlined in this thesis. We analyzed our solution according to the criteria designated in each research area for the evaluation of an approach, including privacy, overhead, and classification performance. We also performed experiments to evaluate our proposed frameworks by using public health datasets. We use the results of our evaluations as feedback for improving our frameworks.

Finally, we also carried out *case studies* based on health datasets for our developed solutions. For privacy-preserving data publishing, we used the heart disease dataset [66], which is among the most popular datasets at the UCI data repository website [71], for anonymization and evaluation of data privacy and utility of our framework [39]. For privacy-preserving distributed machine learning direction, we implemented our framework on Amazon's AWS cloud platform [41]. We used several health-related datasets for experimental evaluation. In particular, we used the heart disease [66] and breast cancer [134] datasets which are among the ten most popular at the UCI repository, and Depresjon (depression in Norwegian) [85] and Psykose [101] datasets from the INTROMAT project.

Based on [115], the motivation for our research is to be of service to society as the results of our works will be used for patients and other individuals in the society. Our research is applied as it aims to find immediate solutions for the problems we have in the healthcare section for the analysis of private data, as we observe in the

INTROMAT project. The majority of our research can be considered analytical as it involves developing new ideas on the basis of reasoning. In these studies, we provide the logic for why and how our proposed framework works. For instance, we developed our optimization-based anonymization and privacy-preserving distributed machine learning frameworks and illustrated the reason why such frameworks should work. However, in several parts of our research, our work relies on experiments and can be considered experimental. In these studies, we mostly rely on the experimental results rather than the logic explaining the reason why our approach should work. For instance, for our proposed augmentation technique for improving the classification performance, we discuss the underlying logic of the algorithm, but we rely on the experimental results to confirm that adopting our approach improves the classification performance.

2.3 Research Design

In this section, we discuss the research design for this thesis. Figure 2.3 shows an overview of our research design.

This research is part of the INTROducing Mental health through Adaptive Technology (INTROMAT) project. The vision in the INTROMAT project is to improve public mental health by adopting innovative and adaptive technologies. Mental illness is a growing concern that accounts for more than one-fifth of the years lived with disability worldwide, higher than other categories of illnesses [183]. Digital technology has been shown to help prevent and treat mental health issues [106, 176]. INTROMAT intends to facilitate effective e-mental health interventions by an interdisciplinary research team.

In the INTROMAT project, several types of data were available that could be used for the discussed ultimate goals in the project, i.e., facilitating effective e-mental health interventions and improving public mental health. Such data could be analyzed, particularly by using machine learning techniques, and the analysis results can be used in the services provided for e-mental health interventions. For instance, in one of the cases, building machine learning models for classifying motor activity signals of depressed and normal individuals was desired in the project [85]. Similarly, classification of motor activity signals collected by wearable devices into two groups of patients diagnosed with schizophrenia and control group using machine learning techniques was desired in the project [101]. In another case, the domain experts were interested in using patients' interaction data with the treatment system developed in the project to evaluate their progress and possibly adapt their tasks on the treatment system based on their interaction data.

Several sets of such data are made open for research and educational purposes, e.g., [25, 26]. However, due to the privacy of the patients and the present regulations for protecting the right of data owners, e.g., GDPR, the access to many other available data in the INTROMAT project for research and development purposes is a challenge [151]. Therefore, we require solutions for analyzing such data while considering privacy concerns. Research Question 1, which addresses the challenges for data analysis based on machine learning techniques in the healthcare domain, is introduced based on such requirements in the INTROMAT project.

Based on the identified requirements and problems, we overviewed the available

Fig. 2.3: Research design

solutions and found two main and closely connected research directions addressing such problems. In one of the directions, the focus is on altering the raw data and publishing it for the analysis, e.g., by using machine learning algorithms. This

direction is referred to as privacy-preserving data publishing in this thesis. In the other direction, the analysis, particularly by using machine learning and data mining algorithms, is performed without publishing the data and when that raw data is stored in a decentralized fashion. This thesis refers to this direction as privacy-preserving distributed machine learning.

In the privacy-preserving data publishing direction, we studied the privacy concerns discussed in this area and the proposed solutions for addressing them. We reviewed the literature related to this research direction and presented research questions to be addressed in this thesis. In particular, we focused on anonymization and several well-known attacks and privacy models. Then, we developed an anonymization solution based on these attacks and privacy models. We evaluated our methods based on publicly available datasets, and in accordance with our evaluation, we worked on improving our anonymization approach. Finally, we published the results of our research in peer-reviewed conferences.

For the privacy-preserving distributed machine learning direction, we studied the literature to find out about the challenges and state-of-the-art solutions for addressing them. We introduce the respective research question in this area to be discussed in this thesis. In particular, we focused on the classification problem, similar to the problems discussed in [85, 101]. We developed our solution based on the state-of-the-art extremely randomized trees algorithm and a secure multiparty computation layer to protect patients' privacy. To evaluate our methods, on the one hand, we described their logic and reason for why these methods work in their corresponding studies. We analyzed our framework, and the analysis results caused us to work to improve it. On the other hand, we used several publicly available healthcare datasets, including the datasets associated with the INTROMAT project, i.e., the Depresjon and Psykose datasets [85, 101], as the benchmarks for evaluating our methods. Moreover, we implemented a proof of concept on Amazon's AWS cloud platform. Finally, the results of our research studies were evaluated and published in several peer-reviewed conferences and one peer-reviewed journal.

CHAPTER 3

RESULTS

As discussed in previous chapters, we made our contribution in two research directions. The first direction is for scenarios in which the data is stored in a centralized location. In this direction, we alter and publish a version of data that preserves privacy, and the analysis is performed on the published data. The other direction is for when the data is distributed among multiple parties. In this direction, we build a machine learning model based on such data while preserving the privacy of data owners. The background and technical details for our contributions are provided in this chapter.

3.1 Privacy-Preserving Data Publishing

In this section, connected to Research Question 2, we consider the problem of privacy-preserving data publishing. In privacy-preserving data publishing, in particular anonymization, the input to our method is the raw data, which can include explicit identifiers, quasi-identifiers, and sensitive and non-sensitive attributes. In addition, the data holder identifies attack models and decides on privacy models that address the attack models. The goal is to generate an anonymized version of data to be published with the data recipient, which is consistent with the privacy models considered. The output of the method is the altered data that is consistent with the considered privacy models. The anonymized data includes an anonymous version of quasi-identifiers, plus sensitive and non-sensitive attributes. The criteria to be considered for privacy-preserving data publishing solutions, as questioned in Research Question 1, are data privacy and data utility which need to be preserved. In this research project, Papers A and B are related to Research Question 2, which is for when the data is stored in one center.

Anonymization methods alter the data to avoid identifying data subjects [99]. Previous studies propose privacy models for anonymization, e.g., k-anonymity [174], l-diversity [131], t-closeness [118], LKC-privacy [140]. The data holder selects a model for anonymization based on the scenario and data utility and privacy requirements. Several methods have been proposed to comply with such privacy models and avoid the associated attacks, i.e., record-linkage and attribute-linkage attacks, e.g., using genetic algorithm, kd-trees algorithm, etc. for achieving anonymity [47, 100, 117].

Our anonymization framework is based on k-anonymity, l-diversity, and t-closeness privacy models which address the record-linkage and attribute-linkage attack models. In the following section, these attack models and privacy models are reviewed using

an example.

3.1.1 *Privacy Models*

In this section, we briefly review the record-linkage and attribute-linkage attack models. Then, we discuss three popular privacy models addressing such attacks, namely, *k*-anonymity, *l*-diversity, and *t*-closeness.

In the *record-linkage* and *attribute-linkage* attack models, we suppose that a version of data after removing the identifier attributes of patients, e.g., name and address, is shared with a data recipient. At the same time, the adversary has access to the data shared with the data recipient. This data contains several attributes through which a patient (record owner) can be identified, i.e., quasi-identifiers, and it is assumed that the adversary has the exact value of these attributes for the victim patient. Finally, there is a sensitive attribute in the data, e.g., HIV, that the adversary is interested in knowing about.

To explain this attack models, we use Tables 3.1a and 3.1b as an example. The 2nd-4th columns are considered as quasi-identifiers and refer to age, the number of children, and the smoking state of the patient (*Yes*/*No*). The 5th column is a sensitive attribute capturing the state of the HIV disease for the patient (*Positive*/*Negative*). Table 3.1a represents shared data after removing the identifier features. Suppose that Table 3.1a is shared with the data recipient. If the adversary knows that the victim is 37 years old, has two children, and smokes, he/she can easily match his/her information to one of the records (record one in Table 3.1a) and identify that the victim is diagnosed with HIV. The record-linkage attack occurs by matching the adversary's information (quasi-identifiers) with published data for identifying the patient's (record owner) sensitive information [84].

The *k-anonymity* privacy model was proposed to address the record-linkage attack model. A dataset is *k*-anonymous when the values of quasi-identifiers for each record are the same as the values for at least *k-1* other records in the data. In this way, the adversary can only match his/her information with at least *k* records. Table 3.1b shows a 3-anonymous version of the same data in Table 3.1a. For instance, in our example in Table 3.1b, if the adversary knows that the victim is 37, has two children, and smokes, he/she can merely match his/her information with a qid group containing the records of three patients, records *1*-*3*.

While the *k*-anonymity model guarantees that a patient is only matched with a qid group, however, this model does not guarantee the protection of patients' privacy against attribute-linkage attacks. That is, *k*-anonymity does not consider the diversity of values for the sensitive attribute in each qid group. In this example, in the first qid group, all the values for the sensitive attribute are *Positive*. Therefore, in the first qid group, the adversary can infer that the victim patient is diagnosed with HIV by matching quasi-identifiers' information. The attribute-linkage attack model occurs in situations where the diversity of values for the sensitive attribute is low. As a result, the adversary may infer the sensitive attribute with high confidence.

To address the attribute-linkage attack, the *l-diversity* model proposes that every qid group should have a least *l* distinct values for the sensitive attribute. For instance, in Table 3.1b, if the adversary matches his/her information with the third qid group,

Index	Quasi Identifier			Sensitive
	Age	Number of Children	Smoke	HIV
1	37	2	Yes	Positive
2	36	0	Yes	Positive
3	40	0	Yes	Positive
4	35	3	Yes	Negative
5	32	1	Yes	Negative
6	34	1	Yes	Negative
7	30	2	No	Positive
8	34	2	No	Negative
9	28	1	No	Negative
10	31	1	No	Negative

(a) Original data

Index	Quasi Identifier			Sensitive
	Age	Number of Children	Smoke	HIV
1	[36-40]	[0-2]	Yes	Positive
2	[36-40]	[0-2]	Yes	Positive
3	[36-40]	[0-2]	Yes	Positive
4	[32-35]	[1-3]	Yes	Negative
5	[32-35]	[1-3]	Yes	Negative
6	[32-35]	[1-3]	Yes	Negative
7	[28-34]	[1-2]	No	Positive
8	[28-34]	[1-2]	No	Negative
9	[28-34]	[1-2]	No	Negative
10	[28-34]	[1-2]	No	Negative

(b) 3-anonymous data

Table 3.1: Patient data tables in original and 3-anonymous formats

he/she can not identify that the patient was diagnosed with HIV for sure because both *Negative* and *Positive* values are in that qid group. However, this does not consider the confidence of the adversary's inference properly. For example, if we have both *Negative* and *Positive* values in all qid groups, we have 2-divers data, but if the proportion of *Positive* values in one qid group is high, the adversary can infer that the patient is diagnosed with HIV with high confidence. The entropy *l*-diversity and recursive *(c,l)*-diversity are proposed to address such issues [131].

The *entropy l-diversity* is one of the existing privacy models to address the distribution of values in the sensitive attribute. A data table meeting the following condition for each qid group is entropy *l*-diverse:

$$-\sum_{s \in S} P(qid, s) \log(P(qid, s)) \geqslant \log(l), \tag{3.1}$$

where *S* is the set of values for sensitive attribute, and *P(qid, s)* is the probability/proportion of value *s* for the sensitive attribute in the qid group.

The entropy *l*-diversity still has several limitations. For instance, if the entropy of values for the sensitive attribute in qid groups is high, the *l* will be high. The entropy is highest when the distribution of values is a uniform distribution. Nevertheless, we prefer the minimum probability for the sensitive value (*Positive* in our example) in the qid group. In our example, we favor as few *Positive*s in the qid groups as possible to lower the confidence of inferring HIV positive for the victim patient. Still, entropy *l*-diversity encourages an equal number of *Positive*s and *Negative*s in the qid groups.

The *recursive (c,l)-diversity* model controls the frequency of values for the sensitive attribute in the qid group. In this model, *c* is a constant greater than zero, $c > 0$. The values for the sensitive attribute *S* are: $s_1, s_2, \ldots, s_m$. The number of occurrence for each value (for the sensitive attribute) in the qid group are: $n_1, n_2, \ldots, n_m$. The number of occurrence for values sorted in a decreasing order are: $r_1, r_2, \ldots, r_m$. If a data table meets $r_1 \leqslant c \sum_{i=l}^{m} r_i$ for each qid group, then the data is recursive *(c,l)*-diverse.

The recursive *(c,l)*-diversity can relax the restrictiveness compared to entropy *l*-diversity. When we have a larger *c*, we can have a larger *l*. Therefore, we can relax the restrictiveness by increasing *c*. This privacy model avoids having a high frequency of highly repeated values (in the dataset for sensitive value) in the qid group. It also forces the less frequent values (in the dataset for sensitive value) to be more frequent in the qid group. However, this may not be desirable in certain scenarios. Many healthcare datasets have sensitive attributes with highly imbalanced values. For instance, in a table of data with 1000 records, we may have merely 20 patients diagnosed with HIV. In our example, by increasing the frequency of a sensitive value (with low frequency in the dataset) in a qid group, the adversary can more confidently infer that the patient is diagnosed with HIV.

The *t-closeness* privacy model proposes having a more similar distribution of values in the sensitive attribute among the qid groups and the whole dataset. In the *t*-closeness model, the maximum distance between these two distributions may not be greater than the threshold *t*. For measuring the distance between probabilistic distributions, one possible metric is as follows:

$$D[P, Q] = \sum_{i=1}^{m} |p_i - q_i|, \tag{3.2}$$

where m is the number of values for the sensitive attribute. $P = \{p_1, p_2, \ldots, p_m\}$ and $Q = \{q_1, q_2, \ldots, q_m\}$ are the distributions of sensitive attribute in the entire dataset and in a particular qid group, respectively. This distance metric (variational distance) does not consider the semantic distance between values. In scenarios where the semantic distance of values is important, we may use other distance measures. For instance, if we have three categories like teacher, police, farmer for occupation, categories are generally not related, but the categories for height, e.g., short, middle, and tall, are semantically related and should be treated accordingly.

3.1.2 Our Approach

For privacy-preserving data publishing, we investigate the possibility of adopting cryptographic algorithms in the context of anonymization of structured data. The basic idea is to map the original dataset to a dataset which is subject to less re-identification risks. Adopting cryptographic algorithms can potentially provide privacy preservation by construction.

For the preservation of privacy, we seek a function to map each unique record of raw data to another unique record, which is different from the raw one and in the same feature space. The anonymized data records must be different enough to prevent identity and attribute attacks. The anonymized data should not allow the possibility for the adversary to map back to the raw data. Therefore, the utilized function for mapping the raw data should not be reversible (without access to the private cryptographic key), or in other words, should be one-way, for those with whom the anonymized data will be shared.

Cryptography fulfills the privacy objectives by construction. Mapping a number to another unique number through one-way functions is the main purpose of cryptography. Therefore, by such intrinsic features of cryptographic algorithms, we can make sure of the preservation of privacy criterion without taking further actions. Since, after encryption, the values would be meaningless numbers for the adversary, and it is not possible for one without a key to map back to the raw data.

As described earlier the anonymization methods should fulfill two criteria, namely privacy preservation and data utility. Application of cryptographic algorithms guarantees the privacy preservation criterion by construction. The state-of-the-art cryptographic algorithms are robust against adversarial attempts for revealing the values which are encrypted. However, we also need to make sure about the performance of this methodology in regard to the utility of data. We experimentally show that our proposed methodology for anonymization of structured data is also efficient regarding the data utility.

The utility of the data needs to be preserved and this is related to the correlation of attributes and labels in data samples. To ensure this criterion is satisfied after encryption of the dataset, the utility of the data is compared before and after anonymization based on classification performance. If the results for raw and anonymized data are close, then in addition to the preservation of the privacy, there also would be a confidence

about the utility of data. A loss to a limited extent in the utility of data is acceptable as there exists a trade-off between privacy and data utility in data anonymization [64]. We evaluate this approach based on several state-of-the-art cryptographic algorithms on the Adult dataset [29]. We observe minor degradation in terms of classification performance (less than 3% reduction in geometric mean) after the anonymization process, which shows preservation of data utility.

This approach is particularly suited in the context of categorical data without semantic relation. In addition, we assume that the distribution of data is not known to the adversary. The proposed approach is relevant in the context of learning classification models and is not applicable in other applications, e.g., visualization of data. The proposed methodology can have a complementary role in combination with other methods as well.

To address the shortcomings of the discussed approach, we propose an optimization-based anonymization framework for datasets with both categorical and numerical attributes. The proposed framework is based on clustering the data samples in a diversity-aware fashion to reduce the risks of identity and attribute linkage attacks. Our method achieves anonymity by formulating and solving this problem as a constrained optimization problem, by jointly considering the *k*-anonymity, *l*-diversity, and *t*-closeness privacy models, for the first time to the best of our knowledge. In other words, we propose a method for anonymizing data that ensures each record is indistinguishable from, at least, *k-1* other records in the shared data while taking the diversity and frequency of values in the sensitive attribute into consideration. We evaluate our method based on the utility and privacy of data after anonymization in comparison to the original data.

We formulate the anonymization problem in a constrained optimization framework as a clustering problem, where the diversity and frequency of sensitive values are captured and enforced by constraints. We refer to our proposed method as diversity-aware anonymization, where diversity captures both the diversity concept in the *l*-diversity privacy model and the frequency and distribution of sensitive values in the *t*-closeness privacy model. The experimental results show the preservation of utility of data for classification tasks and the privacy properties noted in the discussed models.

Similar to other anonymization techniques, in our method, if the number of Quasi-Identifier attributes increases, the utility of anonymized data will be negatively affected. This method is relevant for structured healthcare datasets and is not designed for time-series data. Our proposed method is designed for scenarios in which data is stored centrally. An interesting future research topic is extending our method for anonymizing data stored in a distributed fashion.

In the following, we describe our method for addressing the attack models discussed in Section 3.1.1. In our method, we consider the indistinguishability of samples in a qid group, proposed in *k*-anonymity, diversity of values in sensitive attributes in qid group, discussed in *l*-diversity, and frequency of sensitive values in qid group in *t*-closeness.

In this method, we suppose that the values for the sensitive attribute are either sensitive or not. In our example, the *Positive* value shows that the patient (record owner) is diagnosed with HIV and is sensitive, while the value *Negative* if known to the adversary causes no consequence to the patient. Therefore, we consider a binary state for the values in the sensitive attribute and distribute them in the qid groups evenly.

Our method clusters the points in the space of quasi-identifiers and shares the center of each cluster (qid group) as the quasi-identifiers' values for each qid group. Each cluster contains k samples and is clustered based on the distance of instances to the cluster center and the number of samples with sensitive values in each cluster.

We adopt the constrained optimization framework to solve the described clustering problem. The classical clustering techniques, e.g., k-means [132], do not fulfill our requirements. First, we need to introduce the constraints to have k samples in each cluster to ensure the indistinguishability property of the k-anonymity model. Second, we need to introduce a constraint for distributing instances with sensitive values evenly among qid groups (clusters) to ensure diversity in the l-diversity and t-closeness models.

The described anonymization problem is formulated in the Mixed-Integer Linear Programming (MILP) framework, as follows:

$$\min_{B,C} \quad \sum_{i=0}^{n_C} \sum_{j=0}^{n_S} |B_{ij} \cdot (X_j - Center_i)| \tag{3.3}$$

$$\text{s.t.} \quad \sum_{i=0}^{n_C} B_{ij} = 1, \quad \forall j \in \{0, \ldots, n_S\} \tag{3.4}$$

$$\sum_{j=0}^{n_S} B_{ij} = \frac{n_S}{n_C} = k, \quad \forall i \in \{0, \ldots, n_C\} \tag{3.5}$$

$$\frac{\left(\sum_{j=0}^{n_S} B_{ij} \cdot X_j\right)}{k} = C_i, \quad \forall i \in \{0, \ldots, n_C\} \tag{3.6}$$

$$\sum_{j=0}^{n_S} B_{ij} \cdot S_j \leqslant \alpha \cdot \frac{\sum_{j=0}^{n_S} S_j}{n_C}, \quad \forall i \in \{0, \ldots, n_C\}, \tag{3.7}$$

where n_C is the number of clusters (qid groups), and n_S is the number of samples to be anonymized. X_j is the vector of quasi-identifiers' values for sample j. B_{ij} indicates if sample j belongs to cluster (qid group) i and it is a Boolean optimization variable. $Center_i$ is the i-th cluster center calculated by k-means algorithm to be used as an initial solution in our method to reduce the complexity of our optimization problem.

The parameter k is the number of samples in each cluster and is equal to $\frac{n_S}{n_C}$. C_i is the center of cluster i which will be optimized during solving this problem. The values of vector C_i will be shared with data recipients, i.e., instead of raw quasi-identifiers' values for i-th qid group. S_j is a Boolean parameter, $S_j \in \{0, 1\}$, that identifies if sample j has a sensitive value. Finally, α is a parameter that controls the restrictiveness of the constraint, i.e., the higher the value of α, the less the restrictions in solving this optimization problem. This parameter is introduced to be able to tune the restriction with respect to diversity in each qid group.

Let us discuss the proposed formulated optimization problem. The $|B_{ij} \cdot (X_j - Center_i)|$ expression in Equation (3.3) is the Manhattan distance [116] of sample j, X_j, and cluster center i, $Center_i$, when the Boolean variable B_{ij} is equal to one. B_{ij} will be equal to one, $B_{ij} = 1$, if sample X_j belongs to cluster i, and it will be zero otherwise. The objective function in Equation 3.3 intends to optimize B_{ij}s to minimize the distance

between samples in cluster i and $Center_i$, for all clusters and samples.

Equations (3.4)-(3.7) are the constraints of our proposed optimization problem:

- The first constraint, in Equation (3.4), forces each sample to belong to only one cluster. This is done by ensuring that B_{ij} is one exactly once for all i. This is formulated by having the sum of B_{ij} for sample j and all clusters equal to one as the constraint. Since B_{ij} is binary, it will be 1 for one and only one cluster.

- The second constraint, in Equation (3.5), forces the number of samples in each cluster to be equal to k. The summation of the number of samples must be equal to k for cluster i. This condition can readily be relaxed to: at least k samples in each cluster. This is formulated by having the sum of B_{ij} for cluster i and all samples equal to k as the constraint. Since B_{ij} is binary, there will be k samples in cluster i.

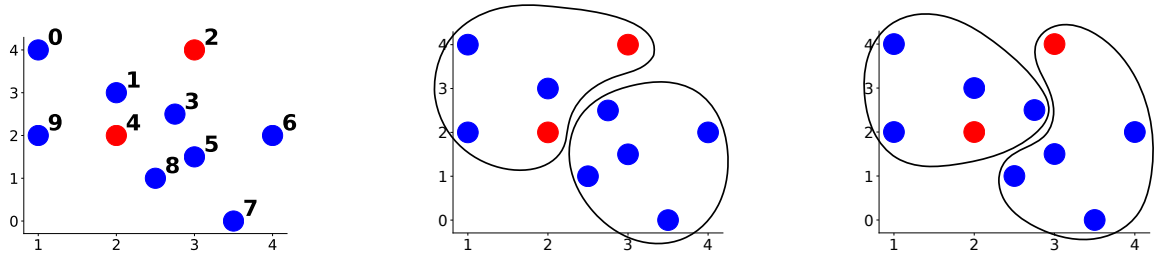
- The third constraint, in Equation (3.6), finds the optimized cluster centers, i.e., C_is. The optimized center for cluster i is the average of all k samples that belong to cluster i. This is formulated by having the sum of all samples, which fall in cluster i (k samples fall in each cluster) divided by k equal to the center of cluster i as the constraint.

- Finally, the last constraint, in Equation (3.7), forces the optimization to distribute the samples with sensitive values ($S_j = 1$) into all clusters. The left-hand side of the constraint is equal to the number of sensitive values in cluster i. The right-hand side is the number of samples with sensitive value divided evenly among the clusters (multiplied by α, which is the parameter for relaxing the hard constraint in our optimization problem). This is formulated by having the number of samples with sensitive value in cluster i to be less than or equal to almost the number of samples with sensitive value in the dataset divided by the number of clusters as the constraint.

After the optimization, we know which sample belongs to which qid group or cluster, based on B matrix. We also know the optimized cluster centers, identified based on the values of C_is. Therefore, the values of sample quasi-identifiers will be replaced by their respective cluster center values. In this way, we obtain a solution that addresses record-linkage and attribute-linkage attack models. We force the samples in the anonymized data to be indistinguishable from *k-1* other samples while considering the diversity of values in the sensitive attribute.

An integer programming problem is an optimization problem in which the variables to be optimized are integers. In case the variables are not limited to integers, our problem is mixed-integer programming. When in the problem, the objective and the constraints are linear, our problem is linear programming [56, 189]. In our optimization problem, the variables B and C are not limited to integers. The problem's objective and constraints are also linear. Therefore, our optimization problem is a mixed-integer linear programming problem and can be solved using solvers supporting MILP problems.

The algorithm for anonymization, which was based on clustering, is described above. The objective and constraints should be created according to the described



(a) Data samples before optimization to find qid groups

(b) Optimization without considering the diversity constraint

(c) Optimization with considering the diversity constraint

Fig. 3.1: Illustrative example for our anonymization method

algorithm and passed to a solver for optimization. CPLEX [11] and GUROBI [10] are two examples of solvers for solving programming problems.

Figure 3.1 presents an example in which the solution in Figure 3.1b merely considers k-anonymity property, while Figure 3.1c considers the diversity of values in the sensitive attribute addressed in l-diversity and t-closeness. The color of the circles shows if the samples contain a sensitive value. If the color is blue, the sample does not have a sensitive value, $S_j = 0$, while a red circle shows having a sensitive value $S_j = 1$.

In Figure 3.1b, samples *0*, *1*, *2*, *4*, *9* fall in the same qid group. The rest of the samples fall in the second group. By sharing the cluster centers for each group, we achieve 5-anonymous data. However, in such a solution, the samples with sensitive values are not evenly distributed. By considering the constraint introduced for the diversity of values in the sensitive attribute, we obtain the solution presented in Figure 3.1c. In this solution, the data is still 5-anonymous, i.e., it has five samples in each cluster. Nevertheless, in this case, sample *2*, falls in the same cluster with *5*, *6*, *7*, *8* to evenly distribute samples with sensitive values.

In Chapter 4, we evaluate our proposed method experimentally based on Heart Disease dataset [66] and consider data utility and data privacy criteria. We discuss the experimental results and the trade-off between data utility and data privacy in anonymization.

3.2 Privacy-Preserving Distributed Machine Learning

The previous section discusses the methods for altering raw data and publishing it. One main drawback of such methods is that they require enough disturbance in data to preserve the privacy of record owners or patients. However, this is not desirable in terms of the utility of the data. In anonymization, particular information that may help the learning algorithms to generate more accurate models can get lost by distortion of raw data.

On the other hand, the majority of anonymization methods are designed for scenarios in which data is collected in one central repository. Nevertheless, in distributed scenarios, we may not be able to send the raw data to a center for the anonymization process. These are among the characteristics and shortcomings of privacy-preserving

data publishing approaches questioned in Research Question 2 that motivate us to also look for alternative solutions for scenarios with requirements exceeding the privacy-preserving data publishing approach capabilities. Therefore, based on the conditions of the problems and our requirements, we need to devise new techniques for privacy-preserving data analysis. Here, we focus on learning from decentralized data without privacy violation.

In this section, connected to Research Question 3, we consider the problem of privacy-preserving distributed machine learning. In privacy-preserving distributed machine learning, we have the training data distributed over several parties, which is the input to our framework. Our assumption is that the data holders do not share their training data but collaborate in the learning process. The goal is to collaboratively build a machine learning model based on all available data without violation of privacy. In addition to privacy, the learned model's performance and the learning process's overhead should be taken into consideration. As questioned in Research Questions 1 and 3, privacy, scalability and overhead, and model's performance are the criteria to be used while proposing and developing a privacy-preserving distributed machine learning framework. The output of our framework is a machine learning model that can be further used for a specific task, e.g., classification. In this research project, Papers C, D, E, and F are related to Research Question 3, which is for when the data is distributed among multiple parties.

Our privacy-preserving distributed machine learning framework is based on the extremely randomized trees algorithm and secure multiparty computation. In the following section, extremely randomized trees algorithm and its characteristics are reviewed, and secure multiparty computation is discussed using an example.

3.2.1 *Underlying Techniques for Our Privacy-Preserving Distributed Machine Learning Framework*

Extremely Randomized Trees (ERT) Algorithm ERT [86] algorithm is a tree-based ensemble learning algorithm that has been widely used for solving classification problems due to its learning performance, robustness to overfitting, and explainability, which are among the characteristics of tree-based ensemble learning algorithms [81, 125, 129]. However, the traditional ERT algorithm operates under the circumstances where the data is stored in a central location. We adapt the ERT algorithm for distributed settings where data is stored and essentially distributed among several parties. In the following we discuss some of the advantages of the ERT algorithm, for its utilization in distributed settings, with respect to other available solutions.

Firstly, since the ERT algorithm is an ensemble learning method, it is robust to tackle overfitting. Ensemble learning methods incorporate weak learners to generate weak classifiers that are independent of other generated classifiers. Therefore, based on Condorcet's jury theorem (1785) [65], the majority vote of this ensemble of learned classifiers predicts better than the vote of an individual classifier, and if we increase the number of classifiers, the accuracy will improve [163]. Therefore, in the ensemble learning method, we generate a collection of classifiers instead of only one, e.g., in distributed ID3 in [75] by Emekçi et al., and finally predict based on the voting result of the learned classifiers. In such ensemble learning methods, randomness parameters

in the learning algorithm cause generating classifiers different from each other. In the ERT algorithm, the randomness of candidate attributes and the splitting point for every decision node in the tree are the randomness parameters as described by Geurts et al. in [86], which result in learning different classifiers. ERT follows the logic of bagging in ensemble learning. Bagging combines the learned classifiers by voting, i.e., it predicts based on the majority vote among learned classifiers. While not increasing the bias, bagging leads to lower variance in our learned model since we are averaging, and lower variance in the learned model reduces the risk of overfitting [81].

Secondly, ERT is tree-based, and tree-based algorithms have been shown to outperform other techniques for the structured data that we are addressing. In [129], the authors report that tree-based learned models usually outperform models learned by standard deep neural networks (e.g., [114, 137]) for tabular-style data. Moreover, in health informatics domain applications, the interpretability of the learned models is advantageous. The patterns that tree-based learned models unveil, particularly in the healthcare domain, can be more useful than the learned model's prediction capability [129]. Tree-based algorithms are more interpretable compared to deep neural networks [125]. This is an advantage for ERT. However, since ERT is an ensemble learning method, and in ensemble methods, instead of learning a one-tree model, e.g., in the ID3 algorithm [159], the algorithm constructs several trees as a model. Hence, this decreases the explainability of such approaches compared to the ID3 algorithm.

Secure Multiparty Computation The studies on the Secure Multiparty Computation (SMC) problem were started by Yao's Millionaires' problem [191]. The Millionaires' problem is a classical problem used for describing the SMC field and is secure two-party computation. In Millionaires' problem, two millionaires want to know which of them is wealthier without revealing the amount of wealth they possess. The secure multiparty computation framework considers the problem of collaborative computation among several parties, each of which hold a secret value; the parties are interested in the result of a computation performed based on their secret values, while they refrain from sharing their secret values with other parties.

The secure multiparty computation can be employed for a wide range of problems, including *healthcare applications* while considering privacy concerns. For instance, the SMC technique can be used for comparing an individual's DNA against a database of DNAs of patients diagnosed with cancer. This comparison can aim to identify if the person whose DNA is being compared against the database is in the high-risk group for a particular type of cancer. On the one hand, DNA information is very sensitive and may not be shared with other institutions or organizations. On the other hand, the benefit from performing this comparison is significant for health and society. Here, the secure multiparty computation can resolve such a dilemma and only return a category of cancer, based on the individual's DNA, without disclosing any information about other DNAs in the database [123]. The same methods for securely comparing, aggregating, multiplying, etc., can be employed in the healthcare domain and its subdomains, e.g., mental health, for various applications.

We provide an example here for explaining secure multiparty computation. Figure 3.2 explains a scenario for an SMC problem. In this problem, as shown in the figure, we have four parties, each holding a secret value. The desired value in this problem is

the result of function F that takes secret values as input and returns a value as output; for instance, the desired value could be the summation of secret values. One simple solution for computing the desired value without sharing secret values with other parties is to share them with a trusted party by everyone. Then, the trusted party can perform the computation and return the result to all parties. Figure 3.3 shows this solution. However, the assumption with respect to the trusted parties is not feasible in many scenarios, so such solutions are not practical. Therefore, based on the type of the computation and the scenarios, we need to devise other solutions to perform the desired collaborative computation in a secure way and without violating privacy.

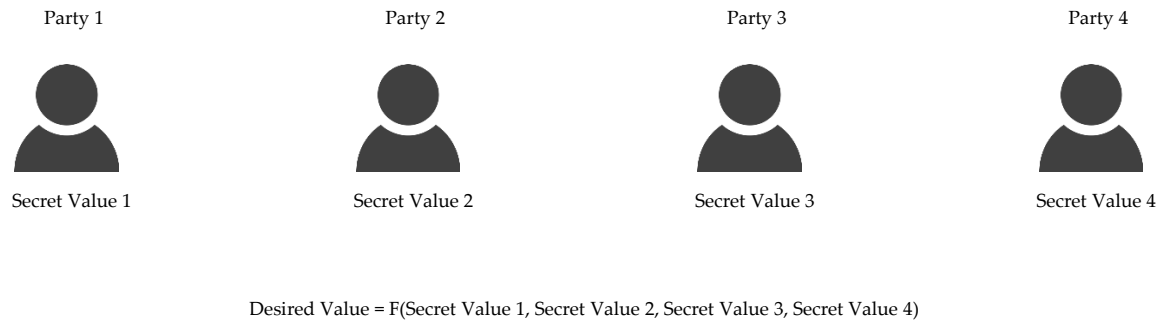


Fig. 3.2: Secure Multiparty Computation

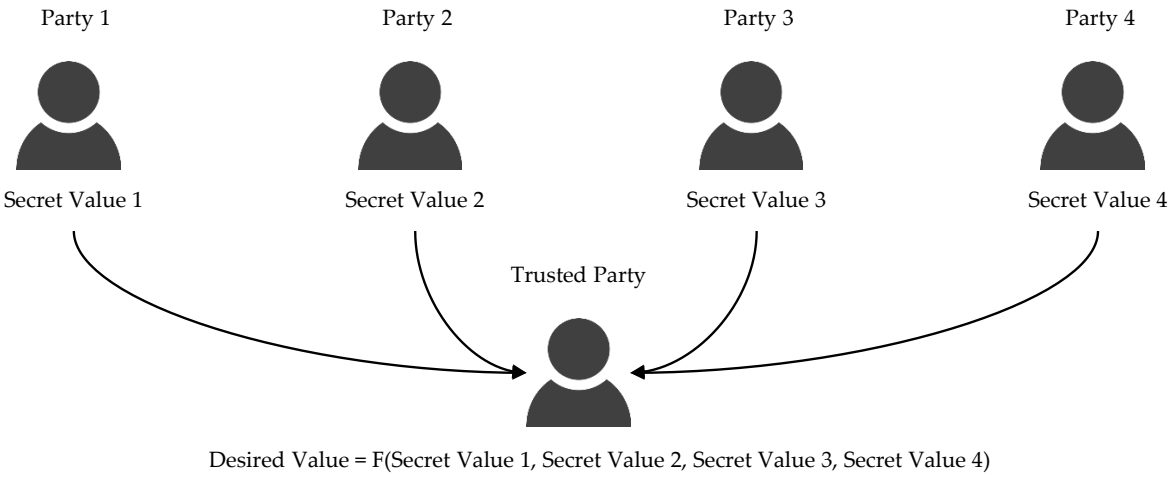


Fig. 3.3: Sharing secret values with a trusted party

Following the above example, and for explaining secure multiparty computation, we describe a simple method for secure aggregation of secret values. In secure aggregation, the desired value is the summation of secret values. Figure 3.4 represents the method for secure aggregation. In this example, we have four parties, each, hold a secret value (S.V.), the parties are interested in the summation of all secret values, i.e., $\sum_{i=1}^{4}$ S.V.i. For securely aggregation the secret values: (i) The first party generates a random mask, aggregates it with its secret value (S.V.1), and sends the result to the next party. (ii) The following parties receive the input, aggregate it with their secret values, and send the result to the next party. The last party sends the result to the first party. (iii) The first party receives the result from the last party, removes its random mask from the result, and informs all parties about the final result.

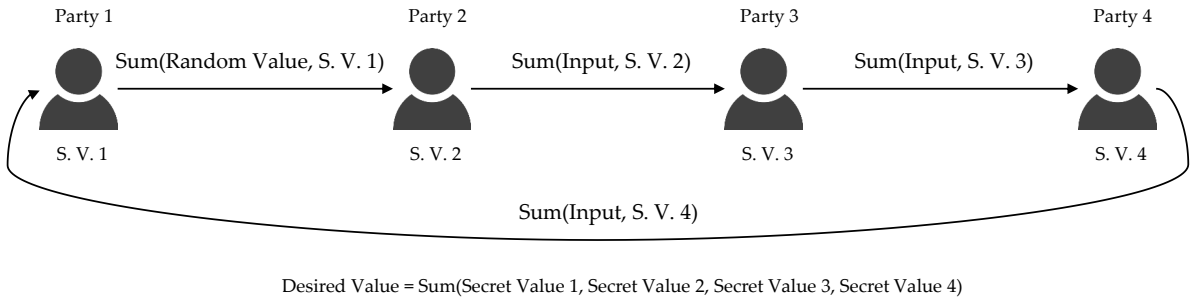


Fig. 3.4: Secure aggregation

In this way, each party cannot identify the secret value of previous parties based on the received information. However, in this method, if two neighbor parties, i.e., the parties before and after a certain party in the ring, collude, they will be able to identify the secret value of the victim party. For instance, if Party 2 reveals the input of Party 3, and at the same time, Party 4 reveals the output of Party 3, then they can reveal the secret value of Party 3. Therefore, the minimum number of colluding parties required for identifying a secret value is two in this method. Moreover, in terms of overhead, for one secure computation operation in this method, each party sends one message and receives one message. Thus, the communication overhead for this method is $2n$, in which n is the number of parties.

In privacy-preserving data mining techniques, two important characteristics of the SMC approach to be selected are the privacy and communication overhead of the approach. In such methods, we seek a higher level of privacy preservation, i.e., a larger minimum number of colluding parties, and lower communication overhead, i.e., a lower number of messages sent and received for each secure operation.

3.2.2 *Our Approach*

In this section, we target the problem of distributed machine learning with multiple data holders, without privacy violation. We assume that the training data is horizontally partitioned, i.e., different records of data are stored on different sources. We consider the classification problem, in which each data record has one category as the target. We consider that data is structured, i.e., it can be stored in spreadsheets, and contains categorical attributes, e.g., gender or mental-disorder history, and numerical attributes, like age or frequency and duration of pathological episodes.

We focus on the class of tree-based algorithms that have been shown to consistently outperform or to be on a par with the other state-of-the-art techniques when it comes to structured data [57, 129]. To learn from such horizontally-partitioned structured data, we propose the privacy-preserving distributed extremely randomized trees (PPD-ERT). We first extend the ERT algorithm [86] for distributed settings to enable learning without explicit sharing of the raw data. We then introduce a secure aggregation technique over the distributed ERT algorithm to avoid any information leakage. We evaluate the proposed solution experimentally and compare the results against the state-of-the-art techniques.

We assume that parties and mediator communications are performed securely, and message passing is performed through secure communication. We also assume

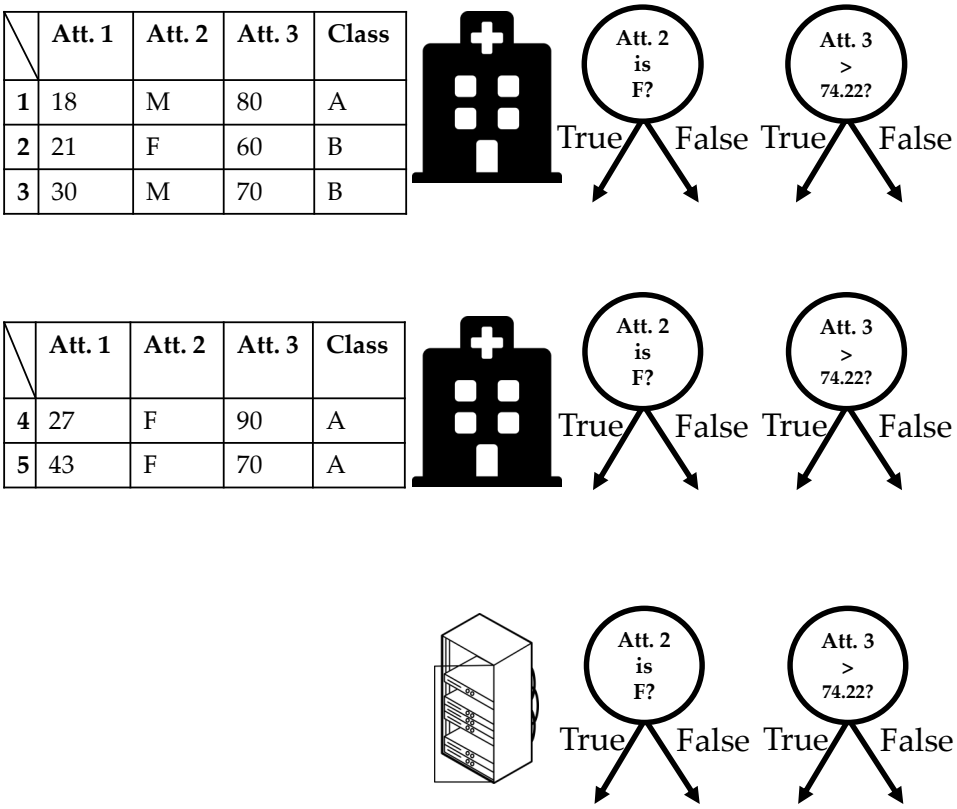
that each party sends and receives messages which are based on correct information. We assume that there is no collusion among the involved parties. Our technique is designed for tree-based learning methods, and its logic may be applicable for extending other tree-based machine learning algorithms in distributed settings.

Here, we provide a simplified example to clarify the procedure of learning for our algorithm without the secure multiparty computation layer. This procedure is shown in Figures 3.5 and 3.6. For the sake of simplicity of presentation, we do not consider the secure aggregation in this example. In the initiation of the learning process, the global random seed, personal random seeds, number and type of data attributes, possible categories or range of data attributes, as well as learning parameters for the algorithm (e.g., k) are shared among all parties. In our example, we have two data holder parties and a mediator. The first and second parties, respectively, hold three and two records for learning as shown in Figure 3.5a. Each record has three attributes (two numerical and one categorical) and one classification label.

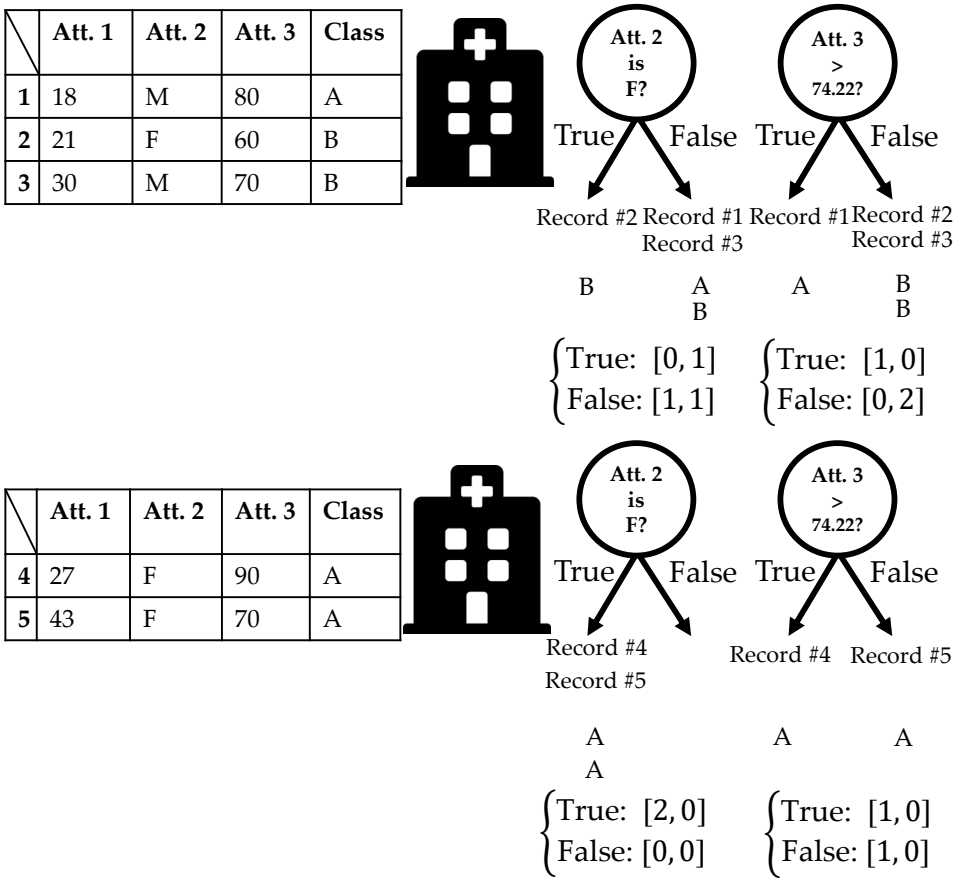
The goal is to learn an ensemble of decision trees from all the records available on the data holder parties based on the privacy-preserving distributed ERT algorithm. The mediator initiates a round of learning a decision tree and repeats it after finishing to have an ensemble of decision trees. At every step in choosing a decision node for our decision tree, each party, including the mediator, generates two random decision nodes based on the global seed. Since all parties use the same seed, they locally generate candidate decision nodes that are similar to the generated decision nodes in other parties. Figure 3.5a demonstrates the local generation of candidate decision nodes for the root of the first decision tree.

In the next step, the parties classify their records with each randomly generated candidate decision node as shown in Figure 3.5b. Several data records fall in the *True* branch (for each candidate decision node), and several data records fall in the *False* branch. Therefore, based on each record's labels (class), we make two vectors for each branch that represent the combination of the labels. For instance, for the first candidate decision node in the first party: the *True* vector is [0, 1], and it means that zero records of this party which belong to the class (label) A, and one record of this party which belongs to the class (label) B fall under *True* branch of this candidate decision node. Thus, each data holder party, for each candidate decision node, generates two vectors representing the records labels' combination (that fall under *True* and *False* branches).

The resulting vectors for each candidate decision node and in all data holder parties should be aggregated and returned to the mediator. Figure 3.6a shows this procedure in which all vectors for the *True* branch of each candidate decision node are aggregated and returned to the mediator, similar to the *False* branch's vectors. At this point, for each candidate decision node, the mediator has the combination of labels for *True* and *False* branches. In addition to deciding about making a leaf or decision node in the decision tree's current position, such vectors determine which candidate decision node is better (has higher score/information gain) and should be chosen. For calculating the score (the information gain here) for a decision node, the combination of labels at each branch is required. The information gain enables the comparison of how decision nodes classify the samples concerning the purity of their labels. This is done by calculating the impurity of labels before and after classification with a decision node. In our example, the second decision node has a higher information gain and is selected.

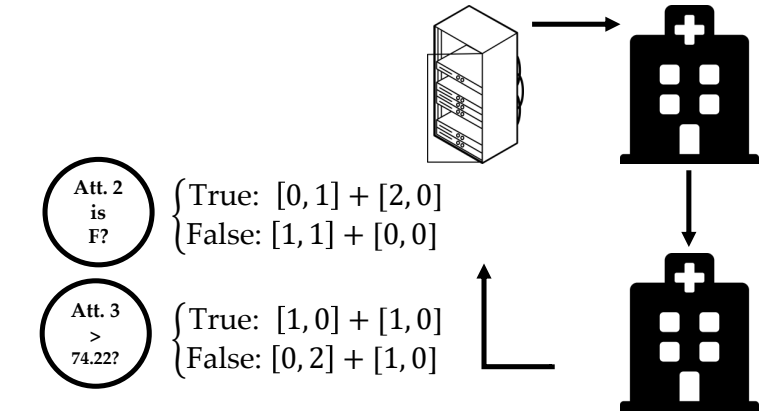


(a) Generating decision nodes randomly

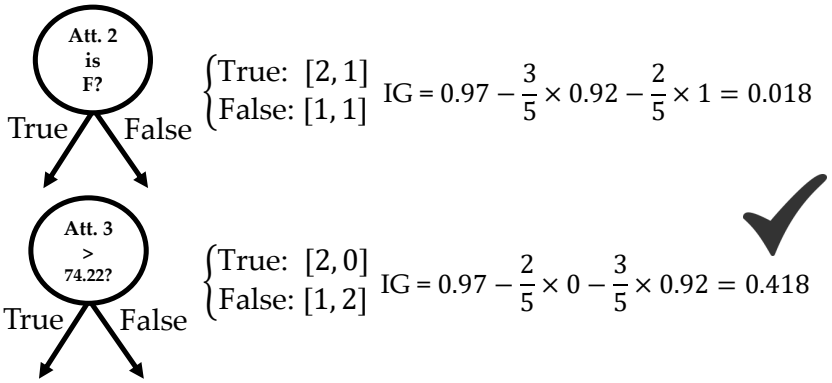


(b) Splitting the data in each computational node

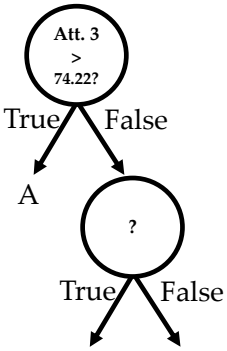
Fig. 3.5: Practical example



IG = The entropy for all labels
− The proportion of labels in the True branch × The entropy of labels in the True branch
− The proportion of labels in the False branch × The entropy of labels in the False branch

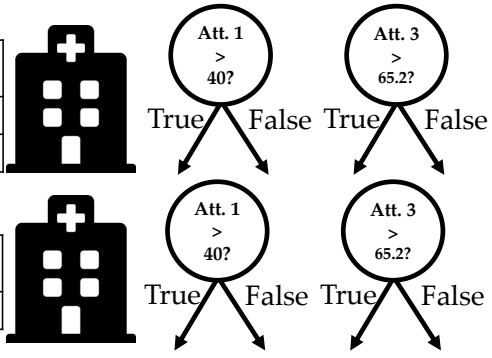


(a) Securely aggregating the results, calculating the scores, and selecting the best option



	Att. 1	Att. 2	Att. 3	Class
2	21	F	60	B
3	30	M	70	B

	Att. 1	Att. 2	Att. 3	Class
5	43	F	70	A



(b) Continuing the same process for the rest of the tree

Fig. 3.6: Practical example

As shown in Figure 3.6b, the second candidate decision node is selected for the decision tree's root. After checking the labels in its *True* branch, $[2, 0]$, we observe that all the records falling in the *True* branch belong to the same class (have the same label: A). Therefore, instead of making a decision node, we make a leaf in the *True* branch. On the other hand, we follow the same procedure of making a decision node for the *False* branch. However, this time, the data holder parties only consider the records that fall in the root's *False* branch, i.e., 2, 3, and 5. We continue the same procedure for the rest of the tree.

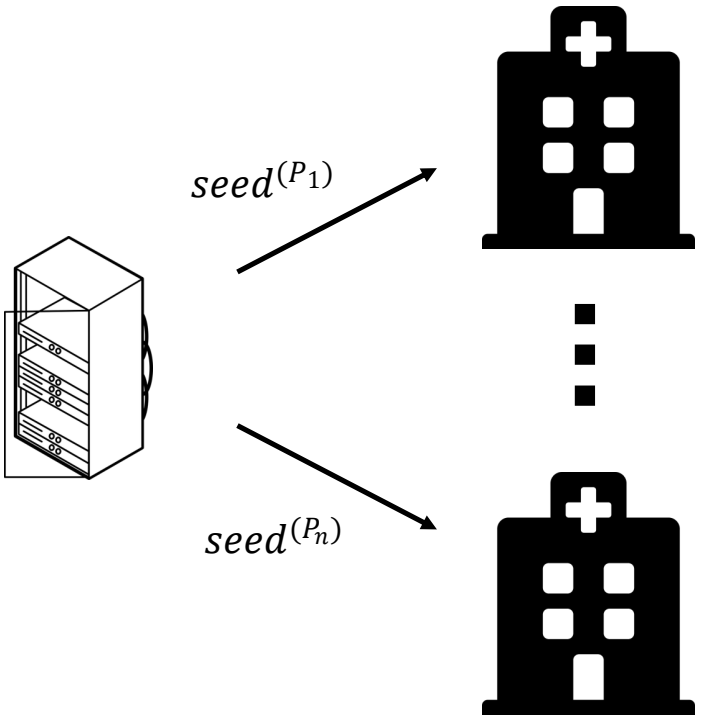
Now we discuss the secure multiparty computation layer in PPD-ERT. We employ an SMC technique in our proposed distributed ERT algorithm to avoid sharing the vectors representing the combination of the data record labels for each candidate decision node and each branch in each data holder party. In addition to the provided privacy by not sharing the raw values of data attributes, which is by construction, adoption of the SMC technique for aggregating the partial results from data holder parties contributes to privacy preservation.

In an extreme example, suppose our data has one sensitive attribute in it, e.g., having conducted transgender surgery before, and each data holder party has only one record on it. Then, sharing the partial results from one party, the vectors representing the combination of data record labels for one candidate decision node, can reveal sensitive information. If the candidate decision node is "whether the record falls into the transgender branch or not," the mediator can infer if that individual with the specified record has conducted transgender surgery. Therefore, to avoid such vulnerabilities, we adopt an SMC technique for aggregating the partial results from the data holder parties.

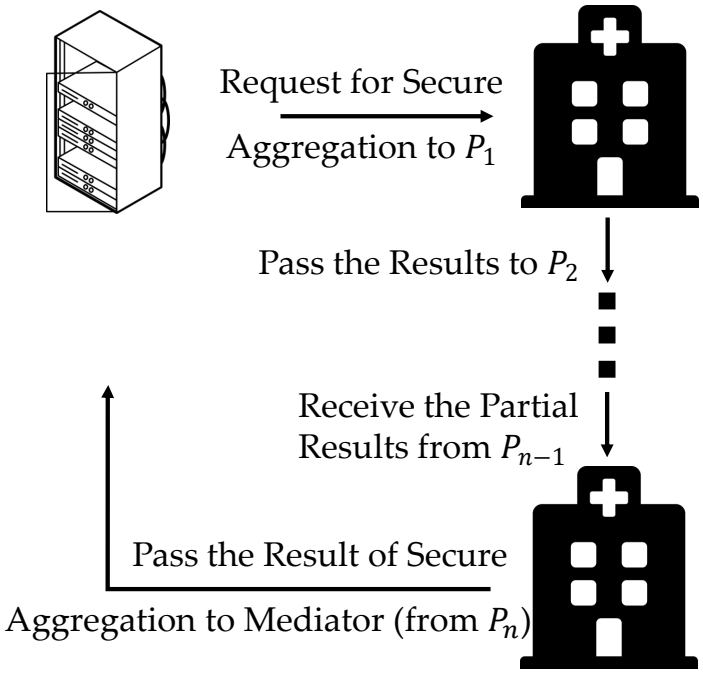
In the adopted SMC technique, shown in Figures 3.7 and 3.8, parties add random integer masks to their partial result vectors' values and pass them to the next party (or mediator). The mediator knows about these random integer masks and subtracts them from the partial results that it receives from the last party. We now describe the proposed technique in detail. The mediator shares a personal random seed with each data holder party through secure communication, to avoid sending and receiving exclusive random numbers between the mediator and each party as shown in Figure 3.7a. In Figures 3.7 and 3.8, superscripts represent the number/ID of the data holder party ($P_1 \leqslant P_m \leqslant P_n$), and subscripts represent the branch (T or F) and class/label of data samples $(1, ..., nc)$.

Then, in the process of learning a decision tree, the mediator sends the request for secure aggregation to the first party, as shown in Figure 3.7b. The party makes calculations described earlier and obtains two resulting vectors for each decision node. Afterward, the party generates random integer masks based on its personal random seed and adds it to the results from the previous step. For each global and the personal random seeds, the states of random function are stored and utilized precisely in every party. If the data holder party receives partial results vectors (*P. R.* in Figure 3.8) from the previous data holder party, then it also aggregates those values to the calculated vector in the previous step. Finally, the party passes its computed vectors to the next party or to the mediator if that party is the last one. Figure 3.8a, shows this procedure in each data holder party.

Finally, when the mediator received these masked aggregated results from the



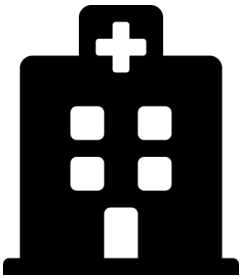
(a) Sharing personal random seeds



(b) Secure aggregation process

Fig. 3.7: Secure aggregation of the results of splitting the data

1. Receive the Partial Results received from previous party (if any):
   - True: $[P.R._{T-1}^{(P_{m-1})}, \ldots, P.R._{T-nc}^{(P_{m-1})}]$
   - False: $[P.R._{F-1}^{(P_{m-1})}, \ldots, P.R._{F-nc}^{(P_{m-1})}]$

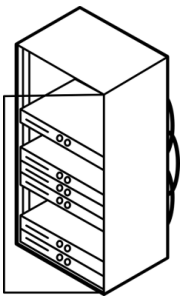


2. Calculate the result of splitting the data :
   - True: $[num_{T-1}^{(P_m)}, \ldots, num_{T-nc}^{(P_m)}]$
   - False: $[num_{F-1}^{(P_m)}, \ldots, num_{F-nc}^{(P_m)}]$
3. Generate random masks based on $seed^{(P_m)}$:
   - True: $[rnd_{T-1}^{(P_m)}, \ldots, rnd_{T-nc}^{(P_m)}]$
   - False: $[rnd_{F-1}^{(P_m)}, \ldots, rnd_{F-nc}^{(P_m)}]$
4. Calculate the result (to be passed) by aggregating:
   - The result of splitting the data
   - Generated random vectors
   - Partial Results

5. Pass the results to the next party (or mediator):
   - True: $[P.R._{T-1}^{(P_m)}, \ldots, P.R._{T-nc}^{(P_m)}]$
   - False: $[P.R._{F-1}^{(P_m)}, \ldots, P.R._{F-nc}^{(P_m)}]$

(a) Responsibilities of each party for secure aggregation

1. Receive the result of secure aggregation (from P_n):
   - True: $[P.R._{T-1}^{(P_n)}, \ldots, P.R._{T-nc}^{(P_n)}]$
   - False: $[P.R._{F-1}^{(P_n)}, \ldots, P.R._{F-nc}^{(P_n)}]$



2. Generate random masks based on $seed^{(P_1)}, \ldots, seed^{(P_n)}$:
   - True: $[rnd_{T-1}^{(P_1)}, \ldots, rnd_{T-nc}^{(P_1)}]$
   - False: $[rnd_{F-1}^{(P_1)}, \ldots, rnd_{F-nc}^{(P_1)}]$

   ⋮

   - True: $[rnd_{T-1}^{(P_n)}, \ldots, rnd_{T-nc}^{(P_n)}]$
   - False: $[rnd_{F-1}^{(P_n)}, \ldots, rnd_{F-nc}^{(P_n)}]$

3. Calculate the desired result of secure aggregation (in order to be used for calculation of score/IG in the ERT algorithm) by subtracting the above generated random masks

(b) Responsibilities of the mediator for calculating the desired result

Fig. 3.8: Secure aggregation of the results of splitting the data

last party. Since the mediator has the personal random seeds, it generates the same random masks as generated on the data holder parties. Then, for further computations for learning the decision tree, the mediator subtracts those random masks from the received result from the last party. At this step, without sharing the partial information about data labels by each data holder party, the mediator has the aggregated vectors representing the combination of data record labels for each branch of each candidate decision node for all parties. Figure 3.8b, shows this procedure in the mediator.

The minimum number of parties required for collusion in order to identify a secret value in this approach, PPD-ERT, is two. PPD-ERT is also limited to datasets without missing values, which is rarely the case in real-world healthcare applications. Therefore, we propose a framework based on PPD-ERT, called *k*-PPD-ERT, that employs an extended SMC technique without increasing the communication overhead. In *k*-PPD-ERT, the minimum number of required colluding parties is raised to *k*, which is a parameter that can be tuned by the user of the algorithm. The *k*-PPD-ERT's secure multiparty computation technique is efficient, and at the same time, similar to Shamir's secret sharing technique [164], is resilient to multiple colluding parties. Our *k*-PPD-ERT framework is designed to handle scenarios in which the dataset contains records with missing values. In the following, we describe our proposed privacy-preserving distributed machine learning framework in detail.

3.2.2.1 *Adaptation of ERT for Distributed Settings*

This section presents the detailed procedure of learning an ensemble of decision trees based on the ERT algorithm in the discussed setting. The pseudocode of the algorithm is also provided for clarity.

Initialization and Start of the Learning Process We have two types of parties in our distributed learning framework. We have a *mediator* that mediates and orchestrates the overall learning process and several *data holder parties* that collaborate with each other and the mediator to learn a classification model. Algorithm 1 and Algorithm 2 show the pseudocodes of the procedures and functions for the mediator and data holder parties, respectively.

(a) **Sharing the Random Seeds**
To start this process, a global seed for the random function is agreed upon among all parties (Algorithm 1, Line 1 and Algorithm 2, Line 1). The global seed is common among the mediator and all data holders. In the ERT algorithm, we have two parameters of randomness for learning a weak classifier. First, we need to randomly select several attributes for the candidate decision nodes, at every step of building our decision tree (Algorithm 1, Line 20 Algorithm 2, Line 20). Second, a random splitting point for every attribute in the candidate decision node is required (Algorithm 1, Line 21, and Lines 23–27, and Algorithm 2, Line 21, and Lines 23–27). The data holder parties and the mediator are required to use the same candidate decision nodes at every step when learning a decision tree. For this purpose, we use the global random seed that all parties, including the mediator, utilize to locally generate these candidate decision nodes (Algorithm 1,

Line 10, and Algorithm 2, Line 14). This is instead of making these randomly-made candidate decision nodes in the mediator and sharing them with all parties for further tasks. Since all parties use a common random seed, i.e., the global random seed, they generate the same candidate decision nodes at every step, without major communication overhead.

In addition, for the secure aggregation of partial results, described further, k selected data holder parties send unique seeds for the random function to other data holders through secure communication (Algorithm 2, Line 2). These random seeds are exclusive and private for each pair of data holder parties.

(b) **Initiate the Process of Learning One Decision Tree**
The privacy-preserving distributed ERT algorithm is an ensemble learning method, therefore, we repeat the process of learning a decision tree for M times, until we have M decision trees (Algorithm 1, Lines 3–4). The number of trees, M, is a parameter tuned by the user to make a trade-off between robustness and overhead. We learn different decision trees every time due to the randomness in ERT. Finally, after repeating the process of learning a decision tree M times, we store the trees in E (Algorithm 1, Line 5). For future prediction, the ensemble of the learned trees, E, will be used.

The Process of Learning One Decision Tree The learning of a decision tree based on the privacy-preserving distributed ERT algorithm is a recursive procedure. The procedure is executed top-down and starts from the root and ends in the leaves. For the root decision node, the *Split_ID* or the ID for the decision node is zero, and there is no previous branch, so the *Branch* input is set to 'None'(Algorithm 1, Line 4).

(a) **Generation of Candidate Decision Nodes**
For building each decision tree, extremely randomized tree, the mediator generates the candidate decision nodes (Algorithm 1, Line 10). The mediator will further select the best decision node among the candidates based on the results received from data holder parties. The candidate decision nodes are generated randomly, based on the global random seed, according to Algorithm 1, Lines 19–27, and Algorithm 2, Lines 19–27. The number of candidate decision nodes, D, is a parameter in the ERT algorithm tuned by the user. D attributes from all possible attributes are selected for candidate decision nodes (Algorithm 1, Line 20, and Algorithm 2, Line 20). Then, each candidate decision node's splitting point is selected (Algorithm 1, Line 21, and Algorithm 2, Line 21). If the attribute is categorical, one random possible category is selected to be checked (Algorithm 1, Lines 24–25, and Algorithm 2, Lines 24–25); otherwise, when the attribute is numerical, a point in the possible range is selected for comparison in the decision node (Algorithm 1, Lines 26–27, and Algorithm 2, Lines 26–27). We assume that all parties already have the possible categories and ranges for each attribute.

Algorithm 1 Mediator

(1) • The global random seed (known to all parties) is set in the mediator
(2) • Wait for data holder parties' connection
(3) **for** $i = 1$ **to** M **do**
(4) | • Generate tree: t_i = *Build_k-PPD-ERT(o, 'None')*
end
(5) $E = \{t_1, t_2, ..., t_M\}$
(6) **Function** *Build_k-PPD-ERT(Split_ID, Branch)*
(7) | • Send *Secret_aggregation(Split_ID, Branch)* request to data holder parties
(8) | • Wait until receiving the results from data holder parties
(9) | • *Sum* = aggregated the received results form data holder parties
(10) | • *Generate_splits()* (based on the global seed)
(11) | **if** *number of classified records is less than* n_{min} ***or*** *labels of the classified records are the same* **then**
(12) | | **return** a leaf label
| **else**
(13) | | • Calculate each split's score (Information Gain) based on *Sum*
(14) | | • Select the split with the highest score.
(15) | | • Inform all parties about the selected split (for *Split_ID*)
(16) | | • Build *tree_T* = *Build_k-PPD-ERT(next Split_ID, 'T')*
(17) | | • Build *tree_F* = *Build_k-PPD-ERT(next Split_ID, 'F')*
(18) | | • Create a node with the selected split, attach *tree_T* and *tree_F* as *T* and *F* subtrees, and **return** the resulting tree.
| **end**
end
(19) **Function** *Generate_splits()*
(20) | • Select D attributes randomly: $\{a_1, ..., a_D\}$
(21) | • Generate D splits: $\{s_1, ..., s_D\}$, where s_i = *Pick_rand_split*(a_i)
(22) | **return** splits $\{s_1, ..., s_D\}$
end
(23) **Function** *Pick_rand_split*(a)
(24) | **if** a *is categorical* **then**
(25) | | **return** a possible category
| **end**
(26) | **if** a *is numerical* **then**
(27) | | **return** a possible value in the min and max range
| **end**
end

Algorithm 2 Data Holder Party

(1) • The global random seed (known to all parties) is set in the data holder party
(2) • Wait for completion of data holder parties initialization. In initialization, k selected data holder parties send their unique seeds to other data holders. In initialization, $SSA_{p_j}^{p_i}$ is sent by party i (i is among the k selected parties) and received by party j
(3) • Connect to the Mediator
(4) **Function** *Secret_aggregation(Split_ID, Branch)*
(5) • $secret\_val^{P_j}$ = *Split_data(Split_ID, Branch)*
(6) • $rand\_sum_{others}^{P_j}$ = Generate and aggregate random masks based the received seeds
(7) **if** *the party,* P_j, *is among* k *selected data holder parties for secure aggregation* **then**
(8) • $rand\_sum_{self}^{P_j}$ = Generate and aggregate random masks based the sent seeds
else
(9) • $rand\_sum_{self}^{P_j} = 0$
end
(10) • *Result* = $secret\_val^{P_j} - rand\_sum_{self}^{P_j} + rand\_sum_{others}^{P_j}$
(11) • Send *Result* to the mediator
end
(12) **Function** *Split_data(Split_ID, Branch)*
(13) • S_{sub} = records in the computational node that should be split based on *Split_ID* and *Branch*
(14) • $\{s_1, ..., s_D\}$ = *Generate_splits()* (based on the global seed)
(15) **for** $i = 1$ **to** D **do**
(16) • Split S_{sub} to two sets (T, F) by s_i
(17) • Append vectors $\{Vec_T, Vec_F\}$ representing the records' labels for each of the above sets to *Result*
end
(18) **return** *Result*
end
(19) **Function** *Generate_splits()*
(20) • Select D attributes randomly: $\{a_1, ..., a_D\}$
(21) • Generate D splits: $\{s_1, ..., s_D\}$, where s_i = *Pick_rand_split*(a_i)
(22) **return** splits $\{s_1, ..., s_D\}$
end
(23) **Function** *Pick_rand_split*(a)
(24) **if** a *is categorical* **then**
(25) **return** a possible category
end
(26) **if** a *is numerical* **then**
(27) **return** a possible value in the min and max range
end
end

(b) **Parties Classify Their Records**

To decide about the candidate decision nodes for each branch, the mediator requires the collective outcome of the classification with candidate decision nodes from all data holders on all their data. By having the combination of data record labels for each branch (*True* and *False*), the mediator can decide if we require a leaf or we need to calculate the score, i.e., information gain (Algorithm 1, Line 11). Information gain captures the extent of samples' purity (concerning their class/category) after splitting and is used as a basis for comparing decision nodes. The mediator sends a request to data holder parties and waits for receiving the result from all parties, which is masked according to the secure aggregation technique described further (Algorithm 1, Lines 7–8). The masked results are two vectors, one for each of the *True* and *False* branches, representing the combination of data record labels after classification with each candidate decision node.

Each party receives *Split_ID* and *Branch* to determine the local records for classification (Algorithm 2, Line 13). Then, the party randomly generates candidate decision nodes based on Lines 19–27 in Algorithm 2 and the global random seed (Algorithm 2, Line 14). Next, it classifies the selected local data based on each candidate decision node and returns the result (Algorithm 2, Lines 15–18).

We describe how each party returns the result to the mediator in the following, using an example. Vec_T represents the combination of labels for the records that fall in the *True* branch, and Vec_F represents the combination of labels for the records that fall in the *False* branch. For instance, if three records with labels A, A, and B fall in the *True* branch of the candidate decision node, and we have three labels, A, B, and C in the dataset, then $Vec_T = [2, 1, 0]$.

(c) **Each Party Sends the Result to the Mediator**
After adopting the secure aggregation protocol described further, each data holder party returns the masked result to the mediator to select the best decision node (or generate a leaf instead of a decision node). For every candidate decision node, the mediator receives and aggregates the results from all parties and obtains two vectors, for *True* and *False* branches, representing the combination of data labels (Algorithm 1, Lines 8–9).

(d) **Mediator Determines the Best Candidate for the Decision Node**
Now that the mediator has the value of *Sum* (Algorithm 1, Line 9), it determines if a decision node or a leaf node is required here in the tree (Algorithm 1, Lines 11). If all labels are the same or if the number of received labels is less than our threshold parameter, the mediator introduces a leaf node (Algorithm 1, Line 12). Otherwise, the mediator calculates the score, i.e., information gain, of each candidate decision node based on the results from data holder parties (Algorithm 1, Line 13). It then selects the candidate decision node with the highest information gain and informs all parties about it (Algorithm 1, Lines 14–15). The selected node will be used to build the tree at the mediator (Algorithm 1, Line 18). This decision is communicated to all data holder parties and is required to select records for classification at every step (Algorithm 2, Line 13).

(e) **The Mediator Initiates Another Round From the First Step**

After selecting the best candidate decision node, the mediator continues the process for each branch of this decision node. Therefore, the same process is performed from the first step, for each of the *True* and *False* branches (Algorithm 1, Lines 16–17). After returning from these recursive calls, the selected subtrees for each branch are returned (Algorithm 1, Lines 12 and 18).

3.2.2.2 *Secure Aggregation of Results From Data Holder Parties*

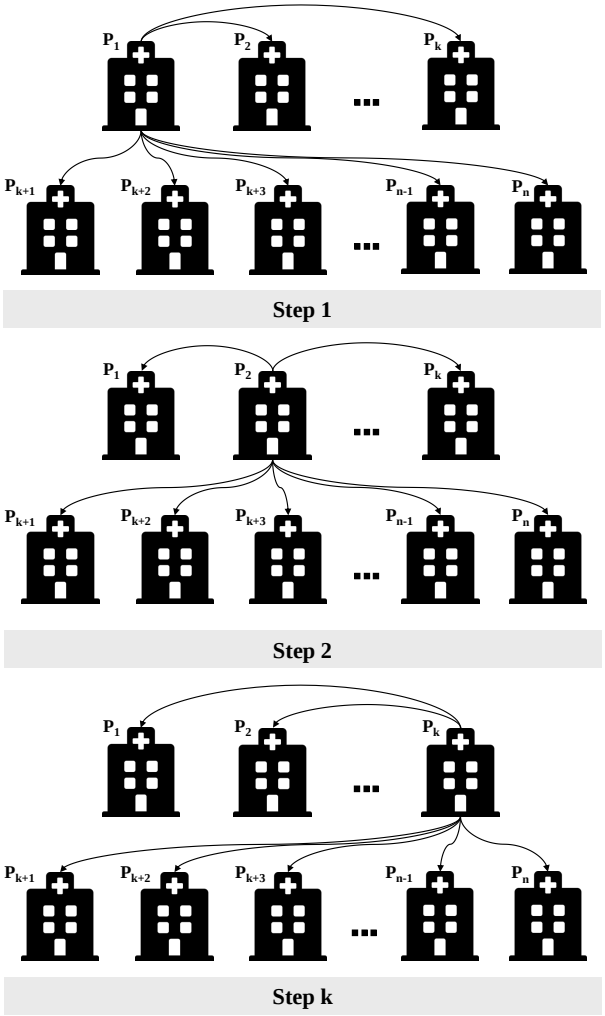
We adopt an SMC technique in our proposed distributed ERT algorithm to avoid sharing the vectors representing the combination of the data record labels for each candidate decision node and each branch in each data holder party. In addition to the provided privacy by not sharing the raw values of data attributes, which is by construction, the adoption of an SMC technique for aggregating the partial results from data holder parties contributes to privacy preservation. In an extreme example, suppose our data has one sensitive attribute in it, e.g., having conducted transgender surgery before, and each data holder party has only one record on it. Then, sharing the partial results from one party, the vectors for the combination of data record labels for each candidate decision node, can reveal sensitive information. If the candidate decision node is "whether the record falls into the transgender branch or not," the mediator can infer if that individual with the specified record has undergone transgender surgery. Therefore, to avoid such vulnerabilities, we adopt an SMC technique to aggregate the partial results from the data holder parties.

The secure aggregation procedure begins with an initialization process. Subsequently, the parties can securely aggregate their secret values through this approach.

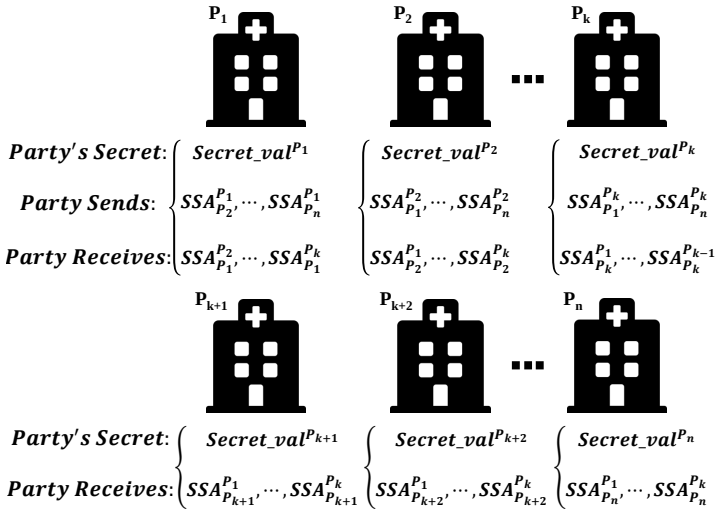
Initialization: In the initialization phase, k selected data holder parties share their unique seeds for the random function with all parties. These seeds are unique and private between each pair of parties. Without loss of generality and for the simplicity of the presentation, we assume that the k selected data holder parties are P_i ($\forall i \in \{1, \ldots, k\}$). Party P_i ($\forall i \in \{1, \ldots, k\}$) sends unique seeds to party P_j ($\forall j \in \{1, \ldots, n \mid i \neq j\}$). Figure 3.9a shows this process.

The seed party P_i shares with party P_j is represented with $SSA_{P_j}^{P_i}$, and it is a unique seed; *SSA* is the short form of Seed for Secure Aggregation. Parties 1 to k, send $n-1$ and receive $k-1$ seeds. Parties $k+1$ to n, receive k seeds. This is shown in Figure 3.9b. Therefore, k parties send $n-1$ and receive $k-1$ messages, and $n-k$ parties send zero and receive k messages. The total communication overhead for initialization is $2k(n-1)$. The communication overhead by adopting this approach is equal to $O(kn)$, which can be adjusted by adapting k based on the sensitivity of the data. If all parties were required to send and receive seed, then, the communication overhead would be equal to $2n(n-1)$. The communication overhead by adopting this approach is equal to $O(n^2)$ [37].

Secure aggregation: In the adopted SMC technique, shown in Figure 3.10, parties add random masks to their partial result vectors and pass them to the mediator. The mediator aggregates the partial results received from all parties. After aggregation, the random masks from all parties cancel each other. We now describe the proposed technique in detail:



(a) The k selected data holder parties sending unique seeds to other data holders



(b) The sent and received seeds after the initialization

Fig. 3.9: Initialization

- **Step 1:** The mediator initiates the secure aggregation process round (Algorithm 1, Line 7). This is shown in Figure 3.10a.

- **Step 2:** Data holder parties generate random masks and aggregate them with their secret values (Algorithm 2, Line 10). This is shown in Figure 3.10b.

  - Parties P_i ($\forall i \in \{1, \ldots, k\}$) generate random masks based on the sent and received seeds (Algorithm 2, Lines 6–9).
  - Parties P_i ($\forall i \in \{k+1, \ldots, n\}$) generate random masks based on received seeds (Algorithm 2, Lines 6–9).

- **Step 3:** In the next step, the parties send the masked results to the mediator (Algorithm 2, Line 11). Then, the mediator receives the results from all parties (Algorithm 1, Line 8). Figure 3.10c shows this.

- **Step 4:** In the last step, the mediator aggregates all the received results to obtain the desired value, i.e., the aggregated secret values from all parties (Algorithm 1, Line 9). This is shown in Figure 3.10d.

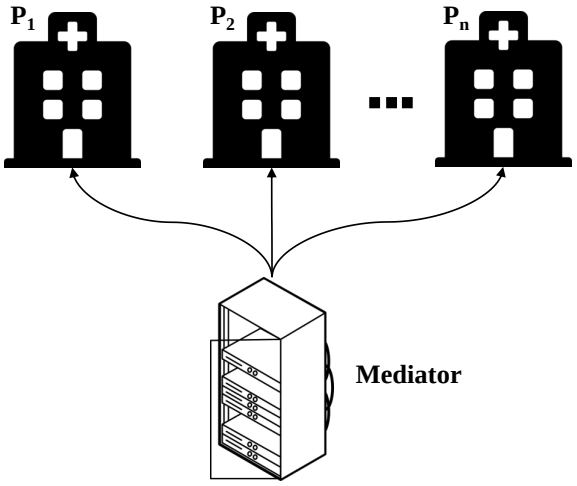
Privacy: We now show that the secret values of the parties are kept private in our proposed protocol. The partial result $Result^{P_i}$, which is shared with the mediator, consists of three components: $secret\_val^{P_i}$, $rnd\_sum^{P_i}_{self}$, and $rnd\_sum^{P_i}_{others}$. The two components, $rnd\_sum^{P_i}_{self}$ and $rnd\_sum^{P_i}_{others}$, mask the secret value.

- For P_i ($\forall i \in \{1, \ldots, k\}$), the value of $rnd\_sum^{P_i}_{self}$ can only be identified by the collusion of $n-1$ parties holding the random seeds for generating the random masks, which are the components of $rnd\_sum^{P_i}_{self}$. At the same time, $rnd\_sum^{P_i}_{others}$ can only be identified by the collusion of $k-1$ parties that generate the components of $rnd\_sum^{P_i}_{others}$. Therefore, the minimum number of colluding parties required to reveal the secrete value of P_i is $n-1$.

- For P_i ($\forall i \in \{k+1, \ldots, n\}$), the value of $rnd\_sum^{P_i}_{self}$ is zero and known to all, and $secret\_val^{P_i}$ is masked by $rnd\_sum^{P_i}_{others}$. However, $rnd\_sum^{P_i}_{others}$ can only be identified by the collusion of k parties that generate the components of $rnd\_sum^{P_i}_{others}$, i.e., the k selected parties for secure aggregation.

In the worst case, i.e., for P_i ($\forall i \in \{k+1, \ldots, n\}$), the k selected parties for secure aggregation are required to collude to identify a secret value; hence, the minimum number of colluding data holder parties is equal to k. Moreover, since only the mediator receives the victim's partial result, the collusion of other parties without the mediator's participation is not possible. Therefore, for identifying a secret value, the collusion of k data holder parties and the mediator is necessary.

Correctness: We also show that the final value of the aggregation of partial results is equal to the aggregation of secret values. The aggregation of all the partial results sent to the mediator is as follows:

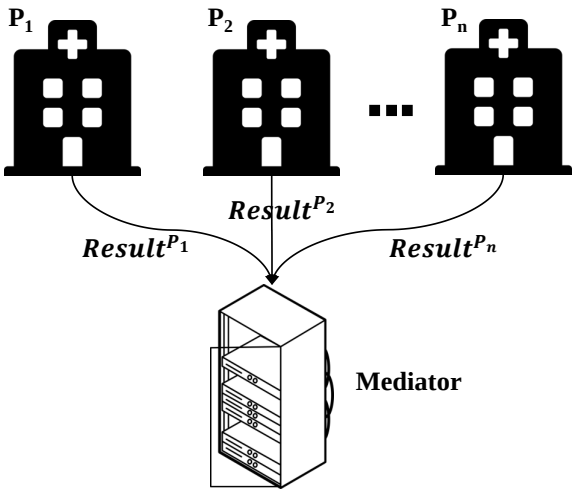


(a) Step 1

Party i

- **For each P_i ($\forall$i $\in \{1, \dots, n\}$): $rand\_sum_{others}^{P_i}$ = Generate and aggregate random masks based on the received seeds**
- **For each P_i ($\forall$i $\in \{1, \dots, k\}$) : $rand\_sum_{self}^{P_i}$ = Generate and aggregate random masks based on the sent seeds**
- **For each P_i ($\forall$i $\in \{k+1, \dots, n\}$): $rand\_sum_{self}^{P_i} = 0$**
- $Result^{P_i} = Secret\_val^{P_i} - rand\_sum_{self}^{P_i} + rand\_sum_{others}^{P_i}$

(b) Step 2



(c) Step 3

Mediator

- Sum = **Recieve** $Result^{P_i}$ ($\forall$**i** $\in \{1, \dots, n\}$) **and aggregate them**

(d) Step 4

Fig. 3.10: Secure aggregation

$$\begin{aligned}\sum_{j=1}^{n} \text{Result}^{P_j} &= \text{secret\_val}^{P_1} - \text{rnd\_sum}^{P_1}_{self} + \text{rnd\_sum}^{P_1}_{others} \\ &\vdots \\ &+ \text{secret\_val}^{P_n} - \text{rnd\_sum}^{P_n}_{self} + \text{rnd\_sum}^{P_n}_{others} \\ &= \sum_{j=1}^{n} \text{secret\_val}^{P_j} - \sum_{j=1}^{n} \text{rnd\_sum}^{P_j}_{self} + \sum_{j=1}^{n} \text{rnd\_sum}^{P_j}_{others}.\end{aligned} \tag{3.8}$$

In addition, we also have the following equations for the data holder parties:

- For P_i ($\forall i \in \{1, \ldots, k\}$), $\text{rnd\_sum}^{P_i}_{self} = \sum_{j=1}^{n} \text{rnd}^{P_i}_{P_j} - \text{rnd}^{P_i}_{P_i}$, where $\text{rnd}^{P_i}_{P_j}$ is the shared random mask between P_i and P_j. On the other hand, $\text{rnd\_sum}^{P_i}_{others} = \sum_{j=1}^{k} \text{rnd}^{P_j}_{P_i} - \text{rnd}^{P_i}_{P_i}$.
- For P_i ($\forall i \in \{k+1, \ldots, n\}$), $\text{rnd\_sum}^{P_i}_{self} = 0$. On the other hand, $\text{rnd\_sum}^{P_i}_{others} = \sum_{j=1}^{k} \text{rnd}^{P_j}_{P_i}$.

Substituting these in Equation 3.8, we obtain:

$$\begin{aligned}\sum_{j=1}^{n} \text{Result}^{P_j} &= \sum_{j=1}^{n} \text{secret\_val}^{P_j} - \sum_{j=1}^{n} \text{rnd\_sum}^{P_j}_{self} + \sum_{j=1}^{n} \text{rnd\_sum}^{P_j}_{others} \\ &= \sum_{j=1}^{n} \text{secret\_val}^{P_j} - \sum_{j=1}^{k} (\sum_{i=1}^{n} \text{rnd}^{P_j}_{P_i} - \text{rnd}^{P_j}_{P_j}) - \sum_{j=k+1}^{n} (0) \\ &+ \sum_{j=1}^{k} (\sum_{i=1}^{k} \text{rnd}^{P_i}_{P_j} - \text{rnd}^{P_j}_{P_j}) + \sum_{j=k+1}^{n} (\sum_{i=1}^{k} \text{rnd}^{P_i}_{P_j}) \\ &= \sum_{j=1}^{n} \text{secret\_val}^{P_j} - \sum_{j=1}^{k} (\sum_{i=1}^{n} \text{rnd}^{P_j}_{P_i}) + \sum_{j=1}^{k} (\text{rnd}^{P_j}_{P_j}) \\ &+ \sum_{j=1}^{k} (\sum_{i=1}^{k} \text{rnd}^{P_i}_{P_j}) - \sum_{j=1}^{k} (\text{rnd}^{P_j}_{P_j}) + \sum_{j=k+1}^{n} (\sum_{i=1}^{k} \text{rnd}^{P_i}_{P_j}) \\ &= \sum_{j=1}^{n} \text{secret\_val}^{P_j} - \sum_{i=1}^{n} (\sum_{j=1}^{k} \text{rnd}^{P_j}_{P_i}) + \sum_{j=1}^{n} (\sum_{i=1}^{k} \text{rnd}^{P_i}_{P_j}) \\ &= \sum_{j=1}^{n} \text{secret\_val}^{P_j}.\end{aligned} \tag{3.9}$$

The above equation shows that the aggregation of partial results from data holder parties is equal to the aggregation of data holder parties' secret values.

As shown above, the correctness and accuracy of our SMC technique do not depend on k or the minimum number of colluding parties. By increasing k, the minimum number of colluding parties required for revealing a secret value increases, which in turn improves the privacy of the method. Increasing k increases the communication overhead in the initialization phase. Therefore, the trade-off is between privacy and communication overhead of the initialization phase.

3.2.2.3 *Handling Missing Values*

Table 3.2: Example of structured data distributed among two parties with missing values

Party	Record	Sex	Height
1	1	M	170
	2	F	155
	3	M	?
2	1	F	?
	2	F	165
	3	M	178

In this section, handling missing values when the data is distributed is explained in the context of our proposed privacy-preserving distributed learning framework, i.e., *k*-PPD-ERT. In the application of distributed learning approaches, particularly in the healthcare domain, we deal with data with missing values. Missing values in a dataset may occur as a result of improper collection of data, refusal of patients to share information, etc. In scenarios where the data is distributed, handling missing values can require a different procedure in comparison to scenarios in which the data is held in one center.

Several approaches can still be used in such scenarios, e.g., deleting records with missing values. However, they might not be helpful in all cases, e.g., where we have a low number of data records or when the percentage of records with missing values is high. Another solution is to replace the missing values in an attribute with the mean/average of the available values in that attribute. This approach avoids deleting data records and is particularly relevant when dealing with smaller datasets with missing values.

For calculating the mean of the available values for an attribute, we require the summation of these values. Due to privacy concerns, data holder parties refrain from sharing the summation of their available values with others. In particular, this is a major privacy concern when each data holder party holds only one record. Therefore, we adopt the approach presented in the secure aggregation section to address this issue, as we merely require the final summation of the available values.

We explain the approach using an example. Suppose we have two parties, and each party holds three records. Table 3.2 represents the data for each party. Each record contains the sex and height of record owners or patients. Two records miss the value for height. Assume that by preserving privacy, we can calculate the summation of available values for the height, i.e., 668 in our example, as well as the summation of the number of records not missing the height value, i.e., 4 in our example. In that case, we can calculate the mean for the height, i.e., 167 in our example.

The summation of the available values and the number of available values are calculated using our secure aggregation method. Finally, the mediator divides the summation of the available values by the number of available values and calculates the mean. Then, the mean is shared with all parties to replace the missing values.

Our technique may also be modified based on the problem settings. For instance, in the above example, suppose the user requires the mean of values for male and female patients separately, i.e., 174 and 160, respectively. Then, our technique can be adjusted

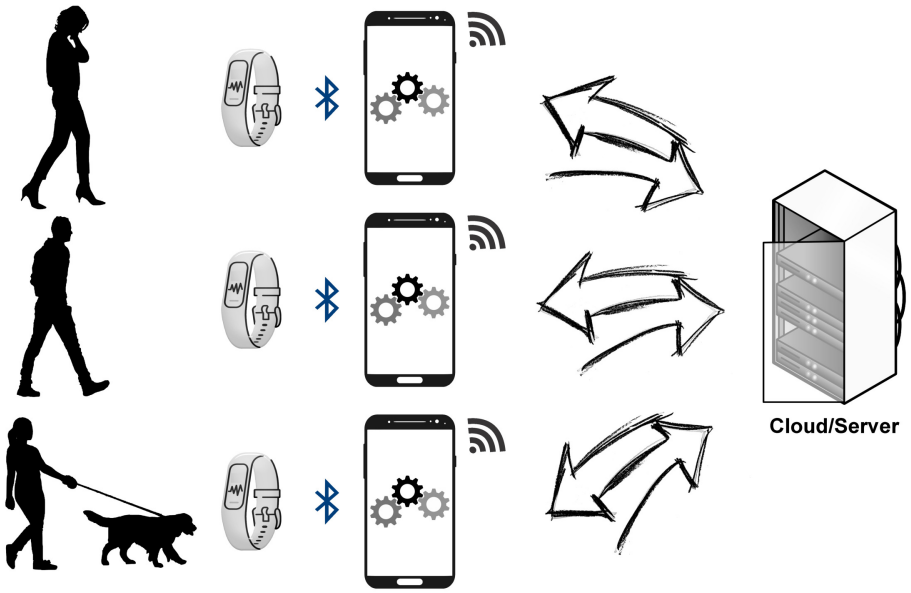


Fig. 3.11: Overall scenario for our privacy-preserving learning

by only securely aggregating the available values belonging to male or female patients.

We use the same technique for categorical attributes, i.e., to calculate the frequencies of categories in one attribute. Then, we may decide how to fill the missing values based on these frequencies. We may decide to replace all values with the most frequent category, i.e., the mode. The missing category can also be drawn randomly based on the distribution of frequencies. Moreover, we may also decide on filling the missing values by jointly considering the frequencies and information from other attributes.

We present the implementation of our framework on Amazon's AWS cloud and evaluate it based on real-world mental health datasets associated with the Norwegian INTROducing Mental health through Adaptive Technology (INTROMAT) project. In Chapter 4, we investigate the classification performance, scalability, and privacy of our privacy-preserving distributed machine learning framework. We compare our proposed framework against the state-of-the-art in terms of the mentioned criteria. We also compare the classification performance of our framework against the centralized ERT and the communication overhead of our framework compared to distributed ERT.

In this research, we investigate the possibility of using our proposed privacy-preserving distributed machine learning framework for sensor data based on the Depresjon (depression in Norwegian) dataset [85], which paves the way for the real-world applications of our approach for wearable technology in the described settings. In our study, we use an approach for augmenting the motor activity signals, which is presented in Paper E. The augmentation addresses several problems with the Depresjon dataset, e.g., imbalanced number of recorded days for individuals in the data, and improves the classification performance. The scenario is shown in Figure 3.11. In Chapter 4, we will evaluate our privacy-preserving extremely randomized trees framework in conjunction with the proposed augmentation technique.

CHAPTER 4

EVALUATION AND DISCUSSION

4.1 Overview of Thesis Contributions

As discussed in previous chapters, the research in this thesis is carried out in two directions. The first direction is for scenarios in which the data is stored in a centralized location. The other direction is for scenarios in which the data is distributed among multiple parties. Accordingly, our contributions in this thesis are twofold and in the discussed directions.

For privacy-preserving data publishing direction, in [36], we investigated the possibility of adopting cryptographic algorithms, which facilitates privacy preservation by construction, in the context of anonymization of structured data. The basic idea is to map the original dataset to a dataset that is subject to less re-identification risk. We evaluated this approach based on several state-of-the-art cryptographic algorithms on the Adult dataset [29]. However, this approach is particularly suited in the context of categorical data without semantic relation.

To address the shortcoming of our previous work [36], in [39], we proposed an optimization-based anonymization framework for datasets with both categorical and numerical attributes. The proposed framework is based on clustering the data samples in a diversity-aware fashion to reduce the risks of identity and attribute linkage attacks. Our method achieves anonymity by formulating and solving this problem as a constrained optimization problem, by jointly considering the k-anonymity, l-diversity, and t-closeness privacy models, for the first time to the best of our knowledge. We evaluated our method based on the utility and privacy of data after anonymization in comparison to the original data.

In the privacy-preserving distributed machine learning direction, we presented a framework based on the Extremely Randomized Trees (ERT) algorithm [86] and Secure Multiparty Computation (SMC) techniques. The ERT algorithm has a competitive performance for structured data, where we have independently meaningful attributes, compared to the existing state-of-the-art techniques, e.g., standard deep neural networks [129]. The ERT algorithm is designed for scenarios in which data is stored on a central repository. We extended the ERT algorithm for distributed settings to reduce the amount of raw data leaving a party, which is necessary where we have privacy and legal concerns [38].

In addition, we proposed a layer of SMC on top of the distributed ERT to preserve the privacy of record owners. In Privacy-Preserving Distributed ERT (PPD-ERT) [38],

we adopt an efficient Secure Multiparty Computation (SMC) technique for secure aggregation of partial results in our approach, which is resilient to two colluding parties in the worst case. Moreover, we employed our proposed technique for learning classification models for the prediction of mental health problems in combination with data augmentation [40].

In *k*-PPD-ERT [37], we improved the SMC layer to be resilient to *k* colluding parties, similar to Shamir's secret sharing technique [164], while keeping the communication overhead in the same order as PPD-ERT. Then, we extended the distributed ERT to show its relevance in settings with limited participation of data holder parties without any major loss in classification performance. Our proposed framework offers the opportunity to make a trade-off among performance, privacy, and overhead.

Finally, in [41], we built upon our previous work [37] and proposed a scalable privacy-preserving framework for distributed machine learning based on the extremely randomized trees algorithm, with linear overhead in the number of parties. We used two popular publicly available healthcare datasets for performance evaluation, i.e., the Heart Disease [66] and the Breast Cancer Wisconsin (Diagnostic) [134] datasets. This data represents medical applications where missing values are present, and our algorithm is designed to handle such scenarios. We presented the implementation of our technique on Amazon's AWS cloud and evaluated it in a real-world setting based on the mental health datasets associated with the Norwegian INTROducing Mental health through Adaptive Technology (INTROMAT) project.

4.2 Evaluation of Contributions Against Research Goal

4.2.1 Privacy-Preserving Data Publishing

In this section, we evaluate our proposed anonymization method experimentally and discuss the experimental results. For evaluation, we consider data utility and data privacy criteria and demonstrate their trade-off [64]. Then, we present and discuss the experimental results.

The data analysis task that is going to be performed on the anonymized data is classification. Therefore, the anonymization method should alter the data to the extent that learning high-performance classification models are possible. We train the learning algorithms on both original and anonymized data to evaluate the anonymization method in terms of data utility preservation. Our method preserves the data utility if the classification model learned from altered data has similar performance compared to the one learned from original data.

On the other hand, the anonymized data should be sufficiently altered to avoid the identification of record owners. In this research, we address the record-linkage and attribute-linkage attack models. We consider the property for making samples indistinguishable in the qid group, discussed in *k*-anonymity privacy model, the diversity of values in sensitive attribute, in *l*-diversity, and the frequency of sensitive values, in *t*-closeness.

There is a trade-off between the utility of data and privacy of data in anonymization methods. On the one hand, we can share no data to preserve patients' privacy, but there will be no utility for the data. On the other hand, we can publish the data in its

original format to maximize the data utility, but the privacy of data subjects is going to be violated. Therefore, in anonymization techniques, we require altering the data to the extent that we establish a trade-off between data utility and privacy [64].

4.2.1.1 Experimental Setup

In our experiments, we use the Heart Disease dataset [66], which is one of the popular datasets publicly available on the UCI repository. We utilize Cleveland's processed dataset [71] to predict the presence of heart disease (presence/absence). The dataset contains 282 complete records, and each belongs to one patient. The data includes 13 attributes which we consider in this work.

Quasi-identifiers are the attributes that the adversary can potentially obtain information about them from other sources. In addition to quasi-identifiers, the sensitive attribute should also be identified. In our experiments, we suppose all 13 attributes are quasi-identifiers. Moreover, we select the Boolean attribute for family history of coronary artery disease as the sensitive attribute.

For evaluation of preservation of utility, we split the dataset into train and test sets. We anonymize the training set using our method with soft constraints and train several classification algorithms based on the resulting data. Then, we measure the classification performance on the test set. We also train the same algorithms on the original data and the data anonymized without considering the diversity constraint and measure the performance of the trained classification models on the test set. The comparison of the classification performance results indicate the utility of anonymized data in our method.

In our experiments, we randomly select 200 samples as the train set and the rest as the test set at each round. We repeat the same process for 1000 rounds and report the average results for classification performance. The algorithms used for learning classification models are Extremely Randomized Trees (ERT), Random Forest, XGBoost, Decision Tree, and linear SVM. The measures used for classification performance are F1-score, Accuracy, and Matthews Correlation Coefficient (MCC).

4.2.1.2 Experimental Results

Table 4.1 shows the classification performance results for three different training sets, i.e., original data, anonymized using our method, and anonymized without considering the diversity constraint. For both anonymization methods k is set to 10. The best performance is for the SVM classification models trained based on the original data. The models trained with ERT and random forest algorithms, which are tree-based ensemble learning methods and have randomness in the algorithms, show a good performance on the original and anonymized data.

The classification results for the original data are at a similar level ($\pm 0.5\%$ due to randomness in the algorithms) or higher than the anonymized data. However, since there is a trade-off between privacy and utility in anonymization [64], we may accept a loss in the utility to obtain privacy. The results in Table 4.1 show that our method preserves the information in data that leads to learning high-performance models. Moreover, the classification performance difference between our method and the approach without considering the diversity is negligible. This indicates that

Table 4.1: Classification performance for trained models on three different versions of Heart Disease dataset (Cleveland) [66, 71]

Algorithm	Original Data			Anonymized Data Without Diversity			Anonymized Data by Our Method		
	F1-score	Accuracy	MCC	F1-score	Accuracy	MCC	F1-score	Accuracy	MCC
ERT	81.0%	81.0%	0.615	81.1%	81.4%	0.625	81.0%	81.4%	0.625
Random Forest	82.5%	82.6%	0.647	80.1%	80.4%	0.603	80.0%	80.3%	0.602
XGBoost	78.9%	79.0%	0.573	74.7%	75.1%	0.493	74.7%	75.1%	0.495
Decision Tree	73.8%	73.8%	0.470	68.9%	69.3%	0.372	69.2%	69.8%	0.382
SVM	83.0%	83.1%	0.656	73.3%	73.3%	0.459	72.8%	72.9%	0.449

introducing the diversity constraint in our method does not significantly affect the data utility.

We now evaluate the privacy preservation of our method in Table 4.2. Here, we set the value of k to 10. This means that if the adversary has the values for quasi-identifiers for one patient, he/she can only map his/her information to 10 records. Therefore, through our method, we avoid record-linkage attacks. Second, our method evenly distributes the samples with sensitive value, i.e., having a family history of coronary artery disease, to qid groups. This weakens the confidence of the adversary's inference for identifying a patient with sensitive value.

Table 4.2: Privacy properties of the anonymized data by our method and the approach without diversity

	No Diversity	Our Method
k in k-anonymity	10	10
l in l-diversity	1	2
l in entropy l-diversity	1	1.64
l and c in recursive (c,l)-diversity	l=1, c $\geqslant$ 1	l=2, c $\geqslant$ 4
D in t-closeness	1.06	0.38

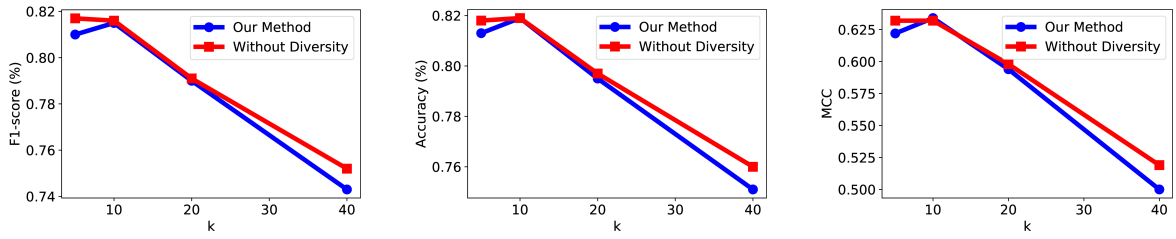
The number of patients with the sensitive value can be different at each round. In our method, in the worst qid group with respect to l-diversity, entropy l-diversity, and recursive (c,l)-diversity, we have two samples with non-sensitive value and eight with the sensitive value. In other words, the proportion of patients with a family history of coronary artery disease in the qid group is 80.0%, which is optimal since the proportion of samples with the sensitive value in the training set at this round was 70.5%. This leads to $l = 2$ in l-diversity, $l = 1.64$ in entropy l-diversity, and $l = 2$ and $c \geqslant 4$ in recursive (c,l)-diversity in Table 4.2. In the worst qid group with respect to the variational distance D in t-closeness, we have six with non-sensitive value and four with the sensitive value, while the proportion of samples with the sensitive value in the dataset at this round was 59.0%. This leads to variational distance $D = 0.38$ in t-closeness.

For the approach without diversity constraint, in the worst qid group with respect to l-diversity, entropy l-diversity, and recursive (c,l)-diversity, we have ten patients with the sensitive value. This leads to $l = 1$ in l-diversity, $l = 1$ in entropy l-diversity, and $l = 1$ and $c \geqslant 1$ in recursive (c,l)-diversity in Table 4.2. This allows the adversary to infer that the patient had a family history of coronary artery disease with 100% confidence. Moreover, in the worst qid group with respect to the variational distance

D in t-closeness, we have nine records with the non-sensitive value and one with the sensitive value. The proportion of samples with the sensitive value in the dataset at this round was 63.0%. This increases the variational distance between the distributions of values in the sensitive attribute in the qid group and the whole dataset to D = 1.06 in Table 4.2.

The results in Table 4.2 demonstrates that by adopting our method, we will have higher l in l-diversity, entropy l-diversity, and recursive (c,l)-diversity. Moreover, the variational distance between the distributions of values in the sensitive attribute for the train set and the qid group is lower in our method. Therefore, regarding the diversity of values in sensitive attributes and the attribute-linkage attack, we observe that introducing the diversity constraint improves patients' privacy.

We also investigate the data privacy and data utility based on different values of k, size of qid groups. For each k, we have 100 rounds that in each we randomly split the data into the train and test sets. The classification performance results are the average results for all rounds. The privacy results are the worst results in all rounds and qid groups. We perform these experiments based on our method and the anonymization approach without the diversity constraint and show the results in Figs. 4.1 and 4.2 for comparison.

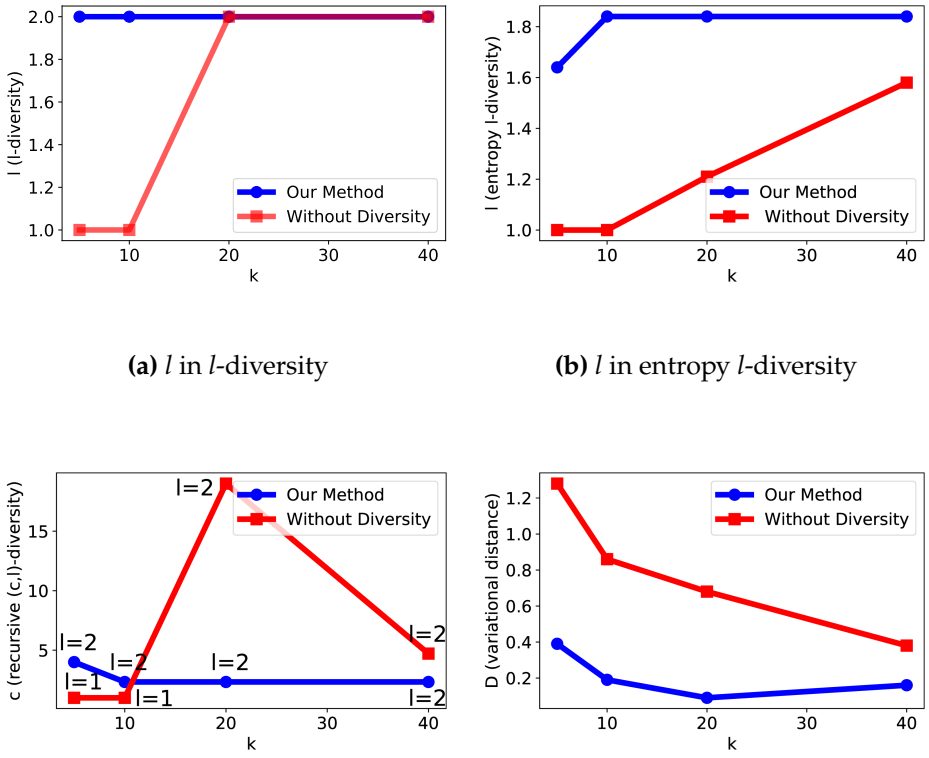


(a) F1-score **(b)** Accuracy **(c)** MCC

Fig. 4.1: The classification performance for anonymized data based on F1-score, Accuracy, and MCC measures for different values of k

Figs. 4.1a-4.1c show the results based on F1-score, Accuracy, and MCC metrics. The patterns in the results show that the higher the qid group size (k), the lower the classification performance. On the other hand, increasing the value of k improves the privacy with respect to the record-linkage attack model. These figures illustrate the trade-off between the privacy and data utility.

The slight increases in the classification performance for the anonymized data by our approach when k is increased from 5 to 10 could be due to the difference in the clustering of samples. When k is larger, satisfying the diversity constraint and distributing samples with the sensitive value to different qid groups for increasing the diversity could be easier. This could lead to creating qid groups with samples that are closer together, which in turn perturbs the data less and increase the classification performance. Moreover, the figures show slight outperformance for trained models based on anonymized data without diversity constraint compared to the trained models based on the data anonymized by our method at certain points. This is due to the fact that in our method, we also consider the attribute-linkage attack model and the privacy models addressing that, which can negatively affect the data utility.



(a) l in l-diversity

(b) l in entropy l-diversity

(c) c in (c,l)-diversity

(d) D in t-closeness

Fig. 4.2: The privacy properties of the data anonymized by our method and the approach without considering the diversity constraint for different values of k

The results in Figs. 4.2a-4.2d exhibit the privacy properties of the anonymized data. Regarding the attribute-linkage attack model, the results display that the data anonymized by our method has higher privacy properties than the anonymized data without diversity constraint. Increasing the value of k significantly improves the diversity and frequency of values in the sensitive attribute, compared to the approach without considering the diversity constraint, but without any major loss in terms of classification performance.

The experimental results show that our method provides privacy protection against record-linkage and attribute-linkage attacks. Furthermore, the utility of the data is retained after anonymization, allowing the learning of high-performance classification models. The slight degradation of utility is the cost for providing patients privacy, which is a common phenomenon in anonymization approaches [64].

4.2.2 *Privacy-Preserving Distributed Machine Learning*

In this section, we evaluate our proposed approach with respect to classification performance, scalability and overhead, and privacy criteria [48].

4.2.2.1 Data

We consider four sets of data for the evaluation in this research. First, we consider two popular publicly available healthcare datasets, i.e., Heart Disease [66] and Breast Cancer Wisconsin (Diagnostic) [134]. For the Heart Disease case, we utilize the processed Cleveland's data [71] to predict the presence or absence of heart disease. In the other case, Wisconsin Diagnostic Breast Cancer (WDBC) data [71] is used to predict breast cancer's diagnosis as benign or malignant.

In addition to the above publicly available datasets, we also consider two mental health detests associated with the INTROMAT project:

- The Depresjon (depression in Norwegian) dataset [85] contains motor activity data from 55 individuals (30 females and 25 males) recorded using an ActiGraph wristband worn on the right wrist. 23 individuals in this dataset have been diagnosed with depression, including both unipolar and bipolar individuals, while the remaining 32 are in the control group. Each individual wore an ActiGraph wristband for an arbitrary number of days, ranging from 5 to 20 days. The condition and control groups were monitored for 291 and 402 days in total, respectively.

- The Psykose dataset [101] contains motor activity data from 54 individuals (23 females and 31 males) recorded using an ActiGraph wristband worn on the right wrist. 22 individuals in this dataset have been diagnosed with schizophrenia, and all used antipsychotic medications, while the remaining 32 are in the control group. Each individual wore an ActiGraph wristband for an arbitrary number of days, ranging from 8 to 20 days. The condition and control groups were monitored for 285 and 402 days in total, respectively.

4.2.2.2 Performance Evaluation Metrics

The performance of the proposed algorithm is evaluated by measuring the F1-score (F1), Accuracy (ACC), and Matthews Correlation Coefficient (MCC), which are defined as follows:

$$F1 = \frac{TP}{TP + 0.5 \cdot (FP + FN)}$$

$$ACC = \frac{TP + TN}{TP + FP + TN + FN}$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where FP, TN, TP and FN definitions are the false positive, true negative, true positive, and false negative, respectively.

4.2.2.3 Evaluation and Results

In this section, we present our evaluation results for classification performance, privacy, overhead, and latency for our proposed framework.

Table 4.3: Classification performance for our proposed method, distributed ID3, and centralized ERT

Dataset	Metric	*k*-PPD-ERT	Distributed ID3	ERT
Heart Disease [66]	Accuracy	80.4%	74.5%	80.4%
	F1-Score	80%	74.3%	80%
Breast Cancer [134]	Accuracy	95.3%	91.3%	95.3%
	F1-Score	95.4%	91.3%	95.4%

Classification Performance for Widely Used Healthcare Datasets To evaluate the classification performance for Heart Disease [66] and Breast Cancer Wisconsin (Diagnostic) [134] datasets, we perform a three-fold cross-validation. We divide the dataset into three parts, and in each round, we use one of the parts as the test set and the rest as the training set and finally report the averaged results. We adopt the F1-score (weighted average) and accuracy as our classification performance metrics. The F1-score is the harmonic mean between the precision and recall metrics, while the accuracy measures the ratio of correctly classified samples. Table 4.3 exhibits the classification performance of our approach, *k*-PPD-ERT, against the distributed ID3 algorithm [75]. We compare our approach against the distributed ID3 [75] since, similar to our approach, it is a state-of-the-art tree-based method that employs SMC techniques for secure aggregation of partial results and addresses classification problems in scenarios where the data is horizontally partitioned. Moreover, the classification performance of the centralized version of ERT is also provided for comparison.

The *k*-PPD-ERT and ERT algorithms follow the same learning procedure. This means that, for both algorithms, the same steps for selecting candidate decision nodes and building the decision tree are followed. In our experiments, we set the same seeds for the random functions and the same learning parameters for both algorithms, e.g., the number of candidate decision nodes. Moreover, the datasets are split into train and test sets in the same way with the same random seed, so these sets are the same for both experiments. Therefore, both algorithms result in the same classification performance, i.e., by following the same procedure, setting the same seeds and parameters, and having the same train and test data.

In our experiments, for our approach, *k*-PPD-ERT, and the ERT algorithm, we learn an ensemble of 25 decision trees. For the number of candidate decision nodes' parameter in the algorithm, we use 5-fold cross-validation on the training set for the model selection (concerning classification performance measured by the F1-score). In the case of the Heart Disease dataset, *k*-PPD-ERT outperforms the distributed ID3 [75] by up to 5.9%. For the Breast Cancer dataset, our approach outperforms the distributed ID3 by up to 4.1%.

Classification Performance for Mental Health Datasets Associated with INTROMAT Project In addition to the widely used public datasets, we also consider the data associated with the INTROMAT project, i.e., Depresjon dataset [85] and Psykose dataset [101]. We use F1-score (weighted average), Accuracy (ACC), and Matthews Correlation Coefficient (MCC) for measuring the classification performance, which are the metrics previously used for performance evaluation on these datasets

Table 4.4: Classification performance (leave one patient out) of different classification algorithms for mental health datasets associated with the INTROMAT project, i.e., Depresjon dataset [85] and Psykose dataset [101]

Algorithms	Depresjon Dataset [85]						Psykose Dataset [101]					
	Augmented Data			Without Augmentation			Augmented Data			Without Augmentation		
	F1-score	ACC	MCC	F1-score	ACC	MCC	F1-score	ACC	MCC	F1-score	ACC	MCC
k-PPD-ERT (Distributed)	76.3%	76.8%	0.518	66.3%	67.0%	0.310	87.9%	88.0%	0.751	81.7%	81.8%	0.623
ID3 (Distributed)	65.1%	65.0%	0.286	65.6%	66.5%	0.296	75.0%	74.8%	0.490	79.3%	79.4%	0.573
ERT (Centralized)	76.3%	76.8%	0.518	66.3%	67.0%	0.310	87.9%	88.0%	0.751	81.7%	81.8%	0.623
Random forest (Centralized)	74.4%	75.1%	0.481	64.3%	64.7%	0.266	90.7%	90.7%	0.807	80.6%	80.7%	0.601
XGBoost (Centralized)	76.2%	76.3%	0.510	64.3%	64.7%	0.265	92.4%	92.5%	0.844	80.7%	80.7%	0.601
Decision Tree (Centralized)	65.7%	65.8%	0.293	60.6%	60.7%	0.191	76.0%	76.0%	0.505	76.1%	76.2%	0.508
Linear SVM (Centralized)	69.5%	69.5%	0.375	68.4%	68.6%	0.349	87.3%	87.2%	0.748	82.8%	82.8%	0.645

Table 4.5: Communication complexity and privacy of different SMC approaches

Approach	Party	Communication		Total Communication (N = number of parties)	Number of Colluding Parties
		Send	Receive		
NOSMC	Data Holders	1	0	$(N-1)\times 1+1\times(N-1)$	1: mediator has the values with no collusion
	Mediator	0	$N-1$		
STSMC	All	2	2	$N\times(2+2)$	2: neighbor parties
k-PPD-ERT	Data Holders	1	0	$(N-1)\times 1+1\times(N-1)$	$k+1$: k data-holder parties and the mediator
	Mediator	0	$N-1$		
Shamir [164]	$k-1$ Parties	N	$N-1$	$N\times(N-1+N-1)+2\times(k-1)$	k parties (k<N)
	One Party	$N-1$	$N-1+k-1$		
	The Rest	$N-1$	$N-1$		

[85, 101]. We consider both the original and augmented data for each dataset. The original data includes the mean and the standard deviation of the activity level along with the proportion of minutes with no activity [85, 101]. The augmented sample reflects the activity level of an individual in a day by locally resampling the raw data from the same individual. The problem related to the difference in the number of recorded days for each individual, which makes the dataset more imbalanced, is addressed by augmentation. Augmentation also addresses the problem of samples with a shorter length, i.e., motor activity signals recorded starting from the middle of the day [40].

We compare our approach against several state-of-the-art machine learning algorithms, including ERT [86], random forest [93], XGBoost [57], Decision Tree [159], and linear SVM algorithm [62]. Table 4.4 shows the classification performance of different algorithms for the INTROMAT data. The results demonstrate that the proposed approach performs on par or better than state-of-the-art techniques. We also compare our approach against the distributed ID3 [75]. For the Depresjon dataset [85], the *k*-PPD-ERT technique outperforms distributed ID3 [75] by 0.7% in terms of F1-score, 0.5% in terms of ACC, and 0.014 in terms of MCC for the original data and by 11.2% in terms of F1-score, 11.8% in terms of ACC, and 0.232 in terms of MCC for the augmented data. For the Psykose dataset [101], the *k*-PPD-ERT technique outperforms distributed ID3 [75] by 2.4% in terms of F1-score, 2.4% in terms of ACC, and 0.05 in terms of MCC for the original data and by 12.9% in terms of F1-score, 13.2% in terms of ACC, and 0.261 in terms of MCC for the augmented data.

Privacy and Overhead of Secure Multi-Party Computation Techniques We now discuss the privacy and communication overhead of our proposed approach. We adopt an SMC technique to avoid direct sharing of the vectors representing the combination of record labels for each candidate decision node with other parties and

the mediator. We compare the communication overhead and privacy of our adopted SMC technique against three other techniques, including the SMC methods employed in [75], i.e., Shamir's technique [164]. Table 4.5 presents the communication overhead and privacy evaluation of each approach. In the table, N is the number of parties, and k is a parameter in *k*-PPD-ERT and Shamir's secret sharing for the minimum number of colluding parties to identify a secret value. The communication overheads in the table are for one round of secure aggregation.

In the first approach (NOSMC), no SMC technique is adopted, and all values are directly shared with the mediator and known to it. This approach has the lowest possible communication cost and number of colluding parties, and, here, it is considered as a baseline. The other approach for the aggregation of partial results is the straightforward SMC (STSMC) approach. In this approach, in the first round, each party aggregates its random mask and its secret value to the received result from the previous party and passes it to the next party, and in the second round, parties subtract their random masks from the aggregated result of the previous round. This method's communication overhead is of the same order as NOSMC, $O(N)$, but it is more robust to collusion. On the other hand, Shamir's secret sharing is an SMC method employed in [75] for secure aggregation. This approach can tolerate the highest number of colluding parties, although it has a high communication overhead, i.e., $O(N^2)$.

Our approach's communication overhead, similar to NOSMC and STSMC, is from order $O(N)$, which is considerably more efficient compared to Shamir's approach with an order of $O(N^2)$. Concerning the number of colluding parties, by adopting our approach, it takes k ($k < N$) data-holder parties and the mediator to collude for identification of the secret values. In our approach, the participation of the mediator for collusion is required to reveal a secret value. The mediator is assumed as an honest party in many scenarios, and in the case of a secret value revelation, we know that the mediator has been involved in the collusion. Shamir's secret sharing requires k ($k < N$) parties to collude for identifying a secret value but suffers from scalability and high communication overhead.

Latency for Our Proof-of-Concept Implementation Finally, we have also implemented our proposed approach on Amazon's AWS cloud to evaluate the latency and scalability of the *k*-PPD-ERT.[1] We consider four scenarios where we change the number of data-holder parties. We consider four datasets, i.e., Heart [66], Breast [134], Depresjon [85], Psykose [101]. For each dataset, the training data (75% of the dataset) is distributed equally among the data-holder parties. The mediator includes a 2 core 2.40 GHz CPU and 512 MB RAM, runs Ubuntu 20.04, and is located in Sweden. The machines in all other locations include a 1 core 2.40 GHz CPU and 512 MB RAM and run Ubuntu 20.04.

The latency results are shown in Figure 4.3. In the first scenario, as shown in Table 4.6, we consider two data-holder parties located in Canada and Germany. Learning one extremely randomized tree through our approach takes $15.9 \pm 1.5, 11.8 \pm 3.5, 3.5 \pm 1.0, 2.4 \pm 0.7$ seconds for Heart, Breast, Depresjon, and Psykose datasets, respectively. In the second scenario, as shown in Table 4.6, we consider five data-holder parties located

[1]The source code of our implementations is available at https://github.com/AminAminifar/kPPDERT_cloud

Table 4.6: The scenarios for our experiments on Amazon's AWS cloud

	Number of data holders	Mediator location	Data holders locations
Scenario 1	2	SE	CA,DE
Scenario 2	5	SE	CA,DE,US,JP,AU
Scenario 3	10	SE	CA,DE,US,JP,AU,SG,IN,KR,FR,EN
Scenario 4	20	SE	CA,DE,US,JP,AU,SG,IN,KR,FR,EN

in Canada, Germany, the United States, Japan, and Australia. Learning one extremely randomized tree through our approach takes $43.5 \pm 4.1, 32.4 \pm 9.6, 9.5 \pm 2.7, 6.6 \pm 2.0$ seconds for Heart, Breast, Depresjon, and Psykose datasets, respectively. In the third scenario, as shown in Table 4.6, we consider ten data-holder parties located in Canada, Germany, the United States, Japan, Australia, Singapore, India, South Korea, France, and England. Learning one extremely randomized tree through our approach takes $43.8 \pm 4.2, 32.6 \pm 9.7, 9.6 \pm 2.7, 6.7 \pm 2.0$ seconds for Heart, Breast, Depresjon, and Psykose datasets, respectively. In the fourth scenario, as shown in Table 4.6, we consider twenty data-holder parties located in Canada, Germany, the United States, Japan, Australia, Singapore, India, South Korea, France, and England, with two parties at each location. Learning one extremely randomized tree through our approach takes $43.6 \pm 4.1, 32.5 \pm 9.7, 9.6 \pm 2.7, 6.8 \pm 2.0$ seconds for Heart, Breast, Depresjon, and Psykose datasets, respectively.

To better understand the reason for the increase and decrease in the latencies reported above and the shape of the graphs in Figure 4.3, it should be noted that the latency depends on the geographical location of the data holders and communication delays. In scenario two, the latency has increased due to the fact that the bottleneck communication distance between the data holders and the mediator is increased. However, the results in scenario three are similar to scenario two because the bottleneck communication distance remains the same. In scenario four, the slight reduction in the latency is due to the fact that we distribute the data among data-holder parties (each party has fewer data samples to process), and the learning process on each party is performed simultaneously and in parallel, similar to big data analysis. These explain the increase of latencies from scenario one to two and the almost flat shapes of the graphs from scenario two to scenario four in Figure 4.3.

Communication Latency of Secure Multi-Party Computation Techniques We also evaluate the communication latency of one secure aggregation round for each SMC approach based on their algorithms, the location of data holders in each scenario, the volume of packets transferred between parties, and the network bandwidth between parties. This shows to what extent adopting each approach can increase the latency.

In this research, we consider the propagation and transmission delays for communication latency [153, 173]. The latency of transferring a packet from P_i to P_j is equal to the sum of propagation and transmission delays and is denoted by $L(P_i, P_j)$. The propagation delay is equal to the distance between parties divided by the velocity of signal propagation, which for unguided transmission through air or space is equal to the speed of light [173]. The transmission delay is equal to the number of bits in the packet divided by the rate of transmission. For transmission delay, we divide the

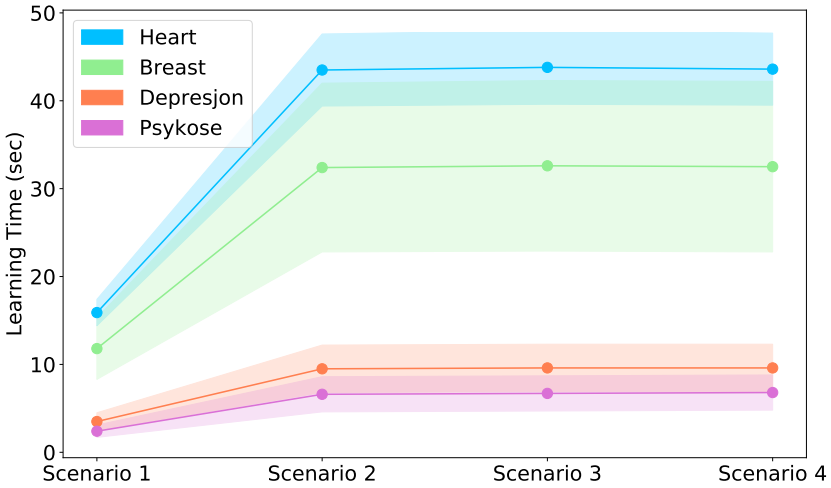


Fig. 4.3: The mean and standard deviation of learning time (ten times performed) of one extremely randomized tree through *k*-PPD-ERT for different datasets in several scenarios on Amazon's AWS cloud

volume of the message to be transferred from P_i to P_j by the bandwidth between these parties.

The network bandwidth between two Amazon machines is measured as 1.05 Mbits/sec using the iPerf tool [14]. When a packet contains two arrays for true and false branches, each including information for five candidate decision nodes for a binary classification task, the volume of each packet is 384 bytes. The volume of the packet depends on the data, i.e., the number of candidate decision nodes and the number of target classes.

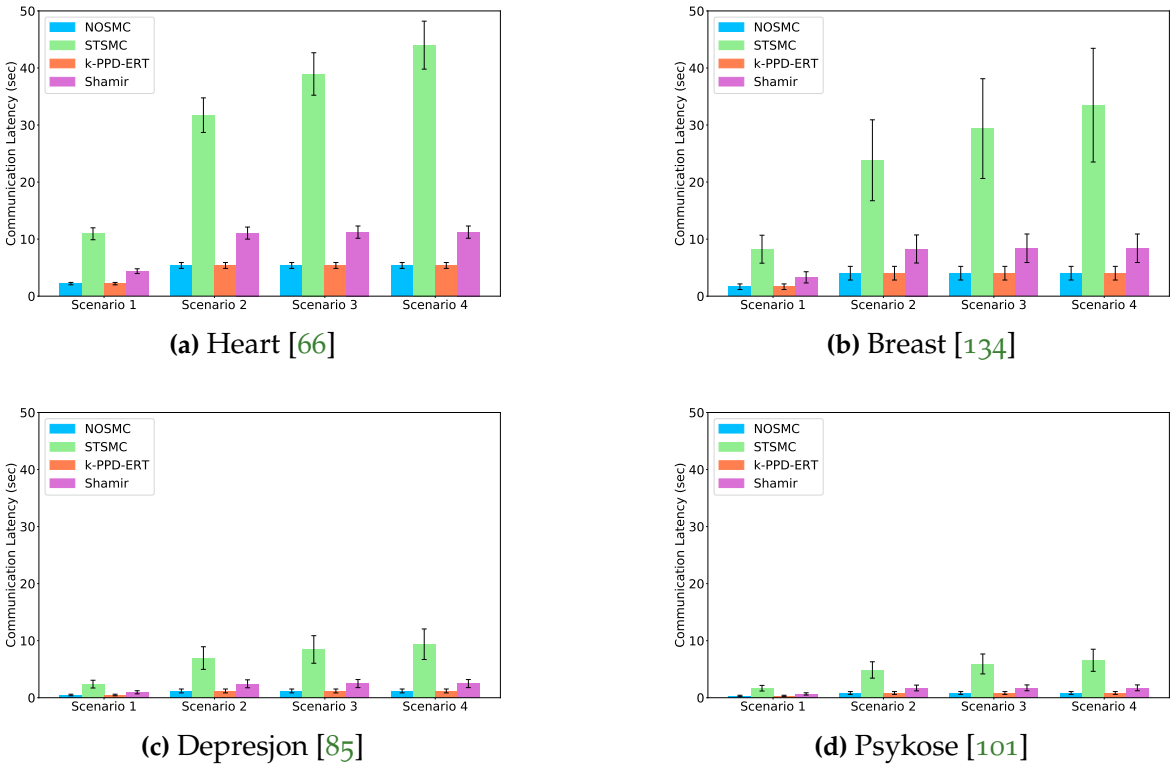
The following are the analysis of communication latency for each method:

- For NOSMC and *k*-PPD-ERT, all parties ($P_i, \forall i \in \{1, \ldots, n\}$) send one message to the mediator (M) in parallel. Since the messages are sent in parallel, the communication latency is equal to the arrival duration of the last message. Therefore, the communication delay is equal to $\max_i L(P_i, M), i \in \{1, \ldots, n\}$.

- For STSMC, we have two loops of message passing between parties in each round, and finally, the first party sends the result to the mediator. Therefore, the communication delay is equal to $2 \cdot (\sum_{i=1}^{n-1} L(P_i, P_{i+1}) + L(P_n, P_1)) + L(P_1, M)$.

- For Shamir, each round of secure aggregation consists of two parts performed sequentially. In the first part, all data-holder parties send one message to $n-1$ parties. When all parties receive these messages, they calculate the intermediate results [75] and send them to the mediator. Therefore, the communication delay is equal to $\max_{i,j} L(P_i, P_j), i, j \in \{i, j \in \{1, \ldots, n\} \mid i \neq j\}$ plus $\max_i L(P_i, M), i \in \{1, \ldots, n\}$.

The number of required secure aggregation operations is also recorded for the experiments in the previous part, i.e., Latency for Our Proof-of-Concept Implementation. The mean and standard deviation of the required number of secure aggregation operations for learning one extremely randomized tree (ten times performed) are $98.8 \pm 9.4, 73.6 \pm 21.9, 22.0 \pm 6.2, 15.4 \pm 4.5$ operations for Heart, Breast, Depresjon, and



(a) Heart [66] **(b)** Breast [134]

(c) Depresjon [85] **(d)** Psykose [101]

Fig. 4.4: The mean and standard deviation of estimated communication latency of different methods for aggregation of secret values in learning one extremely randomized tree (ten times performed) based on different datasets in several scenarios on Amazon's AWS cloud

Psykose datasets, respectively. For estimating the total communication latency of each method for aggregating secret values, the calculated latencies should be multiplied by the number of secure aggregations performed for learning the classification model.

Figure 4.4 shows the mean and standard deviation of communication latency of different methods for aggregation of secret values for each scenario and each dataset. This figure shows that *k*-PPD-ERT has the same communication latency as the NOSMC procedure. Shamir's technique has lower communication latency compared to STSMC, but it still has higher communication latency compared to *k*-PPD-ERT and NOSMC procedures.

It should be noted that the communication latency of these methods should not be confused with the communication overhead presented in Table 4.5. The orders of communication overhead for NOSMC, STSMC, and *k*-PPD-ERT are the same and lower than Shamir's technique. However, since in STSMC, we have two loops of message passing between parties that are performed sequentially, this technique has more delay for a secure aggregation operation. Shamir's technique has two rounds for each SMC operation, and in each round, the message passings are performed in parallel, so it has a lower delay compared to STSMC. For NOSMC and *k*-PPD-ERT, we have one round of message passing that is performed in parallel and has the lowest communication latency.

Finally, we demonstrate that our proposed *k*-PPD-ERT approach provides a solution for the classification of structured data distributed over multiple sources with privacy-preservation considerations, without performance degradation.

4.3 Discussion of Contributions Related to the State-of-the-Art

In this section, we have a review of the literature for our research in both directions. Then, for each direction, we discuss the novelty of our contribution with respect to the related works.

4.3.1 Privacy-Preserving Data Publishing

4.3.1.1 State of the Art

Various approaches for the utilization of data with consideration of privacy and legal concerns are proposed (see Table 4.7). Each approach can be relevant based on the context, problem, and needs. For instance, when we seek the result of data mining analysis, e.g., classification model for prediction, privacy-preserving data mining approaches adopting secure computation techniques are relevant. In particular scenarios, instead of the analysis results, we seek a version of the data that does not violate the privacy of data owners. Depending on the type of data, attack model, and other requirements in the scenario, we may employ statistical disclosure control methods, popular privacy models, e.g., ε-differential privacy [73] or k-anonymity [174], state-of-the-art anonymization methods which are based on generative adversarial networks (GAN) [89], etc.

Table 4.7: Several existing techniques in the state of the art

Approach	Related Studies
Statistical Disclosure Control	[50, 54, 67, 112, 172]
Generative Adversarial Networks (GAN)	[35, 59, 104, 133, 154–156, 190]
Randomization Based	[33, 34, 82, 113, 162]
Optimization Based	[39, 68, 122]
Differential Privacy	[73, 104, 138, 139, 190]
Distributed Anonymization	[102, 103, 105, 130, 141]

Several studies consider encryption algorithms for performing privacy-preserving data mining [52, 53, 88, 90]. These studies mainly employ homomorphic encryption for analysing the data. A comprehensive review of studies in such privacy-preserving data mining approaches for horizontally and vertically partitioned data is provided in [108, 179]. Despite several attempts to improve the computational complexity of homomorphic enrcyption, including somewhat homomorphic encryption and learning-with-error, the existing homomorphic encryption schemes still suffer from extreme computational complexity [147].

The statistical disclosure control (SDC) field studies the approaches for privacy-preserving data publishing of statistical tables or microdata files. Microdata files contain several data records, which consist of multiple values related to one individual, business, etc. Microdata files may contain identifier fields, e.g., name and address, that will not be shared with data recipients to protect the privacy of data owners. Microdata sets, or collections of records including information on individuals, usually are collected through a survey and are products of statistical offices [188]. SDC addresses three types

of disclosure, i.e., identity disclosure, attribute disclosure, and inferential disclosure [61].

Disclosure in SDC is referred to as the undesired linkage of the information to data owners [61, 84]. Identity disclosure is the identification of a data subject from published data. Attribute disclosure occurs if sensitive information about a data subject is revealed through the published data. Inferential disclosure happens when a data subject's information can be more confidently inferred from the released data.

SDC protects data subjects' privacy against the mentioned disclosures by adopting different approaches, from adopting random perturbation methods to introducing limitations on query systems. Several studies [54, 112, 172] consider adding noise to data for privacy protection. Additive noise generally replaces the original sensitive value (v) with a number that is masked with a random value (v+r) and is often used for hiding sensitive numbers in the data [84]. In certain scenarios, the statistical data is released through online query systems. For instance, one SDC solution in such cases is that the system only has access to aggregated data that has been assessed for its sensitivity and disclosure control [61]. Similar to other privacy-preserving data publishing approaches, in SDC, we want to maximize data utility and minimize the risk of privacy violation, which can be contradictory. This is illustrated in Figure 4.5.

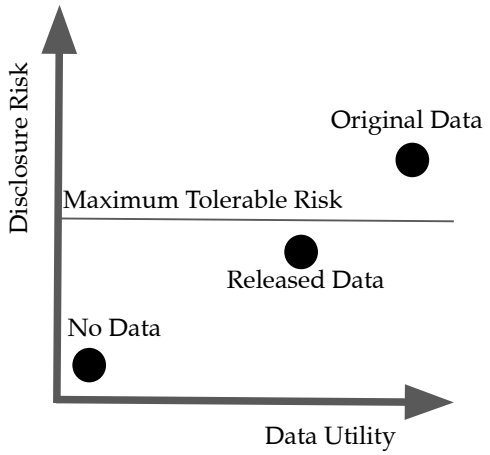


Fig. 4.5: Disclosure risk and data utility objectives [97]

On the other hand, several privacy models address the adversary's probabilistic belief after accessing the published data about the victim's information. This is different from the privacy models that focus on stopping the adversary from linking records, attributes, and tables to a victim data owner. The ε-differential privacy [73] is a privacy model that assures that adding or removing a record from a dataset does not notably affect the outcome of the analysis. Therefore, in such a model, if incorrect data is collected from one data owner, it will not significantly change the result of the anonymization algorithm [84].

Differential privacy is an extensively used statistical method that protects the privacy of data subjects and limits the disclosure of information [74]. Differential privacy techniques usually hide data subjects' information by adding uncertainty in machine learning models [72]. Such techniques also suffer from the trade-off between the data privacy and utility [175]. Moreover, adopting differential privacy techniques and deep

neural networks (DNN) is challenging due to the property of DNNs for memorizing the presence of data samples in the learned models [193].

Recently, generative adversarial network (GAN) models have gained popularity due to their capacity to produce realistic synthetic data in privacy-sensitive applications. DPGAN [190] and PATE-GAN [104] provide differentially private GAN models for medical applications with sensitive data, by adding noise to the model's gradients. However, both PATE-GAN and DPGAN suffer from a significant quality decrease in high-dimensional datasets. Differential privacy techniques introduce a well-known trade-off between performance and privacy level [120], i.e., increasing the noise magnitude improves privacy at the expense of utility loss.

In [59], Choi et al. propose MedGAN to generate high-dimensional synthetic discrete variables by using generative adversarial networks. MedGAN achieves high-performance results for electronic health records and demonstrates the possibility of using synthetic data to preserve patients' privacy. In [155], Pascual et al. show the possibility of generating synthetic epileptic brain activities using generative adversarial networks. Moreover, Pascual et al. in [156] show the possibility of patient re-identification and demonstrate that using synthetic signals produced by the EpilepsyGAN model alleviates the privacy concerns associated with sharing sensitive medical data in the epileptic seizure detection problem.

Anonymization is a popular solution to privacy-preserving data publishing. Anonymization is a process for altering personal data to avoid the identification of the data subjects [99]. In the anonymization field, several attack models are identified and taken into consideration, and privacy models are proposed to address them. For instance, one well-known attack model considered in anonymization is the record linkage model, which is addressed by k-anonymity privacy model [174]. Another important attack model is the attribute linkage model, and this attack is addressed by l-diversity, entropy l-diversity, (c,l)-diversity privacy models [131]. Several methods have been proposed to comply with such privacy models and avoid the associated attacks, i.e., record-linkage and attribute-linkage attacks, e.g., using genetic or kd-trees algorithms to encryption for achieving anonymity [36, 47, 100, 117].

The Mixed-Integer Linear Programming (MILP) frameworks have also been considered for anonymization. In [68], Doka et al. propose the utilization of Mixed-Integer Programming for achieving k-anonymity. Similarly, [122] formulates the anonymization problem in a Mixed-Integer Linear Programming (MILP) framework and achieves k-anonymity based on optimization. This approach uses generalization for anonymization and optimizes the lower and upper bound for each value of quasi-identifiers. Quasi-identifiers are the attributes that the adversary may have information about them, and he/she can use such information to re-identify data subject's records. However, these anonymization methods [68, 122] merely consider k-anonymity and does not prevent the attribute-linkage attack, which is the issue addressed by the l-diversity and t-closeness privacy models.

Finally, in certain scenarios, the data is stored in multiple sites, e.g., several hospitals and medical centers in different cities or countries, and cannot be transferred to one center due to privacy and legal concerns. We seek an anonymized version of data stored on all sites that is consistent with considered privacy models, e.g., k-anonymity. However, the anonymization methods, generally, are designed for scenarios in which

data is held in one center. Thus, we require anonymization methods that support settings in which raw data is distributed. Figure 4.6 shows the distributed (collaborative) anonymization scenario.

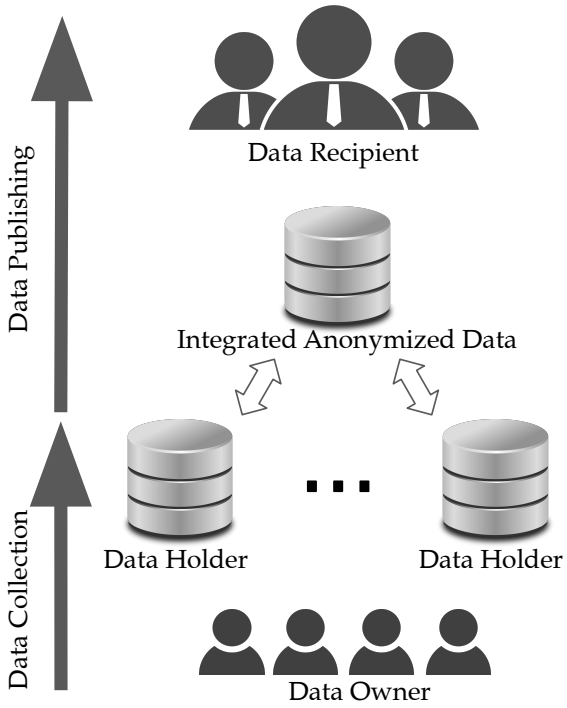


Fig. 4.6: Distributed (collaborative) anonymization

In distributed anonymization, data can be horizontally or vertically partitioned and distributed over multiple data holder parties. Data is horizontally partitioned if each data holder stores a division of all records on it. Data is vertically partitioned if each party holds a part of the records' attributes. Several studies address the problem of anonymization in distributed settings [102, 103, 105, 130, 141]. For instance, in [105], Jurczyk and Xiong propose a method that generates integrated anonymized data based on horizontally partitioned data. Their protocol utilizes secure multiparty computation schemes to minimize the risk of information disclosure among data holders.

4.3.1.2 *Contributions*

For privacy-preserving data publishing, we propose a method to anonymize data to ensure that each record is indistinguishable from, at least, *k-1* other records in the shared data while taking the diversity and frequency of values in the sensitive attribute into consideration. In other words, we propose a method for anonymization of data considering the *k*-anonymity, *l*-diversity, and *t*-closeness privacy models in a unified framework. We formulate the anonymization problem in a constrained optimization framework as a clustering problem, where the diversity and frequency of sensitive values are captured and enforced by constraints. We refer to our proposed method as diversity-aware anonymization, where diversity captures both the diversity concept in the *l*-diversity privacy model and the frequency and distribution of sensitive values in the *t*-closeness privacy model. The experimental results show the preservation of

utility of data for classification tasks and the privacy properties noted in the discussed models.

As discussed above, [68] proposes the utilization of Mixed-Integer Programming for achieving k-anonymity. Similarly, [122] formulates the anonymization problem in a Mixed-Integer Linear Programming (MILP) framework and achieves k-anonymity based on optimization. These anonymization methods [68, 122] merely consider k-anonymity and does not prevent the attribute-linkage attack, which is the issue addressed by the l-diversity and t-closeness privacy models. Therefore, the joint consideration of the k-anonymity, l-diversity, and t-closeness privacy models in such frameworks have not been considered in previous studies.

4.3.2 *Privacy-Preserving Distributed Machine Learning*

4.3.2.1 *State of the Art*

The topic of collaborative learning from distributed data has been discussed in the literature for many years. Several distributed learning techniques have been proposed in the literature [79, 137, 166, 168]. Such techniques contribute to privacy preservation by limiting the amount of data that has to be shared with other parties or transferred to central servers or the cloud.

To explicitly introduce privacy preservation into data mining techniques, several studies [33, 34, 76, 162] adopted randomization techniques to preserve the privacy of individuals. In [33], Agrawal and Srikant present a technique that incorporates noise to raw data before sharing and conducting data mining processes. However, the original values can be estimated based on noise removal techniques. Hence, such techniques do not provide strong security guarantees [96, 110, 111, 181].

In order to perform data mining over data distributed in multiple parties where no private information except the mining results should be disclosed, secure multiparty computation (SMC) is utilized by several studies [69, 108, 179]. The Yao's Millionaires' problem [191], initiated studies about SMC. In SMC, we are usually interested in the result of a computation without knowing about the basic values needed for this computation. Therefore, techniques utilizing SMC usually compute a partial result needed for the learning process without revealing the basic values needed for computation to other parties. Although such methods can provide perfect privacy, the application of SMC can substantially increase communication and computation overheads, leading to efficiency issues, especially when dealing with a high volume of data [75].

Several approaches incorporate encryption techniques for secure computations in the learning tasks [108, 179]. For instance, in [108], it is shown how homomorphic encryption can be used in secure dot protocol, which can be used in particular machine learning algorithms, e.g., for the secure dot-product protocol in SVM algorithm [192]. In another example, homomorphic encryption is used for secure computation in Naïve Bayes Classifier [180]. In another example, commutative encryption can be used for secure union operation, which is required in the association rule mining method [109]. However, the communication and computation overhead of such secure computation techniques are high, which makes them impractical in many real-world

scenarios [157, 181]. Despite attempts to improve the computational complexity of such methods, these schemes still suffer from high computational complexity [147]. In [78], Feigenbaum et al. propose secure multiparty approximation in order to address such issues by accepting a decrease in the accuracy. Nonetheless, their technique for secure and private approximate multiparty computations is still expensive and should be improved to be practical.

In particular, federated learning [46, 114] keeps the data distributed over clients, e.g., mobile devices, and learns a model collaboratively from such decentralized training data. We may also have a central server for the orchestration of the process in this approach. Federated learning observes the data minimization principle in data protection guidelines and limits the privacy risks we face in centralized machine learning. However, the majority of studies in the federated learning domain focus on deep neural network algorithms [107]. In such neural network algorithms, in addition to data-holder parties' contribution, i.e., gradients, sharing model parameters is also a privacy concern. This is due to recent attacks on the neural networks, i.e., membership inference attack [149, 178].

In [186], Wei et al. combine federated learning with differential privacy. The authors propose to add artificial noises to the parameters at the clients' side before model aggregation for providing differential privacy. In [95], Hu et al. propose a federated learning scheme with differential privacy guarantee and evaluate their scheme based on realistic mobile sensing data. In [87], Geyer et al. present a federated optimization algorithm, by considering differential privacy, to hide clients' contributions in the process of learning. This approach requires a sufficiently large number of participating clients to obtain high-performance models. In general, since in differential privacy, we have a trade-off between privacy and data utility, this can degrade the performance of the final machine learning model.

A comprehensive study on security and privacy of federated learning is performed in [143]. The authors extensively discuss the vulnerabilities and threats in federated learning. For instance, one concern in federated learning is data poisoning [49], which is relevant when the attacker incorporates malicious data points to training data in order to maximize the error. In another example, the global model can be used to infer details about training data. Membership Inference attack [149, 178] can be used to determine if a record exists on the training data. Moreover, several techniques and approaches, e.g., secure multiparty computation and differential privacy, for addressing these vulnerabilities and mitigating the identified threats is discussed in the survey.

In the application of federated learning for resource-constraint devices, the communication and computation costs should be considered. In [98], Imteaj et al. surveyed federated learning for resource-constrained IoT devices. This study points out the challenges in applying federated learning on a resource-constrained IoT environment and analyzes the potential solutions for addressing such challenges. The challenges include communication overhead, scalability, energy-efficient training of deep neural networks, limited memory, bandwidth, and energy, heterogeneous hardware, etc.

Federated learning avoids transferring all the raw data to a center for training a machine learning model using conventional centralized methods. In federated optimization for learning a neural network model using the gradient descent algorithm, the clients or parties holding a part of the training data share their gradients with a

central server that orchestrates the learning process. Then, the central server aggregates the received gradients and updates the model parameters based on the received results. In cross-device federated learning, where we have a large number of mobile or IoT devices, communication is often the main bottleneck [107]. In [136], McMahan et al. propose FederatedAveraging (FedAvg) in order to reduce the rounds of communication required for training a deep neural network. First, at each round of communication and update of the global model, FedAvg selects a fraction of clients holding part of the data to participate. Second, the selected clients calculate the gradient and update its parameters locally several times and, then, share the results calculated based on locally updating parameters with the central server. FedAvg addresses the communication cost with such schemes, but this can increase the computation cost on clients' devices. Anyhow, in federated optimization, communication costs dominate compared to central optimization.

In [185], Wang et al. address the problem of training gradient-descent-based machine learning models with data distributed on edge devices. This study decides the best tradeoff between local updates on clients and global updates on the central server to minimize the loss function with a given resource budget. Similar to FedAvg, this approach includes several rounds of local updates on the client device before sharing the client contribution to the central server and global update of network parameters. The local update consumes the computational resources of the client device, and the global update consumes the communication resource of the network. The frequency of global updates is configurable and can occur after one or several local updates. There exist a connection between the frequency of global updates, the performance of the trained machine learning models, and the consumption of resources. This study proposes an algorithm for determining the frequency of global updates to optimize the utilization of available resources.

In [150], Nishio and Yonetani propose FedCS for addressing the challenges related to efficiency in the application of federated learning for clients with limited resources. Similar to FedAvg, in this approach, at every round for updating the parameter on the central server, several clients are randomly asked to download the model from the central server, update them with their local data, and upload it to the central server. However, the limited computational resource of clients delay the updating time, and the limited communication resource delays the upload time. This, in turn, increases the learning time and negatively affects the efficiency of the federated learning process. FedCS addresses this issue by managing the clients based on the condition of their resources. This protocol sets deadlines for download, update, and upload steps for clients. The operator in this protocol, then, selects clients at each round based on the limited time frame and the ability of each client to complete its tasks within the deadline considering its computation and communication resource constraints. This scheme improves efficiency and reduces learning time, as experimentally demonstrated by the authors.

In [121], Li et al. propose FedProx, which is a generalization of FedAvg. FedProx allows variable amounts of work to be performed on client devices depending on the available resources. This is instead of uniform amounts of work performed by each client device proposed by FedAvg. FedProx is particularly practical for resource-constrained federated learning in IoT environments [98].

The majority of previous studies in federated learning domain have focused on deep neural network algorithms. In many applications, tree-based methods can be more accurate than neural networks. Deep neural network algorithms are appropriate solutions when dealing with unstructured data, e.g., for video, audio, and text in [142, 158, 195]. However, the tree-based methods can outperform such algorithms when dealing with structure data, where the data attributes are individually meaningful, and we do not have strong multi-scale structures related to time or space [129]. Therefore, tree-based algorithms are currently being adopted in many applications in which the training data is structured.

Several studies have been done to address the privacy concerns in distributed learning using tree-based data mining techniques. In [181], Vaidya et al. consider the problem of learning decision trees, with random decision trees (RDT) algorithm [77]. They present a technique based on homomorphic encryption and apply it over horizontally and vertically partitioned datasets. However, that approach suffers from extreme computational complexity.

In [124], Lindell and Pinkas propose the utilization of SMC techniques for learning decision trees which is based on ID3 algorithm [159]. In this approach, data is horizontally partitioned among two parties. The number of parties in this method can grow to more than two, but the efficiency and scalability of the technique decrease [157]. Moreover, perturbation techniques may also be used to build approximate decision trees. In [70], collecting all data in one center for mining, from all sources and after applying randomization techniques, leads to decreasing our confidence in the technique's privacy.

Gradient and tree-based algorithms have been employed by several studies in conjunction with strategies related to federated learning [58, 119, 128, 194]. In [194], the authors propose a privacy-preserving distributed data mining method for regression and classification based on the Gradient Boosting Decision Tree (GBDT) algorithm [80]. The trees are trained locally on data-holder parties and passed to the following parties after being modified according to differential privacy requirements [194]. Nevertheless, injecting noise into participants' contribution, model parameters, etc., can increase the learning time and degrade the results of learning due to the trade-off between privacy and data utility [64]. Similarly, in [119], the authors propose a method based on GBDT for distributed scenarios called SimFL. In this framework, each party boosts a number of trees utilizing similarity information using locality-sensitive hashing. However, their privacy model is weaker than secure multiparty computation for improving efficiency, and their model performance is not the same as GBDT but comparable to it [119].

There are other studies that propose tree-based methods that are not gradient-based but are under the name of federated learning, e.g., [127, 177]. In [177], the authors propose a method employing the decision tree algorithm, ID3, that uses the combination of differential privacy and secure multiparty computation for addressing privacy concerns. The model's performance is degraded compared to the performance of the machine learning model in a centralized scenario. In [127], the authors propose a solution based on the random forest algorithm [55, 93]. This method requires a third-party trusted server and employs encryption, which increases the communication and computation overheads [75].

Closely connected to the proposed method in this thesis for the application of

machine learning methods for decentralized data, Emekçi et al. in [75] propose a tree-based method that utilizes a secure multiparty computation technique as an additional layer in their approach to have more confidence about its privacy. Particularly, Shamir's secret sharing [164] is used to aggregate the results received from each party at every step of learning with the ID3 algorithm. The limitation in the incorporation of methods with high communication and computation overheads leads to higher efficiency. However, Shamir's secret sharing technique still introduces major overheads in communication and computation and suffers from the scalability problem.

4.3.2.2 Contributions

In this research direction, we target the problem of learning from data held on multiple sources without explicit sharing of raw information. We assume that the learning data is horizontally partitioned, meaning that different records of data are stored on different sources. We focus on the classification problem and structured health data, which can be stored in spreadsheets. We propose a scalable privacy-preserving framework for distributed machine learning based on the extremely randomized trees algorithm [86], which has a linear overhead in the number of parties and can handle missing values. We refer to our approach as k-PPD-ERT (Privacy-Preserving Distributed Extremely Randomized Trees), in which k is the number of colluding parties in our approach. We use two popular publicly available healthcare datasets for performance evaluation, i.e., the Heart Disease [66] and the Breast Cancer Wisconsin (Diagnostic) [134] datasets. This data represents medical applications where missing values are present, and our algorithm is designed to handle such scenarios. Finally, we present the implementation of our technique on Amazon's AWS cloud and evaluate it in a real-world setting based on the mental health datasets associated with the INTROMAT project [13].

As discussed above, several studies focus on tree-based methods for distributed machine learning which can outperform state-of-the-art deep neural network algorithms when dealing with structure data, where the data attributes are individually meaningful, and we do not have strong multi-scale structures related to time or space [129]. Several studies adopt differential privacy in their approaches, but differential privacy can degrade the performance of the machine learning model due to the trade-off between privacy and data utility [64]. Several studies incorporate inefficient secure computation techniques and homomorphic encryption in the methods, which can substantially increase the communication and computation overheads. We, in particular, propose an efficient privacy-preserving distributed machine learning framework based on the extremely randomized trees and secure multiparty computation, which was not considered in previous studies.

4.4 Discussion of Validity Threats

In this section, we discuss several points and threats to the validity of our proposed techniques.

For our Anonymization framework, one important assumption is that the data records to be published should be truthful. The analysis results directly depend on the published records, and if the records are not accurate, then the results will not

be correct. Therefore, to be able to correctly analyze the data and to produce valid analysis results, the truthfulness of the data records before and after anonymization is necessary.

For our privacy-preserving distributed machine learning framework, we have several assumptions and points that should be taken into consideration before the application.

- Parties should share correct information. In this framework, since only data holder parties are aware of their data values, other parties can not check the correctness of the partial information shared. The resulting model will be negatively affected if data holders send incorrect partial information to the central server or the mediator. This is because the model is trained based on the partial information received from data holders. Therefore, data holder parties should share correct information based on the proposed algorithm for training a high-performance classification model.

- All parties should participate in the learning process according to the algorithm. In our framework, we are using random masks for partial information, and the random mask of each party is canceled by random masks of other parties. Therefore, if a party is supposed to participate in one round, other parties mask their partial information accordingly. If that data holder party does not participate, the result of secure aggregation will not be correct. Hence, if a party is expected to participate in one round for selecting a candidate decision node, then its participation is necessary for training a high-performance classification model.

- The performance of our framework, similar to the ERT algorithm, depends on the data. In case the data is not suitable for learning a high-performance machine learning model with the ERT algorithm, e.g., the data is small, biased, or noisy, then our framework will face the same problem as ERT. Our framework is based on the ERT algorithm and follows the same learning procedure as the ERT algorithm. Our contribution was to extend ERT for learning from decentralized data while protecting the privacy of data owners. Therefore, the data itself plays an important role in training a high-performance machine learning model.

4.5 Reflections on the Research Context

In this section, we discuss how our research contributes to the described context. First, our problem was the privacy issue involved in analyzing healthcare data mainly by using machine learning algorithms. In particular, in the mental health domain context and in connection to the funding project of this thesis, i.e., the INTROMAT project, there exist different sets of data that can not be published and shared with researchers due to privacy concerns. Our proposed solutions address this and facilitate analyzing such sensitive healthcare data.

Using our anonymization technique allow publishing a version of data that addresses record-linkage and attribute-linkage attack models by considering k-anonymity, l-diversity, and t-closeness privacy models. Therefore, when the data holder, in our case hospital, identifies that the above two models of attack threaten the privacy of data

owners, patients in this case, and the above models for protecting privacy suffice their privacy needs, then our anonymization technique can help. By using our technique, an altered version of data could be shared with the data recipient, in INTROMAT case researchers and developers, that can be utilized for analysis purposes.

By using our privacy-preserving distributed machine learning framework, we can learn classification models from decentralized structured healthcare data while protecting patients' privacy. Our framework is designed for scenarios in which data is horizontally partitioned and is distributed among several parties, in our case, hospitals or patients' personal devices. Therefore, in cases where we have the data as described and where only the result of training should be shared, our proposed technique can be used.

The goal of our research in this thesis was to provide a way to analyze existing healthcare data while protecting the privacy of the data owners. Ultimately, this leads to enhancements in healthcare treatment, diagnoses, and decision-making.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Overall Summary of Findings

Artificial intelligence (AI) and its subset Machine Learning (ML) can assist us in increasing the accuracy of decision-making and improving the efficiency and automation of systems. Nowadays, AI made significant progress and is outperforming human experts in in certain domains. Two examples are the classification of rhythms in electrocardiography signals with deep neural networks in [91] and prediction of breast cancer using the AI system presented in [135]; more related studies can be found in [32, 126]. Nevertheless, the application of AI and ML in the healthcare domain raises certain challenges. Security and privacy concerns in the healthcare domain are imperative, and we should carefully consider them in our applications. In data analysis and machine learning applications, we require patients data. However, providing access to such data for analysis purposes can violate the privacy of the patients.

Our main objective, in this study, is to facilitate the possibility of exploiting the sensitive data available in the health domain without privacy violation. That is, to enable performing data mining and machine learning in a privacy-preserving fashion. In particular, the primary motivation for our objective is that this study is a part of the INTROMAT (INtroducing personalized TReatment Of Mental health problems using Adaptive Technology) project, and, in INTROMAT, there is a need for the analysis of sensitive mental health data.

In connection with Research Question 1 and in the context of this research, one of the main challenges is the privacy of data owners or patients in the healthcare domain. On the one hand, if the data in its raw format is published, the adversaries may infer private information about patients even after removing the identifier fields, e.g., by adopting attack models discussed in this thesis. On the other hand, releasing analysis results, e.g., a trained deep neural network model, or neglecting particular details while analyzing the data in a distributed fashion, e.g., not protecting partial information, can pose a threat to the privacy of patients.

We explored the literature and found two primary research directions that address the privacy issue for analyzing healthcare data. The first direction focuses on sharing an altered version of data that can be used for a variety of tasks, e.g., analysis using machine learning techniques or visualization of data. The second direction focuses on analyzing the data, which is decentralized using machine learning techniques. In the first direction, we have two criteria, i.e., data privacy and data utility, for which we

have a trade-off. For the second direction, we have data privacy, model performance, scalability, and overhead criteria that should be considered.

In connection to Research Question 2, the anonymization methods can be considered for publishing health data, for which we have concerns with respect to privacy protection. We have several privacy models that address specified attack models in data publishing. Then, we can anonymize the raw data by altering it based on the privacy model. The problem with such approaches is that we have a trade-off between data privacy and data utility which can degrade the data or pose a threat to the privacy of the data owner. Moreover, such approaches are designed for scenarios in which the data is centralized. One important alternative to address such shortcomings is privacy-preserving distributed machine learning which is the other research direction addressed in this thesis.

In connection to Research Question 3, in scenarios where the health data is distributed among several parties and where we require to train a machine learning model based on such data, privacy-preserving distributed machine learning techniques can be used. In such solutions, privacy, scalability and overhead, and model performance are among the criteria for evaluation. Such techniques usually are designed for a particular machine learning task and are not suitable for other analysis tasks, e.g., calculating the distributions in data or visualization. An alternative solution where we have decentralized health data is distributed anonymization methods. However, anonymization solutions suffer from degradation of data utility due to the data privacy and utility trade-off.

For privacy-preserving data publishing direction, we analyze the data after it is published. Therefore, we require a version of data for publication that preserves the privacy of data owners. Privacy-preserving data publishing methods and approaches are in this direction. In privacy-preserving data publishing, we usually have certain attack models that we try to avoid, e.g., by altering the data before publication. In this thesis, we propose an optimization-based anonymization framework for protecting patients' privacy in publishing datasets that contains categorical and numerical attributes. Our method addresses identity-linkage and attribute-linkage attack models and is based on clustering the data samples in a diversity-aware fashion. The proposed method formulates and solves this problem as a constrained optimization problem and achieves anonymity by jointly considering the k-anonymity, l-diversity, and t-closeness privacy models. We evaluate our framework on popular publicly available structured healthcare data, i.e., the Heart Disease dataset. We show that our optimization-based anonymization framework retains data utility (by evaluation of classification performance) while providing privacy against both record-linkage and attribute-linkage attack models (by evaluation of with respect to k, l, etc. in privacy models).

In the other direction, we perform the analysis without direct access to the data to protect the privacy of data owners. In this direction, the data that is subject to analysis is held in a distributed fashion over multiple parties. Therefore, we require methods that can analyze such data without having direct access to it. In this thesis, we presented a scalable privacy-preserving distributed learning framework along this direction. Our framework is based on Extremely Randomized Trees (ERT) algorithm and Secure Multiparty Computation (SMC) techniques. Furthermore, we extend our framework to be resilient to multiple colluding parties, improve the scalability,

and handle missing values in data. We demonstrate the relevance of the proposed framework by implementing our technique on Amazon's AWS cloud and based on health data, as a proof of concept. We evaluate our proposed technique based on two popular publicly available healthcare datasets, i.e., the Heart Disease and the Breast Cancer Wisconsin (Diagnostic) datasets, and two mental health detests associated with the INTROMAT project, i.e., the Depresjon and Psykose datasets. We show that our framework has the same classification performance as the ERT algorithm but provides privacy in the presence of up to k colluding parties, in a scalable fashion and with a communication overhead of $O(n)$ for a secure aggregation operation.

In conclusion, our research in this thesis was focused on providing the opportunity for utilizing the available healthcare data for analysis while protecting the privacy of data owners. In the long run, this results in improvements in treatment, diagnostics, and decision-making in the healthcare context.

5.2 Directions for Future Work

The topic in this thesis and the research areas of our contributions are broad and can also be combined with other disciplines. Therefore, there are many research subjects in this relation that can be considered for forthcoming studies. In the following, several interesting directions that are relevant to the topic of this thesis are outlined that can be considered in our future works:

- Anonymization methods are designed for scenarios in which data is stored in one center. However, in many real-world applications, the raw data is stored in multiple centers, e.g., at different hospitals or medical centers. Sharing the raw data with one center for anonymizing it can be problematic due to privacy and legal concerns. Therefore, particular anonymization methods are proposed to perform the anonymization in a distributed fashion. One interesting future work is to explore the possibility of extending our anonymization method for distributed scenarios.

- In distributed anonymization, similar to distributed machine learning, we may require to share partial information with other parties for anonymizing the data. This can pose a threat to the privacy of data owners. Therefore, in such scenarios, we may utilize an additional layer of secure multiparty computation (SMC) to protect the privacy of patients. In our future works, we may be able to use our SMC techniques used in PPD-ERT and k-PPD-ERT in distributed anonymization as well.

- We implemented and tested our privacy-preserving distributed machine learning technique on Amazon Web Services to demonstrate its performance. In many applications, the training data is inherently distributed over patients' mobile phones. Our proposed techniques can be employed for learning from such data, while the data does not leave the patient's mobile device for the learning process. One other future work is to implement and use our proposed approach on mobile devices.

- Our privacy-preserving distributed machine learning technique is designed for scenarios in which data is horizontally distributed. One example of a scenario in which data is horizontally distributed is multiple hospitals that hold the same type of data which are collected from different patients at different hospitals. However, in certain scenarios, the training data is vertically partitioned. When the values for different attributes of the same records are stored at multiple places, the data is vertically partitioned. For instance, if the data about age, weight, and height of patients are stored at one hospital and their data for attributes related to a particular type of cancer, e.g., the radius and area of a tumor, is stored in a cancer institute, then, the data is vertically partition. Another future work that can be done is to extend our proposed techniques to be employed in scenarios in which the training data is vertically partitioned.

- One basic assumption in our privacy-preserving distributed machine learning techniques is that data holder parties participate in the learning process by providing correct information. The performance of the learned model is dependant on the correctness of parties' information. In particular scenarios, parties may intentionally or unintentionally collaborate on learning a model by providing incorrect information. One interesting future direction is to address such scenarios. One future research work would be to explore the possibility of detecting such issues and improving the robustness of our proposed techniques against them.

- Regarding the other challenges of using AI in healthcare and in connection with the scope of our INTROMAT project, interoperability is a practical issue in the application of technology in the medical section. There is a vast body of studies in this domain that each of which addresses a particular problem. For example, [144–146, 146, 160, 161] address the interoperability and software architectures in the healthcare domain. In this thesis, we assume that the interoperability is addressed by the data infrastructure manager. One future research area is the interoperability issues involved in the adoption of AI systems in the healthcare domain.

- Regarding the amount of trust that we can put in AI and machines' decision-making, during the past year, there has been a discussion [7, 8, 23, 43], including in the healthcare domain [45]. In many applications, AI can assist human experts in their roles. For instance, if a medical doctor requires to look for tumors in mammogram images, a machine may be used to highlight suspicious areas using segmentation techniques for the medical doctor to double-check. In such scenarios, a human expert makes the final decision to avoid the risk of false negatives.

  Using a machine for decision-making in certain applications can only be beneficial compared to not using them. For example, before the availability of mobile devices for monitoring physiological signals during daily life, individuals could not get helpful recommendations or alerts based on the data that such devices capture and process, but now, such a possibility exists. On the other hand, AI and automatic decision-making have the potential to help humans and are not imposed on us. The application of such technology should be investigated based

on each healthcare scenario. Total delegation of important decision-makings to the machine should be carefully assessed beforehand, e.g., by analyzing the risk based on the scenario [184]. One interesting future research work would be assessing how much trust we can put in AI systems in the healthcare domain.

- Smart wearable devices provide the opportunity to monitor the physiological states and behavioral patterns of patients on a long-term basis and can contribute to early detection and prevention of health disorders, e.g., depression [85]. Examples of wearable devices are Empatica E4 [2], Muse headband [6], ActiGraph [1], fitbit [3], and eGlass [9, 42, 171]. Wearable technologies provide promising solutions for pervasive healthcare with an affordable price by removing time and location constraints [182]. In particular, in [85], Garcia-Ceja et al. propose the detection of depression in individuals based on their motor activity data collected using ActiGraph wristbands. In a similar study, [101], Jakobsen et al. utilize the motor activity data collected using ActiGraph wristbands to detect Schizophrenia in individuals. We plan to consider the data from wearable devices in the future, and we have already considered two datasets associated with the INTROMAT project.

BIBLIOGRAPHY

[1] ActiGraph: medical-grade wearable activity and sleep monitoring solutions for the global research community. https://www.actigraphcorp.com. Accessed: 2020-09-23. 5.2

[2] Empatica E4: a wearable research device that offers real-time physiological data acquisition. https://www.empatica.com/index.html. Accessed: 2020-09-23. 5.2

[3] fitbit. https://www.fitbit.com/no/home. Accessed: 2020-09-23. 5.2

[4] Human Brain Project. https://www.humanbrainproject.eu/en/. Accessed: 2020-09-23. 1.1

[5] Mental health: data and resources. http://www.euro.who.int/en/health-topics/noncommunicable-diseases/mental-health/data-and-resources. Accessed: 2020-09-23. 1.1

[6] muse: the brain sensing headband. http://www.choosemuse.com/. Accessed: 2020-09-23. 5.2

[7] Advancing ai trustworthiness: Updates on responsible ai research. https://www.microsoft.com/en-us/research/blog/advancing-ai-trustworthiness-updates-on-responsible-ai-research/. Accessed: 2022-02-11. 5.2

[8] Building trust in ai. https://www.ibm.com/watson/advantage-reports/future-of-artificial-intelligence/building-trust-in-ai.html. Accessed: 2022-02-11. 5.2

[9] e-Glass: mobile health technologies for real-time detection of epileptic seizures. https://esl.epfl.ch/page-157979-en.html. Accessed: 2020-09-23. 5.2

[10] Gurobi Optimization. https://www.gurobi.com/. Accessed: 2022-02-02. 3.1.2

[11] IBM CPLEX Optimizer. https://www.ibm.com/uk-en/analytics/cplex-optimizer. Accessed: 2022-02-02. 3.1.2

[12] Internet Movie Database (IMDb). https://www.imdb.com/. Accessed: 2020-09-23. 1.1

[13] INTROMAT: introducing mental health through adaptive technology. http://intromat.no/. Accessed: 2020-09-23. 1.1, 4.3.2.2

[14] iperf - the ultimate speed test tool for tcp, udp and sctp. https://iperf.fr/. Accessed: 2021-11-30. 4.2.2.3

[15] London Health Programmes loses unencrypted details of more than 8 million people. https://www.computerweekly.com/news/2240104773/London-Health-Programmes-loses-unencrypted-details-of-more-than-8-million-people. Accessed: 2022-01-20. 2.1.1

[16] Medicare dataset pulled after academics find breach of doctor details possible. https://www.abc.net.au/news/2016-09-29/medicare-pbs-dataset-pulled-over-encryption-concerns/7888686. Accessed: 2020-09-23. 1.1

[17] Missing: Laptop with 8.6million medical records. https://www.thesun.co.uk/archives/news/606395/missing-laptop-with-8-6million-medical-records/. Accessed: 2022-01-20. 2.1.1

[18] National Health Expenditure Accounts (NHEA). https://www.cms.gov/research-statistics-data-and-systems/statistics-trends-and-reports/nationalhealthexpenddata/nationalhealthaccountshistorical. Accessed: 2022-01-24. 1.1

[19] Netflix. https://www.netflix.com/no-en/. Accessed: 2020-09-23. 1.1

[20] NHS laptop loss could put millions of records at risk. https://www.zdnet.com/article/nhs-laptop-loss-could-put-millions-of-records-at-risk/. Accessed: 2022-01-20. 2.1.1

[21] Not so anonymous: Medicare data can be used to identify individual patients, researchers say. https://www.abc.net.au/news/science/2017-12-18/anonymous-medicare-data-can-identify-patients-researchers-say/9267684. Accessed: 2020-09-23. 1.1

[22] Principles of the GDPR. https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/principles-gdpr_en. Accessed: 2020-09-23. 1.1, 1.2

[23] Research collection: Research supporting responsible ai. https://www.microsoft.com/en-us/research/blog/research-collection-research-supporting-responsible-ai/. Accessed: 2022-02-11. 5.2

[24] Research reveals de-identified patient data can be re-identified. http://newsroom.melbourne.edu/news/research-reveals-de-identified-patient-data-can-be-re-identified. Accessed: 2020-09-23. 1.1

[25] The Depresjon Dataset: a motor activity database of depression episodes in unipolar and bipolar patients. https://datasets.simula.no/depresjon/. Accessed: 2021-01-18. 2.3

[26] The Psykose Dataset: a motor activity database of patients with schizophrenia. https://datasets.simula.no/psykose/. Accessed: 2021-01-18. 2.3

[27] The simple process of re-identifying patients in public health records. https://pursuit.unimelb.edu.au/articles/the-simple-process-of-re-identifying-patients-in-public-health-records. Accessed: 2020-09-23. 1.1

[28] UCI Machine Learning Repository: adult data set. https://archive.ics.uci.edu/ml/datasets/adult. Accessed: 2021-08-17. 1.5.1

[29] UCI Machine Learning Repository: adult data set. https://archive.ics.uci.edu/ml/datasets/adult. Accessed: 2021-07-05. 3.1.2, 4.1

[30] Understanding the maths is crucial for protecting privacy. https://pursuit.unimelb.edu.au/articles/understanding-the-maths-is-crucial-for-protecting-privacy. Accessed: 2020-09-23. 1.1

[31] C. C. Aggarwal and S. Y. Philip. *Privacy-preserving data mining: models and algorithms*. Springer Science & Business Media, 2008. 1.5

[32] R. Aggarwal, V. Sounderajah, G. Martin, D. S. Ting, A. Karthikesalingam, D. King, H. Ashrafian, and A. Darzi. Diagnostic accuracy of deep learning in medical imaging: A systematic review and meta-analysis. *NPJ digital medicine*, pages 1–23, 2021. 5.1

[33] R. Agrawal and R. Srikant. Privacy-preserving data mining. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 2000. 4.7, 4.3.2.1

[34] S. Agrawal and J. R. Haritsa. A framework for high-accuracy privacy-preserving mining. In *21st International Conference on Data Engineering*, pages 193–204. IEEE, 2005. 4.7, 4.3.2.1

[35] M. Alzantot, S. Chakraborty, and M. Srivastava. Sensegen: A deep learning architecture for synthetic sensor data generation. In *2017 IEEE International Conference on Pervasive Computing and Communications Workshops*, pages 188–193. IEEE, 2017. 4.7

[36] A. Aminifar, Y. Lamo, K. I Pun, and F. Rabbi. A practical methodology for anonymization of structured health data. In *Proceedings of the 17th Scandinavian Conference on Health Informatics*, pages 127–133, 2019. 4.1, 4.3.1.1

[37] A. Aminifar, F. Rabbi, and Y. Lamo. Scalable privacy-preserving distributed extremely randomized trees for structured data with multiple colluding parties. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2655–2659. IEEE, 2021. 1.5.2, 3.2.2.2, 4.1

[38] A. Aminifar, F. Rabbi, K. I. Pun, and Y. Lamo. Privacy preserving distributed extremely randomized trees. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, pages 1102–1105, 2021. 2.2, 4.1

[39] A. Aminifar, F. Rabbi, V. K. I. Pun, and Y. Lamo. Diversity-aware anonymization for structured health data. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society*, pages 2148–2154. IEEE, 2021. 2.2, 4.1, 4.7

[40] A. Aminifar, F. Rabbi, V. K. I. Pun, and Y. Lamo. Monitoring motor activity data for detecting patients' depression using data augmentation and privacy-preserving distributed learning. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society*, pages 2163–2169. IEEE, 2021. 4.1, 4.2.2.3

[41] A. Aminifar, M. Shokri, F. Rabbi, K. I Pun, and Y. Lamo. Extremely randomized trees with privacy preservation for distributed structured health data. pages 6010–6027. IEEE, 2022. 2.2, 4.1

[42] A. Aminifar, D. Sopic, D. A. Alonso, and R. Zanetti. A wearable system for real-time detection of epileptic seizures, 2020. US Patent App. 16/970,858. 5.2

[43] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016. 5.2

[44] Y. Aono, T. Hayashi, L. Wang, S. Moriai, et al. Privacy-preserving deep learning via additively homomorphic encryption. *IEEE Transactions on Information Forensics and Security*, pages 1333–1345, 2017. 2.1.2

[45] O. Asan, A. E. Bayrak, A. Choudhury, et al. Artificial intelligence and human trust in healthcare: focus on clinicians. *Journal of medical Internet research*, page e15154, 2020. 5.2

[46] S. Baghersalimi, T. Teijeiro, D. Atienza, and A. Aminifar. Personalized real-time federated learning for epileptic seizure detection. *IEEE Journal of Biomedical and Health Informatics*, 2021. 4.3.2.1

[47] R. J. Bayardo and R. Agrawal. Data privacy through optimal k-anonymization. In *21st International conference on data engineering*, pages 217–228. IEEE, 2005. 3.1, 4.3.1.1

[48] E. Bertino, D. Lin, and W. Jiang. A survey of quantification of privacy preserving data mining algorithms. In *Privacy-preserving data mining*, pages 183–205. Springer, 2008. 2.1.2, 4.2.2

[49] B. Biggio, B. Nelson, and P. Laskov. Poisoning attacks against support vector machines. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pages 1467–1474. Omnipress, 2012. 4.3.2.1

[50] A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: the sulq framework. In *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 128–138, 2005. 4.7

[51] S. J. Boccuzzi. Indirect health care costs. In *Cardiovascular health care economics*, pages 63–79. Springer, 2003. 1.1

[52] J. W. Bos, K. Lauter, and M. Naehrig. Private predictive analysis on encrypted medical data. *Journal of biomedical informatics*, pages 234–243, 2014. 4.3.1.1

[53] R. Bost, R. A. Popa, S. Tu, and S. Goldwasser. Machine learning classification over encrypted data. *Cryptology ePrint Archive*, 2014. 4.3.1.1

[54] R. Brand. Microdata protection through noise addition. In *Inference control in statistical databases*, pages 97–116. Springer, 2002. 4.7, 4.3.1.1

[55] L. Breiman. Random forests. *Machine learning*, pages 5–32, 2001. 4.3.2.1

[56] B. Chachuat. Mixed-integer linear programming (milp): Model formulation. http://macc.mcmaster.ca/maccfiles/chachuatnotes/07-MILP-I_handout.pdf. Accessed: 2022–02-20. 3.1.2

[57] T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. Association for Computing Machinery, 2016. 3.2.2, 4.2.2.3

[58] K. Cheng, T. Fan, Y. Jin, Y. Liu, T. Chen, D. Papadopoulos, and Q. Yang. Secureboost: A lossless federated learning framework. *IEEE Intelligent Systems*, pages 87–98, 2021. 4.3.2.1

[59] E. Choi, S. Biswal, B. Malin, J. Duke, W. F. Stewart, and J. Sun. Generating multi-label discrete patient records using generative adversarial networks. In *Machine learning for healthcare conference*, pages 286–305. PMLR, 2017. 4.7, 4.3.1.1

[60] C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin, and M. Y. Zhu. Tools for privacy preserving distributed data mining. *ACM Sigkdd Explorations Newsletter*, pages 28–34, 2002. 2.1.2

[61] Confidentiality and D. A. Committee. Report on statistical disclosure limitation methodology, 2005. 4.3.1.1

[62] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, pages 273–297, 1995. 4.2.2.3

[63] C. Culnane, B. I. Rubinstein, and V. Teague. Health data in an open world. *arXiv preprint arXiv:1712.05627*, 2017. 1.1

[64] J. Davis and O. Osoba. Improving privacy preservation policy in the modern information age. *Health and Technology*, pages 65–75, 2019. 2.1.1, 3.1.2, 4.2.1, 4.2.1.2, 4.2.1.2, 4.3.2.1, 4.3.2.2

[65] N. De Condorcet et al. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Cambridge University Press, 2014. 3.2.1

[66] R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J.-J. Schmid, S. Sandhu, K. H. Guppy, S. Lee, and V. Froelicher. International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American journal of cardiology*, pages 304–310, 1989. 1.5.1, 1.5.2, 2.2, 3.1.2, 4.1, 4.2.1.1, 4.1, 4.2.2.1, 4.3, 4.2.2.3, 4.2.2.3, 4.4a, 4.3.2.2

[67] I. Dinur and K. Nissim. Revealing information while preserving privacy. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 202–210. Association for Computing Machinery, 2003. 4.7

[68] K. Doka, M. Xue, D. Tsoumakos, and P. Karras. k-anonymization by freeform generalization. In *Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security*, pages 519–530. Association for Computing Machinery, 2015. 4.7, 4.3.1.1, 4.3.1.2

[69] W. Du and Z. Zhan. Building decision tree classifier on private data. pages 1–8, 2002. 4.3.2.1

[70] W. Du and Z. Zhan. Using randomized response techniques for privacy-preserving data mining. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 505–510. Association for Computing Machinery, 2003. 4.3.2.1

[71] D. Dua and C. Graff. UCI machine learning repository, 2017. 2.2, 4.2.1.1, 4.1, 4.2.2.1

[72] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Privacy aware learning. *Journal of the ACM*, pages 1–57, 2014. 4.3.1.1

[73] C. Dwork. Differential privacy. In *International Colloquium on Automata, Languages, and Programming*, pages 1–12. Springer, 2006. 4.3.1.1, 4.7, 4.3.1.1

[74] C. Dwork. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pages 1–19. Springer, 2008. 4.3.1.1

[75] F. Emekçi, O. Sahin, D. Agrawal, and A. El Abbadi. Privacy preserving decision tree learning over multiple parties. *Data & Knowledge Engineering*, pages 348–361, 2007. 2.1.2, 3.2.1, 4.2.2.3, 4.2.2.3, 4.2.2.3, 4.2.2.3, 4.3.2.1

[76] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy preserving mining of association rules. *Information Systems*, pages 343–364, 2004. 4.3.2.1

[77] W. Fan, H. Wang, P. S. Yu, and S. Ma. Is random model better? on its accuracy and efficiency. In *Third IEEE International Conference on Data Mining*, pages 51–58. IEEE, 2003. 4.3.2.1

[78] J. Feigenbaum, Y. Ishai, T. Malkin, K. Nissim, M. J. Strauss, and R. N. Wright. Secure multiparty computation of approximations. In *International Colloquium on Automata, Languages, and Programming*, pages 927–938. Springer, 2001. 4.3.2.1

[79] F. Forooghifar, A. Aminifar, and D. Atienza. Resource-aware distributed epilepsy monitoring using self-awareness from edge to cloud. *IEEE transactions on biomedical circuits and systems*, pages 1338–1350, 2019. 2.1.2, 4.3.2.1

[80] J. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001. 4.3.2.1

[81] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*. Springer series in statistics New York, 2001. 3.2.1

[82] W. Fuller. Masking procedures for microdata disclosure. *Journal of Official Statistics*, pages 383–406, 1993. 4.7

[83] B. C. Fung, K. Wang, R. Chen, and P. S. Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys*, pages 1–53, 2010. 2.1.1, 2.1.1

[84] B. C. Fung, K. Wang, A. W.-C. Fu, and S. Y. Philip. *Introduction to privacy-preserving data publishing: Concepts and techniques*. CRC Press, 2010. 1.5, 2.1.1, 3.1.1, 4.3.1.1, 4.3.1.1

[85] E. Garcia-Ceja, M. Riegler, P. Jakobsen, J. Tørresen, T. Nordgreen, K. J. Oedegaard, and O. B. Fasmer. Depresjon: a motor activity database of depression episodes in unipolar and bipolar patients. In *Proceedings of the 9th ACM multimedia systems conference*, pages 472–477, 2018. 2.2, 2.3, 3.2.2.3, 4.2.2.1, 4.2.2.3, 4.4, 4.2.2.3, 4.4c, 5.2

[86] P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine learning*, pages 3–42, 2006. 3.2.1, 3.2.2, 4.1, 4.2.2.3, 4.3.2.2

[87] R. C. Geyer, T. Klein, and M. Nabi. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557*, 2017. 4.3.2.1

[88] R. Gilad-Bachrach, N. Dowlin, K. Laine, K. Lauter, M. Naehrig, and J. Wernsing. Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy. In *International Conference on Machine Learning*, pages 201–210. PMLR, 2016. 4.3.1.1

[89] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2014. 4.3.1.1

[90] T. Graepel, K. Lauter, and M. Naehrig. Ml confidential: Machine learning on encrypted data. In *International Conference on Information Security and Cryptology*, pages 1–21. Springer, 2012. 4.3.1.1

[91] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature medicine*, pages 65–69, 2019. 1.1, 5.1

[92] M. Hartman, A. B. Martin, B. Washington, A. Catlin, N. H. E. A. Team, et al. National health care spending in 2020: Growth driven by federal spending in response to the covid-19 pandemic: National health expenditures study examines us health care spending in 2020. *Health Affairs*, pages 13–25, 2022. 1.1

[93] T. Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, pages 278–282. IEEE, 1995. 4.2.2.3, 4.3.2.1

[94] W. House. Consumer data privacy in a networked world: A framework for protecting privacy and promoting innovation in the global digital economy. *White House, Washington, DC*, 2012. 1.1

[95] R. Hu, Y. Guo, H. Li, Q. Pei, and Y. Gong. Personalized federated learning with differential privacy. *IEEE Internet of Things Journal*, pages 9530–9539, 2020. 4.3.2.1

[96] Z. Huang, W. Du, and B. Chen. Deriving private information from randomized data. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 37–48. Association for Computing Machinery, 2005. 4.3.2.1

[97] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, R. Lenz, J. Longhurst, E. S. Nordholt, G. Seri, and P. Wolf. Handbook on statistical disclosure control. *ESSnet on Statistical Disclosure Control*, 2010. 4.5

[98] A. Imteaj, U. Thakker, S. Wang, J. Li, and M. H. Amini. A survey on federated learning for resource-constrained iot devices. *IEEE Internet of Things Journal*, 2021. 4.3.2.1

[99] Health informatics — Pseudonymization. Standard, International Organization for Standardization, Geneva, CH, 2017. 3.1, 4.3.1.1

[100] V. S. Iyengar. Transforming data to satisfy privacy constraints. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 279–288. Association for Computing Machinery, 2002. 3.1, 4.3.1.1

[101] P. Jakobsen, E. Garcia-Ceja, L. A. Stabell, K. J. Oedegaard, J. O. Berle, V. Thambawita, S. A. Hicks, P. Halvorsen, O. B. Fasmer, and M. A. Riegler. Psykose: A motor activity database of patients with schizophrenia. In *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems*, pages 303–308. IEEE, 2020. 2.2, 2.3, 4.2.2.1, 4.2.2.3, 4.4, 4.2.2.3, 4.4d, 5.2

[102] W. Jiang and C. Clifton. Privacy-preserving distributed k-anonymity. In *IFIP Annual Conference on Data and Applications Security and Privacy*, pages 166–177. Springer, 2005. 4.7, 4.3.1.1

[103] W. Jiang and C. Clifton. A secure distributed framework for achieving k-anonymity. *The VLDB journal*, pages 316–333, 2006. 4.7, 4.3.1.1

[104] J. Jordon, J. Yoon, and M. Van Der Schaar. Pate-gan: Generating synthetic data with differential privacy guarantees. In *International conference on learning representations*, 2018. 4.7, 4.3.1.1

[105] P. Jurczyk and L. Xiong. Distributed anonymization: Achieving privacy for both data subjects and data providers. In *IFIP Annual Conference on Data and Applications Security and Privacy*, pages 191–207. Springer, 2009. 4.7, 4.3.1.1

[106] S. Kahlon, P. Lindner, and T. Nordgreen. Virtual reality exposure therapy for adolescents with fear of public speaking: a non-randomized feasibility and pilot study. *Child and adolescent psychiatry and mental health*, page 47, 2019. 1.2, 2.3

[107] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019. 4.3.2.1

[108] M. Kantarcioglu. A survey of privacy-preserving methods across horizontally partitioned data. In *Privacy-preserving data mining*, pages 313–335. Springer, 2008. 2.1.2, 4.3.1.1, 4.3.2.1

[109] M. Kantarcioglu and C. Clifton. Privacy-preserving distributed mining of association rules on horizontally partitioned data. *IEEE transactions on knowledge and data engineering*, pages 1026–1037, 2004. 4.3.2.1

[110] M. Kantarcioglu and J. Vaidya. An architecture for privacy-preserving mining of client information. In *Proceedings of the IEEE international conference on Privacy, security and data mining-Volume 14*, pages 37–42. Citeseer, 2002. 4.3.2.1

[111] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar. On the privacy preserving properties of random data perturbation techniques. In *Third IEEE international conference on data mining*, pages 99–106. IEEE, 2003. 4.3.2.1

[112] J. J. Kim. A method for limiting disclosure in microdata based on random noise and transformation. In *Proceedings of the section on survey research methods*, pages 303–308. American Statistical Association Alexandria, VA, 1986. 4.7, 4.3.1.1

[113] J. J. Kim, W. E. Winkler, et al. Masking microdata files. In *Proceedings of the Survey Research Methods Section, American Statistical Association*, pages 114–119. Citeseer, 1995. 4.7

[114] J. Konečnỳ, H. B. McMahan, D. Ramage, and P. Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*, 2016. 2.1.2, 3.2.1, 4.3.2.1

[115] C. R. Kothari. *Research methodology: Methods and techniques*. New Age International, 2004. 2.2

[116] E. F. Krause. Taxicab geometry. *The Mathematics Teacher*, pages 695–706, 1973. 3.1.2

[117] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In *22nd International conference on data engineering*, pages 25–25. IEEE, 2006. 3.1, 4.3.1.1

[118] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering*, pages 106–115. IEEE, 2007. 3.1

[119] Q. Li, Z. Wen, and B. He. Practical federated gradient boosting decision trees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4642–4649, 2020. 4.3.2.1

[120] T. Li and N. Li. On the tradeoff between privacy and utility in data publishing. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 517–526. Association for Computing Machinery, 2009. 4.3.1.1

[121] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, pages 429–450, 2020. 4.3.2.1

[122] Y. Liang and R. Samavi. Optimization-based k-anonymity algorithms. *Computers & Security*, page 101753, 2020. 4.7, 4.3.1.1, 4.3.1.2

[123] Y. Lindell. Secure multiparty computation. *Commun. ACM*, pages 86–96, 2020. 3.2.1

[124] Y. Lindell and B. Pinkas. Privacy preserving data mining. *Journal of cryptology*, 2002. 2.1.2, 4.3.2.1

[125] Z. C. Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, pages 31–57, 2018. 3.2.1

[126] X. Liu, L. Faes, A. U. Kale, S. K. Wagner, D. J. Fu, A. Bruynseels, T. Mahendiran, G. Moraes, M. Shamdas, C. Kern, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The lancet digital health*, pages e271–e297, 2019. 5.1

[127] Y. Liu, Y. Liu, Z. Liu, Y. Liang, C. Meng, J. Zhang, and Y. Zheng. Federated forest. *IEEE Transactions on Big Data*, 2020. 4.3.2.1

[128] Y. Liu, Z. Ma, X. Liu, S. Ma, S. Nepal, R. Deng, and K. Ren. Boosting privately: Federated extreme gradient boosting for mobile crowdsensing. In *2020 IEEE 40th International Conference on Distributed Computing Systems*, pages 1–11. IEEE, 2020. 4.3.2.1

[129] S. Lundberg, G. Erion, H. Chen, A. DeGrave, J. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S. Lee. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, pages 56–67, 2020. 3.2.1, 3.2.2, 4.1, 4.3.2.1, 4.3.2.2

[130] Z. Luo, S. Chen, and Y. Li. A distributed anonymization scheme for privacy-preserving recommendation systems. In *2013 IEEE 4th International Conference on Software Engineering and Service Science*, pages 491–494. IEEE, 2013. 4.7, 4.3.1.1

[131] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, pages 3–es, 2007. 3.1, 3.1.1, 4.3.1.1

[132] J. MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, pages 281–297. University of California Press, 1967. 3.1.2

[133] M. Malekzadeh, R. G. Clegg, A. Cavallaro, and H. Haddadi. Mobile sensor data anonymization. In *Proceedings of the international conference on internet of things design and implementation*, pages 49–58. Association for Computing Machinery, 2019. 4.7

[134] O. L. Mangasarian, W. N. Street, and W. H. Wolberg. Breast cancer diagnosis and prognosis via linear programming. *Operations Research*, pages 570–577, 1995. 1.5.2, 2.2, 4.1, 4.2.2.1, 4.3, 4.2.2.3, 4.2.2.3, 4.4b, 4.3.2.2

[135] S. M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G. S. Corrado, A. Darzi, et al. International evaluation of an ai system for breast cancer screening. *Nature*, pages 89–94, 2020. 5.1

[136] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, pages 1273–1282. PMLR, 2017. 2.1.2, 4.3.2.1

[137] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017. 3.2.1, 4.3.2.1

[138] N. Mohammed, D. Alhadidi, B. C. Fung, and M. Debbabi. Secure two-party differentially private data release for vertically partitioned data. *IEEE transactions on dependable and secure computing*, pages 59–71, 2013. 4.7

[139] N. Mohammed, R. Chen, B. C. Fung, and P. S. Yu. Differentially private data release for data mining. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 493–501. Association for Computing Machinery, 2011. 4.7

[140] N. Mohammed, B. C. Fung, P. C. Hung, and C.-k. Lee. Anonymizing healthcare data: a case study on the blood transfusion service. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 1285–1294. Association for Computing Machinery, 2009. 3.1

[141] N. Mohammed, B. C. Fung, P. C. Hung, and C.-K. Lee. Centralized and distributed anonymization for high-dimensional healthcare data. *ACM Transactions on Knowledge Discovery from Data*, pages 1–33, 2010. 4.7, 4.3.1.1

[142] A. Moradi Vartouni, M. Shokri, and M. Teshnehlab. Auto-threshold deep svdd for anomaly-based web application firewall. 2021. 4.3.2.1

[143] V. Mothukuri, R. M. Parizi, S. Pouriyeh, Y. Huang, A. Dehghantanha, and G. Srivastava. A survey on security and privacy of federated learning. *Future Generation Computer Systems*, pages 619–640, 2021. 4.3.2.1

[144] S. K. Mukhiya, U. Ahmed, F. Rabbi, K. I Pun, and Y. Lamo. Adaptation of idpt system based on patient-authored text data using nlp. In *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems*, pages 226–232. IEEE, 2020. 5.2

[145] S. K. Mukhiya, F. Rabbi, K. I Pun, and Y. Lamo. An architectural design for self-reporting e-health systems. In *2019 IEEE/ACM 1st International Workshop on Software Engineering for Healthcare*, pages 1–8. IEEE, 2019. 5.2

[146] S. K. Mukhiya, J. D. Wake, Y. Inal, and Y. Lamo. Adaptive systems for internet-delivered psychological treatments. *IEEE Access*, pages 112220–112236, 2020. 5.2

[147] M. Naehrig, K. Lauter, and V. Vaikuntanathan. Can homomorphic encryption be practical? In *Proceedings of the 3rd ACM workshop on Cloud computing security workshop*, pages 113–124. Association for Computing Machinery, 2011. 4.3.1.1, 4.3.2.1

[148] A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy*, pages 111–125. IEEE, 2008. 1.1

[149] M. Nasr, R. Shokri, and A. Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE symposium on security and privacy*, pages 739–753. IEEE, 2019. 4.3.2.1

[150] T. Nishio and R. Yonetani. Client selection for federated learning with heterogeneous resources in mobile edge. In *ICC 2019-2019 IEEE International Conference on Communications*, pages 1–7. IEEE, 2019. 4.3.2.1

[151] T. Nordgreen, F. Rabbi, J. Torresen, Y. S. Skar, F. Guribye, Y. Inal, E. Flobakk, J. D. Wake, S. K. Mukhiya, A. Aminifar, et al. Challenges and possible solutions in cross-disciplinary and cross-sectorial research teams within the domain of e-mental health. *Journal of Enabling Technologies*, pages 241–251, 2021. 2.3

[152] W. H. Organization et al. Mental and neurological disorders, fact sheet no. 265. december 2001, 2002. 1.1

[153] A. Pahlevan. Multi-objective system-level management of modern green data centers. Technical report, EPFL, 2019. 4.2.2.3

[154] N. Park, M. Mohammadi, K. Gorde, S. Jajodia, H. Park, and Y. Kim. Data synthesis based on generative adversarial networks. *arXiv preprint arXiv:1806.03384*, 2018. 4.7

[155] D. Pascual, A. Aminifar, D. Atienza, P. Ryvlin, and R. Wattenhofer. Synthetic epileptic brain activities using generative adversarial networks. *arXiv preprint arXiv:1907.10518*, 2019. 4.7, 4.3.1.1

[156] D. Pascual, A. Amirshahi, A. Aminifar, D. Atienza, P. Ryvlin, and R. Wattenhofer. Epilepsygan: Synthetic epileptic brain activities with privacy preservation. pages 2435–2446. IEEE, 2020. 4.7, 4.3.1.1

[157] B. Pinkas. Cryptographic techniques for privacy-preserving data mining. *ACM Sigkdd Explorations Newsletter*, pages 12–19, 2002. 2.1.2, 4.3.2.1

[158] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S. Chang, and T. Sainath. Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing*, pages 206–219, 2019. 4.3.2.1

[159] J. R. Quinlan. Induction of decision trees. *Machine learning*, pages 81–106, 1986. 3.2.1, 4.2.2.3, 4.3.2.1

[160] F. Rabbi and Y. Lamo. Development of an e-mental health infrastructure for supporting interoperability and data analysis. pages 59–66, 2019. 5.2

[161] F. Rabbi, Y. Lamo, and W. MacCaull. A model based slicing technique for process mining healthcare information. In *International Conference on Systems Modelling and Management*, pages 73–81. Springer, 2020. 5.2

[162] S. J. Rizvi and J. R. Haritsa. Maintaining data privacy in association rule mining. In *VLDB'02: Proceedings of the 28th International Conference on Very Large Databases*, pages 682–693. Elsevier, 2002. 4.7, 4.3.2.1

[163] L. Rokach. *Pattern classification using ensemble methods*. World Scientific, 2010. 3.2.1

[164] A. Shamir. How to share a secret. *Communications of the ACM*, pages 612–613, 1979. 3.2.2, 4.1, 4.5, 4.2.2.3, 4.3.2.1

[165] M. Shokri, S. H. Khasteh, and A. Aminifar. Adaptive fuzzy watkins: A new adaptive approach for eligibility traces in reinforcement learning. *International Journal of Fuzzy Systems*, pages 1443–1454, 2019. 1.4

[166] R. Shokri and V. Shmatikov. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1310–1321. Association for Computing Machinery, 2015. 4.3.2.1

[167] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, pages 484–489, 2016. 1.1

[168] V. Smith, S. Forte, C. Ma, M. Takáč, M. I. Jordan, and M. Jaggi. Cocoa: A general framework for communication-efficient distributed optimization. *The Journal of Machine Learning Research*, pages 8590–8638, 2017. 4.3.2.1

[169] D. Sopic, A. Aminifar, A. Aminifar, and D. Atienza. Real-time classification technique for early detection and prevention of myocardial infarction on wearable devices. In *Biomedical Circuits and Systems Conference, 2017 IEEE*, pages 1–4. IEEE, 2017. 1.4, 2.1.2

[170] D. Sopic, A. Aminifar, A. Aminifar, and D. Atienza. Real-time event-driven classification technique for early detection and prevention of myocardial infarction on wearable systems. *IEEE transactions on biomedical circuits and systems*, pages 1–11, 2018. 1.4, 2.1.2

[171] D. Sopic, A. Aminifar, and D. Atienza. e-glass: A wearable system for real-time detection of epileptic seizures. In *2018 IEEE International Symposium on Circuits and Systems*, pages 1–5. IEEE, 2018. 5.2

[172] N. Spruill. The confidentiality and analytic usefulness of masked business microdata. *Proceedings of the Section on Survey Research Methods, 1983*, pages 602–607, 1983. 4.7, 4.3.1.1

[173] W. Stallings. *Data and computer communications*. Pearson Education India, 2007. 4.2.2.3

[174] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, pages 557–570, 2002. 1.1, 1.1, 3.1, 4.3.1.1, 4.3.1.1

[175] A. Team et al. Learning with privacy at scale. *Apple Machine Learning Journal*, pages 1–25, 2017. 4.3.1.1

[176] N. Titov, B. Dear, O. Nielssen, L. Staples, H. Hadjistavropoulos, M. Nugent, K. Adlam, T. Nordgreen, K. H. Bruvik, A. Hovland, et al. Icbt in routine care: a descriptive analysis of successful clinics in five countries. *Internet interventions*, pages 108–115, 2018. 1.2, 2.3

[177] S. Truex, N. Baracaldo, A. Anwar, T. Steinke, H. Ludwig, R. Zhang, and Y. Zhou. A hybrid approach to privacy-preserving federated learning. In *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security*, pages 1–11. Association for Computing Machinery, 2019. 4.3.2.1

[178] S. Truex, L. Liu, M. Gursoy, L. Yu, and W. Wei. Demystifying membership inference attacks in machine learning as a service. *IEEE Transactions on Services Computing*, pages 2073–2089, 2019. 4.3.2.1

[179] J. Vaidya. A survey of privacy-preserving methods across vertically partitioned data. In *Privacy-preserving data mining*, pages 337–358. Springer, 2008. 2.1.2, 4.3.1.1, 4.3.2.1

[180] J. Vaidya and C. Clifton. Privacy preserving naive bayes classifier for vertically partitioned data. In *Proceedings of the 2004 SIAM international conference on data mining*, pages 522–526. SIAM, 2004. 4.3.2.1

[181] J. Vaidya, B. Shafiq, W. Fan, D. Mehmood, and D. Lorenzi. A random decision tree framework for privacy-preserving data mining. *IEEE transactions on dependable and secure computing*, pages 399–411, 2013. 4.3.2.1

[182] U. Varshney. Pervasive healthcare and wireless health monitoring. *Mobile Networks and Applications*, pages 113–127, 2007. 5.2

[183] D. Vigo, G. Thornicroft, and R. Atun. Estimating the true global burden of mental illness. *The Lancet Psychiatry*, pages 171–178, 2016. 1.2, 2.3

[184] D. Vose. *Risk analysis: a quantitative guide*. John Wiley & Sons, 2008. 5.2

[185] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan. Adaptive federated learning in resource constrained edge computing systems. *IEEE Journal on Selected Areas in Communications*, pages 1205–1221, 2019. 4.3.2.1

[186] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. Quek, and H. V. Poor. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security*, pages 3454–3469, 2020. 4.3.2.1

[187] K. West. Patient medical information at risk from stolen computers. *Missouri medicine*, pages 10–12, 2014. 2.1.1

[188] L. Willenborg and T. De Waal. *Statistical disclosure control in practice*. Springer Science & Business Media, 1996. 4.3.1.1

[189] L. A. Wolsey and G. L. Nemhauser. *Integer and Combinatorial Optimization*. 3.1.2

[190] L. Xie, K. Lin, S. Wang, F. Wang, and J. Zhou. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739*, 2018. 4.7, 4.3.1.1

[191] A. C.-C. Yao. How to generate and exchange secrets. In *27th Annual Symposium on Foundations of Computer Science*, pages 162–167. IEEE, 1986. 3.2.1, 4.3.2.1

[192] H. Yu, X. Jiang, and J. Vaidya. Privacy-preserving svm using nonlinear kernels on horizontally partitioned data. In *Proceedings of the 2006 ACM symposium on Applied computing*, pages 603–610, 2006. 4.3.2.1

[193] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, pages 107–115, 2021. 4.3.1.1

[194] L. Zhao, L. Ni, S. Hu, Y. Chen, P. Zhou, F. Xiao, and L. Wu. Inprivate digging: Enabling tree-based distributed data mining with differential privacy. In *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*, pages 2087–2095. IEEE, 2018. 4.3.2.1

[195] C. Zhuang, T. She, A. Andonian, M. Mark, and D. Yamins. Unsupervised learning from video with deep neural embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9563–9572, 2020. 4.3.2.1

Part II

ARTICLES

PAPER A

SCIENTIFIC PAPER I: A PRACTICAL METHODOLOGY FOR ANONYMIZATION OF STRUCTURED HEALTH DATA

Amin Aminifar, Yngve Lamo, Violet Ka I Pun, and Fazle Rabbi

In Proceedings of the 17th Scandinavian Conference on Health Informatics (SHI), Linköping University Electronic Press. 2019.

A Practical Methodology for Anonymization of Structured Health Data

Amin Aminifar[1], Yngve Lamo[1], Ka I Pun[1,2], and Fazle Rabbi[3]
[1]Western Norway University of Applied Sciences, Bergen, Norway {amin.aminifar, yngve.lamo, ka.i.pun}@hvl.no
[2]University of Oslo, Oslo, Norway
[3]University of Bergen, Bergen, Norway {fazle.rabbi}@uib.no

Abstract
Hospitals, as data custodians, have the need to share a version of the data in hand with external research institutes for analysis purposes. For preserving the privacy of the patients, anonymization methods are employed to produce a modified version of data for publishing; these methodologies shall not reveal the patient's information while maintaining the utility of data. In this article, we propose a practical methodology for anonymization of structured health data based on cryptographic algorithms, which preserves the privacy by construction. Our initial experimental results indicate that the methodology might outperform the existing solutions by retaining the utility of data.

Keywords
Anonymization, privacy-preserving data sharing, structured health data, data mining, cryptography.

1 INTRODUCTION

Hospitals, nowadays, are increasingly collecting data from patients as it allows to provide better treatment and precise diagnosis. Analyzing such data by sharing it with researchers can be useful for society. However, the shared data should not compromise the privacy of the individuals. Removing the identifier fields like name and address, is not enough for preserving privacy from certain attacks, e.g., linking attack [1]. Such attacks can re-identify the individuals and reveal specific information based on the raw data. One solution to this is that the data custodians, e.g., hospitals, anonymize such data before sharing.

1.1 Anonymization

Having access to high-quality data is a necessity for medical and pharmaceutical experts and researchers for facilitating decision making. Sharing healthcare data can benefit several parties, including hospitals, medical and pharmaceutical researchers outside the hospital, patients, and data mining researchers. Hospitals, more precisely, medical experts and researchers, can make use of the result of data analysis performed by external research centers. Medical practitioners and pharmaceutical researchers outside the hospital need the data for analysis leading to informed decision making. Patients, indirectly through this, will receive better services from hospitals and medical centers outside the hospital. Finally, data mining researchers will have access to real health data and use them as benchmarks for their methods. However, raw health data contains patients' sensitive information and can compromise their privacy. Therefore, health data holders are looking for anonymization techniques that prepare the health data for release, while keeping the quality of data and preserving the privacy of patients.

Patients consider hospitals as trustworthy entities, so they are willing to share their data with hospitals. Nevertheless, this trust is not transitive to other entities such as research centers outside the hospitals. Many believe that removing specific identifying information including name, telephone, and social security number, is sufficient for releasing the data. As several previous studies show [1, 2], merely removing the identifier fields is deficient for preserving the privacy of individuals. Sweeney [1] shows, an adversary by having limited information from an individual, say from another dataset, can match other attributes, called quasi-identifiers (QID), and reidentify the individual. Three prominent examples about this are provided in [1, 3-6, 7].

At some points, hospitals, instead of analyzing the data by themselves and sharing the analysis results, e.g., statistics or classifiers, need to share the data with external research centers, e.g., universities and pharmaceutical companies, in order to make use of other professional resources outside. Therefore, they should share the data with external researchers specialist in data analysis. Moreover, having the data give much freedom to external research centers for data analysis. Frequent requests from hospitals for providing statistical information and fine-tuning the data mining results is not feasible [2].

1.2 Motivational Example

Hospitals are considered to be the trusted party, and thus have access to the raw data. However, they, in general have limited resources for some specific data analyses. Therefore, it is common to delegate the analysis process to external research institutions. To preserve privacy of individuals, data should be anonymized in the hospitals, and only anonymized data can be shared with external institutions or released to the public. Note that any party external to hospitals can be the adversary, as illustrated in Figure 1.

After analyzing the published data, the results will be released to the hospital, which can be, for instance, a discriminator function as the outcome of the learning from anonymized data. With this function, the hospital can

classify new raw records as follows: firstly, the new record should be anonymized in the same way as the published data anonymized; secondly, the new anonymized record can be passed to discriminator function, shared by the external institutions, for classification. In this way, hospitals can make use of services outside without compromising the privacy of their patients.

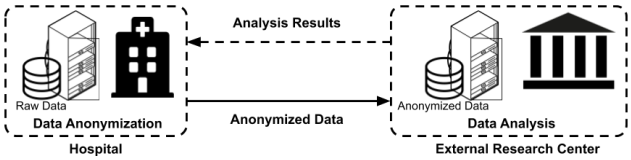


Figure 1 medical data anonymization and analysis.

In this paper, we propose a methodology to anonymize structured health data based on cryptographic algorithms and without assumptions on the characteristics of the encryption method. Adopting cryptographic algorithms guarantees privacy preservation by construction. Moreover, the comparison results of the data utility between raw and anonymized data generated based on our proposed methodology and the existing methods are promising. The proposed methodology can have a complementary role in combination with previous methods as well.

The organization of the rest of this article is as follows. In Section 2, a short review of previous methods for anonymization of the structured data is provided. Section 3 addresses the proposed approach for anonymization, along with providing some preliminary information. Section 4 presents the necessary information and settings concerning the experiments. Section 5 is devoted to the evaluation and experimental results. Finally, in Section 6, conclusions and future research directions are provided.

2 RELATED WORKS

For research purposes, data custodians need to release a version of data in a way that individuals cannot be re-identified. Statistical and multi-level databases are among the other approaches for addressing these kinds of needs. Despite the assumption made in [1], statistical disclosure control [8] is an active research area for addressing today's needs to provide accurate information while protecting the privacy of the various parties involved [9, 10]. On the other hand, anonymization techniques are between other solutions in this regard. For sharing the data records, microdata, in anonymization, we try to irreversibly alter the personal data until the re-identification of data subjects is no longer possible [11].

Anonymization methods provide a new class of acceptable solutions to this problem. Typically, anonymization techniques for structured data make use of generalization method. More specifically, such techniques modify or generalized the data records components in a way that a data record is hardly distinguishable from others. Some important related studies are k-anonymity [1], l-diversity [12], t-closeness [13], and LKC-privacy [2]. To date, k-anonymity remains the most widely known privacy model for anonymization during the past two decades. To thwart privacy threats, k-anonymity privacy model generalizes and suppresses data record components or features into equivalence groups so that any record is indistinguishable from at least k other data records [14, 2]. However, in this method, when the dimensionality of data is high, most of the data must be generalized or suppressed for achieving k-anonymity; this negatively affects the utility of data and degrades it [2]. Other methods try to rectify the issue, for instance, by imposing limitations on the problem, such as the supposition of limited knowledge of the adversary about the patient. For example, in the LKC-privacy model, the adversary is supposed to have only the values for a part of the QID attributes of the victim's record, L attributes [2].

The proposed approach in this study described in Section 3 tries to provide a solution for the above problem, i.e., anonymization of structured data. The problem here is the same as the one described in the above research studies, while we formally define the problem in Section 3. The proposed approach of this study for the solution is completely different from that provided in the previous studies. This study investigates the application of cryptographic algorithms, which is distinguishing from previous works. The majority of previous studies consider performing machine learning over homomorphically-encrypted data [15-18], while in this paper we do not make such assumptions.

3 METHOD

In this section, we first define the anonymization problem and then propose a practical solution to this problem. Two main concerns for data anonymization is privacy preservation and data utility, discussed in the following subsection. There is often an inherent trade-off between these two metrics. At one extreme, all data can be released, for maximizing the utility, and as a result, violate the privacy entirely. On the other extreme, releasing no data can maximize privacy; however, there would be no data utility [14]. The proposed methodology in this section provides an approach for addressing this problem, which is based on cryptography for data anonymization.

3.1 Problem Definition

In the following two subsections we discuss the two criteria for this problem. We define the anonymization problem as guaranteeing the privacy while maximizing the utility of the data for the statistical and machine learning data analysis.

3.1.1 Privacy Preservation

This section explains the privacy threats for sharing the raw information through an example; there exist two types of privacy concerns, namely identity linkage and attribute linkage. Table 1 shows the raw patient data. The raw data does not have the identifier features but is still vulnerable to the violation of privacy. Education, sex, and age are quasi-identifying attributes [1]. Disorder is the sensitive feature that the adversary does not know about the victim patient and tries to infer it. Finally, there exists one class for every record in the dataset.

Based on the following assumptions about the adversary, there are two types of privacy concerns to address. As mentioned in Introduction, the adversary is assumed to have anonymous data for all the patients. Moreover, the adversary has parts of the victim patient's record, in its raw format; this information is part of or all the quasi-identifying attributes and is only for one patient. The extent

ID	Quasi-identifier (QID)			Sensitive	Class
	Education	Sex	Age	Disorder	
1	BSc	F	40	Depression	cat. #1
2	MSc	M	53	ADHD	cat. #1
3	HS-grad	F	40	Depression	cat. #2
4	PhD	F	31	Social Anxiety	cat. #1
5	MSc	M	31	Bipolar	cat. #2

Table 1 An example of raw data table.

of adversary's information about the victim patient is assumed differently in different studies. For instance, in [1] the author for k-anonymity model assumes that the adversary has all the values for quasi-identifying attributes, but in [2] in LKC-privacy model limits the adversary's information to only the values of L number of the quasi-identifying attributes. Finally, the adversary does not know about the sensitive information of the victim and is willing to infer it. Accordingly, hospitals face two common privacy concerns [2] described below:

- **Identity Disclosure:** If the record is highly specific, matching the records with the victim's information is simple, which lead to the inference of the patient's sensitive information. For instance, in Table 1, the raw data table, if the adversary knows that the victim's education and age are 'MSc' and '31', respectively, then s/he confidently identifies that record number 5 is the victim's and infers that the victim's disorder is 'Bipolar'.
- **Attribute Disclosure:** If with some quasi-identifying attributes, the sensitive value happens repeatedly, it makes the inference of the sensitive value easy, although the accurate data record of the victim is not identifiable. For instance, in Table 1, the raw data table, if the adversary knows that the victim's sex and age are 'F' and '40', respectively, then, s/he can match the victim's information to records number 1 and 3. However, since both sensitive values for record number 1 and 3 are the same, 'Depression', then, the adversary can infer with 100% confidence that the victim's disorder is 'Depression'.

3.1.2 Utility of Data

To make sure that the anonymization method is not degrading the utility of the data, a comparison of the utility of raw data with the anonymous data is essential. The classification performance is a valid criterion for making a comparison between the utility of data before and after anonymization. Since the main concern of this study is sharing the data for data mining purposes, the difference between the classification performance for the raw and anonymized data shows the excellence and efficiency of the algorithm.

Information gain [19] is another criterion that indicates how much a method may degrade or improve the data quality for every feature of the data individually. Information gain was first introduced for decision trees and is based on the information entropy [20]. Nevertheless, since it does not consider the correlation and combination of the attributes, it is not as reliable as the classification performance criterion.

3.2 The Anonymization Method

For the preservation of privacy, we seek a function to map each unique record of raw data to another unique record, different from the raw record and in the same feature space. The anonymized data records must be different enough to prevent identity and attribute attacks. The anonymized data must not allow the possibility for the adversary to map back to the raw data. Therefore, the utilized function for mapping the raw data must not be reversible, or in other words must be one-way, for those with whom the anonymized data will be shared.

Cryptography fulfills the privacy objectives by construction. Mapping a number to another unique number through one-way functions is the main purpose of cryptography. Therefore, by such intrinsic features of cryptographic algorithms, we can make sure of the preservation of privacy criterion without taking further actions. Since, after encryption, the values would be meaningless numbers for the adversary, and it is not possible for one without a key to map back to the raw data.

Due to the objective of this study for anonymization of the structured health data containing categorical and numerical features, encryption is entirely feasible. Since in both cases there are numbers, more precisely category numbers and numerical values, which are mapped to other numbers. The sensitive attribute is not an exception and is encrypted as well. Normalization of data is the second phase of anonymization. Normalization, in addition to the positive impact on learning, reinforces preserving the privacy as this is a hashing phase after encryption.

As described earlier the anonymization methods should fulfill two criteria, namely privacy preservation and data utility. Application of cryptographic algorithms guarantees the privacy preservation criterion by construction. However, we also need to make sure about the performance of this methodology in regard to the utility of data. In this study, we experimentally show that our proposed methodology for anonymization of structured data is also efficient regarding the data utility.

The utility of the data needs to be preserved and this is related to the correlation of attributes and labels in data samples and the algebraic distance of samples from each other. To ensure satisfying this criterion after encryption and normalization of the dataset, the utility of the data is compared before and after anonymization based on two measurements described previously in this section. If the results for raw and anonymized data are close, then in addition to the preservation of the privacy, there also would be a confidence about the utility of data. A loss to a limited extent in the utility of data is acceptable as there exists a trade-off between privacy and data utility in data anonymization [14].

4 EVALUATION SETUP

4.1 Dataset for Evaluation of the Methodology

Adult dataset [21] is the de facto benchmark for evaluation of anonymization models [2, 12, 22-27]. In this dataset, the samples belong to two different classes; the rates of the positive and negative classes are 76.07% and 23.93%. The total number of records is 48842 (train=32561, test=16281), and the train and test sets were separated when

shared. Each record has 14 attributes, including eight categorical and six numerical ones. Furthermore, the dataset contains missing values. This study considers all the attributes as QID, although it is possible to suppose part of them as QID, like in [2] which considers marital-status as sensitive and others as QID attributes.

4.2 Encryption Algorithms

For the evaluation of the proposed approach, four cryptographic algorithms, including two from symmetric and two from asymmetric encryption systems, are considered. The symmetric algorithms are Advanced Encryption Standard (AES) and Data Encryption Standard (DES); the input and output data and key size for each is 128 and 64 bits, respectively. The Asymmetric algorithms are RivestShamirAdleman (RSA) and ElGamal, which both are also homomorphic over multiplication. The key size for each is 2048 and 1024 bits, respectively. All the keys are generated randomly for every iteration of experiments, based on the toolbox.

4.3 Comparison with K-Anonymity

In order to evaluate the results of our methodology, a comparison between the results of the proposed and former methods of anonymization is necessary. K-anonymity is one of the most popular privacy models. In [28], the authors propose Mondrian for obtaining k-anonymity. This study considers this work for anonymizing the data based on the k-anonymity model for comparison with the proposed methodology. The corresponding parameters for these methods are k, set of QID, and the mode of the algorithm, which can be either relaxed or strict. In the experiments, k is set to 10 and QID are set to all the attributes, and the results for both relaxed and strict modes are provided.

4.4 Utility Measure

Two measures employed here for evaluation of data utility are information gain and classification performance. Information gain is based on information entropy and is being used to evaluate how well an attribute alone predicts the classes for samples in comparison to other attributes. In other words, every attribute is used to categorize samples, then the information entropy of the classes of the categorized samples are calculated. The lower the entropy of the samples' classes in each category of samples categorized based on that specific attribute, the higher the information gain of that attribute. The loss of information gain after anonymization can indicate the extent of deterioration of data. However, since this measure does not consider the combination of attributes, it is not as reliable as classification performance. For calculation of classification performance, we used the geometric mean of the ratios of correctly classified samples to the number of samples in that particular class. Geometric mean is the only correct average for normalized measurement [29].

5 EVALUATION RESULTS

To evaluate the efficiency of our proposed methodology, the Adult dataset [21] is anonymized with the proposed methodology by this paper. Afterward, the information gain and classification performance for raw and anonymized data are calculated and recorded for comparison and evaluation. The closer the results of raw and anonymized data the higher our confidence to the anonymization methodology regarding the preservation of data utility.

As mentioned earlier, after one level of encryption, we need to normalize the data in order to obtain the anonymized data. The normalization method used for our experiments is min-max normalization:

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}, \quad (1)$$

where x_{new} is the normalized value of x, the encrypted number, and x_{min} and x_{max} are respectively minimum and maximum values of the corresponding column in the matrix of encrypted numbers.

Furthermore, for more certainty, the experiments for every method iterates for ten times, and the average results are measured. In every, iteration the key for encryption algorithms are generated separately and randomly, to ensure the classification results are independent of the keys.

5.1 Information Gain

The encryption is particularly useful when the attribute is numerical since, concerning the learning results, encryption of the number of categories is similar to mapping each specific category number to another random number specific for that category; therefore, for such attributes, encryption is not a necessary process. However, in this study's experiments, we encrypted all the attributes and normalized the data afterward. Before and after anonymization by this methodology, the information gain of categorical attributes always remains the same, because of the characteristics of this measure, so there would be no points in reporting them here.

Table 2 presents the information gain of the numerical attributes of raw and anonymized datasets; the results are from the average for ten independent iterations. The results in this table show that our anonymization methodology does not reduce the information gain of the numerical attribute unless in attributes 1 and 13, albeit negligible. Considering the information gain, the proposed methodology preserves the utility of data to a considerable extent.

5.2 Classification Performance

In addition to the anonymization with the proposed methodology of this paper, for comparison, we also anonymized the Adult dataset with Mondrian multidimensional k-anonymity approach [28]. Then, the results of these methods, along with the raw dataset, are used for learning a classification function. The learning algorithm used in this research is the random forest algorithm [30]. The training and testing sets for the raw data and anonymized data based on our proposed methodology are the same as published in [21]. However, for Mondrian multidimensional k-anonymity approach for every iteration, we take 70% of randomly shuffled data as the training set and the remaining 30% as the testing set; splitting the train and test sets for learning and evaluation in this setting is conventional and valid, considering the studies in the field [31].

Table 3 exhibits the classification performance based on the geometric mean measure, i.e., geometric mean of the ratios of correctly classified samples to the number of samples in

DATASET	INFORMATION GAIN					
	Attribute 1	Attribute 3	Attribute 5	Attribute 11	Attribute 12	Attribute 13
Raw Data	0.09754	0	0.09328	0.11452	0.05072	0.05814
Anonymized Data (RSA Alg.)	0.096839	0	0.093379	0.118778	0.051108	0.057001
Anonymized Data (ElGamal Alg.)	0.097563	0	0.093507	0.118503	0.05157	0.056479
Anonymized Data (DES Alg.)	0.096581	0	0.093452	0.118688	0.051163	0.05713
Anonymized Data (AES Alg.)	0.096755	0	0.093434	0.118512	0.051061	0.057325

Table 2 Information Gain for numerical attributes of the Adult dataset [21] before and after anonymization.

that particular class, for raw and anonymized data obtained adopting several methods. All the results in Table 3 are the average of the results of ten independent iterations. The information gain table provided in this article is calculated using WEKA software [32]. The difference between the classification performance of anonymized data based on our methodology and the raw data is less than 3%; our proposed methodology, however, outperforms Mondrian multidimensional k-anonymity regarding classification performance for adult dataset as the results show that the geometric mean measure for our anonymization approach, in the worst case, is higher for at least 5%.

Dataset	Geometric Mean (%)
Raw Data	75.37
Anonymized Data (K-Anonymity Mondrian [21], Relaxed, K=10, QI = Attribute 1-14)	67.87
Anonymized Data (K-Anonymity Mondrian [21], Strict, K=10, QI = Attribute 1-14)	68.08
Anonymized Data (RSA Alg.)	73.30
Anonymized Data (ElGamal Alg.)	73.59
Anonymized Data (DES Alg.)	73.22
Anonymized Data (AES Alg.)	73.57

Table 3 Classification performance based on geometric mean for all methods for Adult dataset [21].

The results in Tables 2 and 3 show that our proposed methodology only deteriorates the data to a negligible extent depending on the application; this is justifiable as there exists a cost for preserving the privacy of individuals. A comparison between the classification results of the anonymized data obtained by our proposed methodology and Mondrian multidimensional k-anonymity approach, in Table 3, indicates that our methodology outperforms theirs as the prediction results, with the same learning algorithm, are more accurate. Moreover, the results suggest that maintaining the utility of data is not dependent on a specific cryptographic algorithm.

Comparisons of two data utility measures for raw and anonymized data show that this methodology preserves the relations of values in the data table to a considerable extent. Therefore, analyses dependent on the relations of the data attributes to each other, and the labels are feasible and supported, e.g., learning tasks through machine learning algorithms. Such analyses are not dependent on the exact values in raw data since the anonymization changes the range of values for each attribute. The anonymized data is a matrix of numbers, likewise to the raw data, and it can be used the same way as the raw data. Moreover, regarding the privacy concerns described in the Problem Definition Section, if one manages to change the values in the raw data until the adversary cannot map it back to the original values, then the desired purpose is achieved. Using cryptographic algorithms for anonymization along with the fundamental property of these algorithms, i.e., mapping numbers by one-way injective functions, dismisses the described privacy concerns, in other words, matching data values from what the adversary has and what is published as anonymized data is not possible.

6 CONCLUSION

In this study, we investigated the approach of anonymizing the structured health data by utilizing cryptographic algorithms, which is, to the best of our knowledge, the first application of these algorithms in anonymization. Anonymization methods must fulfill two criteria, namely privacy preservation and data utility. We evaluated the presented methodology on the de facto benchmark dataset for anonymization. The results are promising and indicate that such an approach may be employed in real-world applications by the healthcare sector. However, similar to the majority of anonymization techniques, our proposed methodology impacts the quality of data mining results, even though we have shown that this degradation is less than the previous works in the data anonymization domain. This methodology is particularly practical for anonymizing the data for data mining applications. For future works, the applicability of this approach may be investigated for unstructured types of health data, e.g., physiological signals. Moreover, automatic de-identification of clinical notes and overcoming the particular challenges is another closely related research area that can be tied up with natural language processing [33, 34]. Further studies on the field mentioned above would be analogous to this study and worthwhile.

7 ACKNOWLEDGMENTS

This research is supported by INTROMAT (INtroducing personalized TReatment Of Mental health problems using Adaptive Technology) Project [35], funded by the Norwegian Research Council (259293/o70). The paper is partially supported by SIRIUS: Centre for Scalable Data Access.

8 REFERENCES

[1] Sweeney, L., 2002. k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05), pp.557-570.

[2] Mohammed, N., Fung, B., Hung, P.C. and Lee, C.K., 2009, June. Anonymizing healthcare data: a case study on the blood transfusion service. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1285-1294). ACM.

[3] Health Data in an Open World. (2019, July 17). Retrieved from https://arxiv.org/ftp/arxiv/papers/1712/1712.05627.pdf

[4] Research reveals de-identified patient data can be re-identified. (2019, July 17). Retrieved from https://about.unimelb.edu.au/newsroom/news/2017/december/research-reveals-de-identified-patient-data-can-be-re-identified

[5] The simple process of re-identifying patients in public health records. (2019, July 17). Retrieved from https://pursuit.unimelb.edu.au/articles/the-simple-process-of-re-identifying-patients-in-public-health-records

[6] Understanding the maths is crucial for protecting privacy. (2019, July 17). Retrieved from https://pursuit.unimelb.edu.au/articles/understanding-the-maths-is-crucial-for-protecting-privacy

[7] Narayanan, A. and Shmatikov, V., 2008. Robust de-anonymization of large datasets (how to break anonymity of the Netflix prize dataset). University of Texas at Austin.

[8] Willenborg, L. and De Waal, T., 1996. Statistical disclosure control in practice. Springer Science & Business Media.

[9] Domingo-Ferrer, J. and Montes, F. eds., 2018. Privacy in Statistical Databases: UNESCO Chair in Data Privacy, International Conference, PSD 2018, Valencia, Spain, September 26–28, 2018, Proceedings. Springer.

[10] Fienberg, S.E. and van der Linden, W.J., Statistics for Social and Behavioral Sciences.

[11] International Organization for Standardization: ISO 25237:2017 Health informatics – Pseudonymization.

[12] Machanavajjhala, A., Gehrke, J., Kifer, D. and Venkitasubramaniam, M., 2006, April. l-diversity: Privacy beyond k-anonymity. In 22nd International Conference on Data Engineering (ICDE'06) (pp. 24-24). IEEE.

[13] Li, N., Li, T. and Venkatasubramanian, S., 2007, April. t-closeness: Privacy beyond k-anonymity and l-diversity. In 2007 IEEE 23rd International Conference on Data Engineering (pp. 106-115). IEEE.

[14] Davis, J.S. and Osoba, O., 2019. Improving privacy preservation policy in the modern information age. Health and Technology, 9(1), pp.65-75.

[15] Bost, R., Popa, R.A., Tu, S. and Goldwasser, S., 2015, February. Machine learning classification over encrypted data. In NDSS (Vol. 4324, p. 4325).

[16] Graepel, T., Lauter, K. and Naehrig, M., 2012, November. ML confidential: Machine learning on encrypted data. In International Conference on Information Security and Cryptology (pp. 1-21). Springer, Berlin, Heidelberg.

[17] Bos, J.W., Lauter, K. and Naehrig, M., 2014. Private predictive analysis on encrypted medical data. Journal of biomedical informatics, 50, pp.234-243.

[18] Gilad-Bachrach, R., Dowlin, N., Laine, K., Lauter, K., Naehrig, M. and Wernsing, J., 2016, June. Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy. In International Conference on Machine Learning (pp. 201-210).

[19] Quinlan, J.R., 1986. Induction of decision trees. Machine learning, 1(1), pp.81-106.

[20] Shannon, C.E., 1948. A mathematical theory of communication. Bell system technical journal, 27(3), pp.379-423.

[21] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[22] Bayardo, R.J. and Agrawal, R., 2005, April. Data privacy through optimal k-anonymization. In 21st International conference on data engineering (ICDE'05) (pp. 217-228). IEEE.

[23] Fung, B.C., Wang, K. and Philip, S.Y., 2007. Anonymizing classification data for privacy preservation. IEEE transactions on knowledge and data engineering.

[24] Iyengar, V.S., 2002, July. Transforming data to satisfy privacy constraints. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 279-288). ACM.

[25] Mohammed, N., Fung, B., Wang, K. and Hung, P.C., 2009, March. Privacy-preserving data mashup. In Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology (pp. 228-239). ACM.

[26] Wang, K. and Fung, B., 2006, August. Anonymizing sequential releases. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 414-423). ACM.

[27] Wang, K., Fung, B.C. and Philip, S.Y., 2007. Handicapping attacker's confidence: an alternative to k-anonymization. Knowledge and Information Systems.

[28] LeFevre, K., DeWitt, D.J. and Ramakrishnan, R., 2006, April. Mondrian multidimensional k-anonymity. In ICDE (Vol. 6, p. 25).

[29] Fleming, P.J. and Wallace, J.J., 1986. How not to lie with statistics: the correct way to summarize benchmark results. Communications of the ACM, 29(3), pp.218-221.

[30] Ho, T.K., 1995, August. Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition (Vol. 1, pp. 278-282). IEEE.

[31] MITCHELL, T. M. (2017). Machine learning. New York, McGraw Hill.

[32] Witten, I.H., Frank, E., Hall, M.A. and Pal, C.J., 2016. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

[33] Deleger, L., Molnar, K., Savova, G., Xia, F., Lingren, T., Li, Q., Marsolo, K., Jegga, A., Kaiser, M., Stoutenborough, L. and Solti, I., 2013. Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. Journal of the American Medical Informatics Association, 20(1).

[34] Liu, Z., Tang, B., Wang, X. and Chen, Q., 2017. De-identification of clinical notes via recurrent neural network and conditional random field. Journal of biomedical informatics, 75.

[35] INTROMAT (INtroducing personalized TReatment Of Mental health problems using Adaptive Technology). (2019, July 17). Retrieved from https://intromat.no/

PAPER B

SCIENTIFIC PAPER II: DIVERSITY-AWARE ANONYMIZATION FOR STRUCTURED HEALTH DATA

Amin Aminifar, Fazle Rabbi, Violet Ka I Pun, and Yngve Lamo

The 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 2021.

Diversity-Aware Anonymization for Structured Health Data

Amin Aminifar[1], Fazle Rabbi[1,2], Violet Ka I Pun[1,3], and Yngve Lamo[1]

***Abstract*— Patients' health data are captured by local hospital facilities, which has the potential for data analysis. However, due to privacy and legal concerns, local hospital facilities are unable to share the data with others which makes it difficult to apply data analysis and machine learning techniques over the health data. Analysis of such data across hospitals can provide valuable information to health professionals. Anonymization methods offer privacy-preserving solutions for sharing data for analysis purposes. In this paper, we propose a novel method for anonymizing and sharing data that addresses the record-linkage and attribute-linkage attack models. Our proposed method achieves anonymity by formulating and solving this problem as a constrained optimization problem which is based on the *k*-anonymity, *l*-diversity, and *t*-closeness privacy models. The proposed method has been evaluated with respect to the utility and privacy of data after anonymization in comparison to the original data.**

I. Introduction

Patients' data is private and may contain sensitive information, e.g., information about a health condition. Such data may not be shared with other parties in their raw format due to privacy and legal concerns [1], [2]. However, such data may be required for analysis purposes to provide value to medical experts and utilized for analysis by adopting privacy-preserving data mining or privacy-preserving data sharing approaches depending on the particular application and scenario.

Privacy-preserving data mining techniques perform the analysis without direct access to the data. Several approaches adopt homomorphic encryption techniques for learning tasks [3], [4]. However, such methods suffer from communication and computation overhead and are not always practical [5]. Several state-of-the-art techniques modify and adapt algorithms for learning from distributed data without sharing data and sacrificing privacy [6], [7], [8], [9], [10]. Nevertheless, each algorithm should be extended to support privacy-preserving distributed learning. Moreover, learning a classification model from data is not the only objective in particular scenarios, and a version of data may be required to be published, e.g., for medical expert inspection and visualization.

Privacy-preserving data sharing techniques share an altered version of data for analysis. Several studies add noise to data and perturb it before sharing [11], [12], [13]. However, the utility of data will be negatively affected by the perturbation of data. On the other hand, privacy will not be preserved if the noise added is not sufficient. Moreover, noise removal approaches pose a threat to the privacy of such methods [14], [15]. Several studies adopt neural networks and generative adversarial networks (GAN) [16] for altering the data before sharing [2], [17], [18], [19]. Such approaches mainly focus on particular time-series data and data in wearable devices' applications [20], [21], [22], [23].

Anonymization methods also alter the data to avoid identifying data subjects in such datasets [24]. Previous studies proposed several privacy models for anonymization, e.g., *k*-anonymization [25], *l*-diversity [26], *t*-closeness [27], *LKC*-privacy [28]. The data holder selects a model based on the scenario, utility, and privacy requirements. Several methods have been proposed to comply with such privacy models and avoid the associated attacks, i.e., record-linkage and attribute-linkage attacks, e.g., using genetic algorithms to kd-trees algorithms for generalization and achieving anonymity [29], [30], [31], [32], [33].

In particular, [34] proposes the utilization of Mixed-Integer Programming for achieving *k*-anonymity. Similarly, [35] formulates the anonymization problem in a Mixed-Integer Linear Programming (MILP) framework and achieves *k*-anonymity based on optimization. This approach uses generalization for anonymization and optimizes the lower and upper bound for each value of quasi-identifiers, which are the attributes that the adversary may have information about for identification. However, these anonymization methods [34], [35] merely consider *k*-anonymity and does not prevent the attribute-linkage attack, which is the issue addressed by the *l*-diversity and *t*-closeness privacy models. Therefore, the joint consideration of the *k*-anonymity, *l*-diversity, and *t*-closeness privacy models in such frameworks have not been considered to date.

In this paper, we propose a method to anonymize data to ensure that each record is indistinguishable from, at least, *k-1* other records in the shared data while taking the diversity and frequency of values in the sensitive attribute into consideration. In other words, we propose a method for anonymization of data considering the *k*-anonymity, *l*-diversity, and *t*-closeness privacy models in a unified framework. We formulate the anonymization problem in a constrained optimization framework as a clustering problem, where the diversity and frequency of sensitive values are captured and enforced by constraints. We refer to our proposed method as diversity-aware anonymization, where diversity captures both the diversity concept in the *l*-diversity privacy model and the frequency and distribution of sensitive values in the *t*-closeness privacy model. The experimental results show the preservation of utility of data for classification tasks and the privacy properties noted in the discussed models.

The rest of this paper is organized as follows: Section II covers the background with respect to *k*-anonymity,

[1]Western Norway University of Applied Sciences, Bergen, Norway firstname.lastname@hvl.no
[2]University of Bergen, Bergen, Norway
[3]University of Oslo, Oslo, Norway

l-diversity, *t*-closeness, and their corresponding attack models. We formulate our proposed anonymization method in the constrained optimization framework in Section III. Section IV provides the experimental results for evaluation of our method. Section V concludes our paper.

II. Background

In this section, we briefly review the record-linkage and attribute-linkage attack models. In addition, we discuss three popular privacy models addressing such attacks, i.e., *k*-anonymity, *l*-diversity, and *t*-closeness.

In the record-linkage and attribute-linkage attack models, we suppose that a version of data after removing the identifier attributes of patients, e.g., name and address, is shared with a data recipient. At the same time, the adversary has access to the data shared with the data recipient. This data contains several attributes through which a patient (record owner) can be identified, i.e., quasi-identifiers, and it is assumed that the adversary has the exact value of these attributes for the victim patient. Finally, there is a sensitive attribute in the data, e.g., family history for a health pathology, that the adversary is interested in knowing about.

To explain this attack models, we use Tables Ia and Ib as an example. The 2nd-4th columns are considered as quasi-identifiers and refer to age, the number of children, and the smoking state of the patient (*Yes*/*No*). The 5th column is a sensitive attribute capturing the state of the HIV disease for the patient (*Positive*/*Negative*). Table Ia represents shared data after removing the identifier features. Suppose that Table Ia is shared with the data recipient. If the adversary knows that the victim is 37 years old, has two children, and smokes, he/she can easily match his/her information to one of the records (record one in Table Ia) and identify that the victim is diagnosed with HIV. The record-linkage attack occurs by matching the adversary's information (quasi-identifiers) with published data for identifying the patient's (record owner) sensitive information [36].

The *k*-anonymity privacy model was proposed to address the record-linkage attack model. A dataset is *k*-anonymous when the values of quasi-identifiers for each record are the same as the values for at least *k-1* other records in the data. In this way, the adversary can only match his/her information with at least *k* records. Table Ib shows a 3-anonymous version of the same data in Table Ia. For instance, in our example in Table Ib, if the adversary knows that the victim is 37, has two children, and smokes, he/she can merely match his/her information with a qid group containing the records of three patients, records *1-3*.

While the *k*-anonymity model guarantees that a patient is only matched with a qid group, however, this model does not guarantee the protection of patients' privacy against attribute-linkage attacks. That is, *k*-anonymity does not consider the diversity of values for the sensitive attribute in each qid group. In this example, in the first qid group, all the values for the sensitive attribute are *Positive*. Therefore, in the first qid group, the adversary can infer that the victim patient is diagnosed with HIV by matching quasi-identifiers' information. The attribute-linkage attack model occurs in situations where the diversity of values for the sensitive attribute is low. As a result, the adversary may infer the sensitive attribute with high confidence.

To address the attribute-linkage attack, the *l*-diversity model proposes that every qid group should have a least *l* distinct values for the sensitive attribute. For instance, in Table Ib, if the adversary matches his/her information with the third qid group, he/she can not identify that the patient was diagnosed with HIV for sure because both *Negative* and *Positive* values are in that qid group. However, this does not consider the confidence of the adversary's inference properly. For example, if we have both *Negative* and *Positive* values in all qid groups, we have 2-divers data, but if the proportion of *Positive* values in one qid group is high, the adversary can infer that the patient is diagnosed with HIV with high confidence. The entropy *l*-diversity and recursive *(c,l)*-diversity are proposed to address such issues [26].

Entropy *l*-diversity is one of the existing privacy models to address the distribution of values in the sensitive attribute. A data table meeting the following condition for each qid group is entropy *l*-diverse:

$$-\sum_{s\in S} P(qid,s)\log(P(qid,s)) \geq \log(l), \qquad (1)$$

where *S* is the set of values for sensitive attribute, and *P(qid, s)* is the probability/proportion of value *s* for the sensitive attribute in the qid group.

The entropy *l*-diversity still has several limitations. For instance, if the entropy of values for the sensitive attribute in qid groups is high, the *l* will be high. The entropy is highest when the distribution of values is a uniform distribution. Nevertheless, we prefer the minimum probability for the sensitive value (*Positive* in our example) in the qid group. In our example, we favor as few *Positive*s in the qid groups as possible to lower the confidence of inferring HIV positive for the victim patient. Still, entropy *l*-diversity encourages an equal number of *Positive*s and *Negative*s in the qid groups.

Recursive *(c,l)*-diversity controls the frequency of values for the sensitive attribute in the qid group. In this model, c is a constant greater than zero, $c > 0$. The values for the sensitive attribute *S* are: $s_1, s_2, \ldots, s_m$. The number of occurrence for each value (for the sensitive attribute) in the qid group are: $n_1, n_2, \ldots, n_m$. The number of occurrence for values sorted in a decreasing order are: $r_1, r_2, \ldots, r_m$. If a data table meets $r_1 \leq c\sum_{i=l}^{m} r_i$ for each qid group, then the data is recursive *(c,l)*-diverse.

The recursive *(c,l)*-diversity can relax the restrictiveness compared to entropy *l*-diversity. When we have a larger c, we can have a larger l. Therefore, we can relax the restrictiveness by increasing c. This privacy model avoids having a high frequency of highly repeated values (in the dataset for sensitive value) in the qid group. It also forces the less frequent values (in the dataset for sensitive value) to be more frequent in the qid group. However, this may not be desirable in certain scenarios. Many healthcare datasets have sensitive attributes with highly imbalanced values. For

Index	Quasi Identifier			Sensitive
	Age	Number of Children	Smoke	HIV
1	37	2	Yes	Positive
2	36	0	Yes	Positive
3	40	0	Yes	Positive
4	35	3	Yes	Negative
5	32	1	Yes	Negative
6	34	1	Yes	Negative
7	30	2	No	Positive
8	34	2	No	Negative
9	28	1	No	Negative
10	31	1	No	Negative

(a) Original data

Index	Quasi Identifier			Sensitive
	Age	Number of Children	Smoke	HIV
1	[36-40]	[0-2]	Yes	Positive
2	[36-40]	[0-2]	Yes	Positive
3	[36-40]	[0-2]	Yes	Positive
4	[32-35]	[1-3]	Yes	Negative
5	[32-35]	[1-3]	Yes	Negative
6	[32-35]	[1-3]	Yes	Negative
7	[28-34]	[1-2]	No	Positive
8	[28-34]	[1-2]	No	Negative
9	[28-34]	[1-2]	No	Negative
10	[28-34]	[1-2]	No	Negative

(b) 3-anonymous data

TABLE I: Patient data tables in original and 3-anonymous formats

instance, in a table of data with 1000 records, we may have merely 20 patients diagnosed with HIV. In our example, by increasing the frequency of a sensitive value (with low frequency in the dataset) in a qid group, the adversary can more confidently infer that the patient is diagnosed with HIV.

The *t*-closeness privacy model proposes having a more similar distribution of values in the sensitive attribute among the qid groups and the whole dataset. In the *t*-closeness model, the maximum distance between these two distributions may not be greater than the threshold *t*. For measuring the distance between probabilistic distributions, one possible metric is as follows:

$$D[P,Q] = \sum_{i=1}^{m} |p_i - q_i|, \tag{2}$$

where m is the number of values for the sensitive attribute. $P = \{p_1, p_2, \ldots, p_m\}$ and $Q = \{q_1, q_2, \ldots, q_m\}$ are the distributions of sensitive attribute in the entire dataset and in a particular qid group, respectively. This distance metric (variational distance) does not consider the semantic distance between values. In scenarios where the semantic distance of values is important, we may use other distance measures.

In this paper, we propose a method for anonymization of data by jointly considering the *k*-anonymity, *l*-diversity, and *t*-closeness privacy models in a unified framework.

III. Approach

In this section, we describe our method for addressing the attack models discussed in Section II. In our method, we consider the indistinguishability of samples in a qid group, proposed in *k*-anonymity, diversity of values in sensitive attributes in qid group, discussed in *l*-diversity, and frequency of sensitive values in qid group in *t*-closeness.

In this method, we suppose that the values for the sensitive attribute are either sensitive or not. In our example, the *Positive* value shows that the patient (record owner) is diagnosed with HIV and is sensitive, while the value *Negative* if known to the adversary causes no consequence to the patient. Therefore, we consider a binary state for the values in the sensitive attribute and distribute them in the qid groups evenly.

Our method clusters the points in the space of quasi-identifiers and shares the center of each cluster (qid group) as the quasi-identifiers' values for each qid group. Each cluster contains k samples and is clustered based on the distance of instances to the cluster center and the number of samples with sensitive values in each cluster.

We adopt the constrained optimization framework to solve the described clustering problem. The classical clustering techniques do not fulfill our requirements. First, we need to introduce the constraints to have k samples in each cluster to ensure the indistinguishability property of the *k*-anonymity model. Second, we need to introduce a constraint for distributing instances with sensitive values evenly among qid groups (clusters) to ensure diversity in the *l*-diversity and *t*-closeness models.

The described anonymization problem is formulated in the Mixed-Integer Linear Programming (MILP) framework, as follows:

$$\min_{B,C} \quad \sum_{i=0}^{n_C} \sum_{j=0}^{n_S} |B_{ij} \cdot (X_j - Center_i)| \tag{3}$$

$$\text{s.t.} \quad \sum_{i=0}^{n_C} B_{ij} = 1, \quad \forall j \in \{0, \ldots, n_S\} \tag{4}$$

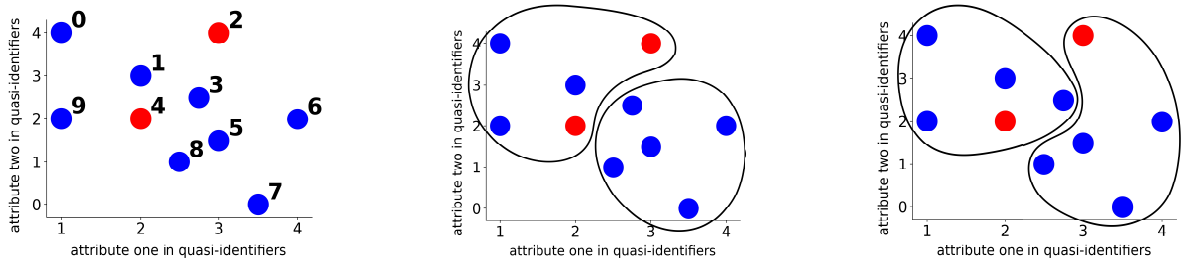
$$\sum_{j=0}^{n_S} B_{ij} = \frac{n_S}{n_C} = k, \quad \forall i \in \{0, \ldots, n_C\} \tag{5}$$

$$\frac{\left(\sum_{j=0}^{n_S} B_{ij} \cdot X_j\right)}{k} = C_i, \quad \forall i \in \{0, \ldots, n_C\} \tag{6}$$

$$\sum_{j=0}^{n_S} B_{ij} \cdot S_j \leq \alpha \cdot \frac{\sum_{j=0}^{n_S} S_j}{n_C}, \quad \forall i \in \{0, \ldots, n_C\}, \tag{7}$$

where n_C is the number of clusters (qid groups), and n_S is the number of samples to be anonymized. X_j is the vector of quasi-identifiers' values for sample j. B_{ij} indicates if sample j belongs to cluster (qid group) i and it is a Boolean optimization variable. $Center_i$ is the i-th cluster center calculated by k-means algorithm to be used as an initial solution in our method to reduce the complexity of our optimization problem.

The parameter k is the number of samples in each cluster and is equal to $\frac{n_S}{n_C}$. C_i is the center of cluster i which will be optimized during solving this problem. The values of



(a) Data samples before optimization to find qid groups

(b) Optimization without considering the diversity constraint

(c) Optimization with considering the diversity constraint

Fig. 1: Illustrative example for our anonymization method

vector C_i will be shared with data recipients, i.e., instead of raw quasi-identifiers' values for *i*-th qid group. S_j is a Boolean parameter, $S_j \in \{0, 1\}$, that identifies if sample *j* has a sensitive value. Finally, α is a parameter that controls the restrictiveness of the constraint, i.e., the higher the value of α, the less the restrictions in solving this optimization problem. This parameter is introduced to be able to tune the restriction with respect to diversity in each qid group.

Let us discuss the proposed formulated optimization problem. The $|B_{ij} \cdot (X_j - Center_i)|$ expression in Eq. (3) is the Manhattan distance of sample *j*, X_j, and cluster center *i*, $Center_i$, when the Boolean variable B_{ij} is equal to one. B_{ij} will be equal to one, $B_{ij} = 1$, if sample X_j belongs to cluster *i*, and it will be zero otherwise. The objective function in Eq. 3 intends to optimize B_{ij}s to minimize the distance between samples in cluster *i* and $Center_i$, for all clusters and samples.

Eqs. (4)-(7) are the constraints of our proposed optimization problem:

- The first constraint, in Eq. (4), forces each sample to belong to only one cluster. This is done by ensuring that B_{ij} is one exactly once for all *i*.
- The second constraint, in Eq. (5), forces the number of samples in each cluster to be equal to *k*. The summation of the number of samples must be equal to *k* for cluster *i*. This condition can readily be relaxed to: at least *k* samples in each cluster.
- The third constraint, in Eq. (6), finds the optimized cluster centers, i.e., C_is. The optimized center for cluster *i* is the average of all *k* samples that belong to cluster *i*.
- Finally, the last constraint, in Eq. (7), forces the optimization to distribute the samples with sensitive values ($S_j = 1$) into all clusters. The left-hand side of the constraint is equal to the number of sensitive values in cluster *i*. The right-hand side is the number of samples with sensitive value divided evenly among the clusters (multiplied by α, which is the parameter for relaxing the hard constraint in our optimization problem).

After the optimization, we know which sample belongs to which qid group or cluster, based on *B* matrix. We also know the optimized cluster centers, identified based on the values of C_is. Therefore, the values of sample quasi-identifiers will be replaced by their respective cluster center values. In this way, we obtain a solution that addresses record-linkage and attribute-linkage attack models. We force the samples in the anonymized data to be indistinguishable from *k-1* other samples while considering the diversity of values in the sensitive attribute.

Fig. 1 presents an example in which the solution in Fig. 1b merely considers *k*-anonymity property, while Fig. 1c considers the diversity of values in the sensitive attribute addressed in *l*-diversity and *t*-closeness. The color of the circles shows if the samples contain a sensitive value. If the color is blue, the sample does not have a sensitive value, $S_j = 0$, while a red circle shows having a sensitive value $S_j = 1$.

In Fig. 1b, samples *0, 1, 2, 4, 9* fall in the same qid group. The rest of the samples fall in the second group. By sharing the cluster centers for each group, we achieve 5-anonymous data. However, in such a solution, the samples with sensitive values are not evenly distributed. By considering the constraint introduced for the diversity of values in the sensitive attribute, we obtain the solution presented in Fig. 1c. In this solution, the data is still 5-anonymous, i.e., it has five samples in each cluster. Nevertheless, in this case, sample *2*, falls in the same cluster with *5, 6, 7, 8* to evenly distribute samples with sensitive values.

IV. Evaluation and Discussion

In this section, we evaluate the proposed method experimentally and discuss the experimental results. For evaluation, we consider data utility and data privacy criteria and demonstrate their trade-off [39]. Then, we present and discuss the experimental results.

In this paper, the data analysis task that is going to be performed on the anonymized data is classification. Therefore, the anonymization method should alter the data to the extent that learning high-performance classification models are possible. We train the learning algorithms on both original and anonymized data to evaluate the anonymization method in terms of data utility preservation. Our method preserves the data utility if the classification model learned from altered data has similar performance compared to the one learned from original data.

TABLE II: Classification performance for trained models on three different versions of Heart Disease dataset (Cleveland) [37], [38]

Algorithm	Original Data			Anonymized Data Without Diversity			Anonymized Data by Our Method		
	F1-score	Accuracy	MCC	F1-score	Accuracy	MCC	F1-score	Accuracy	MCC
ERT	81.0%	81.0%	0.615	81.1%	81.4%	0.625	81.0%	81.4%	0.625
Random Forest	82.5%	82.6%	0.647	80.1%	80.4%	0.603	80.0%	80.3%	0.602
XGBoost	78.9%	79.0%	0.573	74.7%	75.1%	0.493	74.7%	75.1%	0.495
Decision Tree	73.8%	73.8%	0.470	68.9%	69.3%	0.372	69.2%	69.8%	0.382
SVM	83.0%	83.1%	0.656	73.3%	73.3%	0.459	72.8%	72.9%	0.449

On the other hand, the anonymized data should be sufficiently altered to avoid the identification of record owners. In this paper, we address the record-linkage and attribute-linkage attack models. We consider the property for making samples indistinguishable in the qid group, discussed in *k*-anonymity privacy model, the diversity of values in sensitive attribute, in *l*-diversity, and the frequency of sensitive values, in *t*-closeness.

There is a trade-off between the utility of data and privacy of data in anonymization methods. On the one hand, we can share no data to preserve patients' privacy, but there will be no utility for the data. On the other hand, we can publish the data in its original format to maximize the data utility, but the privacy of data subjects is going to be violated. Therefore, in anonymization techniques, we require altering the data to the extent that we establish a trade-off between data utility and privacy [39].

A. Experimental Setup

In our experiments, we use the Heart Disease dataset [37], which is one of the popular datasets publicly available on the UCI repository. We utilize Cleveland's processed dataset [38] to predict the presence of heart disease (presence/absence). The dataset contains 282 complete records, and each belongs to one patient. The data includes 13 attributes which we consider in this work.

Quasi-identifiers are the attributes that the adversary can potentially obtain information about them from other sources. In addition to quasi-identifiers, the sensitive attribute should also be identified. In our experiments, we suppose all 13 attributes are quasi-identifiers. Moreover, we select the Boolean attribute for family history of coronary artery disease as the sensitive attribute.

For evaluation of preservation of utility, we split the dataset into train and test sets. We anonymize the training set using our method with soft constraints and train several classification algorithms based on the resulting data. Then, we measure the classification performance on the test set. We also train the same algorithms on the original data and the data anonymized without considering the diversity constraint and measure the performance of the trained classification models on the test set. The comparison of the classification performance results indicate the utility of anonymized data in our method.

In our experiments, we randomly select 200 samples as the train set and the rest as the test set at each round. We repeat the same process for 1000 rounds and report the average results for classification performance. The algorithms used for learning classification models are Extremely Randomized Trees (ERT), Random Forest, XGBoost, Decision Tree, and linear SVM. The measures used for classification performance are F1-score, Accuracy, and Matthews Correlation Coefficient (MCC).

B. Experimental Results

Table II shows the classification performance results for three different training sets, i.e., original data, anonymized using our method, and anonymized without considering the diversity constraint. For both anonymization methods k is set to 10.

The classification results for the original data are at a similar level ($\pm 0.5\%$ due to randomness in the algorithms) or higher than the anonymized data. However, since there is a trade-off between privacy and utility in anonymization [39], we may accept a loss in the utility to obtain privacy. The results in Table II show that our method preserves the information in data that leads to learning high-performance models. Moreover, the classification performance difference between our method and the approach without considering the diversity is negligible. This indicates that introducing the diversity constraint in our method does not significantly affect the data utility.

We now evaluate the privacy preservation of our method in Table III. Here, we set the value of k to 10. This means that if the adversary has the values for quasi-identifiers for one patient, he/she can only map his/her information to 10 records. Therefore, through our method, we avoid record-linkage attacks. Second, our method evenly distributes the samples with sensitive value, i.e., having a family history of coronary artery disease, to qid groups. This weakens the confidence of the adversary's inference for identifying a patient with sensitive value.

The number of patients with the sensitive value can be different at each round. In our method, in the worst qid group with respect to *l*-diversity, entropy *l*-diversity, and recursive *(c,l)*-diversity, we have two samples with non-sensitive value and eight with the sensitive value. In other words, the proportion of patients with a family history of coronary artery disease in the qid group is 80.0%, which is optimal since the proportion of samples with the sensitive value in the training set at this round was 70.5%. This leads to $l = 2$ in *l*-diversity, $l = 1.64$ in entropy *l*-diversity, and $l = 2$ and $c \geq 4$ in recursive *(c,l)*-diversity in Table III. In the worst qid group with respect to the variational distance D in *t*-closeness, we have six with non-sensitive value and four with the sensitive value, while the proportion of samples with

TABLE III: Privacy properties of the anonymized data by our method and the approach without diversity

	No Diversity	Our Method
k in k-anonymity	10	10
l in l-diversity	1	2
l in entropy l-diversity	1	1.64
l and c in recursive (c,l)-diversity	l=1, $c \geq 1$	l=2, $c \geq 4$
D in t-closeness	1.06	0.38

the sensitive value in the dataset at this round was 59.0%. This leads to variational distance $D = 0.38$ in t-closeness.

For the approach without diversity constraint, in the worst qid group with respect to l-diversity, entropy l-diversity, and recursive (c,l)-diversity, we have ten patients with the sensitive value. This leads to $l = 1$ in l-diversity, $l = 1$ in entropy l-diversity, and $l = 1$ and $c \geq 1$ in recursive (c,l)-diversity in Table III. This allows the adversary to infer that the patient had a family history of coronary artery disease with 100% confidence. Moreover, in the worst qid group with respect to the variational distance D in t-closeness, we have nine records with the non-sensitive value and one with the sensitive value. The proportion of samples with the sensitive value in the dataset at this round was 63.0%. This increases the variational distance between the distributions of values in the sensitive attribute in the qid group and the whole dataset to $D = 1.06$ in Table III.

The results in Table III demonstrates that by adopting our method, we will have higher l in l-diversity, entropy l-diversity, and recursive (c,l)-diversity. Moreover, the variational distance between the distributions of values in the sensitive attribute for the train set and the qid group is lower in our method. Therefore, regarding the diversity of values in sensitive attributes and the attribute-linkage attack, we observe that introducing the diversity constraint improves patients' privacy.

We also investigate the data privacy and data utility based on different values of k, size of qid groups. For each k, we have 100 rounds that in each we randomly split the data into the train and test sets. The classification performance results are the average results for all rounds. The privacy results are the worst results in all rounds and qid groups. We perform these experiments based on our method and the anonymization approach without the diversity constraint and show the results in Figs. 2 and 3 for comparison.

Figs. 2a-2c show the results based on F1-score, Accuracy, and MCC metrics. The patterns in the results show that the higher the qid group size (k), the lower the classification performance. On the other hand, increasing the value of k improves the privacy with respect to the record-linkage attack model. These figures illustrate the trade-off between the privacy and data utility.

The results in Figs. 3a-3d exhibit the privacy properties of the anonymized data. Regarding the attribute-linkage attack model, the results display that the data anonymized by our method has higher privacy properties than the anonymized data without diversity constraint. Increasing the value of k significantly improves the diversity and frequency of values in the sensitive attribute, compared to the approach without considering the diversity constraint, but without any major loss in terms of classification performance.

The experimental results show that our method provides privacy against record-linkage and attribute-linkage attacks. Furthermore, the utility of the data is retained after anonymization, allowing learning of high-performance classification models. The slight degradation of utility is the cost for providing patients privacy, which is a common phenomenon in anonymization approaches [39].

V. Conclusion

In this paper, we have proposed a method for obtaining anonymized data by ensuring that data samples are indistinguishable in qid groups while considering the diversity and frequency of values in the sensitive attribute. Our method is based on constrained optimization and clustering of the samples into qid groups by jointly considering the k-anonymity, l-diversity, and t-closeness privacy models. The evaluation results show that the proposed method retains data utility while reducing the privacy concerns related to data sharing.

ACKNOWLEDGMENT

This research is supported by INTROducing Mental health through Adaptive Technology (INTROMAT) project. The paper is partially supported by SIRIUS: Centre for Scalable Data Access.

References

[1] S. D. Lustgarten, Y. L. Garrison, M. T. Sinnard, and A. W. Flynn, "Digital privacy in mental healthcare: current issues and recommendations for technology use," *Current Opinion in Psychology*, 2020.

[2] D. Pascual, A. Amirshahi, A. Aminifar, D. Atienza, P. Ryvlin, and R. Wattenhofer, "Epilepsygan: Synthetic epileptic brain activities with privacy preservation," *IEEE Transactions on Biomedical Engineering*, 2020.

[3] M. Kantarcioglu, "A survey of privacy-preserving methods across horizontally partitioned data," in *Privacy-preserving data mining*. Springer, 2008.

[4] J. Vaidya, "A survey of privacy-preserving methods across vertically partitioned data," in *Privacy-preserving data mining*. Springer, 2008.

[5] J. Vaidya, B. Shafiq, W. Fan, D. Mehmood, and D. Lorenzi, "A random decision tree framework for privacy-preserving data mining," *IEEE transactions on dependable and secure computing*, 2013.

[6] A. Aminifar, F. Rabbi, K. I. Pun, and Y. Lamo, "Privacy preserving distributed extremely randomized trees," in *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, 2021.

[7] A. Aminifar, F. Rabbi, and Y. Lamo, "Scalable privacy-preserving distributed extremely randomized trees for structured data with multiple colluding parties," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021.

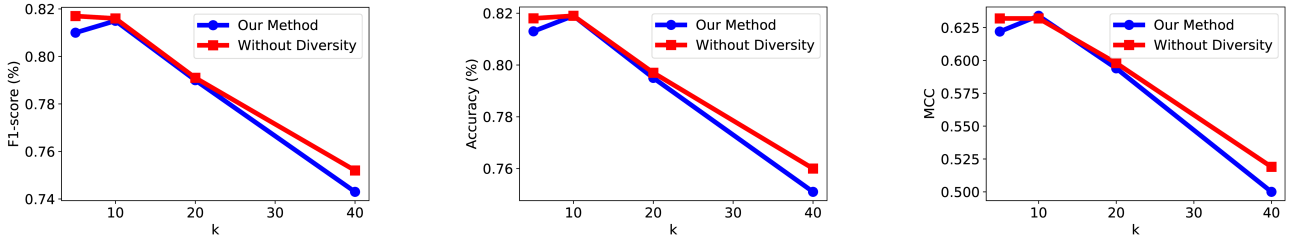
[8] J. Konečnỳ, H. B. McMahan, D. Ramage, and P. Richtárik, "Federated optimization: Distributed machine learning for on-device intelligence," *arXiv preprint arXiv:1610.02527*, 2016.

[9] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, *et al.*, "Communication-efficient learning of deep networks from decentralized data," *arXiv preprint arXiv:1602.05629*, 2016.

[10] S. Baghersalimi, T. Teijeiro, D. Atienza, and A. Aminifar, "Personalized real-time federated learning for epileptic seizure detection," *IEEE Journal of Biomedical and Health Informatics*, 2021.

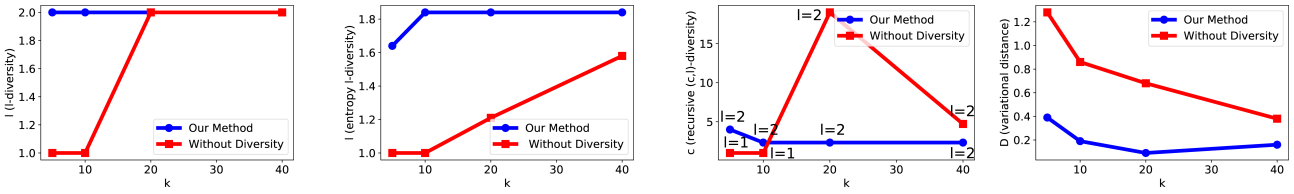
[11] R. Agrawal and R. Srikant, "Privacy-preserving data mining," in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 2000.

[12] S. Agrawal and J. R. Haritsa, "A framework for high-accuracy privacy-preserving mining," in *21st International Conference on Data Engineering (ICDE'05)*. IEEE, 2005.



(a) F1-score (b) Accuracy (c) MCC

Fig. 2: The classification performance for anonymized data based on F1-score, Accuracy, and MCC measures for different values of k



(a) l in l-diversity (b) l in entropy l-diversity (c) c in (c,l)-diversity (d) D in t-closeness

Fig. 3: The privacy properties of the data anonymized by our method and the approach without considering the diversity constraint for different values of k

[13] S. J. Rizvi and J. R. Haritsa, "Maintaining data privacy in association rule mining," in *VLDB'02: Proceedings of the 28th International Conference on Very Large Databases*. Elsevier, 2002.

[14] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, "On the privacy preserving properties of random data perturbation techniques," in *Third IEEE international conference on data mining*. IEEE, 2003.

[15] Z. Huang, W. Du, and B. Chen, "Deriving private information from randomized data," in *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, 2005.

[16] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *arXiv preprint arXiv:1406.2661*, 2014.

[17] M. Malekzadeh, R. G. Clegg, A. Cavallaro, and H. Haddadi, "Mobile sensor data anonymization," in *Proceedings of the international conference on internet of things design and implementation*, 2019.

[18] M. Alzantot, S. Chakraborty, and M. Srivastava, "Sensegen: A deep learning architecture for synthetic sensor data generation," in *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE, 2017.

[19] N. Park, M. Mohammadi, K. Gorde, S. Jajodia, H. Park, and Y. Kim, "Data synthesis based on generative adversarial networks," *arXiv preprint arXiv:1806.03384*, 2018.

[20] D. Sopic, A. Aminifar, A. Aminifar, and D. Atienza, "Real-time event-driven classification technique for early detection and prevention of myocardial infarction on wearable systems," *IEEE transactions on biomedical circuits and systems*, 2018.

[21] ——, "Real-time classification technique for early detection and prevention of myocardial infarction on wearable devices," in *2017 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 2017.

[22] A. Aminifar, P. Eles, and Z. Peng, "Optimization of message encryption for real-time applications in embedded systems," *IEEE Transactions on Computers*, 2017.

[23] F. Forooghifar, A. Aminifar, and D. Atienza, "Resource-aware distributed epilepsy monitoring using self-awareness from edge to cloud," *IEEE transactions on biomedical circuits and systems*, 2019.

[24] "Health informatics — Pseudonymization," International Organization for Standardization," Standard, 2017.

[25] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2002.

[26] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2007.

[27] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *2007 IEEE 23rd International Conference on Data Engineering*. IEEE, 2007, pp. 106–115.

[28] N. Mohammed, B. C. Fung, P. C. Hung, and C.-k. Lee, "Anonymizing healthcare data: a case study on the blood transfusion service," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009, pp. 1285–1294.

[29] V. S. Iyengar, "Transforming data to satisfy privacy constraints," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2002.

[30] R. J. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymization," in *21st International conference on data engineering (ICDE'05)*. IEEE, 2005, pp. 217–228.

[31] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Mondrian multidimensional k-anonymity," in *22nd International conference on data engineering (ICDE'06)*. IEEE, 2006.

[32] A. Majeed, F. Ullah, and S. Lee, "Vulnerability-and diversity-aware anonymization of personally identifiable information for improving user privacy and utility of publishing data," *Sensors*, 2017.

[33] A. Aminifar, Y. Lamo, K. I. Pun, and F. Rabbi, "A practical methodology for anonymization of structured health data," in *Proceedings of the 17th Scandinavian Conference on Health Informatics*, 2019.

[34] K. Doka, M. Xue, D. Tsoumakos, and P. Karras, "k-anonymization by freeform generalization," in *Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security*, 2015.

[35] Y. Liang and R. Samavi, "Optimization-based k-anonymity algorithms," *Computers & Security*, 2020.

[36] B. C. Fung, K. Wang, A. W.-C. Fu, and S. Y. Philip, *Introduction to privacy-preserving data publishing: Concepts and techniques*. CRC Press, 2010.

[37] R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J. Schmid, S. Sandhu, K. Guppy, S. Lee, and V. Froelicher, "International application of a new probability algorithm for the diagnosis of coronary artery disease," *The American journal of cardiology*, 1989.

[38] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[39] J. S. Davis and O. Osoba, "Improving privacy preservation policy in the modern information age," *Health and Technology*, 2019.

PAPER C

SCIENTIFIC PAPER III: PRIVACY PRESERVING DISTRIBUTED EXTREMELY RANDOMIZED TREES

Amin Aminifar, Fazle Rabbi, Violet Ka I Pun, and Yngve Lamo

In Proceedings of the 36th Annual ACM Symposium on Applied Computing. 2021.

Privacy Preserving Distributed Extremely Randomized Trees*

Amin Aminifar[1], Fazle Rabbi[1,2], Ka I Pun[1,3], and Yngve Lamo[1]
[1]Western Norway University of Applied Sciences, [2]University of Bergen, [3]University of Oslo
{amin.aminifar,fazle.rabbi,ka.i.pun,yngve.lamo}@hvl.no

ABSTRACT

Applying machine learning and data mining algorithms over data distributed in multiple sources is challenging. One complication is to perform data analysis without compromising personal information, which is a primary concern in healthcare applications. Another issue involves communication overhead incurred from the transfer of raw data from one party to others for conducting centralized data mining. In healthcare applications, we are particularly interested in running data mining algorithms over big data without disclosing sensitive information about data subjects due to privacy and legal concerns. In this paper, we consider the classification problem and show how the Extremely Randomized Trees (ERT) algorithm could be adapted for settings where (structured) data is distributed over multiple sources. We propose the Privacy-Preserving Distributed ERT approach for privacy-preserving utilization of the ERT algorithm in a distributed setting. To the best of our knowledge, this is the first application of the ERT algorithm in the distributed setting, with privacy consideration (without sharing the raw data or intermediate training values), without any loss in classification performance.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; *Distributed artificial intelligence*; *Cooperation and coordination*; **Machine learning**; *Machine learning algorithms*;

KEYWORDS

Distributed Learning, Extremely Randomized Trees, Privacy-Preserving Data Mining, Structured Data

ACM Reference Format:
Amin Aminifar[1], Fazle Rabbi[1,2], Ka I Pun[1,3], and Yngve Lamo[1]. 2021. Privacy Preserving Distributed Extremely Randomized Trees. In *Proceedings of ACM SAC Conference (SAC'21).* ACM, New York, NY, USA, Article 4, 4 pages. https://doi.org/10.1145/3412841.3442110

1 INTRODUCTION

In many real-world applications, such as in healthcare systems, data is inherently distributed over an arbitrary number of sources instead of being stored in a central database. It is not straightforward to apply data mining algorithms in situations where distributed data cannot be transferred to a central location due to communication overheads, as well as privacy concerns. Figure 1 shows one such scenario and environment for this problem. The figure illustrates a setting where hospitals need to apply data mining methods for extracting useful patterns from patients' data. Although individual hospital information systems may be able to locally store health information and perform data mining with their limited resources, it is a necessity to share health information across hospitals to fully exploit the learning capacity of the data mining techniques. However, this is a challenging task due to privacy and legal concerns. Hospitals often need to comply with privacy regulations that restrict sharing health information about patients with other parties [13, 16, 19]. A similar problem exists when the data is stored on patients' personal devices, such as mobile phones or wearable devices with limited resources [8, 21–23]. How can we utilize large amount of healthcare data stored in an arbitrary number of sources for data mining without disclosing the private information of the subjects? In this paper, we address this problem by developing a novel approach for privacy-preserving data mining over distributed (structured) healthcare information.

Traditionally, it was assumed that all sources holding part of the data may share their information with a trusted party. However, sharing sensitive data with trusted parties is not a feasible assumption in many scenarios. In order to address the privacy concern, one solution would be to perturb data and share it. However, perturbation-based solutions do not provide absolute data privacy and utility because the privacy will not be preserved if the perturbation is not sufficient and the data utility will decrease if the perturbation is not controlled precisely [4, 26]. Similarly, anonymization techniques, e.g., [1, 14, 17, 24], share an altered version of data to prevent the re-identification of data subjects [10]. Nevertheless, there is always a trade-off between data privacy and utility in these techniques [4]. Therefore, such techniques have limited applicability. Moreover, communication and computational overheads would still be a problem for the approaches we discussed above, especially when dealing with large scale data.

There exist several data mining algorithms that utilize the indirect use of raw data. One such approach is the cryptographic technique and secure multi-party computation method for conducting privacy-preserving data mining [5, 11, 25]. However, they are inefficient when dealing with big data, due to extreme communication/computation costs [26]. Other techniques have been proposed to address communication/computational overheads of the stated privacy-preserving data mining algorithms, e.g., [7, 12, 18]. These solutions provide privacy as well as efficiency w.r.t. communication and computational overheads. Nevertheless, the data mining algorithms should be modified, depending on the possibility to support applications in distributed settings, which may negatively affect the machine learning model's performance.

*Produces the permission block, and copyright information

SAC'21, March 22-March 26, 2021, Gwangju, South Korea

ACM ISBN 978-1-4503-8104-8/21/03…$15.00
https://doi.org/10.1145/3412841.3442110

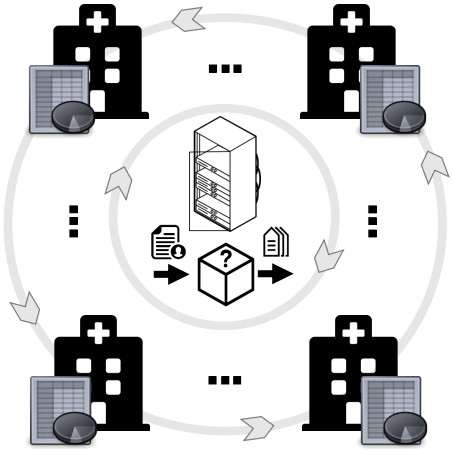


Figure 1: Overview of the environment for learning from structured data distributed over several parties

In this paper, we target the problem of learning from multiple data holders, without explicit sharing of the raw healthcare information. We assume that the learning data is horizontally partitioned, i.e., different records of data are stored on different sources. We consider the classification problem, in which each data record has one category as the target. We consider that data is structured, i.e., it can be stored in spreadsheets, and contains categorical attributes, e.g., gender or mental-disorder history, and numerical attributes, like age or frequency and duration of pathological episodes. We focus on the class of tree-based algorithms that have been shown to consistently outperform or to be on a par with the other state-of-the-art techniques when it comes to structured data [3, 15]. To learn from such horizontally-partitioned structured data, we propose privacy-preserving distributed extremely randomized trees (PPD-ERT). We first extend the ERT algorithm [9] to a distributed setting, to enable learning without explicit sharing of the raw data. We then introduce a secure aggregation technique over the distributed ERT algorithm to avoid any information leakage. We evaluate the proposed solution experimentally and compare the results against the state-of-the-art techniques.

2 PRIVACY PRESERVING DISTRIBUTED ERT

This section presents the proposed solution which is based on Extremely Randomized Trees (ERT) [9] algorithm, and discusses the procedure of learning an ensemble of decision trees based on the ERT algorithm in the discussed settings. Our main contributions w.r.t. the traditional ERT algorithm are:

- We extended ERT to the distributed setting.
- We employed a security layer by utilizing SMC techniques.

2.1 Initialization and Initiation

In the initialization phase, the mediator starts the process of leaning. The mediator initiates and mediates the overall learning process. It begins with sharing the global and personal random seeds with data holder parties. The mediator will then repeatedly learn decision trees based on our privacy-preserving distributed ERT algorithm.

In the ERT algorithm, we have two parameters of randomness for learning a weak classifier. First, we need to randomly select several attributes, among all possible data attributes, for selecting candidate decision nodes at every step of building our decision tree. Second, a random splitting point for every attribute in the candidate decision node is required. The data holder parties and the mediator are required to have the same candidate decision nodes at every step of learning a decision tree. Therefore, instead of making these randomly-made candidate decision nodes in the mediator and sharing them with all parties for further tasks, we share a common random seed that all parties, including the mediator, use to locally generate these candidate decision nodes. Since all parties use a common random seed, i.e., the global random seed, they generate the same candidate decision nodes at every step, without any communication overhead. Moreover, for the secure aggregation of partial results, described further in Section 2.3, each data holder party and the mediator share a personal random seed. These random seeds are exclusive and private for each data holder party. Only the data holder party and the mediator have access to this personal random seed.

2.2 The Process of Learning One Decision Tree

The learning of a decision tree based on the privacy-preserving distributed ERT algorithm is a recursive procedure, which is executed top-down, starting from the root and ending at the leaves.

The mediator generates the candidate decision nodes, for building the decision tree, after receiving the results from the data holder parties to select the best candidate among them. The candidate decision nodes are generated randomly based on the global random seed. Several attributes from the dataset's possible attributes are selected for candidate decision nodes. Then, each candidate decision node's splitting points are selected. We assume that all parties already have the possible categories and ranges for each attribute.

To decide the candidate decision nodes for each branch, the mediator requires the collective outcome of the classification with candidate decision nodes from all data holders on all their data. By having the combination of data record labels for each branch, the mediator can both decide if we require a leaf at that place or if we should calculate the information gain. The mediator sends a request to the first data holder party and waits for receiving the aggregated result from the last party through secure aggregation described in Section 2.3. The aggregated results are two vectors, one for each branch, representing the combination of data record labels after classification with each candidate decision node.

Having the aggregated results, the mediator determines if a decision node is required for that place in the tree. If all the labels are the same or if the number of received labels is less than the threshold parameter in the ERT algorithm, the mediator puts a label on that place, as a leaf. Otherwise, the mediator calculates the information gain of each candidate decision node based on the results from data holder parties. It then selects the candidate decision node with the highest information gain and informs all parties about this. The selected node will be used to build the tree at the mediator. After selecting the best decision node candidate, the same process is performed for each of the branches.

This process leads to learning a single decision tree; we repeat the same process for having an ensemble of decision trees.

2.3 Secure Aggregation of Results From Parties

We adopt an SMC technique in our proposed distributed ERT algorithm to avoid sharing the vectors representing the combination of the data record labels for each candidate decision node and each branch in each data holder party. In addition to the provided privacy by not sharing the raw values of data attributes, which is by construction, adoption of the SMC technique for aggregating the partial results from data holder parties contributes to privacy preservation. In an extreme example, suppose our data has one sensitive attribute in it, e.g., having previously conducted transgender surgery, and each data holder party has only one record on it. Then, sharing the partial results from one party, i.e., the vectors representing the combination of data record labels for one candidate decision node, can reveal sensitive information. If the candidate decision node is "whether the record falls into the transgender branch or not," the mediator can infer if that individual with the specified record has conducted transgender surgery. Therefore, to avoid such vulnerabilities, we adopt an SMC technique for aggregating the partial results from the data holder parties. We consider privacy among collaborating parties, but we assume no active external adversaries.

We now describe the proposed technique. The mediator shares a personal random seed with each data holder party through secure communication, to avoid sending and receiving exclusive random numbers between the mediator and each party.

Then, in the process of learning a decision tree, the mediator sends the request for secure aggregation to the first party. The party makes calculations described earlier and obtains two resulting vectors for each decision node. Afterwards, the party generates random integer masks based on its personal random seed and adds it to the results from the previous step. If the data holder party receives partial results vectors from the previous data holder party, then it also aggregates those values to the calculated vector in the previous step. Eventually, the party passes its outcome to the next party or mediator if that party is the last one.

Finally, the mediator receives the masked aggregated results from the last party. Since the mediator has the personal random seeds, it generates the same random masks as generated on the data holder parties. Then, the mediator subtracts those random masks from the received masked aggregated result. At this step, without sharing the partial information about data labels by each data holder party, the mediator has the aggregated vectors representing the combination of data record labels for each branch of each candidate decision node for all parties.

3 EVALUATION AND DISCUSSION

In this section, we evaluate our proposed approach w.r.t. classification performance, scalability and overhead, and privacy criteria [2]. We compare our approach with [7] since, similar to our approach, it is a tree-based method, employing SMC techniques for secure aggregation of partial results, to address classification problems in scenarios where data is horizontally partitioned.

Table 1: Comparison of Classification Performance

		Distributed Approaches		Centralized Approaches	
Dataset	*Metric*	*PPD-ERT*	*Distributed ID3* [7]	*ERT* [9]	*ID3*[20]
Multiple Features	*Accuracy*	98.3%	88%	98.3%	93.5%
	F1-Score	98.3%	*Not Reported*	98.3%	93.5%
Nursery	*Accuracy*	98.1%	95.7%	98.1%	99.5%
	F1-Score	95.3%	*Not Reported*	95.3%	79.2%

First, the privacy-preserving distributed ERT algorithm basically breaks the task of the centralized ERT algorithm into several parts distributed on different nodes but does not introduce any negative impact on performance by construction. Secondly, the SMC technique adopted to introduce privacy does not change the result of aggregation as opposed to the existing differential privacy techniques. The resulting vectors, representing the combination of record labels for each branch, aggregated securely by the described SMC technique, yields the same results as aggregation without adopting any SMC techniques. Therefore, the classification performance of our privacy-preserving distributed ERT remains the same as the centralized ERT. However, the proposed approach in [7] suffers from a decline in classification performance caused by its underlying learning algorithm, i.e., the ID3 algorithm.

We now evaluate the classification performance of our proposed approach. Similar to [7], we utilize Multiple Features and Nursery datasets [6] and use 2/3 of the data for learning and the rest for the test. We adopt the F1-Score and accuracy as our classification performance metrics. The accuracy of the proposed approach in [7] is also reported here for comparison. For the Multiple Features dataset, since the number of records for each class is the same, the accuracy is a proper metric for evaluating the classification performance. However, since the Nursery dataset is imbalanced, the accuracy is not a reliable measure; hence, we also consider the F1-Score. Table 1 compares the classification performance of our approach PPD-ERT with the one in [7], with their best setting where 128 parties are collaborating. Moreover, the classification performance of centralized versions of ERT [9] and ID3 [20] algorithms, i.e., the underlying standard learning algorithms for PPD-ERT and the proposed approach in [7], are also provided for comparison.

In our experiments, on the PPD-ERT, and the ERT algorithm, we learn an ensemble of 25 decision trees. For the number of candidate decision nodes' parameter in the algorithm, we used 5-fold cross-validation for the model selection (concerning classification performance measured by the F1-Score). For the Multiple Features dataset, we generate 65 candidate decision nodes (proportionate to 10% of the number of features in the dataset) at every step, and for the Nursery dataset, eight candidate decision nodes (proportionate to 90% of the number of features in the dataset) are generated. The results in Table 1 for PPD-ERT, ERT, and ID3 are the average of 10 rounds of learning and evaluation. In the case of the Multiple Features dataset, the PPD-ERT algorithm outperforms the proposed technique in [7] by 10.3%. For the Nursery dataset, the PPD-ERT outperforms the method in [7] by 2.4%. However, in the case of the Nursery dataset, since the data is imbalanced, using the accuracy metric may lead to misleading results. When considering the F1-Score metric, which is a reliable metric even for imbalanced datasets, the simple ID3 algorithm that always outperforms the

Table 2: Communication Complexity of Different SMC Approaches

Approach	Party	Communication		Total Communication
		Send	Receive	
NOSMC	Data Holders	1	0	(n-1)×1 + 1×(n-1)
	Mediator	0	n-1	
PPD-ERT	All	1	1	n×(1+1)

proposed method in [7] shows 16.1% lower performance compared to the PPD-ERT approach.

We now discuss the privacy and overhead of our proposed approach. We adopt an SMC technique to avoid direct sharing of the vectors, representing the combination of record labels for each candidate decision node, with other parties and the mediator. We compare the communication overhead and privacy of our adopted SMC technique against the NOSMC approach. Table 2 presents the communication overhead of both methods. In the table, n is the number of parties, and the communication overheads in the table are for one round of secure aggregation.

In the first approach (NOSMC), no SMC technique is adopted, and all the values are directly shared with the mediator and known to it. This approach has the lowest possible communication cost and one colluding parties, and is considered as a baseline. On the one hand, our approach's communication overhead is from order $O(n)$, which is from the same order as NOSMC. On the other hand, our technique offers interesting privacy features compared to NOSMC. Firstly, it takes three parties (or two parties in case the data holder party is the first or last) for collusion. Secondly, one of the colluding parties needs to be the mediator, which can be assumed as an honest party in many scenarios. In the case of a secret value revelation, we know that the mediator has been involved in the collusion.

We demonstrate that our proposed PPD-ERT approach provides a solution to classification of structured data distributed over multiple sources with privacy-preservation consideration. In particular, our approach does not negatively affect the classification performance compared to the centralized ERT algorithm.

4 CONCLUSION

In this paper, we have extended the ERT algorithm to ensure privacy in a distributed setting, where data is held by several parties. In our proposed algorithm, on the one hand, the data holders do not share their data values with other parties for learning. On the other hand, the required partial-information from data holders, the combination of labels after splitting their records by candidate decision nodes, which has a low risk of revealing important information, is securely aggregated to minimize the likelihood of inference of sensitive information by an adversary. We have evaluated our proposed algorithm extensively and demonstrated its efficiency in terms of prediction performance, scalability and overheads, as well as privacy. We show that our approach outperforms the state-of-the-art distributed ID3 by up to 10.3% in terms of classification performance while ensuring scalability and privacy.

ACKNOWLEDGMENTS

This research is supported by INTROducing Mental health through Adaptive Technology (INTROMAT) project. The paper is partially supported by SIRIUS: Centre for Scalable Data Access.

REFERENCES

[1] A Aminifar, Y Lamo, KI Pun, and F Rabbi. 2019. A Practical Methodology for Anonymization of Structured Health Data. In *Proceedings of the 17th Scandinavian Conference on Health Informatics*.
[2] E Bertino, D Lin, and W Jiang. 2008. A survey of quantification of privacy preserving data mining algorithms. In *Privacy-preserving data mining*. Springer.
[3] T Chen and C Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*.
[4] JS Davis and O Osoba. 2019. Improving privacy preservation policy in the modern information age. *Health and Technology* (2019).
[5] W Du and Z Zhan. 2002. Building decision tree classifier on private data. (2002).
[6] D Dua and C Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml
[7] F Emekçi, OD Sahin, D Agrawal, and A El Abbadi. 2007. Privacy preserving decision tree learning over multiple parties. *Data & Knowledge Engineering* (2007).
[8] F Forooghifar, A Aminifar, and D Atienza. 2019. Resource-aware distributed epilepsy monitoring using self-awareness from edge to cloud. *IEEE transactions on biomedical circuits and systems* (2019).
[9] P Geurts, D Ernst, and L Wehenkel. 2006. Extremely randomized trees. *Machine learning* (2006).
[10] ISO 25237:2017 2017. *Health informatics — Pseudonymization.* Standard. International Organization for Standardization, Geneva, CH.
[11] M Kantarcioglu. 2008. A survey of privacy-preserving methods across horizontally partitioned data. In *Privacy-preserving data mining*. Springer.
[12] J Konečnỳ, HB McMahan, D Ramage, and P Richtárik. 2016. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527* (2016).
[13] P Kumar and HJ Lee. 2012. Security issues in healthcare applications using wireless medical sensor networks: A survey. *sensors* (2012).
[14] N Li, T Li, and S Venkatasubramanian. 2007. t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering*. IEEE, 106–115.
[15] SM Lundberg, G Erion, H Chen, A DeGrave, JM Prutkin, B Nair, R Katz, J Himmelfarb, N Bansal, and SI Lee. 2020. From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence* (2020).
[16] SD Lustgarten, YL Garrison, MT Sinnard, and AWP Flynn. 2020. Digital privacy in mental healthcare: current issues and recommendations for technology use. *Current Opinion in Psychology* (2020).
[17] A Machanavajjhala, D Kifer, J Gehrke, and M Venkitasubramaniam. 2007. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)* (2007).
[18] HB McMahan, E Moore, D Ramage, S Hampson, et al. 2016. Communication-efficient learning of deep networks from decentralized data. *arXiv preprint arXiv:1602.05629* (2016).
[19] D Pascual, A Amirshahi, A Aminifar, D Atienza, P Ryvlin, and R Wattenhofer. 2020. EpilepsyGAN: Synthetic Epileptic Brain Activities with Privacy Preservation. In *IEEE Transactions on Biomedical Engineering*.
[20] JR Quinlan. 1986. Induction of decision trees. *Machine learning* (1986).
[21] A Saeed, FD Salim, T Ozcelebi, and J Lukkien. 2020. Federated Self-Supervised Learning of Multi-Sensor Representations for Embedded Intelligence. *IEEE Internet of Things Journal* (2020).
[22] D Sopic, A Aminifar, A Aminifar, and D Atienza. 2017. Real-time classification technique for early detection and prevention of myocardial infarction on wearable devices. In *2017 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE.
[23] D Sopic, A Aminifar, A Aminifar, and D Atienza. 2018. Real-time event-driven classification technique for early detection and prevention of myocardial infarction on wearable systems. *IEEE transactions on biomedical circuits and systems* (2018).
[24] L Sweeney. 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* (2002).
[25] J Vaidya. 2008. A survey of privacy-preserving methods across vertically partitioned data. In *Privacy-preserving data mining*. Springer.
[26] J Vaidya, B Shafiq, W Fan, D Mehmood, and D Lorenzi. 2013. A random decision tree framework for privacy-preserving data mining. *IEEE transactions on dependable and secure computing* (2013).

PAPER D

SCIENTIFIC PAPER IV: SCALABLE PRIVACY-PRESERVING DISTRIBUTED EXTREMELY RANDOMIZED TREES FOR STRUCTURED DATA WITH MULTIPLE COLLUDING PARTIES

Amin Aminifar, Fazle Rabbi, and Yngve Lamo

In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2021.

SCALABLE PRIVACY-PRESERVING DISTRIBUTED EXTREMELY RANDOMIZED TREES FOR STRUCTURED DATA WITH MULTIPLE COLLUDING PARTIES

Amin Aminifar[1] *Fazle Rabbi*[1,2] *Yngve Lamo*[1]

[1]Western Norway University of Applied Sciences, [2]University of Bergen

ABSTRACT

Today, in many real-world applications of machine learning algorithms, the data is stored on multiple sources instead of at one central repository. In many such scenarios, due to privacy concerns and legal obligations, e.g., for medical data, and communication/computation overhead, for instance for large scale data, the raw data cannot be transferred to a center for analysis. Therefore, new machine learning approaches are proposed for learning from the distributed data in such settings. In this paper, we extend the distributed Extremely Randomized Trees (ERT) approach w.r.t. privacy and scalability. First, we extend distributed ERT to be resilient w.r.t. the number of colluding parties in a scalable fashion. Then, we extend the distributed ERT to improve its scalability without any major loss in classification performance. We refer to our proposed approach as k-PPD-ERT or Privacy-Preserving Distributed Extremely Randomized Trees with k colluding parties.

Index Terms— Distributed Learning, Privacy-Preserving Data Mining, Extremely Randomized Trees, Secure Multiparty Computation, Structured Data

1. INTRODUCTION

A basic assumption in traditional data mining algorithms is that all training data are stored in one data center where mining algorithms run. However, this assumption is not practical in many of today's real-world applications. Today, data is generated and stored on various machines, often located in distributed places. For example, health data is generated and stored at various hospitals, health service providers, and patients' personal devices. Such raw data cannot be shared with a data mining center due to privacy and legal concerns [1, 2]. At the same time, if each party performs mining on its limited data, the performance of the resulting model will largely be subordinate to the performance of a model that can be learned from all the data. Therefore, new mining approaches are required to learn from data distributed across multiple sources while maintaining privacy.

The learning from distributed data in privacy-preserving fashion have been extensively studied over the past decades. The first category of solutions is based on sharing raw data with a trusted third party, which might not be practical in certain scenarios since individuals' privacy cannot be protected from that party [3]. On the other hand, several studies have focused on perturbation-based solutions, e.g., [4–8], to address this issue by adding noise to the data before sharing it. While perturbing the data improves privacy, it also reduces the data utility. Moreover, noise removal techniques cast doubt on the privacy of such approaches [9, 10]. In addition, several anonymization methods, e.g., [11, 12], have been proposed to alter data values, by adopting techniques such as generalization (in k-anonymization [13]) or encryption of data values (in [14]), to avoid reidentification of data subjects [15], e.g., through linking attack [13]. However, in such perturbation-based and anonymization techniques, there is a trade-off between data utility and privacy, which make them impractical in certain scenarios.

Existing literature on data mining over distributed platforms incorporate approaches based on cryptographic and secure multiparty computing techniques [16–20]. However, such methods significantly increase communication and computing overhead, making them inefficient and impractical for many real-world scenarios, where we have large-scale data or limited communication and computing features, e.g., in mobile phones or resource-limited wearable devices [21–24]. Several state-of-the-art solutions, such as [3, 25, 26], aim to address learning in distributed settings in terms of reducing communication and computational overheads. This is because the complexity and scalability of the approach, along with the quality of data mining results and privacy, are among the three primary metrics for evaluating privacy-preserving data mining algorithms [27].

In this paper, we focus on the Extremely Randomized Trees (ERT) algorithm [28], which has a competitive performance for structured data, where we have independently meaningful attributes, compared to the existing state-of-the-art techniques, e.g., standard deep neural networks [29]. We consider the ERT algorithm in the distributed setting to reduce the amount of raw data leaving a party and privacy concerns [30]. We extend this distributed ERT framework in order to improve its scalability and privacy. We adopt an efficient Secure Multiparty Computation (SMC) technique for secure aggregation of partial results in our approach, which is resilient to multiple colluding parties, similar to Shamir's secret sharing technique [31]. We further propose a practical implementation of our proposed framework to reduce its overhead and improve its scalability. Moreover, we extend our proposed framework for efficient handling of large scale data and where only a subset of the parties participate in the process of learning. Our proposed framework offers the opportunity to make a trade-off among performance, privacy, and overhead.

This research is supported by INTROducing Mental health through Adaptive Technology (INTROMAT) project.

2. BACKGROUND

Extremely Randomized Trees (ERT) is a tree-based ensemble supervised learning method [28]. This approach is robust to overfitting since it follows the logic of bagging, i.e., it generates an ensemble of different weak classifiers and finally classifies based on a majority vote among these classifiers. The randomness parameters for generating distinctive weak classifiers are data attributes and splitting points for generating candidate decision nodes.

This paper considers the distributed ERT framework, which is adapted for learning classifier models from structured data, with categorical/numerical attributes and categorical labels, distributed over an arbitrary number of sources. In such a setting, the training data is horizontally partitioned and distributed over multiple sources, i.e., different records are stored on different data holder parties. The raw data cannot be shared with a central server for mining due to privacy and legal concerns. Therefore, the distributed ERT learns from the data without direct access to it and merely by partial and limited information from parties that hold the training data.

Distributed ERT iteratively learns an ensemble of decision trees. Learning a decision tree requires selecting a decision node at each step. The selection of decision nodes is performed based on the information gain. Information gain is a measure/score that indicates how well a decision node, compared to others, classifies the data samples to have more pure sets of samples at every branch of the decision node considering samples' labels. To calculate the information gain, the classification results of candidate decision nodes are required (from all data holder parties). Therefore, in distributed ERT, every data holder party classifies its records with the randomly generated decision nodes and obtains partial results (two vectors representing the combination/mixture of record labels fall into True and False branches). The aggregation of such partial results from all data holder parties enables the calculation of scores/information gains.

The direct sharing of such partial results to other parties puts the privacy of data subjects at risk. For instance, assuming the party holds only one record, if the candidate decision node classifies the data based on a sensitive attribute, e.g., suffering from a mental disorder, then the partial result indicates if the data subject falls under a certain category. For calculating the score, only the aggregation of partial results is required. In distributed ERT, each party aggregates its partial results to the previous party's received result and sends it to the next party. Although this technique is more efficient compared to the employed techniques in similar studies [3], e.g., Shamir's secret sharing technique [31], the number of colluding parties to reveal a secret value, in the worst case, is one.

In this study, we extend the distributed ERT framework and the secure aggregation protocol to be resilient to k colluding parties, where k is determined by the user. We further propose an efficient implementation for our framework, which is scalable and robust for large scale data w.r.t. the participation of a subset of data holder parties.

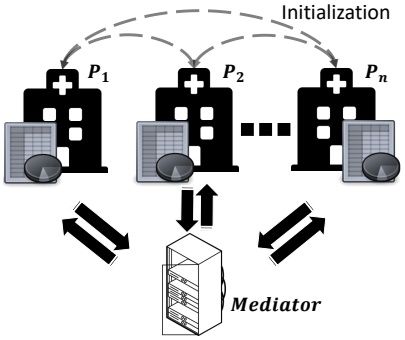


Fig. 1: Overall scenario for our privacy-preserving learning

3. APPROACH

In this section, we explain the proposed k-PPD-ERT algorithm. Section 3.1 describes the adopted secure aggregation technique for k-PPD-ERT. In Section 3.2, we explain how we can improve the scalability of the approach to learn from large scale data.

3.1. Privacy in the Presence of k Colluding Parties

Figure 1 illustrates the overall scenario for the proposed privacy-preserving learning framework. In the initialization phase of the k-PPD-ERT algorithm, each data holder party shares two seeds for the random function to other data holder parties (and receives two in return from each data holder party). The first seed (Seed for Selection of Parties, SSP) is unique for each sender but common for receivers, but the second seed (Seed for Secure Aggregation, SSA) is unique for each sender and receiver couple.

We suppose that the number of data holder parties is n. Therefore, after this initialization procedure, party m, P_m (where $1 \leq m \leq n$), receives two sets of $n-1$ seeds from other data holder parties ($\{SSP_{all}^{P_1}, \ldots, SSP_{all}^{P_n}\}$ and $\{SSA_{P_m}^{P_1}, \ldots, SSA_{P_m}^{P_n}\}$) and holds the seeds which were sent to other parties ($SSP_{all}^{P_m}$ and $\{SSA_{P_1}^{P_m}, \ldots, SSA_{P_n}^{P_m}\}$). Moreover, the secret value of party m is denoted by $secret\_val^{P_m}$.

The responsibilities of party m in one round of secure aggregation is explained in the following steps:

(a) **Identifying the k parties that participate in the secure aggregation for P_m:**
Party P_m uses $SSP_{all}^{P_m}$, in its random function, to identify which parties participate in secure aggregation for P_m, i.e., by randomly generating the party indices. Then, P_m generates random masks based on the *sent SSA* seeds ($\{SSA_{P_1}^{P_m}, \ldots, SSA_{P_n}^{P_m}\}$) of selected parties and aggregates them. It stores the result of aggregation in $rnd\_sum_{self}^{P_m}$.

(b) **Identifying the parties for which P_m participate in the secure aggregation:**
Party P_m uses its *received SSP* seeds ($\{SSP_{all}^{P_1}, \ldots, SSP_{all}^{P_n}\}$) to identify the parties with whose *received SSA* seeds, P_m must generate random masks. Then,

P_m generates random masks based on the *received SSA* seeds ($\{SSA^{P_1}_{P_m}, \ldots, SSA^{P_n}_{P_m}\}$) of selected parties and aggregates them. It stores the result of aggregation in $rnd\_sum^{P_m}_{others}$.

(c) **Aggregation and transfer of partial results ($P.R.$) to the mediator:**
Party P_m calculates $P.R.^{P_m}$ as follows: $P.R.^{P_m} = secret\_val^{P_m} - rnd\_sum^{P_m}_{self} + rnd\_sum^{P_m}_{others}$. Then, P_m sends $P.R.^{P_m}$ to the mediator.

The mediators calculates the desired result (aggregation of secret values) by aggregating all received partial results.

Privacy: We now show that the secret values of parties are kept private in our proposed protocol. The partial result $P.R.^{P_m}$, which is shared with the mediator consists of three components: $secret\_val^{P_m}$, $rnd\_sum^{P_m}_{self}$, and $rnd\_sum^{P_m}_{others}$. The two components, $rnd\_sum^{P_m}_{self}$ and $rnd\_sum^{P_m}_{others}$, mask the secret value. The value of $rnd\_sum^{P_m}_{self}$ can only be identified by collusion of k parties holding the random seeds for generating the random masks, which are the components of $rnd\_sum^{P_m}_{self}$. At the same time, $rnd\_sum^{P_m}_{others}$ can only be identified by collusion of k (potentially) other parties which generate the components of $rnd\_sum^{P_m}_{others}$. In the worst case, the k parties involved in $rnd\_sum^{P_m}_{self}$ and $rnd\_sum^{P_m}_{others}$ may be the same; hence, the minimum number of colluding data holder parties equals to k. Moreover, since the mediator receives the victim's partial result, the collusion of other parties without the mediator's participation is ineffective. Therefore, for identifying a secret value, the collusion of k data holder parties and the mediator is necessary.

Correctness: We also show that the final value of aggregation of partial results is equal to the aggregation of secret values. Without loss of generality we consider $k = n - 1$. The aggregation of all partial results sent to the mediator is as follows:

$$\begin{aligned}\sum_{j=1}^{n} P.R.^{P_j} &= secret\_val^{P_1} - rnd\_sum^{P_1}_{self} + rnd\_sum^{P_1}_{others} \\ &\vdots \\ &+ secret\_val^{P_n} - rnd\_sum^{P_n}_{self} + rnd\_sum^{P_n}_{others} \\ &= \sum_{j=1}^{n} secret\_val^{P_j} - \sum_{j=1}^{n} rnd\_sum^{P_j}_{self} + \sum_{j=1}^{n} rnd\_sum^{P_j}_{others}.\end{aligned} \quad (1)$$

Based on (a), $rnd\_sum^{P_m}_{self} = \sum_{i=1}^{n} rnd^{P_m}_{P_i} - rnd^{P_m}_{P_m}$, where $rnd^{P_m}_{P_i}$ is the shared random mask between P_m and P_i. On the other hand, based on (b), $rnd\_sum^{P_m}_{others} = \sum_{i=1}^{n} rnd^{P_i}_{P_m} - rnd^{P_m}_{P_m}$. Substituting these two equations in equation 1, we obtain:

$$\begin{aligned}\sum_{j=1}^{n} P.R.^{P_j} &= \sum_{j=1}^{n} secret\_val^{P_j} - \sum_{j=1}^{n} rnd\_sum^{P_j}_{self} + \sum_{j=1}^{n} rnd\_sum^{P_j}_{others} \\ &= \sum_{j=1}^{n} secret\_val^{P_j} - \sum_{j=1}^{n} (\sum_{i=1}^{n} rnd^{P_j}_{P_i} - rnd^{P_j}_{P_j}) \\ &+ \sum_{j=1}^{n} (\sum_{i=1}^{n} rnd^{P_i}_{P_j} - rnd^{P_j}_{P_j}) = \sum_{j=1}^{n} secret\_val^{P_j}.\end{aligned} \quad (2)$$

The above equations show that the aggregation of partial results from data holder parties is equal to the aggregation of data holder parties' secret values.

3.2. Efficient Handling of Large Scale Data

In distributed ERT, all the data holder parties participate in (collaborate on) the process of selecting the best decision node/leaf at every round of the algorithm. However, in order to efficiently handle large scale data sets and reduce the communication/computation overheads, in k-PPD-ERT, only a subset of data holder parties participate in the process of learning at every round. The probability of participation of each party in the learning process at each round is a parameter that is set by the user.

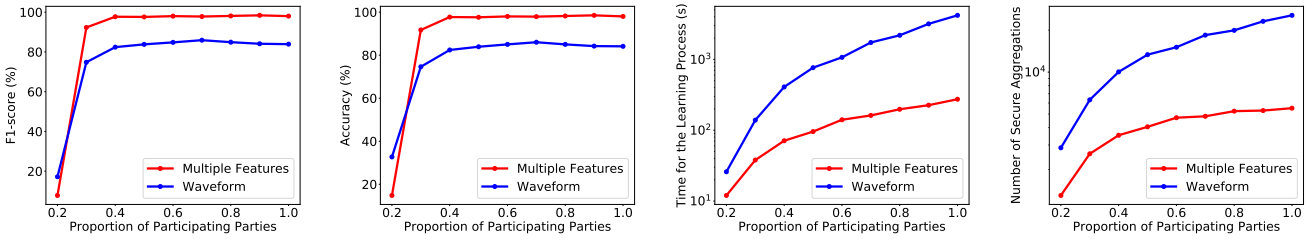
The algorithm uses the aggregation of data holder parties' partial results to calculate the candidate decision nodes' score/information gain. In certain rounds, the result of this aggregation is used to select a leaf for the tree. In k-PPD-ERT, when not all the parties participate in the aggregation process, the result of aggregation changes. However, in Section 4, we experimentally show that this technique does not lead to a major loss in the classification performance of our learned models.

Random participation of data holders in the described process changes the result of aggregation and, consequently, the learning. However, the learning results are not noticeably affected (shown experimentally in Section 4). The randomness in the participation of data holder parties, similar to the randomness in the generation of candidate decision nodes in the distributed ERT, is another source of randomness in our approach. Introducing another source of randomness in ensemble learning methods while keeping the algorithm's ability to generate weak classifiers is in accordance with the logic of bagging.

To determine which parties participate at each round, the mediator shares a common random seed (Seed for Participating Parties, SPP) with all data holder parties. Therefore, by using this seed and the constant probability of participation, every party determines the participating parties in that round of secure aggregation (for selecting the best candidate decision node/leaf). Then, each participating party picks its k peer parties for secure aggregation based on the available parties in that round, determined by SPP.

4. EVALUATION AND DISCUSSION

In this section, first, we evaluate the adopted secure aggregation technique. We compare our technique with distributed ERT and Shamir's techniques. These secure aggregation techniques are evaluated based on the communication cost in one round of secure aggregation and the minimum number of parties that need to participate in collusion in order to identify a secret value. Then, we examine the limited participation of data holder parties in the process of selecting the best candidate decision node/leaf. We evaluate the classification performance and the scalability of k-PPD-ERT offered by adopting this approach.



(a) F1-score w.r.t. proportion of participating parties (b) Accuracy w.r.t. proportion of participating parties (c) Elapsed learning time w.r.t. proportion of participating parties (d) Secure aggregations w.r.t. proportion of participating parties

Fig. 2: Analysis of the classification performance, the elapsed learning time, and number of secure aggregations for learning based on different proportions of participating parties in the learning process

Table 1 exhibits and compares communication costs (in one round of secure aggregation) and the minimum number of parties necessary to collude for identifying a secret value. According to the table, the communication complexity of k-PPD-ERT has the lowest order, while offering the highest minimum number of colluding parties for identifying a secret value. The communication complexity of the k-PPD-ERT technique is $O(n)$, similar to the distributed ERT, while this equals to $O(n^2)$ for Shamir's technique. On the other hand, the minimum number of colluding parties for k-PPD-ERT is k data holder parties plus the mediator, which is the highest. Therefore, the k-PPD-ERT's secure aggregation technique offers privacy with multiple colluding parties, while preserving the algorithm's scalability.

Table 1: Scalability and privacy comparison against existing techniques

Approach	Party	Communication (N is the number of parties)			Min Number of Colluding Parties
		Send	Receive	Total (All N parties)	
Distributed ERT	All	1	1	$2N$	1
k-PPD-ERT	Data Holders	1	0	$2(N-1)$	$k+1$ $(k<N)$
	Mediator	0	$N-1$		
Shamir [31]	k-1 Parties	N	$N-1$	$2(N^2-N+k-1)$	k $(k<N)$
	One Party	$N-1$	$N+k-2$		
	The Rest	$N-1$	$N-1$		

In k-PPD-ERT's secure aggregation technique, the total number of send and receive messages in k-PPD-ERT is independent of k, so we can always set k to $n-1$. This does not introduce any cost, concerning the communication, in our algorithm.

We now evaluate data holder parties' limited participation at every round of a selecting decision node/leaf. To investigate this feature, we use Multiple Features [32] and Waveform Database Generator (Version 1) [33] datasets, and allocate $2/3$ of each dataset for learning and the rest for testing. We distribute the training data evenly among ten parties. The mediator learns an ensemble of 25 decision trees by k-PPD-ERT in every experiment. We repeat the learning process for situations in which the proportion of participating parties at every round of selecting the best decision node/leaf is $0.2, 0.3, 0.4, ..., 1$. Figure 2 visualizes the results of these experiments. In every experiment, we record: the classification performance, shown in Figure 2a and 2b, the elapsed time for learning process, in Figure 2c, and the number of required secure aggregations for the learning process, in Figure 2d. The Y-axis in Figure 2c and 2d has a logarithmic scale (because of the differences in the magnitude of results for Multiple Features and Waveform datasets).

On the one hand, the results in Figure 2a and 2b show that random participation of only 40% of data holder parties at each round leads to high classification performance. The difference in classification performance for 40% of participation and more (even when all parties participate, similar to distributed ERT) is negligible. Furthermore, in some experiments with data holders' limited participation, we obtain models with higher classification performance. The logic behind bagging and the introduced source of randomness in k-PPD-ERT may explain these improvements.

On the other hand, the results in Figure 2c and 2d show improvements concerning the scalability when fewer data holders participate in learning at each round. Figure 2c shows the decrease of elapsed time for learning a model by reducing the number of participating parties. In addition, Figure 2d exhibits the continuous growth of secure aggregation rounds by increasing the number of parties that participate in different rounds of selecting a decision node/leaf for our decision trees.

The results in Figure 2 show that our algorithm's scalability improves by limiting the number of data holder parties that participate in every round of selection of a decision node/leaf. However, the learning performance and its resulting models will not have any noticeable loss.

5. CONCLUSION

In this paper, we consider the distributed ERT framework and extend it by adopting a secure aggregation technique that is resilient to the collusion of up to k data holder parties and the mediator. We further proposed a scalable implementation for our framework, which is efficient w.r.t. the communication overhead. Moreover, we investigated the efficient handling of large scale data with the limited participation of data holder parties at every round of the learning process. Our evaluation demonstrates the privacy preservation and resilience of the proposed framework w.r.t. the number of colluding parties and its scalability and robustness for large scale data w.r.t. the participation of a subset of data holder parties.

6. REFERENCES

[1] Samuel D Lustgarten, Yunkyoung L Garrison, Morgan T Sinnard, and Anthony WP Flynn, "Digital privacy in mental healthcare: current issues and recommendations for technology use," *Current Opinion in Psychology*, 2020.

[2] Damian Pascual, Alireza Amirshahi, Amir Aminifar, David Atienza, Philippe Ryvlin, and Roger Wattenhofer, "Epilepsygan: Synthetic epileptic brain activities with privacy preservation," *IEEE Transactions on Biomedical Engineering*, 2020.

[3] Fatih Emekçi, Ozgur D Sahin, Divyakant Agrawal, and Amr El Abbadi, "Privacy preserving decision tree learning over multiple parties," *Data & Knowledge Engineering*, 2007.

[4] Rakesh Agrawal and Ramakrishnan Srikant, "Privacy-preserving data mining," in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 2000.

[5] Shipra Agrawal and Jayant R Haritsa, "A framework for high-accuracy privacy-preserving mining," in *21st International Conference on Data Engineering (ICDE'05)*. IEEE, 2005.

[6] Alexandre Evfimievski, Ramakrishnan Srikant, Rakesh Agrawal, and Johannes Gehrke, "Privacy preserving mining of association rules," *Information Systems*, 2004.

[7] Shariq J Rizvi and Jayant R Haritsa, "Maintaining data privacy in association rule mining," in *VLDB'02: Proceedings of the 28th International Conference on Very Large Databases*. Elsevier, 2002.

[8] Rakesh Agrawal, Ramakrishnan Srikant, Johannes Gehrke, and Alexandre Evfimievski, "Privacy preserving mining of association rules," *Information systems*, 2004.

[9] Hillol Kargupta, Souptik Datta, Qi Wang, and Krishnamoorthy Sivakumar, "On the privacy preserving properties of random data perturbation techniques," in *Third IEEE international conference on data mining*. IEEE, 2003.

[10] Zhengli Huang, Wenliang Du, and Biao Chen, "Deriving private information from randomized data," in *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, 2005.

[11] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2007.

[12] Noman Mohammed, Benjamin CM Fung, Patrick CK Hung, and Cheuk-kwong Lee, "Anonymizing healthcare data: a case study on the blood transfusion service," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009.

[13] Latanya Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2002.

[14] Amin Aminifar, Yngve Lamo, Ka I Pun, and Fazle Rabbi, "A practical methodology for anonymization of structured health data," in *Proceedings of the 17th Scandinavian Conference on Health Informatics*, 2019.

[15] "Health informatics — Pseudonymization," Standard, International Organization for Standardization, 2017.

[16] Chris Clifton, Murat Kantarcioglu, Jaideep Vaidya, Xiaodong Lin, and Michael Y Zhu, "Tools for privacy preserving distributed data mining," *ACM Sigkdd Explorations Newsletter*, 2002.

[17] Yehuda Lindell and Benny Pinkas, "Privacy preserving data mining.," *Journal of cryptology*, 2002.

[18] Murat Kantarcioglu, "A survey of privacy-preserving methods across horizontally partitioned data," in *Privacy-preserving data mining*. Springer, 2008.

[19] Jaideep Vaidya, "A survey of privacy-preserving methods across vertically partitioned data," in *Privacy-preserving data mining*. Springer, 2008.

[20] Yoshinori Aono, Takuya Hayashi, Lihua Wang, Shiho Moriai, et al., "Privacy-preserving deep learning via additively homomorphic encryption," *IEEE Transactions on Information Forensics and Security*, 2017.

[21] Benny Pinkas, "Cryptographic techniques for privacy-preserving data mining," *ACM Sigkdd Explorations Newsletter*, 2002.

[22] Dionisije Sopic, Amin Aminifar, Amir Aminifar, and David Atienza, "Real-time classification technique for early detection and prevention of myocardial infarction on wearable devices," in *2017 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 2017.

[23] Dionisije Sopic, Amin Aminifar, Amir Aminifar, and David Atienza, "Real-time event-driven classification technique for early detection and prevention of myocardial infarction on wearable systems," *IEEE transactions on biomedical circuits and systems*, 2018.

[24] Farnaz Forooghifar, Amir Aminifar, and David Atienza, "Resource-aware distributed epilepsy monitoring using self-awareness from edge to cloud," *IEEE transactions on biomedical circuits and systems*, 2019.

[25] Jakub Konečnỳ, H Brendan McMahan, Daniel Ramage, and Peter Richtárik, "Federated optimization: Distributed machine learning for on-device intelligence," *arXiv preprint arXiv:1610.02527*, 2016.

[26] H Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, et al., "Communication-efficient learning of deep networks from decentralized data," *arXiv preprint arXiv:1602.05629*, 2016.

[27] Elisa Bertino, Dan Lin, and Wei Jiang, "A survey of quantification of privacy preserving data mining algorithms," in *Privacy-preserving data mining*. Springer, 2008.

[28] Pierre Geurts, Damien Ernst, and Louis Wehenkel, "Extremely randomized trees," *Machine learning*, 2006.

[29] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee, "From local explanations to global understanding with explainable ai for trees," *Nature machine intelligence*, 2020.

[30] Amin Aminifar, Fazle Rabbi, Ka I Pun, and Yngve Lamo, "Privacy preserving distributed extremely randomized trees," in *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, 2021.

[31] Adi Shamir, "How to share a secret," *Commun. ACM*, 1979.

[32] "UCI Machine Learning Repository: multiple features data set," https://archive.ics.uci.edu/ml/datasets/Multiple+Features, Accessed: 2021-02-09.

[33] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen, *Classification and regression trees*, CRC press, 1984.

PAPER E

SCIENTIFIC PAPER V: MONITORING MOTOR ACTIVITY DATA FOR DETECTING PATIENTS' DEPRESSION USING DATA AUGMENTATION AND PRIVACY-PRESERVING DISTRIBUTED LEARNING

Amin Aminifar, Fazle Rabbi, Violet Ka I Pun, and Yngve Lamo

The 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 2021.

Monitoring Motor Activity Data for Detecting Patients' Depression Using Data Augmentation and Privacy-Preserving Distributed Learning

Amin Aminifar[1], Fazle Rabbi[1,2], Violet Ka I Pun[1,3], and Yngve Lamo[1]

Abstract— **Wearable devices are currently being considered to collect personalized physiological information, which is lately being used to provide healthcare services to individuals. One application is detecting depression by utilization of motor activity signals collected by the ActiGraph wearable wristbands. However, to develop an accurate classification model, we require to use a sufficient volume of data from several subjects, taking the sensitivity of such data into account. Therefore, in this paper, we present an approach to extract classification models for predicting depression based on a new augmentation technique for motor activity data in a privacy-preserving fashion. We evaluate our approach against the state-of-the-art techniques and demonstrate its performance based on the mental health datasets associated with the Norwegian INTROducing Mental health through Adaptive Technology (INTROMAT) Project.**

I. Introduction

Mental health disorders are the primary contributor to chronic diseases in Europe [1]. Twenty-five percent of people develop at least one mental or behavioral disorder in their life [2]. Depression is the most prevalent among mental health disorders and is expected to increase in the following years [3], [4], [5]. Therefore, addressing and controlling depression is necessary for society as it affects individuals' physical, emotional, and economic aspects [6].

Wearable devices provide the opportunity to monitor patients on a long-term basis to detect and prevent health disorders in earlier stages [7], [8], [9], [10]. Wearable technologies offer pervasive healthcare solutions at an affordable price by removing time and location restrictions [11]. The data collected by such devices has attracted a lot of attention for mental health applications [12]. One such application is detecting depression in patients based on motor activity data collected from ActiGraph wristband [13]. The motor activity is captured by the accelerometry signals acquired by the ActiGraph wristband. Figure 1 explains a scenario for the analysis of sensor data generated by wearable devices. In this figure, the activity signal of each individual is collected by a wristband, and is transferred to the personal mobile phone. Then, the raw data may be preprocessed and prepared for the analysis task locally on the phone or analyzed in a distributed fashion [14], [15].

Monitoring mental health and, in particular, depression by using signals collected by wearable devices involve several challenges. Firstly, sharing healthcare data for analysis purposes is not always feasible due to privacy and legal concerns [16], [17], [18], [19]. In particular, privacy and security are among the most concerning challenges in real-time health monitoring using mobile health technologies [20], [21], [22]. Privacy-preserving data sharing, e.g., [23], [24], [25], and privacy-preserving data mining [26], [27], [28], [29], [30], [15] approaches offer a solution to data analysis without the raw data leaving the individuals' devices. Secondly, although there is a connection between mental health problems and disturbance in internal biological systems, relations between mood and physiological signals are not well-identified [31], [13]. Therefore, finding the correlation between physiological signals and mental health problems is challenging.

This paper addresses analyzing motor activity data collected by the ActiGraph wristband. We use the Depresjon (depression in Norwegian) dataset [1] [13] which contains motor activity signals of patients from control (non-depressed) and condition (depressed) groups. Our goal is to predict depression in patients based on such data. Previous studies [13] have considered a feature-based approach for the detection of depression. However, as we show in this paper, the prediction performance may be improved by further exploiting the information carried in the signals (beyond the basic statistical attributes, e.g., the mean and standard deviation).

In this paper, we propose an augmentation approach for generating new records from the Depresjon dataset [13] to improve the classification performance. In other words, our approach produces new data records from the raw data in order to use them for the learning and evaluation process. We show that adopting our augmentation approach leads to learning classification models with higher performance, i.e., up to 7.9% higher F1-score, 8.2% higher Accuracy, and 0.169 higher Matthews Correlation Coefficient. However, the motor activity raw data that is required for the analysis is generated on each patient's wearable device and inherently distributed. Such data cannot be transferred to a center for further analysis due to personal and/or legal privacy concerns (e.g., to infer mental health status from the data). To address this privacy issue, we investigate the possibility of using our recently proposed privacy-preserving distributed machine learning approach, PPD-ERT [33], for sensor data based on the Depresjon dataset, which paves the way for the real-world applications of our approach for wearable technology in the described settings.

The remainder of this article is structured as follows: The approach and details about generating records from the

[1]Western Norway University of Applied Sciences, Bergen, Norway firstname.lastname@hvl.no
[2]University of Bergen, Bergen, Norway
[3]University of Oslo, Oslo, Norway

[1]The Depresjon dataset is publicly available at [32], and is collected within the INTROMAT project.

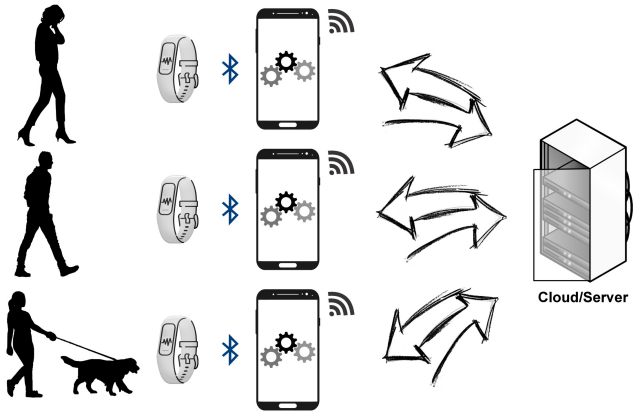


Fig. 1: Analysis of sensor data generated by wearable devices

Depresjon dataset are described in Section II. The evaluation of the approach and experimental results are presented in Section III. Finally, Section IV concludes the paper.

II. Approach

This section presents our approach to detecting depression based on the motor activity data. We, first, describe our approach for generating data records from the Depresjon dataset. Then, we discuss how PPD-ERT [33] is utilized for privacy-preserving distributed learning of classification models from generated data records and in the context of real-world wearable devices.

Let us first give an overview of the Depresjon dataset. The dataset consists of motor activity data of 55 patients (30 and 25 for females and males) collected by ActiGraph wristband worn at the right wrist of patients. In this dataset, 23 of the patients are diagnosed with depression, including both unipolar and bipolar patients, and the remaining 32 belong to the control group. Each patient wore the ActiGraph wristband for an arbitrary number of days, between 5 to 20 days. The total number of days for the condition group is 291 days, and for the control group is 402.

The recorded values (samples) for each patient in each minute are proportionate to the quantity, duration, and the strength of the patient's movements. Each patient has at least a sample value greater than or equal to zero for every minute of a day. It should also be noted that, on the first day for each patient, the recording started in the middle of the day. We refer to the data for each day of each patient as a record. Each record consists of several sample values (or samples in short).

The authors in [13] proposed the application of the mean and the standard deviation of the activity level along with the proportion of minutes with no activity in a day as the data attributes for depression classification. In addition, a normalization between zero and one is performed for attribute values. Therefore, this approach leads to only 693 records (291 for the condition group and 402 for the control group), one for each day in the raw data.

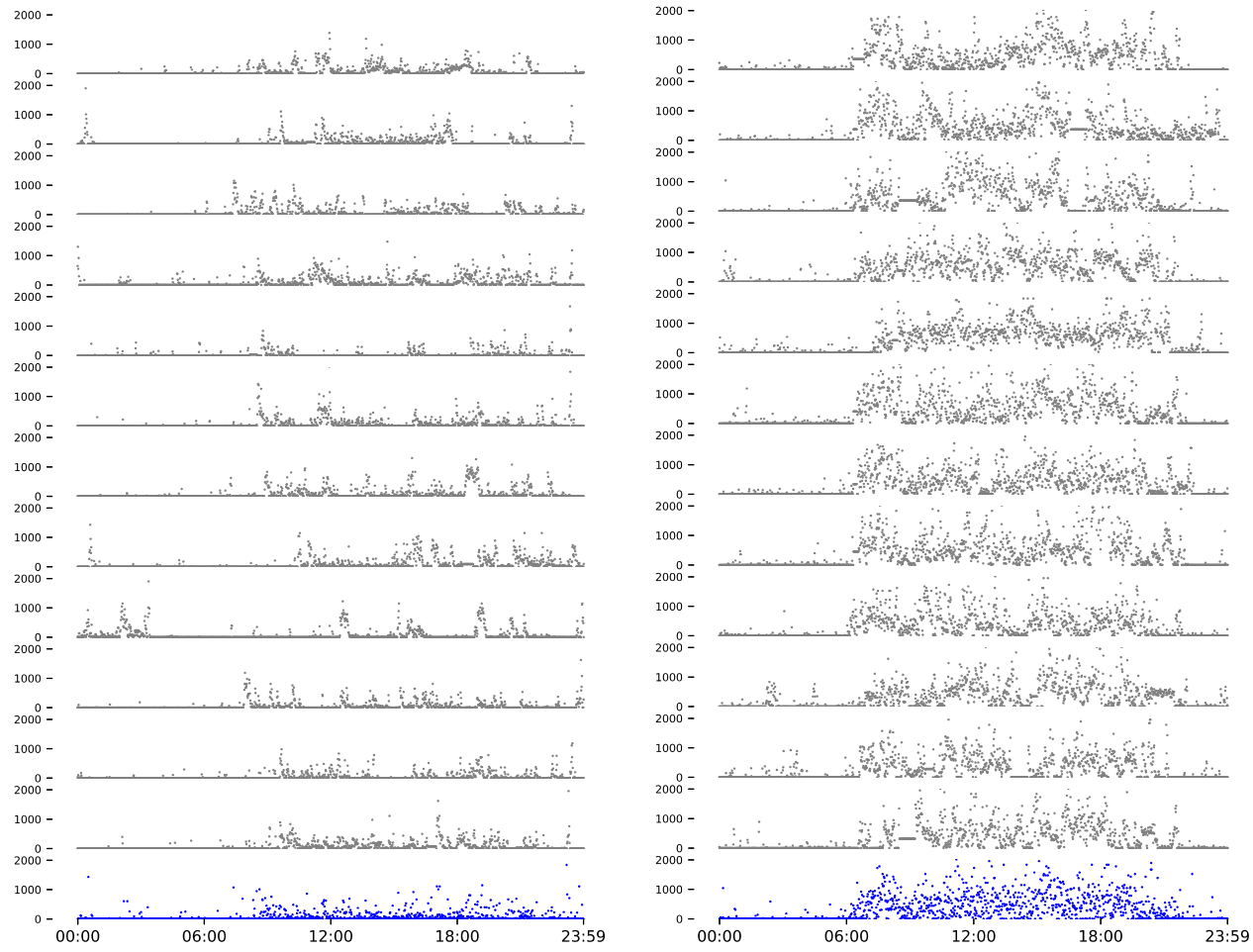
Although adopting the proposed approach in [13] extracts a representation of the raw data that results in a fair classification performance, it may still lead to suboptimal results. In this dataset, the total number of recorded data for patients is limited, i.e., only 693 days. Therefore, if we generate one record for each day, the volume of the data that the algorithm is trained on will be small, which in turn limits the detection performance. Moreover, the motor activity signal on certain days are shorter, where we do not have a recorded sample for every minute of the day. In this way, the mean, standard deviation, and the proportion of zero activity for that data are affected and will be very different from the days with complete recording. On the other hand, the number of recorded days for each patient is different. We have less than one week of recorded activity for some of the patients, while we have almost three weeks for some others. Therefore, the approach presented in [13] makes the data more imbalanced, which may eventually lead to poor classification performance.

This paper adopts a data augmentation approach for generating data records from the original Depresjon dataset. Data augmentation is a functional approach for increasing the diversity and volume of data by augmenting records at random [34], [35]. The majority of machine learning algorithms, e.g., deep neural networks, learn higher performance classification models when they are trained on larger datasets. Moreover, data augmentation can lead to better generalization and robustness by learning models invariant to the transformation of the data, e.g., learning an object classifier model that can classify objects correctly even if the images are rotated.

By adopting a data augmentation approach, we generate an equal number of records for each patient. All the generated records will have a unique size equal to the number of minutes in a day. For each patient, we generate n records, where n can be adjusted based on the user needs. The length of each record, l, is equal to the number of minutes in a day, i.e., $l = 1440$ (60×24), representing the patient activity level in one day.

Let us denote the set of all samples for patient i by S_i and define it as: $S_i = \{s_{ijk} \in R_{ij}, \forall j, k\}$. R_{ij} captures the j-th record of patient i and s_{ijk} is the sample k in record R_{ij}. For every minute t during the day, we check the available samples for this patient and for this specific time in the day, e.g., t = *12:00*. The recorded samples for different days of this patient around this particular time, i.e., $t \pm \delta$, are the candidates for being selected as the new (augmented) sample for this timestamp, where $2 \cdot \delta$ is the duration of this interval. The parameter δ determines the time interval within which we acquire the augmented sample.

$\hat{R}_{ij}$ captures the j-th augmented record of patient i and is defined as: $\hat{R}_{ij} = [\hat{s}_{ij1}, ..., \hat{s}_{ijl}]$. $\hat{s}_{ijk}$ denotes the k-th sample for the generated record $\hat{R}_{ij}$, where $1 \leq k \leq l$. $\hat{s}_{ijk}$ is the sample at time t and is selected at random from set S_i and in the time interval $[t - \delta, t + \delta]$. This is formally defined as $\hat{s}_{ijk} \in \{s \in S_i | t - \delta \leq t(s) \leq t + \delta\}$, where $t(s)$ is the time of sample s. This process is repeated until we have n



(a) Subject 20 from condition group

(b) Subject 20 from control group

Fig. 2: The figure presents two examples of generated records based on the raw data for individuals in condition and control groups. Each signal/record showed by gray dots represents one recorded day of the patients. The last signal shown by blue dots represents the generated record by our proposed approach ($l = 1440$ and $\delta = 10$). The examples show the correspondence of raw data and generated records and that the subject in the condition group usually has less activity, after the sleeping time, compared to the one from the control group.

records for each patient.[2]

The augmented record reflects the patients' activity level in a day and is proportionate to the original data since its samples for all timestamps (t) are randomly chosen from samples for the close timestamps ($t \pm \delta$) in the reported days. Therefore, the approach preserves the changes in the activity level of the patients in the data. This is particularly important, as studies found evidence that suggests a relationship between decreased daytime motor activity and increased nighttime activity and a depressive state, compared to healthy individuals [36], [13]. In similar studies, the decreased motor activity and more diversity in the activity level are reported for patients suffering from bipolar depression [37]. That being said, this means that preservation of activity level changes during the day for a patient is one of the main advantages of our augmentation approach.

Figure 2 shows the generation of records from two patients' raw data. The horizontal axis represents the time in a day, and the vertical axis shows the activity level. All patient's activity levels at different timestamps are shown in the figure by gray bubbles, and the blue bubbles are the samples for the generated record by our augmentation technique. Figures 2a and 2b show the raw signals/records (twelve days) and the augmented record for Subject 20 from the condition group and Subject 20 from the control group, respectively. The figure shows the association of the generated record and the raw data. In the intervals that the patient usually has a low level of activity, e.g., from midnight to the morning, the generated record also shows a low level of activity and vice versa.

The augmented dataset can then be used by the machine learning algorithms for the detection of depression. The raw data generated from each patient's activity are stored on

[2]The source code of our approach is available at https://github.com/AminAminifar/dataprep

TABLE I: Classification performance (leave one patient out) of different classification algorithms for the approach in [13] and the generated data records based on our approach

Algorithms		Our approach			Approach in [13]		
		F1-score	ACC	MCC	F1-score	ACC	MCC
Distributed	PPD-ERT	**76.3%**	**76.8%**	**0.518**	66.3%	67.0%	0.310
	Distributed ID3	65.1%	65.0%	0.286	65.6%	66.5%	0.296
Centralized	ERT	76.3%	76.8%	0.518	66.3%	67.0%	0.310
	Random forest	74.4%	75.1%	0.481	64.3%	64.7%	0.266
	XGBoost	76.2%	76.3%	0.510	64.3%	64.7%	0.265
	Decision Tree	65.7%	65.8%	0.293	60.6%	60.7%	0.191
	Linear SVM	69.5%	69.5%	0.375	68.4%	68.6%	0.349

patients' personal devices. Due to privacy and legal issues, such data cannot be transferred to a center for analysis in such healthcare applications. A practical solution in such situations is performing analysis through privacy-preserving distributed data analysis methods. Therefore, here we adopt our proposed PPD-ERT algorithms [33], [38] for analyzing the Depresjon dataset and learning the classification model. The ensemble learning procedure adopted by PPD-ERT reduces the risks of overfitting.

The described approach for generating data instances (augmenting data) is compatible with our privacy-preserving distributed methods. This is because the new records are generated merely based on one patient's raw data and are independent of other patients' data. Therefore, each patient generates the instances on its own device locally. Then, the generated records are the data that is used for training the PPD-ERT algorithm. By employing the PPD-ERT approach, we learn high-performance classification models without sharing raw data or sensitive information. The learned models will then be used for detecting depression by each individual.

III. Evaluation and Discussion

In this section, we evaluate our proposed augmentation technique for motor activity data. We consider several classification algorithms to assess and compare the results obtained from our proposed approach and the approach in [13]. Moreover, we use our recently proposed method, PPD-ERT [33], for the described problem, i.e., detection of depression in patients based on motor activity data, to investigate the possibility of applying this method for such data from wearable devices.

The objective here is to learn classification models to detect depression based on the motor activity signals collected by the ActiGraph wristband. The trained model will later be used to detect depression in individuals based on their activity levels. The target categories for classification are two, i.e., normal/control category and depressed/condition category.

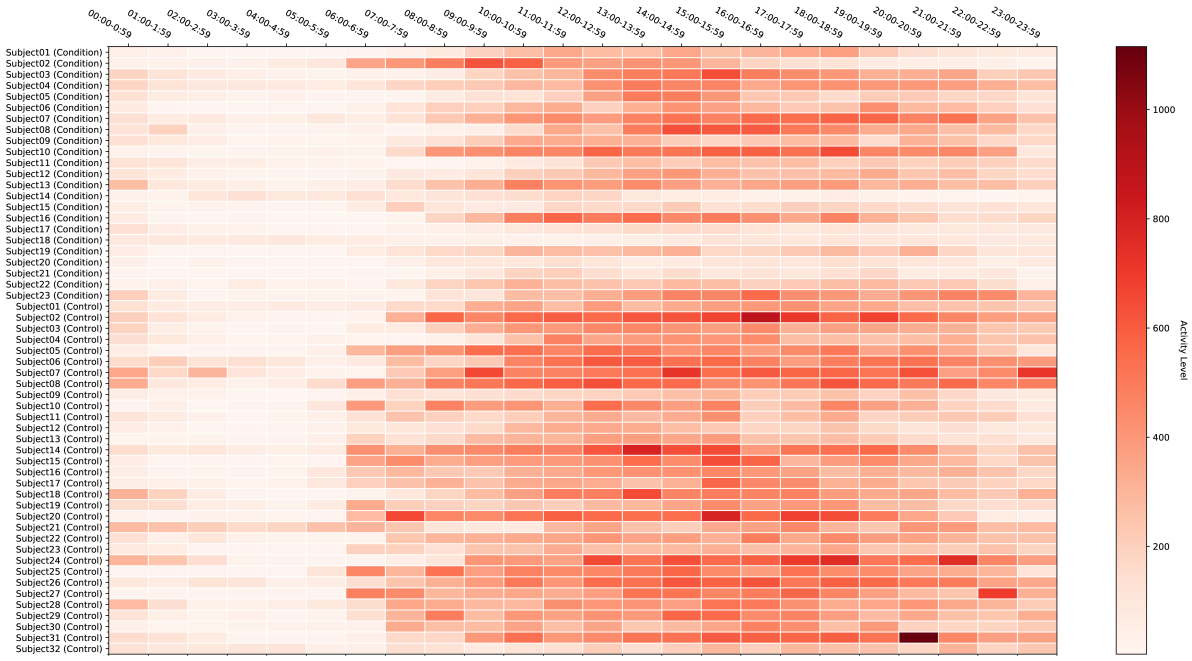
As described in Section II, [13] proposes using a dataset (obtained from original data) which contains three attributes (i.e., mean, standard deviation, and zero activity ratio) and one label for each record, and each record represents one day of collected data for one patient. This is while our approach generates records that contain a sample for each minute during the day, i.e., 1440 attributes for each record ($l = 1440$). In our experiments, we generate 100 records for each patient ($n = 100$). Every record is generated based on the samples collected at different days of a patient's collected signals. Each timestamp's sample for the record is selected among the available samples in 10 minutes time span around it ($\delta = 10$). Therefore, in both approaches, each generated record belongs to one and only one patient. This provides the possibility for leave-one-patient-out evaluation, which in turn enables the adoption of our privacy-preserving distributed learning framework.

In our experiments, we measure the classification performance of several learning algorithms on data generated based on the two approaches, with leave-one-patient-out evaluation. We perform the leave-one-out evaluation for each patient, where the target patient's data is considered as the test set and the remaining data from other patients is considered as the training set. We use F1-score (weighted average), Accuracy (ACC), and Matthews Correlation Coefficient (MCC) to measure the quality of classification, which are the metrics used for performance evaluation on this dataset [13].

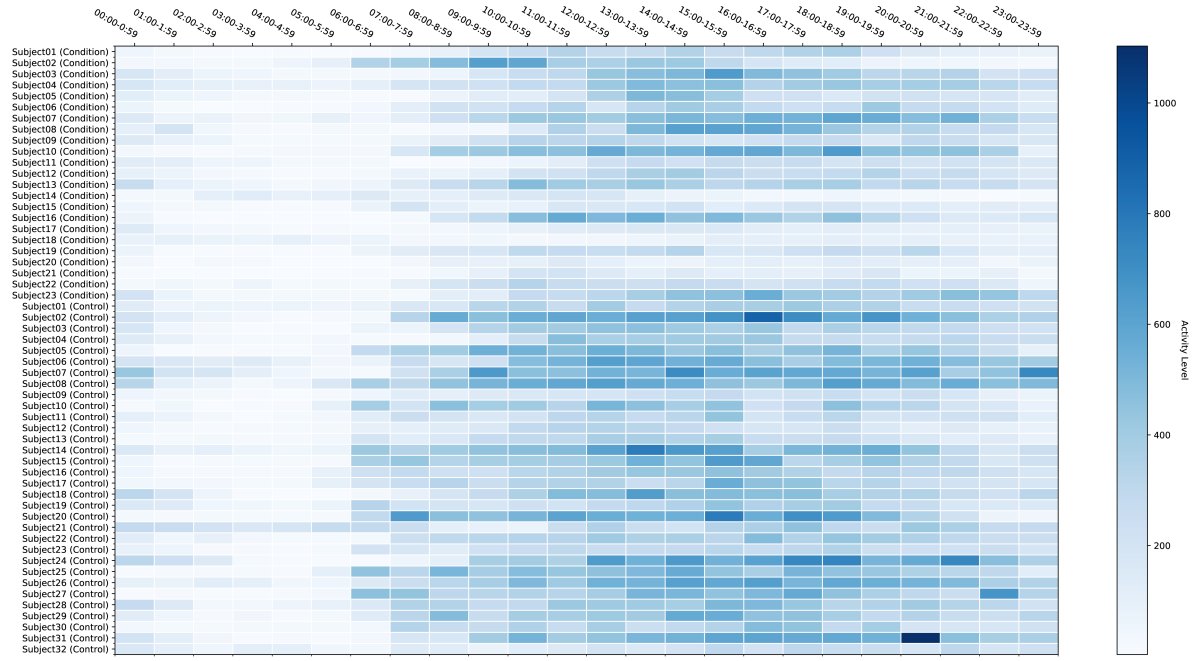
We perform the learning process on both the data with attributes proposed in [13] and generated data records by our approach, based on five centralized and two privacy-preserving distributed machine learning algorithms. Table I exhibits these results.

The results show a substantial improvement in the classification performance by employing our approach for generating data records from raw data. Particularly, tree-based ensemble learning approaches, i.e., PPD-ERT, ERT [39], random forest [40], and XGBoost [41], present more accurate results when trained on data generated by our augmentation approach. This is while training on the data with attributes proposed in [13] yields the best results when employing the linear SVM algorithm [42]. Comparing the best results for both approaches shows that applying our approach leads to learning more accurate classification models, i.e., models with 7.9% higher F1-score, 8.2% higher ACC, and 0.169 higher MCC. The PPD-ERT and ERT algorithms follow the same learning procedure and have the same classification performance [33].

Figures 3a and 3b show the heat-map for the raw and generated data, respectively. Each box represents the average activity level of one patient in one-hour intervals in a day. For the raw data, Figure 3a represents the average activity level



(a) Heat-map for the raw data



(b) Heat-map for the augmented data

Fig. 3: Heat-map for averaged activity level in one-hour intervals for each patient

of each patient based on all recorded days for him/her. Figure 3b shows the heat-map for the average activity of patients based on the generated data by adopting our approach.

The heat-maps of activity level based on the raw dataset and the generated dataset in Figures 3a and 3b are visually similar. This similarity explains the association of the generated records and the original dataset. In order to measure the similarity between the generated data by our approach and the raw data, we calculate the relative difference among the corresponding values for each cell, averaged over the entire heat-map. The average relative difference is calculated as follow:

$$D[R, A] = \frac{1}{n \cdot m} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{|r_{ij} - a_{ij}|}{r_{ij}}, \qquad (1)$$

where n is the number of patients in each group, and m is the number of one-hour intervals in a day. $R = \{r_{11}, r_{12}, \ldots, r_{nm}\}$ is the set of average activity level of one patient in one-hour intervals in a day calculated from raw dataset. The r_{ij} is the average activity level of patient i in one-hour interval j in the raw data. Moreover, $A = \{a_{11}, a_{12}, \ldots, a_{nm}\}$ is the set of average activity level from augmented dataset. The average activity level of patient i in one-hour interval j in the augmented data is captured by a_{ij}. The value of D for the condition group is 3.3%. The value of D for the control group is equal to 3.5%.

In summary, the evaluation results in this section indicate the preservation of the activity-level information in the augmented data for the detection of depression from motor activity data. Our experimental results show that modern techniques, e.g., tree-based ensemble learning algorithms, learn more accurate classifier models given such extensive information compared to learning from the few basic statistical attributes in previous studies.

IV. Conclusion

In this paper, we propose an approach based on data augmentation to analyze the Depresjon dataset and improve the performance of detecting depression in subjects. We introduced an approach for augmenting data records from the Depresjon dataset, which leads to higher detection performance when employing modern learning algorithms. Employing our approach leads to learning more accurate models with up to 7.9% higher F1-score, 8.2% higher ACC, and 0.169 higher MCC. Moreover, the possibility of employing privacy-preserving data analysis for such data is investigated. We demonstrate the possibility of using our privacy-preserving distributed data analysis technique, PPD-ERT, for wearable devices/sensors to ensure the preservation of the privacy of sensitive information for the patients in the context of depression and mental health disorders.

ACKNOWLEDGMENT

This research is supported by INTROducing Mental health through Adaptive Technology (INTROMAT) project. The paper is partially supported by SIRIUS: Centre for Scalable Data Access.

References

[1] "Mental health: data and resources," https://www.euro.who.int/en/health-topics/noncommunicable-diseases/mental-health/data-and-resources, accessed: 2021-02-17.

[2] W. H. Organization *et al.*, "Mental and neurological disorders," in *Mental and neurological disorders*, 2001.

[3] M. Olfson, B. G. Druss, and S. C. Marcus, "Trends in mental health care among children and adolescents," *New England Journal of Medicine*, 2015.

[4] G. V. Polanczyk, G. A. Salum, L. S. Sugaya, A. Caye, and L. A. Rohde, "Annual research review: A meta-analysis of the worldwide prevalence of mental disorders in children and adolescents," *Journal of child psychology and psychiatry*, 2015.

[5] J. M. Twenge, "Time period and birth cohort differences in depressive symptoms in the us, 1982–2013," *Social Indicators Research*, 2015.

[6] M. A. Vammen, S. Mikkelsen, Å. M. Hansen, J. P. Bonde, M. B. Grynderup, H. Kolstad, L. Kærlev, O. Mors, R. Rugulies, and J. F. Thomsen, "Emotional demands at work and the risk of clinical depression: a longitudinal study in the danish public sector," *Journal of occupational and environmental medicine*, 2016.

[7] D. Sopic, A. Aminifar, A. Aminifar, and D. Atienza, "Real-time classification technique for early detection and prevention of myocardial infarction on wearable devices," in *2017 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 2017.

[8] ——, "Real-time event-driven classification technique for early detection and prevention of myocardial infarction on wearable systems," *IEEE transactions on biomedical circuits and systems*, 2018.

[9] D. Sopic, A. Aminifar, and D. Atienza, "e-glass: A wearable system for real-time detection of epileptic seizures," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2018, pp. 1–5.

[10] F. Forooghifar, A. Aminifar, L. Cammoun, I. Wisniewski, C. Ciumas, P. Ryvlin, and D. Atienza, "A self-aware epilepsy monitoring system for real-time epileptic seizure detection," *Mobile Networks and Applications*, pp. 1–14, 2019.

[11] U. Varshney, "Pervasive healthcare and wireless health monitoring," *Mobile Networks and Applications*, 2007.

[12] F. J. Penedo and J. R. Dahn, "Exercise and well-being: a review of mental and physical health benefits associated with physical activity," *Current opinion in psychiatry*, 2005.

[13] E. Garcia-Ceja, M. Riegler, P. Jakobsen, J. Tørresen, T. Nordgreen, K. J. Oedegaard, and O. B. Fasmer, "Depresjon: a motor activity database of depression episodes in unipolar and bipolar patients," in *Proceedings of the 9th ACM multimedia systems conference*, 2018.

[14] F. Forooghifar, A. Aminifar, and D. Atienza, "Resource-aware distributed epilepsy monitoring using self-awareness from edge to cloud," *IEEE transactions on biomedical circuits and systems*, 2019.

[15] S. Baghersalimi, T. Teijeiro, D. Atienza, and A. Aminifar, "Personalized real-time federated learning for epileptic seizure detection," *IEEE Journal of Biomedical and Health Informatics*, 2021.

[16] M. Malekzadeh, R. G. Clegg, A. Cavallaro, and H. Haddadi, "Mobile sensor data anonymization," in *Proceedings of the international conference on internet of things design and implementation*, 2019.

[17] D. Pascual, A. Aminifar, D. Atienza, P. Ryvlin, and R. Wattenhofer, "Synthetic epileptic brain activities using generative adversarial networks," *Machine Learning for Health (ML4H) at Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[18] D. Pascual, A. Amirshahi, A. Aminifar, D. Atienza, P. Ryvlin, and R. Wattenhofer, "Epilepsygan: Synthetic epileptic brain activities with privacy preservation," *IEEE Transactions on Biomedical Engineering*, 2020.

[19] S. D. Lustgarten, Y. L. Garrison, M. T. Sinnard, and A. W. Flynn, "Digital privacy in mental healthcare: current issues and recommendations for technology use," *Current Opinion in Psychology*, 2020.

[20] A. Aminifar, P. Eles, and Z. Peng, "Optimization of message encryption for real-time applications in embedded systems," *IEEE Transactions on Computers*, 2017.

[21] A. Aminifar, "Minimal adversarial perturbations in mobile health applications: The epileptic brain activity case study," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 1205–1209.

[22] ——, "Universal adversarial perturbations in epileptic seizure detection," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–6.

[23] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2002.
[24] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2007.
[25] A. Aminifar, Y. Lamo, K. I. Pun, and F. Rabbi, "A practical methodology for anonymization of structured health data," in *Proceedings of the 17th Scandinavian Conference on Health Informatics*, 2019.
[26] M. Kantarcioglu, "A survey of privacy-preserving methods across horizontally partitioned data," in *Privacy-preserving data mining*. Springer, 2008.
[27] J. Vaidya, "A survey of privacy-preserving methods across vertically partitioned data," in *Privacy-preserving data mining*. Springer, 2008.
[28] J. Konečnỳ, H. B. McMahan, D. Ramage, and P. Richtárik, "Federated optimization: Distributed machine learning for on-device intelligence," *arXiv preprint arXiv:1610.02527*, 2016.
[29] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, *et al.*, "Communication-efficient learning of deep networks from decentralized data," *arXiv preprint arXiv:1602.05629*, 2016.
[30] M. Malekzadeh, B. Hasircioglu, N. Mital, K. Katarya, M. E. Ozfatura, and D. Gündüz, "Dopamine: Differentially private federated learning on medical data," *arXiv e-prints*, pp. arXiv–2101, 2021.
[31] E. M. Marco, E. Velarde, R. Llorente, and G. Laviola, "Disrupted circadian rhythm as a common player in developmental models of neuropsychiatric disorders," *Neurotoxin Modeling of Brain Disorders—Life-long Outcomes in Behavioral Teratology*, 2015.
[32] "The Depresjon Dataset," https://datasets.simula.no/depresjon/, accessed: 2021-07-30.
[33] A. Aminifar, F. Rabbi, K. I. Pun, and Y. Lamo, "Privacy preserving distributed extremely randomized trees," in *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, 2021.
[34] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation strategies from data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.
[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, 2012.
[36] C. Burton, B. McKinstry, A. S. Tătar, A. Serrano-Blanco, C. Pagliari, and M. Wolters, "Activity monitoring in patients with depression: a systematic review," *Journal of affective disorders*, 2013.
[37] J. Scott, G. Murray, C. Henry, G. Morken, E. Scott, J. Angst, K. R. Merikangas, and I. B. Hickie, "Activation in bipolar disorders: a systematic review," *JAMA psychiatry*, 2017.
[38] A. Aminifar, F. Rabbi, and Y. Lamo, "Scalable privacy-preserving distributed extremely randomized trees for structured data with multiple colluding parties," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021.
[39] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning*, 2006.
[40] T. K. Ho, "Random decision forests," in *Proceedings of 3rd international conference on document analysis and recognition*. IEEE, 1995.
[41] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016.
[42] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, 1995.

PAPER F

SCIENTIFIC PAPER VI: EXTREMELY RANDOMIZED TREES WITH PRIVACY PRESERVATION FOR DISTRIBUTED STRUCTURED HEALTH DATA

Amin Aminifar, Matin Shokri, Fazle Rabbi, Violet Ka I Pun, and Yngve Lamo

In IEEE Access. 2022.

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 15, 2021, accepted December 31, 2021, date of publication January 11, 2022, date of current version January 18, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3141709

Extremely Randomized Trees With Privacy Preservation for Distributed Structured Health Data

AMIN AMINIFAR[1], MATIN SHOKRI[2], FAZLE RABBI[1,3], VIOLET KA I. PUN[1,4], AND YNGVE LAMO[1]

[1]Department of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, 5063 Bergen, Norway
[2]Faculty of Computer Engineering, K. N. Toosi University of Technology, Tehran 19697-64499, Iran
[3]Department of Information Science and Media Studies, University of Bergen, 5007 Bergen, Norway
[4]Department of Informatics, University of Oslo, 0315 Oslo, Norway

Corresponding author: Amin Aminifar (amin.aminifar@hvl.no)

This work was supported by the INTROducing Mental health through the Adaptive Technology (INTROMAT) Project by the Norwegian Research Council (NFR) under Grant 259293.

ABSTRACT Artificial intelligence and machine learning have recently attracted considerable attention in the healthcare domain. The data used by machine learning algorithms in healthcare applications is often distributed over multiple sources, for instance, hospitals or patients' personal devices. One main difficulty lies in analyzing such data without compromising patients' privacy and personal data, which is a primary concern in healthcare applications. Therefore, in these applications, we are interested in running machine learning algorithms over distributed data without disclosing sensitive information about the data subjects. In this paper, we propose a distributed extremely randomized trees algorithm for learning from distributed data with privacy preservation. We present the implementation of our technique (which we refer to as *k*-PPD-ERT) on a cloud platform and demonstrate its performance based on medical data, including Heart Disease, Breast Cancer, and mental health datasets (Depresjon and Psykose datasets) associated with the Norwegian INTROducing Mental health through Adaptive Technology (INTROMAT) project.

INDEX TERMS Distributed learning, extremely randomized trees, privacy-preserving machine learning, structured health data, federated machine learning.

I. INTRODUCTION

Artificial intelligence (AI) and automated decision-making have the potential to improve accuracy and efficiency in healthcare applications. In particular, AI is proven to outperform medical experts in certain domains. Two examples are the classification of rhythms in electrocardiography signals with deep neural networks in [1] and prediction of breast cancer using the AI system presented in [2]; more related studies can be found in [3], [4]. However, the application of AI and machine learning for automated decision-making in healthcare comes with challenges, such as security and privacy. For instance, a patient's privacy is violated if sharing his/her medical data with a third-party data recipient reveals that he/she has a medical condition. This becomes more challenging considering that, in healthcare systems, the data could be distributed over a number of sources rather than being stored in a central database.

In distributed settings, hospitals need to apply data mining methods to extract useful patterns from patients' data. Although hospitals may individually be able to use their limited resources and locally stored health information to perform data mining, the use of available health information across several hospitals leads to obtaining more valuable and accurate information. However, this is a challenging task due to privacy and legal concerns. Hospitals often need to comply with privacy regulations that restrict sharing health information about patients with other parties, e.g., other hospitals, family doctors, and specialists [5], [6]. A similar problem exists when the data is distributed over patients' personal devices, such as mobile phones or wearable devices [7]–[11].

Traditionally, it was assumed that all sources holding part of the data might share their information with a trusted party. However, such an assumption, i.e., putting this level of trust

The associate editor coordinating the review of this manuscript and approving it for publication was Liangxiu Han.

in a third party, is not feasible in every scenario because the privacy of data sources cannot be protected from the third party [12]. In order to address the privacy concern, one solution would be to perturb the data before sharing it. However, perturbation-based solutions have limitations in satisfying data privacy and data utility requirements [13], [14]. This is because the utility of the data will decrease if the perturbation is not precisely controlled, and the privacy will not be preserved if the perturbation is not sufficient [14]. Similarly, anonymization techniques, e.g., [15]–[20], share an altered version of data to prevent the re-identification of data subjects [21]. Moreover, methods providing differential privacy [22] share data while preserving the privacy of individuals by adding noise. Nevertheless, in these techniques, there is always a trade-off between data privacy and data utility [13].

Previous studies also consider cryptographic techniques and secure multi-party computation methods for conducting privacy-preserving data mining [23]–[25]. However, such approaches are inefficient, mainly when dealing with large-scale data, due to considerable communication and computation costs [14]. Several techniques, e.g., [12], [26], [27], have been proposed to address these types of overheads in the privacy-preserving machine learning algorithms and to improve their efficiency.

In this paper, we target the problem of learning from data held on multiple sources without explicit sharing of raw information. We assume that the learning data is horizontally partitioned, meaning that different records of data are stored on different sources. We focus on the classification problem and structured health data, which can be stored in spreadsheets. We build upon our previous work [28] and propose a scalable privacy-preserving framework for distributed machine learning based on the extremely randomized trees algorithm, which has a linear overhead in the number of parties and can handle missing values. We refer to our approach as k-PPD-ERT (Privacy-Preserving Distributed Extremely Randomized Trees), in which k is the number of colluding parties in our approach. We use two popular publicly available healthcare datasets for performance evaluation, i.e., the Heart Disease [29] and the Breast Cancer Wisconsin (Diagnostic) [30] datasets. This data represents medical applications where missing values are present, and our algorithm is designed to handle such scenarios. Finally, we present the implementation of our technique on Amazon's AWS cloud and evaluate it in a real-world setting based on the mental health datasets associated with the Norwegian INTROducing Mental health through Adaptive Technology (INTROMAT) project [31].

The remainder of this paper is organized as follows. Section II reviews the state of the art of distributed privacy-preserving machine learning techniques to address the discussed problem. Section III covers the background related to the extremely randomized trees algorithm and secure multi-party computation. In Section IV, we illustrate our proposed k-PPD-ERT method, which is an adaptation and extension of the ERT algorithm for distributed settings. Section V illustrates the distributed extremely randomized trees algorithm through a small example. In Section VI, we evaluate the performance, overhead and privacy of the proposed technique. Section VII serves as the conclusion of this article.

II. STATE OF THE ART

The topic of collaborative learning from distributed data has been discussed in the literature for many years. A wide range of distributed learning techniques has been proposed in the literature that do not explicitly consider privacy aspects [26], [32]–[34]. Nevertheless, such techniques indirectly contribute to privacy preservation by limiting the amount of data that has to be shared with other parties or transferred to central servers or the cloud.

Randomization has been adopted in several studies [35]–[38] to preserve the privacy of individuals in data mining techniques. For instance, a technique that incorporates noise into raw data before sharing and performing data mining processes is proposed in [35]. However, the original values can be estimated using noise removal techniques. Hence, such techniques do not provide strong privacy guarantees [14], [39]–[41].

Secure multi-party computation (SMC) has been employed in several studies [12], [23]–[25], [42], [43] to perform data mining over data distributed in multiple parties, where no private information except the mining results should be disclosed. In SMC, we are interested in the result of a computation without knowing the secret values required for this computation. Therefore, techniques utilizing SMC usually compute intermediate results in the learning process without revealing the secret to other parties. Although such methods can satisfy the privacy requirements, the incorporation of inefficient secure computation techniques and homomorphic encryption in the method can substantially increase the communication and computation overheads. This leads to issues related to efficiency, particularly when we have a large number of parties or when we are dealing with a high volume of data [12].

Cryptographic methods have been adopted by several studies [23], [24], [44] for achieving privacy [14]. These methods address classification, clustering, anomaly detection, etc., by employing different data mining algorithms [45]–[48]. Nevertheless, such techniques usually suffer from communication and computation overheads and are impractical when dealing with large-scale data [49].

Federated learning has been proposed to collaboratively train a model, with the orchestration of one party, while keeping the training data decentralized [26], [32], [50]. Several systematic literature reviews of the state-of-the-art federated machine learning techniques are performed in [51]–[53]. The majority of previous studies in this domain have focused on deep neural network algorithms. In such neural network algorithms, in addition to data-holder parties' contribution, i.e., gradients, sharing model parameters is also a privacy concern.

This is due to recent attacks on the neural networks, i.e., membership inference attack [54], [55]. For addressing privacy concerns, previous studies adopt differential privacy [22] in their methods [56]–[58]. However, differential privacy can degrade the performance of the machine learning model due to the trade-off between privacy and data utility [13].

In many applications, tree-based methods can be more accurate than neural networks. Deep neural network algorithms are appropriate solutions when dealing with unstructured data, e.g., for video, audio, and text in [59]–[61]. However, the tree-based methods can outperform such algorithms when dealing with structure data, where the data attributes are individually meaningful, and we do not have strong multi-scale structures related to time or space [62]. Therefore, tree-based algorithms are currently being adopted in many applications in which the training data is structured.

Tree-based machine learning techniques have been investigated in conjunction with privacy concerns and distributed learning in several studies [12], [14], [42], [63]. In [14], the authors consider the problem of learning decision trees, with Random Decision Trees (RDT) algorithm [63]. They present a technique based on homomorphic encryption and apply it for horizontally and vertically partitioned datasets. However, this approach suffers from high computational complexity. In [42], the authors propose the utilization of SMC techniques for learning decision trees based on the ID3 algorithm [64]. In this approach, the data is horizontally partitioned and distributed among two parties. The number of parties in this method can be increased to more than two, but the efficiency and scalability of the technique decrease [49]. Moreover, perturbation techniques may also be used to build approximate decision trees. In [65], the authors propose the application of Randomized Response techniques to disguise the data before transferring it to a center for learning decision trees based on their modified ID3 algorithm. Nevertheless, transferring the entire data from all sources to one center, even after applying randomization techniques, undermines our confidence in the technique's privacy.

Gradient and tree-based algorithms have been employed by several studies in conjunction with strategies related to federated learning [66]–[69]. In [68], the authors propose a privacy-preserving distributed data mining method for regression and classification based on the Gradient Boosting Decision Tree (GBDT) algorithm [70]. The trees are trained locally on data-holder parties and passed to the following parties after being modified according to differential privacy requirements [68]. Nevertheless, injecting noise into participants' contribution, model parameters, etc., can increase the learning time and degrade the results of learning due to the trade-off between privacy and data utility [13]. Similarly, in [69], the authors propose a method based on GBDT for distributed scenarios called SimFL. In this framework, each party boosts a number of trees utilizing similarity information using locality-sensitive hashing. However, their privacy model is weaker than secure multi-party computation for improving efficiency, and their model performance is not the same as GBDT but comparable to it [69].

There are other studies that propose *tree-based methods that are not gradient-based* but are under the name of federated learning, e.g., [71], [72]. In [72], the authors propose a method employing the decision tree algorithm, ID3, that uses the combination of differential privacy and secure multi-party computation for addressing privacy concerns. The model's performance is degraded compared to the performance of the machine learning model in a centralized scenario. In [71], the authors propose a solution based on the random forest algorithm [73], [74]. This method requires a third-party trusted server and employs encryption, which increases the communication and computation overheads [12].

Closely connected to this work, the authors in [12] propose a tree-based method that utilizes a secure multi-party computation technique as an additional layer in their approach to have more confidence about its privacy. Particularly, Shamir's secret sharing [75] is used to aggregate the results received from each party at every step of learning with the ID3 algorithm. The limitation in the incorporation of methods with high communication and computation overheads leads to higher efficiency. However, Shamir's secret sharing technique still introduces major overheads in communication and computation and suffers from the scalability problem.

In our preliminary study [76], we have considered the problem of privacy-preserving machine learning using the extremely randomized trees algorithm, which is only robust to two colluding parties (in the worst-case scenario). We extend this idea to k colluding parties in [28]. However, this approach suffers from quadratic complexity in the worst-case scenario, i.e., $O(n^2)$, and is limited to datasets without missing values, which is rarely a case in real-world healthcare applications. In this work, we addressed these problems and proposed a scalable privacy-preserving distributed extremely randomized trees framework, with $O(kn)$ complexity, where k can be adjusted based on the sensitivity of the data. We implement our technique on Amazon's AWS cloud and evaluate it in a real-world setting based on the mental health datasets associated with the Norwegian INTROducing Mental health through Adaptive Technology (INTROMAT) project.

III. BACKGROUND

In this section, we present a brief overview of the extremely randomized trees (ERT) algorithm and secure multi-party computation (SMC), which provide the basis for our privacy-preserving distributed machine learning framework.

A. THE ERT ALGORITHM

ERT [77] is a tree-based ensemble learning algorithm that has been widely used for solving classification problems due to its learning performance and robustness to overfitting, which are among the characteristics of tree-based ensemble learning algorithms [62], [78], [79]. However, the traditional ERT algorithm is used when the data is stored in a central location.

We adapt the ERT algorithm for distributed settings where data is stored and essentially distributed among several parties. In the following, we discuss some of the advantages of the ERT algorithm compared to other available solutions for its utilization in distributed settings.

First, since the ERT algorithm is an ensemble learning method, it is robust in tackling overfitting. Ensemble learning methods incorporate weak learners to generate weak classifiers that are independent of other generated classifiers. Therefore, based on Condorcet's jury theorem (1785) [80], the majority vote of this ensemble of learned classifiers predicts better than the vote of an individual classifier, and if we increase the number of classifiers, the accuracy improves [81]. Therefore, in the ensemble learning method, we generate a collection of classifiers instead of only one, e.g., in [12], and finally predict based on the voting result of the learned classifiers. In such ensemble learning methods, randomness parameters in the learning algorithm cause generating classifiers different from each other. In the ERT algorithm, the randomness of candidate attributes and the splitting point for every decision node in the tree are the randomness parameters [77], which result in learning different classifiers. The ERT approach follows the logic of bagging in ensemble learning. Bagging combines the learned classifiers by voting, i.e., it predicts based on the majority vote among the learned classifiers. While not increasing the bias, bagging leads to lower variance in our learned model since we are averaging, and the lower variance in the learned model reduces the risk of overfitting [78].

Second, ERT is tree-based, and tree-based algorithms have been shown to outperform other techniques for structured data that we are addressing. In [62], the authors report that for tabular-style data where the data attributes are individually meaningful and where we do not have strong multi-scale structures related to time or space, learned models from tree-based algorithms usually outperform models learned by standard deep neural networks, e.g., [26], [32]. Moreover, in the health domain's applications, the interpretability of the learned models is advantageous. The patterns that tree-based learned models unveil, particularly in the healthcare domain, may be more useful than the prediction capability of the learned model [62]. Tree-based algorithms are more interpretable compared to deep neural networks [79]. This is an advantage for ERT. However, since ERT is an ensemble learning method, and in ensemble methods, instead of learning a model with a single tree, e.g., in the ID3 algorithm [64], the algorithm constructs several trees as a model. Hence, this decreases the explainability of such approaches compared to the ID3 algorithm.

B. SECURE MULTI-PARTY COMPUTATION

The secure multi-party computation framework, initiated by Yao's Millionaires' problem [82], considers the problem of collaborative computation among several parties, each of which holds a secret value. The parties are interested in the result of a computation performed based on their secret values, while they refrain from sharing their secret values with other parties.

A simple solution for computing the desired value without sharing secret values with other parties is to share them with a party that is trusted by everyone. The trusted party can then perform the computation and return the result to all parties. However, the assumption of trusted parties is not feasible in many scenarios because the privacy of parties with secret values cannot be protected from the third party, so such solutions are not practical. Therefore, based on the type of the computation and the scenarios, we need to devise other solutions to perform the desired collaborative computation in a secure way and without violating privacy.

To illustrate SMC, we describe a simple method for secure aggregation of secret values. Figure 1 represents the method for secure aggregation. In this example, we have four parties, each holding a secret value ($S.V.$), and the parties are interested in the summation of all secret values, i.e., $\sum_{i=1}^{4} S.V._i$. For securely aggregating the secret values:

(i) The first party generates a random mask, aggregates it with its secret value ($S.V._1$), and sends the result to the next party.
(ii) The following parties receive the input, aggregate it with their secret values, and send the result to the next party. The last party sends the result to the first party.
(iii) The first party receives the result from the last party, removes its random mask from the result, and informs all parties about the final result.

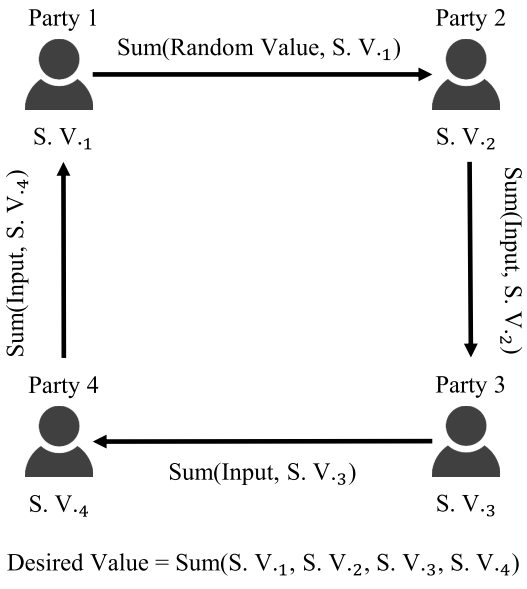


FIGURE 1. Secure aggregation.

In this way, each party cannot identify the secret value of the previous parties based on the received information. However, in this method, if two neighboring parties, i.e., the parties before and after a certain party in the ring, collude, they will be able to identify the secret value of the victim party. For instance, if Party 2 reveals the input of Party 3, and at

the same time, Party 4 reveals the output of Party 3, then they can reveal the secret value of Party 3. Therefore, the minimum number of colluding parties required for identifying a secret value is two in this method. Moreover, in terms of overhead, for one secure computation operation in this method, each party sends one message and receives one message. Thus, the communication overhead for this method is $2n$, in which n is the number of parties.

IV. PRIVACY-PRESERVING DISTRIBUTED EXTREMELY RANDOMIZED TREES

This section presents the proposed solution, which is based on the extremely randomized trees (ERT) algorithm and the secure multi-party computation (SMC) scheme. As mentioned in Section I, we refer to our approach as k-PPD-ERT, where k is the number of colluding parties in our approach in the secure aggregation process. Note that k is a parameter that can be tuned based on the privacy requirements. The algorithm preserves privacy since, on the one hand, the algorithm is distributed and the raw data is not directly shared, and, on the other hand, the partial information is aggregated using a secure multi-party computation technique. Finally, our proposed framework is based on the ERT or Extremely Randomized Trees algorithm in [77].

A. ADAPTATION OF ERT FOR DISTRIBUTED SETTINGS

This section presents the detailed procedure of learning an ensemble of decision trees based on the ERT algorithm in the discussed setting. The pseudocode of the algorithm is also provided for clarity.

1) INITIALIZATION AND START OF THE LEARNING PROCESS

We have two types of parties in our distributed learning framework. We have a *mediator* that mediates and orchestrates the overall learning process and several *data-holder parties* that collaborate with each other and the mediator to learn a classification model. Algorithm 1 and Algorithm 2 show the pseudocodes of the procedures and functions for the mediator and data-holder parties, respectively.

(a) **Sharing the Random Seeds**
To start this process, a global seed for the random function is agreed upon among all parties (Algorithm 1, Line 1 and Algorithm 2, Line 1). The global seed is common among the mediator and all data holders. In the ERT algorithm, we have two parameters of randomness for learning a weak classifier. First, we need to randomly select several attributes for the candidate decision nodes, at every step of building our decision tree (Algorithm 1, Line 24 Algorithm 2, Line 25). Second, a random splitting point for every attribute in the candidate decision node is required (Algorithm 1, Line 25, and Lines 28–35, and Algorithm 2, Line 26, and Lines 29–36). The data-holder parties and the mediator are required to use the same candidate decision nodes at every step when learning a decision tree. For this purpose, we use the global random seed that all parties, including the mediator, utilize to locally generate these candidate decision nodes (Algorithm 1, Line 11, and Algorithm 2, Line 17). This is instead of making these randomly-made candidate decision nodes in the

Algorithm 1 Mediator

1 • The global random seed (known to all parties) is set in the mediator
2 • Wait for data-holder parties' connection
3 **for** $i = 1$ **to** M **do**
4 | • Generate tree: t_i = *Build_k-PPD-ERT(0, 'None')*
5 **end**
6 $E = \{t_1, t_2, \ldots, t_M\}$
7 **Function** *Build_k-PPD-ERT(Split_ID, Branch)*
8 | • Send *Secret_aggregation(Split_ID, Branch)* request to data-holder parties
9 | • Wait until receiving the results from data-holder parties
10 | • *Sum* = aggregated the received results form data-holder parties
11 | • *Generate_splits()* (based on the global seed)
12 | **if** *number of classified records is less than* n_{min} ***or*** *labels of the classified records are the same* **then**
13 | | **return** a leaf label
14 | **else**
15 | | • Calculate each split's score (Information Gain) based on *Sum*
16 | | • Select the split with the highest score.
17 | | • Inform all parties about the selected split (for *Split_ID*)
18 | | • Build *tree_T* = *Build_k-PPD-ERT(next Split_ID, 'T')*
19 | | • Build *tree_F* = *Build_k-PPD-ERT(next Split_ID, 'F')*
20 | | • Create a node with the selected split, attach *tree_T* and *tree_F* as T and F subtrees, and **return** the resulting tree.
21 | **end**
22 **end**
23 **Function** *Generate_splits()*
24 | • Select D attributes randomly: $\{a_1, \ldots, a_D\}$
25 | • Generate D splits: $\{s_1, \ldots, s_D\}$, where s_i = *Pick_rand_split*(a_i)
26 | **return** splits $\{s_1, \ldots, s_D\}$
27 **end**
28 **Function** *Pick_rand_split(a)*
29 | **if** *a is categorical* **then**
30 | | **return** a possible category
31 | **end**
32 | **if** *a is numerical* **then**
33 | | **return** a possible value in the min and max range
34 | **end**
35 **end**

Algorithm 2 Data-Holder Party

- The global random seed (known to all parties) is set in the data-holder party
- Wait for completion of data-holder parties initialization. In initialization, k selected data-holder parties send their unique seeds to other data holders. In initialization, $SSA_{p_j}^{p_i}$ is sent by party i (i is among the k selected parties) and received by party j
- Connect to the Mediator

Function *Secret_aggregation(Split_ID, Branch)*
- $secret\_val^{P_j} = Split\_data(Split\_ID, Branch)$
- $rand\_sum_{others}^{P_j}$ = Generate and aggregate random masks based the received seeds
- **if** *the party,* P_j*, is among* k *selected data-holder parties for secure aggregation* **then**
  - $rand\_sum_{self}^{P_j}$ = Generate and aggregate random masks based the sent seeds
- **else**
  - $rand\_sum_{self}^{P_j} = 0$
- **end**
- $Result = secret\_val^{P_j} - rand\_sum_{self}^{P_j} + rand\_sum_{others}^{P_j}$
- Send *Result* to the mediator

end

Function *Split_data(Split_ID, Branch)*
- S_{sub} = records in the computational node that should be split based on *Split_ID* and *Branch*
- $\{s_1, \ldots, s_D\} = Generate\_splits()$ (based on the global seed)
- **for** $i = 1$ **to** D **do**
  - Split S_{sub} to two sets (T, F) by s_i
  - Append vectors $\{Vec_T, Vec_F\}$ representing the records' labels for each of the above sets to *Result*
- **end**
- **return** *Result*

end

Function *Generate_splits()*
- Select D attributes randomly: $\{a_1, \ldots, a_D\}$
- Generate D splits: $\{s_1, \ldots, s_D\}$, where $s_i = Pick\_rand\_split(a_i)$
- **return** splits $\{s_1, \ldots, s_D\}$

end

Function *Pick_rand_split(a)*
- **if** *a is categorical* **then**
  - **return** a possible category
- **end**
- **if** *a is numerical* **then**
  - **return** a possible value in the min and max range
- **end**

end

mediator and sharing them with all parties for further tasks. Since all parties use a common random seed, i.e., the global random seed, they generate the same candidate decision nodes at every step, without major communication overhead.

In addition, for the secure aggregation of partial results, described further in Section IV-B, k selected data-holder parties send unique seeds for the random function to other data holders through secure communication (Algorithm 2, Line 2). These random seeds are exclusive and private for each pair of data-holder parties.

(b) **Initiate the Process of Learning One Decision Tree**
The privacy-preserving distributed ERT algorithm is an ensemble learning method, therefore, we repeat the process of learning a decision tree for M times, until we have M decision trees (Algorithm 1, Lines 3–5). The number of trees, M, is a parameter tuned by the user to make a trade-off between robustness and overhead. We learn different decision trees every time due to the randomness in ERT. Finally, after repeating the process of learning a decision tree M times, we store the trees in E (Algorithm 1, Line 6). For future prediction, the ensemble of the learned trees, E, will be used.

2) THE PROCESS OF LEARNING ONE DECISION TREE

The learning of a decision tree based on the privacy-preserving distributed ERT algorithm is a recursive procedure. The procedure is executed top-down and starts from the root and ends in the leaves. For the root decision node, the *Split_ID* or the ID for the decision node is zero, and there is no previous branch, so the *Branch* input is set to '*None*'(Algorithm 1, Line 4).

(a) **Generation of Candidate Decision Nodes**
For building each decision tree, extremely randomized tree, the mediator generates the candidate decision nodes (Algorithm 1, Line 11). The mediator will further select the best decision node among the candidates based on the results received from data-holder parties. The candidate decision nodes are generated randomly, based on the global random seed, according to Algorithm 1, Lines 23–35, and Algorithm 2, Lines 24–36. The number of candidate decision nodes, D, is a parameter in the ERT algorithm tuned by the user. D attributes from all possible attributes are selected for candidate decision nodes (Algorithm 1, Line 24, and Algorithm 2, Line 25). Then, each candidate decision node's splitting point is selected (Algorithm 1, Line 25, and Algorithm 2, Line 26). If the attribute is categorical, one random possible category is selected to be checked (Algorithm 1, Lines 29–31, and Algorithm 2, Lines 30–32); otherwise, when the attribute is numerical, a point in the possible range is selected for comparison in the decision node (Algorithm 1, Lines 32–34, and Algorithm 2, Lines 33–35). We assume that all parties already have the possible categories and ranges for each attribute.

(b) **Parties Classify Their Records**
To decide about the candidate decision nodes for each branch, the mediator requires the collective outcome of

the classification with candidate decision nodes from all data holders on all their data. By having the combination of data record labels for each branch (*True* and *False*), the mediator can decide if we require a leaf or we need to calculate the score, i.e., information gain (Algorithm 1, Line 12). Information gain captures the extent of samples' purity (concerning their class/category) after splitting and is used as a basis for comparing decision nodes. The mediator sends a request to data-holder parties and waits for receiving the result from all parties, which is masked according to the secure aggregation technique described in Section IV-B (Algorithm 1, Lines 8–9). The masked results are two vectors, one for each of the *True* and *False* branches, representing the combination of data record labels after classification with each candidate decision node.

Each party receives *Split_ID* and *Branch* to determine the local records for classification (Algorithm 2, Line 16). Then, the party randomly generates candidate decision nodes based on Lines 24–36 in Algorithm 2 and the global random seed (Algorithm 2, Line 17). Next, it classifies the selected local data based on each candidate decision node and returns the result (Algorithm 2, Lines 18–22).

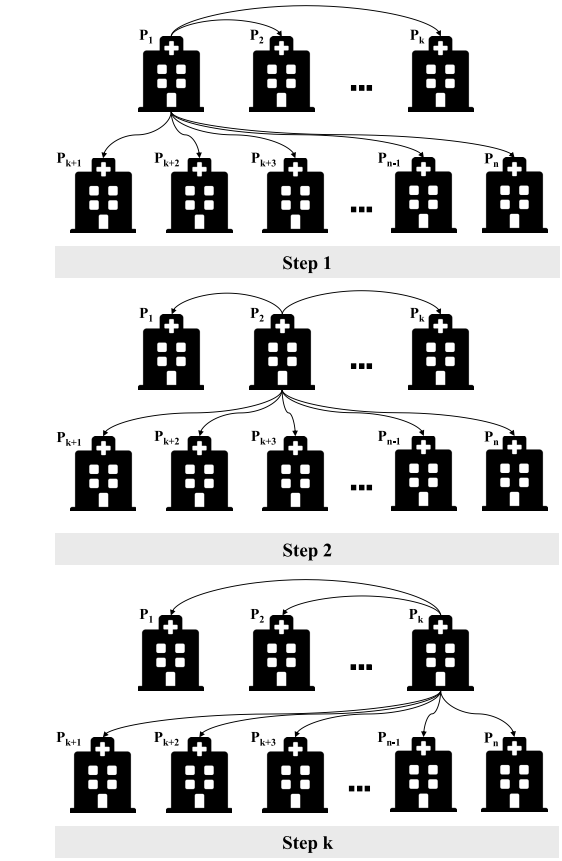
We describe how each party returns the result to the mediator in the following, using an example. Vec_T represents the combination of labels for the records that fall in the *True* branch, and Vec_F represents the combination of labels for the records that fall in the *False* branch. For instance, if three records with labels *A*, *A*, and *B* fall in the *True* branch of the candidate decision node, and we have three labels, *A*, *B*, and *C* in the dataset, then $Vec_T = [2, 1, 0]$.

(c) **Each Party Sends the Result to the Mediator**

After adopting the secure aggregation protocol described in Section IV-B, each data-holder party returns the masked result to the mediator to select the best decision node (or generate a leaf instead of a decision node). For every candidate decision node, the mediator receives and aggregates the results from all parties and obtains two vectors, for *True* and *False* branches, representing the combination of data labels (Algorithm 1, Lines 9–10).

(d) **Mediator Determines the Best Candidate for the Decision Node**

Now that the mediator has the value of *Sum* (Algorithm 1, Line 10), it determines if a decision node or a leaf node is required here in the tree (Algorithm 1, Lines 12). If all labels are the same or if the number of received labels is less than our threshold parameter, the mediator introduces a leaf node (Algorithm 1, Line 13). Otherwise, the mediator calculates the score, i.e., information gain, of each candidate decision node based on the results from data-holder parties (Algorithm 1, Line 15). It then selects the candidate decision node with the highest information gain and informs all parties about it (Algorithm 1, Lines 16–17). The selected node will be used to build the tree at the mediator (Algorithm 1, Line 20). This decision is communicated to all data-holder parties and is required to select records for classification at every step (Algorithm 2, Line 16).



(a) The k selected data-holder parties sending unique seeds to other data holders

(b) The sent and received seeds after the initialization

FIGURE 2. Initialization.

(e) **The Mediator Initiates Another Round From the First Step**

After selecting the best candidate decision node, the mediator continues the process for each branch of this

decision node. Therefore, the same process is performed from the first step, for each of the *True* and *False* branches (Algorithm 1, Lines 18–19). After returning from these recursive calls, the selected subtrees for each branch are returned (Algorithm 1, Lines 13 and 20).

B. SECURE AGGREGATION OF RESULTS FROM DATA-HOLDER PARTIES

We adopt an SMC technique in our proposed distributed ERT algorithm to avoid sharing the vectors representing the combination of the data record labels for each candidate decision node and each branch in each data-holder party. In addition to the provided privacy by not sharing the raw values of data attributes, which is by construction, the adoption of an SMC technique for aggregating the partial results from data-holder parties contributes to privacy preservation. In an extreme example, suppose our data has one sensitive attribute in it, e.g., having conducted transgender surgery before, and each data-holder party has only one record on it. Then, sharing the partial results from one party, the vectors for the combination of data record labels for each candidate decision node, can reveal sensitive information. If the candidate decision node is "whether the record falls into the transgender branch or not," the mediator can infer if that individual with the specified record has undergone transgender surgery. Therefore, to avoid such vulnerabilities, we adopt an SMC technique to aggregate the partial results from the data-holder parties.

The secure aggregation procedure begins with an initialization process. Subsequently, the parties can securely aggregate their secret values through this approach.

1) INITIALIZATION

In the initialization phase, k selected data-holder parties share their unique seeds for the random function with all parties. These seeds are unique and private between each pair of parties. Without loss of generality and for the simplicity of the presentation, we assume that the k selected data-holder parties are P_i $(\forall i \in \{1, \ldots, k\})$. Party P_i $(\forall i \in \{1, \ldots, k\})$ sends unique seeds to party P_j $(\forall j \in \{1, \ldots, n \mid i \neq j\})$. Figure 2a shows this process.

The seed party P_i shares with party P_j is represented with $SSA_{P_j}^{P_i}$, and it is a unique seed; *SSA* is the short form of Seed for Secure Aggregation. Parties *1* to k, send $n-1$ and receive $k-1$ seeds. Parties $k+1$ to n, receive k seeds. This is shown in Figure 2b. Therefore, k parties send $n-1$ and receive $k-1$ messages, and $n-k$ parties send zero and receive k messages. The total communication overhead for initialization is $2k(n-1)$. The communication overhead by adopting this approach is equal to $O(kn)$, which can be adjusted by adapting k based on the sensitivity of the data. If all parties were required to send and receive seed, then, the communication overhead would be equal to $2n(n-1)$. The communication overhead by adopting this approach is equal to $O(n^2)$ [28].

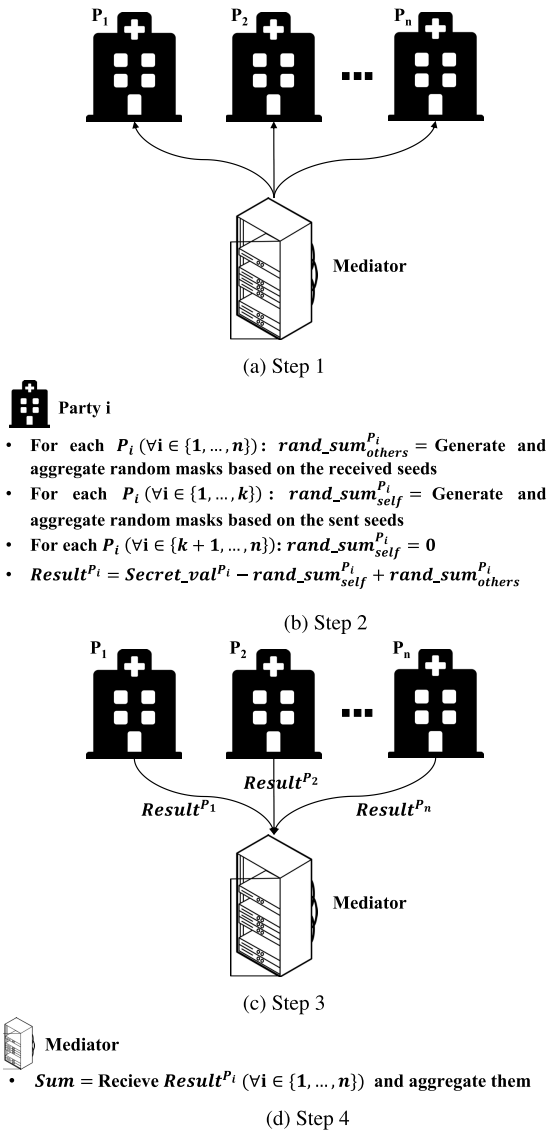


FIGURE 3. Secure aggregation.

2) SECURE AGGREGATION

In the adopted SMC technique, shown in Figure 3, parties add random masks to their partial result vectors and pass them to the mediator. The mediator aggregates the partial results received from all parties. After aggregation, the random masks from all parties cancel each other. We now describe the proposed technique in detail:

- **Step 1:** The mediator initiates the secure aggregation process round (Algorithm 1, Line 8). This is shown in Figure 3a.
- **Step 2:** Data-holder parties generate random masks and aggregate them with their secret values (Algorithm 2, Line 12). This is shown in Figure 3b.
  - Parties P_i $(\forall i \in \{1, \ldots, k\})$ generate random masks based on the sent and received seeds (Algorithm 2, Lines 6–11).
  - Parties P_i $(\forall i \in \{k+1, \ldots, n\})$ generate random masks based on received seeds (Algorithm 2, Lines 6–11).

- **Step 3:** In the next step, the parties send the masked results to the mediator (Algorithm 2, Line 13). Then, the mediator receives the results from all parties (Algorithm 1, Line 9). Figure 3c shows this.
- **Step 4:** In the last step, the mediator aggregates all the received results to obtain the desired value, i.e., the aggregated secret values from all parties (Algorithm 1, Line 10). This is shown in Figure 3d.

3) PRIVACY

We now show that the secret values of the parties are kept private in our proposed protocol. The partial result $Result^{P_i}$, which is shared with the mediator, consists of three components: $secret\_val^{P_i}$, $rnd\_sum^{P_i}_{self}$, and $rnd\_sum^{P_i}_{others}$. The two components, $rnd\_sum^{P_i}_{self}$ and $rnd\_sum^{P_i}_{others}$, mask the secret value.

- For P_i ($\forall i \in \{1, \ldots, k\}$), the value of $rnd\_sum^{P_i}_{self}$ can only be identified by the collusion of $n-1$ parties holding the random seeds for generating the random masks, which are the components of $rnd\_sum^{P_i}_{self}$. At the same time, $rnd\_sum^{P_i}_{others}$ can only be identified by the collusion of $k-1$ parties that generate the components of $rnd\_sum^{P_i}_{others}$. Therefore, the minimum number of colluding parties required to reveal the secrete value of P_i is $n-1$.
- For P_i ($\forall i \in \{k+1, \ldots, n\}$), the value of $rnd\_sum^{P_i}_{self}$ is zero and known to all, and $secret\_val^{P_i}$ is masked by $rnd\_sum^{P_i}_{others}$. However, $rnd\_sum^{P_i}_{others}$ can only be identified by the collusion of k parties that generate the components of $rnd\_sum^{P_i}_{others}$, i.e., the k selected parties for secure aggregation.

In the worst case, i.e., for P_i ($\forall i \in \{k+1, \ldots, n\}$), the k selected parties for secure aggregation are required to collude to identify a secret value; hence, the minimum number of colluding data-holder parties is equal to k. Moreover, since only the mediator receives the victim's partial result, the collusion of other parties without the mediator's participation is not possible. Therefore, for identifying a secret value, the collusion of k data-holder parties and the mediator is necessary.

4) CORRECTNESS

We also show that the final value of the aggregation of partial results is equal to the aggregation of secret values. The aggregation of all the partial results sent to the mediator is as follows:

$$\begin{aligned}
&\sum_{j=1}^{n} Result^{P_j} \\
&= secret_val^{P_1} - rnd_sum^{P_1}_{self} + rnd_sum^{P_1}_{others} \\
&\quad \vdots \\
&\quad + secret_val^{P_n} - rnd_sum^{P_n}_{self} + rnd_sum^{P_n}_{others} \\
&= \sum_{j=1}^{n} secret_val^{P_j} - \sum_{j=1}^{n} rnd_sum^{P_j}_{self} + \sum_{j=1}^{n} rnd_sum^{P_j}_{others}.
\end{aligned} \tag{1}$$

In addition, we also have the following equations for the data-holder parties:

- For P_i ($\forall i \in \{1, \ldots, k\}$), $rnd\_sum^{P_i}_{self} = \sum_{j=1}^{n} rnd^{P_i}_{P_j} - rnd^{P_i}_{P_i}$, where $rnd^{P_i}_{P_j}$ is the shared random mask between P_i and P_j. On the other hand, $rnd\_sum^{P_i}_{others} = \sum_{j=1}^{k} rnd^{P_j}_{P_i} - rnd^{P_i}_{P_i}$.
- For P_i ($\forall i \in \{k+1, \ldots, n\}$), $rnd\_sum^{P_i}_{self} = 0$. On the other hand, $rnd\_sum^{P_i}_{others} = \sum_{j=1}^{k} rnd^{P_j}_{P_i}$.

Substituting these in Equation 1, we obtain:

$$\begin{aligned}
&\sum_{j=1}^{n} Result^{P_j} \\
&= \sum_{j=1}^{n} secret_val^{P_j} - \sum_{j=1}^{n} rnd_sum^{P_j}_{self} + \sum_{j=1}^{n} rnd_sum^{P_j}_{others} \\
&= \sum_{j=1}^{n} secret_val^{P_j} - \sum_{j=1}^{k}\Big(\sum_{i=1}^{n} rnd^{P_j}_{P_i} - rnd^{P_j}_{P_j}\Big) - \sum_{j=k+1}^{n} (0) \\
&\quad + \sum_{j=1}^{k}\Big(\sum_{i=1}^{k} rnd^{P_i}_{P_j} - rnd^{P_j}_{P_j}\Big) + \sum_{j=k+1}^{n}\Big(\sum_{i=1}^{k} rnd^{P_i}_{P_j}\Big) \\
&= \sum_{j=1}^{n} secret_val^{P_j} - \sum_{j=1}^{k}\Big(\sum_{i=1}^{n} rnd^{P_j}_{P_i}\Big) + \sum_{j=1}^{k}\big(rnd^{P_j}_{P_j}\big) \\
&\quad + \sum_{j=1}^{k}\Big(\sum_{i=1}^{k} rnd^{P_i}_{P_j}\Big) - \sum_{j=1}^{k}\big(rnd^{P_j}_{P_j}\big) + \sum_{j=k+1}^{n}\Big(\sum_{i=1}^{k} rnd^{P_i}_{P_j}\Big) \\
&= \sum_{j=1}^{n} secret_val^{P_j} - \sum_{i=1}^{n}\Big(\sum_{j=1}^{k} rnd^{P_j}_{P_i}\Big) + \sum_{j=1}^{n}\Big(\sum_{i=1}^{k} rnd^{P_i}_{P_j}\Big) \\
&= \sum_{j=1}^{n} secret_val^{P_j}.
\end{aligned} \tag{2}$$

The above equation shows that the aggregation of partial results from data-holder parties is equal to the aggregation of data-holder parties' secret values.

As shown above, the correctness and accuracy of our SMC technique do not depend on k or the minimum number of colluding parties. By increasing k, the minimum number of colluding parties required for revealing a secret value increases, which in turn improves the privacy of the method. Increasing k increases the communication overhead in the initialization phase. Therefore, the trade-off is between privacy and communication overhead of the initialization phase.

C. HANDLING MISSING VALUES

In this section, handling missing values when the data is distributed is explained in the context of our proposed privacy-preserving distributed learning framework, i.e., k-PPD-ERT. In the application of distributed learning approaches, particularly in the healthcare domain, we deal with data with missing values. Missing values in a dataset may occur as a result of improper collection of data, refusal of

TABLE 1. Example of structured data distributed among two parties with missing values.

Party	Record	Sex	Height
1	1	M	170
	2	F	155
	3	M	?
2	1	F	?
	2	F	165
	3	M	178

patients to share information, etc. In scenarios where the data is distributed, handling missing values can require a different procedure in comparison to scenarios in which the data is held in one center.

Several approaches can still be used in such scenarios, e.g., deleting records with missing values. However, they might not be helpful in all cases, e.g., where we have a low number of data records or when the percentage of records with missing values is high. Another solution is to replace the missing values in an attribute with the mean/average of the available values in that attribute. This approach avoids deleting data records and is particularly relevant when dealing with smaller datasets with missing values.

For calculating the mean of the available values for an attribute, we require the summation of these values. Due to privacy concerns, data-holder parties refrain from sharing the summation of their available values with others. In particular, this is a major privacy concern when each data-holder party holds only one record. Therefore, we adopt the approach presented in Section IV-B to address this issue, as we merely require the final summation of the available values.

We explain the approach using an example. Suppose we have two parties, and each party holds three records. Table 1 represents the data for each party. Each record contains the sex and height of record owners or patients. Two records miss the value for height. Assume that by preserving privacy, we can calculate the summation of available values for the height, i.e., 668 in our example, as well as the summation of the number of records not missing the height value, i.e., 4 in our example. In that case, we can calculate the mean for the height, i.e., 167 in our example.

The summation of the available values and the number of available values are calculated using our secure aggregation method. Finally, the mediator divides the summation of the available values by the number of available values and calculates the mean. Then, the mean is shared with all parties to replace the missing values.

Our technique may also be modified based on the problem settings. For instance, in the above example, suppose the user requires the mean of values for male and female patients separately, i.e., 174 and 160, respectively. Then, our technique can be adjusted by only securely aggregating the available values belonging to male or female patients.

We use the same technique for categorical attributes, i.e., to calculate the frequencies of categories in one attribute. Then, we may decide how to fill the missing values based on these frequencies. We may decide to replace all values with the most frequent category, i.e., the mode. The missing category can also be drawn randomly based on the distribution of frequencies. Moreover, we may also decide on filling the missing values by jointly considering the frequencies and information from other attributes.

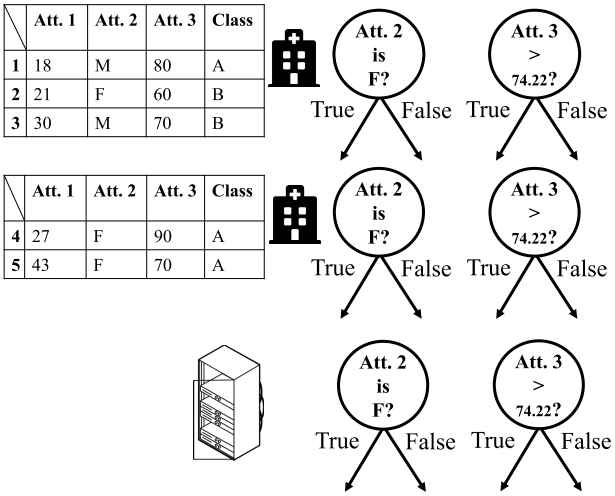
V. ILLUSTRATIVE EXAMPLE

In this section, we provide an illustrative example to clarify the procedure of learning for our algorithm. This procedure is shown in Figure 4. For the sake of simplicity of the presentation, we do not consider the secure aggregation in this section. In the learning process initiation, the global random seed, secure aggregation's random seeds, number and type of data attributes, possible categories or range of data attributes, and learning parameters for the algorithm are shared among all parties. In our example, we have two data-holder parties and a mediator. The first and second parties hold three and two training data records, respectively, as shown in Figure 4a. Each record has three attributes (two numerical and one categorical) and one label.

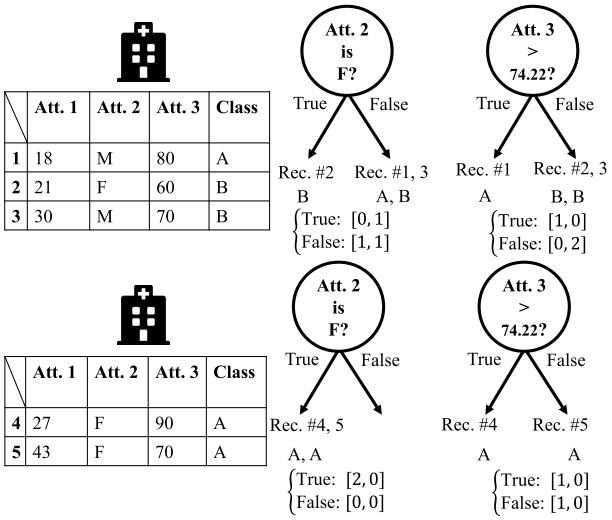
The goal is to learn an ensemble of decision trees from all the records available on the data-holder parties based on our algorithm. The mediator initiates a round of learning a decision tree and, after finishing the procedure for learning one tree, repeats it to have an ensemble of decision trees. At every step of choosing a decision node for the decision tree, each party, including the mediator, generates two random decision nodes based on the global seed. Since all parties use the same seed, they locally generate candidate decision nodes that are similar to the generated decision nodes in other parties. Figure 4a shows the local generation of the candidate decision nodes for the first decision tree's root.

In the next step, the parties classify their records using each randomly generated candidate decision node, as shown in Figure 4b. Several data records fall under the *True* branch (for each candidate decision node) and several fall under the *False* branch. Therefore, based on the records' labels (classes), we make two vectors for each branch that represent the combination of the labels. For instance, for the first candidate decision node in the first party: the *True* vector is [0, 1], and it means that zero records of this party belonging to class (label) *A*, and one record of this party belonging to class (label) *B* fall under the *True* branch of this candidate decision node. Thus, each data-holder party, for each candidate decision node, generates two vectors representing the combination of records labels (that fall under *True* and *False* branches).

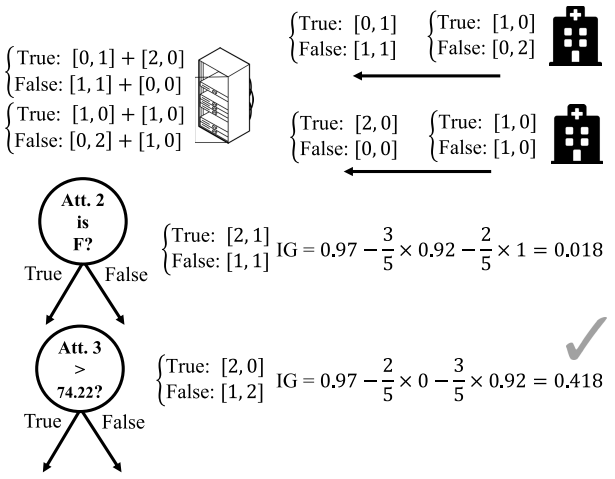
The resulting vectors for each candidate decision node and in all data-holder parties should be returned to the mediator and, then, be aggregated there. Figure 4c shows this procedure, in which all vectors for the *True* and *False* branches of each candidate decision node are returned to the mediator. At this point, for each candidate decision node, the mediator has the combination of labels for the *True* and *False* branches. In addition to deciding on making a leaf or decision node in the decision tree's current position, such vectors determine



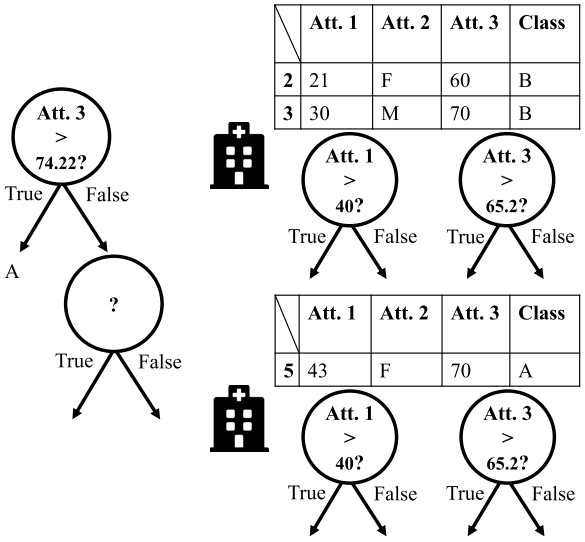
(a) Generating decision nodes randomly



(b) Splitting the data in each data-holder party



(c) Aggregating the results, calculating the scores



(d) Continuing the same process for the rest of the tree

FIGURE 4. **Illustrative example.**

which candidate decision node has a higher score/information gain and should be selected. For calculating the score (information gain) for a decision node, the combination of labels at each branch is required. In our example, the second decision node has a higher information gain and is selected.

As shown in Figure 4d, the second candidate decision node is selected for the root of the decision tree. After checking the labels in its *True* branch, [2, 0], we observe that all the records falling in the *True* branch belong to the same class (have the same label: *A*). Therefore, instead of making a decision node, we make a leaf in the *True* branch. We follow the same procedure of making a decision node for the *False* branch. However, this time, the data-holder parties only consider the records that fall in the root's *False* branch, i.e., 2, 3, and 5. We continue the same procedure for the rest of the tree.

VI. EVALUATION AND DISCUSSION

In this section, we evaluate our proposed approach with respect to classification performance, scalability and overhead, and privacy criteria [83].

A. DATA

We consider two sets of data for the evaluation in this paper. First, we consider two popular publicly available healthcare datasets, i.e., Heart Disease [29] and Breast Cancer Wisconsin (Diagnostic) [30]. For the Heart Disease case, we utilize the processed Cleveland's data [84] to predict the presence or absence of heart disease. In the other case, Wisconsin Diagnostic Breast Cancer (WDBC) data [84] is used to predict breast cancer's diagnosis as benign or malignant.

In addition to the above publicly available datasets, we also consider two mental health detests associated with

the Norwegian INTROMAT (INTROducing Mental health through Adaptive Technology) project:

- The Depresjon dataset [85] contains motor activity data from 55 individuals (30 females and 25 males) recorded using an ActiGraph wristband worn on the right wrist. 23 individuals in this dataset have been diagnosed with depression, including both unipolar and bipolar individuals, while the remaining 32 are in the control group. Each individual wore an ActiGraph wristband for an arbitrary number of days, ranging from 5 to 20 days. The condition and control groups were monitored for 291 and 402 days in total, respectively.
- The Psykose dataset [86] contains motor activity data from 54 individuals (23 females and 31 males) recorded using an ActiGraph wristband worn on the right wrist. 22 individuals in this dataset have been diagnosed with schizophrenia, and all used antipsychotic medications, while the remaining 32 are in the control group. Each individual wore an ActiGraph wristband for an arbitrary number of days, ranging from 8 to 20 days. The condition and control groups were monitored for 285 and 402 days in total, respectively.

B. PERFORMANCE EVALUATION METRICS

The performance of the proposed algorithm is evaluated by measuring the F1-score ($F1$), Accuracy (ACC), and Matthews Correlation Coefficient (MCC), which are defined as follows:

$$F1 = \frac{TP}{TP+0.5\cdot(FP+FN)}$$
$$ACC = \frac{TP+TN}{TP+FP+TN+FN}$$
$$MCC = \frac{TP\cdot TN - FP\cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

where FP, TN, TP and FN definitions are the false positive, true negative, true positive, and false negative, respectively.

C. EVALUATION AND RESULTS

1) CLASSIFICATION PERFORMANCE FOR WIDELY USED HEALTHCARE DATASETS

To evaluate the classification performance for Heart Disease [29] and Breast Cancer Wisconsin (Diagnostic) [30] datasets, we perform a three-fold cross-validation. We divide the dataset into three parts, and in each round, we use one of the parts as the test set and the rest as the training set and finally report the averaged results. We adopt the F1-score (weighted average) and accuracy as our classification performance metrics. The F1-score is the harmonic mean between the precision and recall metrics, while the accuracy measures the ratio of correctly classified samples. Table 2 exhibits the classification performance of our approach, k-PPD-ERT, against the distributed ID3 algorithm [12]. We compare our approach against the distributed ID3 [12] since, similar to our approach, it is a state-of-the-art tree-based method that employs SMC techniques for secure aggregation of partial results and addresses classification problems in scenarios where the data is horizontally partitioned. Moreover, the classification performance of the centralized version of ERT is also provided for comparison.

TABLE 2. Classification performance for our proposed method, distributed ID3, and centralized ERT.

Dataset	Metric	k-PPD-ERT	Distributed ID3	ERT
Heart Disease [29]	Accuracy	80.4%	74.5%	80.4%
	F1-Score	80%	74.3%	80%
Breast Cancer [30]	Accuracy	95.3%	91.3%	95.3%
	F1-Score	95.4%	91.3%	95.4%

The k-PPD-ERT and ERT algorithms follow the same learning procedure. This means that, for both algorithms, the same steps for selecting candidate decision nodes and building the decision tree are followed. In our experiments, we set the same seeds for the random functions and the same learning parameters for both algorithms, e.g., the number of candidate decision nodes. Moreover, the datasets are split into train and test sets in the same way with the same random seed, so these sets are the same for both experiments. Therefore, both algorithms result in the same classification performance, i.e., by following the same procedure, setting the same seeds and parameters, and having the same train and test data.

In our experiments, for our approach, k-PPD-ERT, and the ERT algorithm, we learn an ensemble of 25 decision trees. For the number of candidate decision nodes' parameter in the algorithm, we use 5-fold cross-validation on the training set for the model selection (concerning classification performance measured by the F1-score). In the case of the Heart Disease dataset, k-PPD-ERT outperforms the distributed ID3 [12] by up to 5.9%. For the Breast Cancer dataset, our approach outperforms the distributed ID3 by up to 4.1%.

2) CLASSIFICATION PERFORMANCE FOR MENTAL HEALTH DATASETS ASSOCIATED WITH INTROMAT PROJECT

In addition to the widely used public datasets, we also consider the data associated with the Norwegian INTROMAT (INTROducing Mental health through Adaptive Technology) project, i.e., Depresjon dataset [85] and Psykose dataset [86]. We use F1-score (weighted average), Accuracy (ACC), and Matthews Correlation Coefficient (MCC) for measuring the classification performance, which are the metrics used for performance evaluation on these datasets [85], [86]. We consider both the original and augmented data for each dataset. The original data includes the mean and the standard deviation of the activity level along with the proportion of minutes with no activity [85], [86]. The augmented sample reflects the activity level of an individual in a day by locally resampling the raw data from the same individual. The problem related to the difference in the number of recorded days for each individual, which makes the dataset more imbalanced, is addressed by augmentation. Augmentation also addresses

TABLE 3. Classification performance (leave one patient out) of different classification algorithms for mental health datasets associated with the Norwegian INTROMAT project, i.e., Depresjon dataset [85] and Psykose dataset [86].

Algorithms	Depresjon Dataset [85]						Psykose Dataset [86]					
	Augmented Data			Without Augmentation			Augmented Data			Without Augmentation		
	F1-score	ACC	MCC	F1-score	ACC	MCC	F1-score	ACC	MCC	F1-score	ACC	MCC
k-PPD-ERT (Distributed)	76.3%	76.8%	0.518	66.3%	67.0%	0.310	87.9%	88.0%	0.751	81.7%	81.8%	0.623
ID3 (Distributed)	65.1%	65.0%	0.286	65.6%	66.5%	0.296	75.0%	74.8%	0.490	79.3%	79.4%	0.573
ERT (Centralized)	76.3%	76.8%	0.518	66.3%	67.0%	0.310	87.9%	88.0%	0.751	81.7%	81.8%	0.623
Random forest (Centralized)	74.4%	75.1%	0.481	64.3%	64.7%	0.266	90.7%	90.7%	0.807	80.6%	80.7%	0.601
XGBoost (Centralized)	76.2%	76.3%	0.510	64.3%	64.7%	0.265	92.4%	92.5%	0.844	80.7%	80.7%	0.601
Decision Tree (Centralized)	65.7%	65.8%	0.293	60.6%	60.7%	0.191	76.0%	76.0%	0.505	76.1%	76.2%	0.508
Linear SVM (Centralized)	69.5%	69.5%	0.375	68.4%	68.6%	0.349	87.3%	87.2%	0.748	82.8%	82.8%	0.645

TABLE 4. Communication complexity and privacy of different SMC approaches.

Approach	Party	Communication		Total Communication (N = number of parties)	Number of Colluding Parties
		Send	Receive		
NOSMC	Data Holders	1	0	$(N-1)\times 1+1\times(N-1)$	1: mediator has the values with no collusion
	Mediator	0	$N-1$		
STSMC	All	2	2	$N\times(2+2)$	2: neighbor parties
k-PPD-ERT	Data Holders	1	0	$(N-1)\times 1+1\times(N-1)$	$k+1$: k data-holder parties and the mediator
	Mediator	0	$N-1$		
Shamir [75]	$k-1$ Parties	N	$N-1$	$N\times(N-1+N-1)+2\times(k-1)$	k parties (k<N)
	One Party	$N-1$	$N-1+k-1$		
	The Rest	$N-1$	$N-1$		

the problem of samples with a shorter length, i.e., motor activity signals recorded starting from the middle of the day [87].

We compare our approach against several state-of-the-art machine learning algorithms, including ERT [77], random forest [73], XGBoost [88], Decision Tree [64], and linear SVM algorithm [89]. Table 3 shows the classification performance of different algorithms for the INTROMAT data. The results demonstrate that the proposed approach performs on par or better than state-of-the-art techniques. We also compare our approach against the distributed ID3 [12]. For the Depresjon dataset [85], the k-PPD-ERT technique outperforms distributed ID3 [12] by 0.7% in terms of F1-score, 0.5% in terms of ACC, and 0.014 in terms of MCC for the original data and by 11.2% in terms of F1-score, 11.8% in terms of ACC, and 0.232 in terms of MCC for the augmented data. For the Psykose dataset [86], the k-PPD-ERT technique outperforms distributed ID3 [12] by 2.4% in terms of F1-score, 2.4% in terms of ACC, and 0.05 in terms of MCC for the original data and by 12.9% in terms of F1-score, 13.2% in terms of ACC, and 0.261 in terms of MCC for the augmented data.

3) PRIVACY AND OVERHEAD OF SECURE MULTI-PARTY COMPUTATION TECHNIQUES

We now discuss the privacy and overhead of our proposed approach. We adopt an SMC technique to avoid direct sharing of the vectors representing the combination of record labels for each candidate decision node with other parties and the mediator. We compare the communication overhead and privacy of our adopted SMC technique against three other techniques, including the SMC methods employed in [12], i.e., Shamir's technique [75]. Table 4 presents the communication overhead and privacy evaluation of each approach. In the table, N is the number of parties, and k is a parameter in k-PPD-ERT and Shamir's secret sharing for the minimum number of colluding parties to identify a secret value. The communication overheads in the table are for one round of secure aggregation.

In the first approach (NOSMC), no SMC technique is adopted, and all values are directly shared with the mediator and known to it. This approach has the lowest possible communication cost and number of colluding parties, and, here, it is considered as a baseline. The other approach for the aggregation of partial results is the straightforward SMC (STSMC) approach. In this approach, in the first round, each party aggregates its random mask and its secret value to the received result from the previous party and passes it to the next party, and in the second round, parties subtract their random masks from the aggregated result of the previous round. This method's communication overhead is of the same order as NOSMC, $O(N)$, but it is more robust to collusion. On the other hand, Shamir's secret sharing is an SMC method employed in [12] for secure aggregation. This approach can tolerate the highest number of colluding parties, although it has a high communication overhead, i.e., $O(N^2)$.

Our approach's communication overhead, similar to NOSMC and STSMC, is from order $O(N)$, which is considerably more efficient compared to Shamir's approach with an order of $O(N^2)$. Concerning the number of colluding parties, by adopting our approach, it takes k ($k < N$) data-holder parties and the mediator to collude for identification of the secret values. In our approach, the participation of the mediator for

TABLE 5. The scenarios for our experiments on Amazon's AWS cloud.

	Number of data holders	Mediator location	Data holders locations
Scenario 1	2	SE	CA,DE
Scenario 2	5	SE	CA,DE,US,JP,AU
Scenario 3	10	SE	CA,DE,US,JP,AU,SG,IN,KR,FR,EN
Scenario 4	20	SE	CA,DE,US,JP,AU,SG,IN,KR,FR,EN

collusion is required to reveal a secret value. The mediator is assumed as an honest party in many scenarios, and in the case of a secret value revelation, we know that the mediator has been involved in the collusion. Shamir's secret sharing requires k ($k < N$) parties to collude for identifying a secret value but suffers from scalability and high communication overhead.

4) LATENCY FOR OUR PROOF-OF-CONCEPT IMPLEMENTATION

Finally, we have also implemented our proposed approach on Amazon's AWS cloud to evaluate the latency and scalability of the k-PPD-ERT.[1] We consider four scenarios where we change the number of data-holder parties. We consider four datasets, i.e., Heart [29], Breast [30], Depresjon [85], Psykose [86]. For each dataset, the training data (75% of the dataset) is distributed equally among the data-holder parties. The mediator includes a 2 core 2.40 GHz CPU and 512 MB RAM, runs Ubuntu 20.04, and is located in Sweden. The machines in all other locations include a 1 core 2.40 GHz CPU and 512 MB RAM and run Ubuntu 20.04.

The latency results are shown in Figure 5. In the first scenario, as shown in Table 5, we consider two data-holder parties located in Canada and Germany. Learning one extremely randomized tree through our approach takes 15.9 ± 1.5, 11.8 ± 3.5, 3.5 ± 1.0, 2.4 ± 0.7 seconds for Heart, Breast, Depresjon, and Psykose datasets, respectively. In the second scenario, as shown in Table 5, we consider five data-holder parties located in Canada, Germany, the United States, Japan, and Australia. Learning one extremely randomized tree through our approach takes 43.5 ± 4.1, 32.4 ± 9.6, 9.5 ± 2.7, 6.6 ± 2.0 seconds for Heart, Breast, Depresjon, and Psykose datasets, respectively. In the third scenario, as shown in Table 5, we consider ten data-holder parties located in Canada, Germany, the United States, Japan, Australia, Singapore, India, South Korea, France, and England. Learning one extremely randomized tree through our approach takes 43.8 ± 4.2, 32.6 ± 9.7, 9.6 ± 2.7, 6.7 ± 2.0 seconds for Heart, Breast, Depresjon, and Psykose datasets, respectively. In the fourth scenario, as shown in Table 5, we consider twenty data-holder parties located in Canada, Germany, the United States, Japan, Australia, Singapore, India, South Korea, France, and England, with two parties at each location. Learning one extremely randomized tree through our approach takes 43.6 ± 4.1, 32.5 ± 9.7, 9.6 ± 2.7, 6.8 ± 2.0 seconds for Heart, Breast, Depresjon, and Psykose datasets, respectively.

[1]The source code of our implementations is available at https://github.com/AminAminifar/kPPDERT_cloud

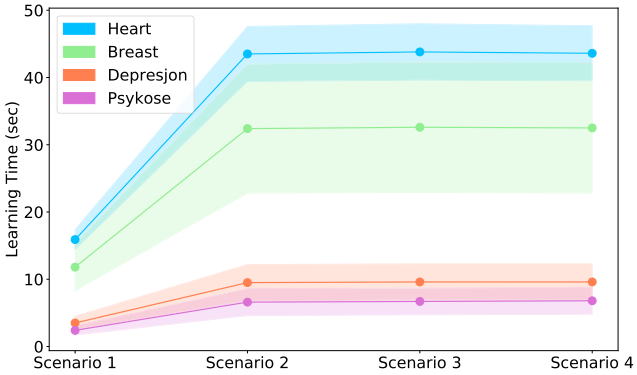


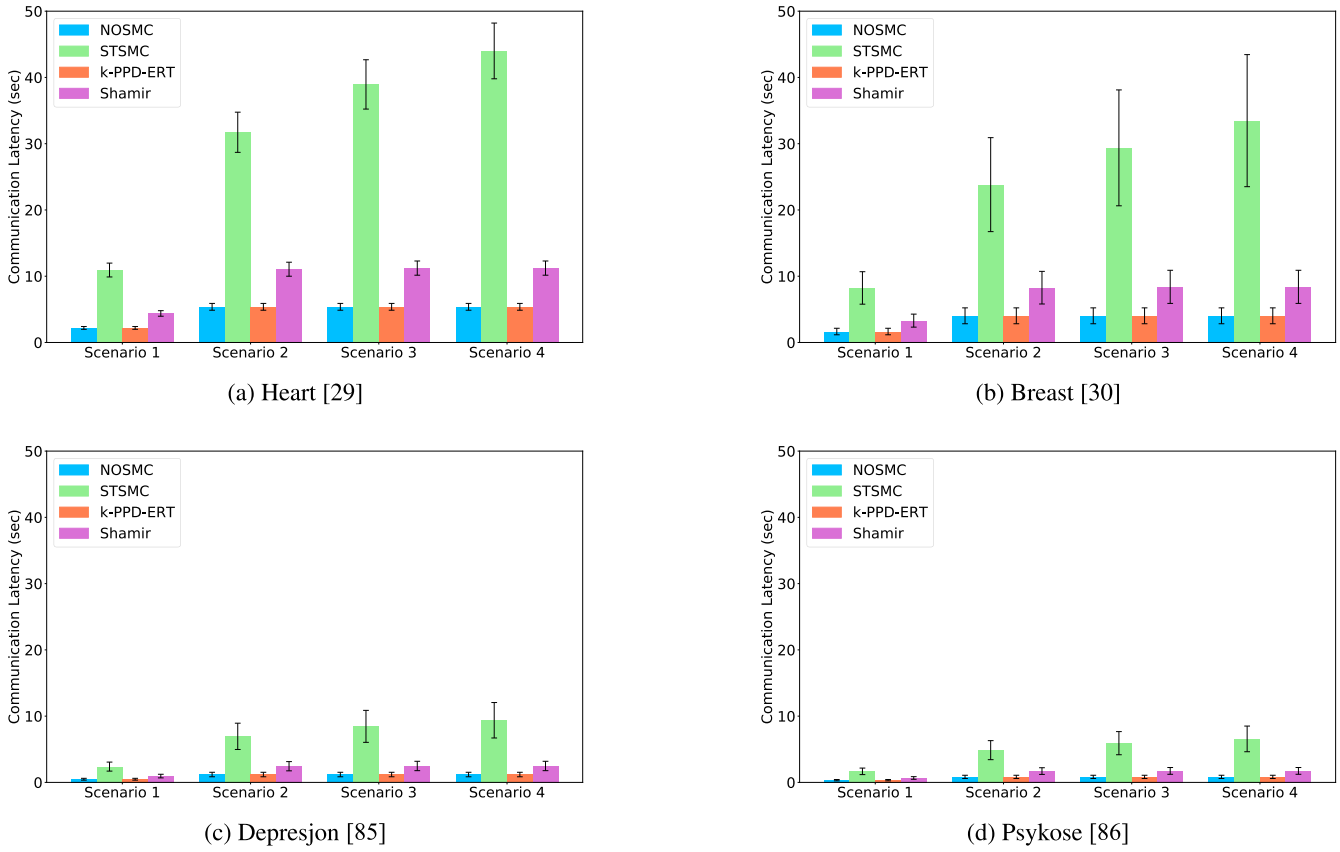
FIGURE 5. The mean and standard deviation of learning time (ten times performed) of one extremely randomized tree through k-PPD-ERT for different datasets in several scenarios on Amazon's AWS cloud.

To better understand the reason for the increase and decrease in the latencies reported above and the shape of the graphs in Figure 5, it should be noted that the latency depends on the geographical location of the data holders and communication delays. In scenario two, the latency has increased due to the fact that the bottleneck communication distance between the data holders and the mediator is increased. However, the results in scenario three are similar to scenario two because the bottleneck communication distance remains the same. In scenario four, the slight reduction in the latency is due to the fact that we distribute the data among data-holder parties (each party has fewer data samples to process), and the learning process on each party is performed simultaneously and in parallel, similar to big data analysis. These explain the increase of latencies from scenario one to two and the almost flat shapes of the graphs from scenario two to scenario four in Figure 5.

5) COMMUNICATION LATENCY OF SECURE MULTI-PARTY COMPUTATION TECHNIQUES

We also evaluate the communication latency of one secure aggregation round for each SMC approach based on their algorithms, the location of data holders in each scenario, the volume of packets transferred between parties, and the network bandwidth between parties. This shows to what extent adopting each approach can increase the latency.

In this paper, we consider the propagation and transmission delays for communication latency [90], [91]. The latency of transferring a packet from P_i to P_j is equal to the sum of propagation and transmission delays and is denoted by $L(P_i, P_j)$. The propagation delay is equal to the distance between parties divided by the velocity of signal propagation, which for unguided transmission through air or space is equal to the speed of light [90]. The transmission delay is equal to the number of bits in the packet divided by the rate of transmission. For transmission delay, we divide the volume of the message to be transferred from P_i to P_j by the bandwidth between these parties.



(a) Heart [29]

(b) Breast [30]

(c) Depresjon [85]

(d) Psykose [86]

FIGURE 6. **The mean and standard deviation of estimated communication latency of different methods for aggregation of secret values in learning one extremely randomized tree (ten times performed) based on different datasets in several scenarios on Amazon's AWS cloud.**

The network bandwidth between two Amazon machines is measured as 1.05 Mbits/sec using the iPerf tool [92]. When a packet contains two arrays for true and false branches, each including information for five candidate decision nodes for a binary classification task, the volume of each packet is 384 bytes. The volume of the packet depends on the data, i.e., the number of candidate decision nodes and the number of target classes.

The following are the analysis of communication latency for each method:

- For NOSMC and k-PPD-ERT, all parties (P_i, $\forall i \in \{1, \ldots, n\}$) send one message to the mediator (M) in parallel. Since the messages are sent in parallel, the communication latency is equal to the arrival duration of the last message. Therefore, the communication delay is equal to $\max_i L(P_i, M), i \in \{1, \ldots, n\}$.
- For STSMC, we have two loops of message passing between parties in each round, and finally, the first party sends the result to the mediator. Therefore, the communication delay is equal to $2 \cdot (\sum_{i=1}^{n-1} L(P_i, P_{i+1}) + L(P_n, P_1)) + L(P_1, M)$.
- For Shamir, each round of secure aggregation consists of two parts performed sequentially. In the first part, all data-holder parties send one message to $n-1$ parties. When all parties receive these messages, they calculate the intermediate results [12] and send them to the mediator. Therefore, the communication delay is equal to $\max_{i,j} L(P_i, P_j), i, j \in \{i, j \in \{1, \ldots, n\} \mid i \neq j\}$ plus $\max_i L(P_i, M), i \in \{1, \ldots, n\}$.

The number of required secure aggregation operations is also recorded for the experiments in Section VI-C4. The mean and standard deviation of the required number of secure aggregation operations for learning one extremely randomized tree (ten times performed) are 98.8 ± 9.4, 73.6 ± 21.9, 22.0 ± 6.2, 15.4 ± 4.5 operations for Heart, Breast, Depresjon, and Psykose datasets, respectively. For estimating the total communication latency of each method for aggregating secret values, the calculated latencies should be multiplied by the number of secure aggregations performed for learning the classification model.

Figure 6 shows the mean and standard deviation of communication latency of different methods for aggregation of secret values for each scenario and each dataset. This figure shows that k-PPD-ERT has the same communication latency as the NOSMC procedure. Shamir's technique has lower communication latency compared to STSMC, but it still has higher communication latency compared to k-PPD-ERT and NOSMC procedures.

It should be noted that the communication latency of these methods should not be confused with the communication overhead presented in Table 4. The orders of communication overhead for NOSMC, STSMC, and k-PPD-ERT are the

same and lower than Shamir's technique. However, since in STSMC, we have two loops of message passing between parties that are performed sequentially, this technique has more delay for a secure aggregation operation. Shamir's technique has two rounds for each SMC operation, and in each round, the message passings are performed in parallel, so it has a lower delay compared to STSMC. For NOSMC and *k*-PPD-ERT, we have one round of message passing that is performed in parallel and has the lowest communication latency.

Finally, we demonstrate that our proposed *k*-PPD-ERT approach provides a solution for the classification of structured data distributed over multiple sources with privacy-preservation considerations, without performance degradation.

VII. CONCLUSION

In this paper, we present the privacy-preserving distributed extremely randomized trees algorithm for learning without privacy concerns in the healthcare domain. We have evaluated our proposed algorithm extensively using two popular structured healthcare datasets and two mental health datasets associated with the Norwegian INTROducing Mental health through Adaptive Technology (INTROMAT) project. Our approach outperforms the state of the art in distributed tree-based models by up to 11.2% in terms of F1-score, 11.8% in terms of ACC, and 0.232 in terms of MCC for the Depresjon augmented dataset, and by up to 12.9% in terms of F1-score, 13.2% in terms of ACC, and 0.261 in terms of MCC for the Psykose augmented dataset. Moreover, we present the implementation of our technique on Amazon's AWS cloud, as a proof of concept, to evaluate the latency and scalability of our framework. The proposed algorithm has linear overhead with respect to the number of parties and can also handle datasets with missing values. We demonstrated our framework's efficiency in terms of prediction performance, scalability, and overheads, as well as privacy. The proposed framework provides the possibility of developing high-quality and accurate machine learning models without privacy concerns and is expected to contribute to a better healthcare system in the long term. As future work, we plan to explore the possibility of extending the proposed framework to settings where the parties do not follow the honest-but-curious security model, which is beyond the scope of this work.

REFERENCES

[1] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature Med.*, vol. 25, no. 1, pp. 65–69, Jan. 2019.

[2] S. McKinney *et al.*, "International evaluation of an AI system for breast cancer screening," *Nature*, vol. 577, pp. 89–94, Jan. 2020.

[3] X. Liu, L. Faes, A. U. Kale, S. K. Wagner, D. J. Fu, A. Bruynseels, T. Mahendiran, G. Moraes, M. Shamdas, C. Kern, J. R. Ledsam, M. K. Schmid, K. Balaskas, E. J. Topol, L. M. Bachmann, P. A. Keane, and A. K. Denniston, "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis," *Lancet Digit. Health*, vol. 1, no. 6, pp. e271–e297, Oct. 2019.

[4] R. Aggarwal, V. Sounderajah, G. Martin, D. S. W. Ting, A. Karthikesalingam, D. King, H. Ashrafian, and A. Darzi, "Diagnostic accuracy of deep learning in medical imaging: A systematic review and meta-analysis," *npj Digit. Med.*, vol. 4, no. 1, p. 65, Dec. 2021.

[5] S. D. Lustgarten, Y. L. Garrison, M. T. Sinnard, and A. W. Flynn, "Digital privacy in mental healthcare: Current issues and recommendations for technology use," *Current Opinion Psychol.*, vol. 36, pp. 25–31, Dec. 2020.

[6] D. Pascual, A. Amirshahi, A. Aminifar, D. Atienza, P. Ryvlin, and R. Wattenhofer, "EpilepsyGAN: Synthetic epileptic brain activities with privacy preservation," *IEEE Trans. Biomed. Eng.*, vol. 68, no. 8, pp. 2435–2446, Aug. 2021.

[7] A. Saeed, F. D. Salim, T. Ozcelebi, and J. Lukkien, "Federated self-supervised learning of multisensor representations for embedded intelligence," *IEEE Internet Things J.*, vol. 8, no. 2, pp. 1030–1040, Jan. 2021.

[8] F. Forooghifar, A. Aminifar, and D. Atienza, "Resource-aware distributed epilepsy monitoring using self-awareness from edge to cloud," *IEEE Trans. Biomed. Circuits Syst.*, vol. 13, no. 6, pp. 1338–1350, Dec. 2019.

[9] D. Sopic, A. Aminifar, A. Aminifar, and D. Atienza, "Real-time event-driven classification technique for early detection and prevention of myocardial infarction on wearable systems," *IEEE Trans. Biomed. Circuits Syst.*, vol. 12, no. 5, pp. 982–992, Oct. 2018.

[10] D. Sopic, A. Aminifar, A. Aminifar, and D. Atienza, "Real-time classification technique for early detection and prevention of myocardial infarction on wearable devices," in *Proc. IEEE Biomed. Circuits Syst. Conf. (BioCAS)*, Oct. 2017, pp. 1–4.

[11] R. Zanetti, A. Arza, A. Aminifar and D. Atienza, "Real-time EEG-based cognitive workload monitoring on wearable devices," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 1, pp. 265–277, Jan. 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9464276, doi: 10.1109/TBME.2021.3092206.

[12] F. Emekci, O. D. Sahin, D. Agrawal, and A. El Abbadi, "Privacy preserving decision tree learning over multiple parties," *Data Knowl. Eng.*, vol. 63, no. 2, pp. 348–361, Nov. 2007.

[13] J. S. Davis and O. Osoba, "Improving privacy preservation policy in the modern information age," *Health Technol.*, vol. 9, no. 1, pp. 65–75, Jan. 2019.

[14] J. Vaidya, B. Shafiq, W. Fan, D. Mehmood, and D. Lorenzi, "A random decision tree framework for privacy-preserving data mining," *IEEE Trans. Dependable Secure Comput.*, vol. 11, no. 5, pp. 399–411, Sep. 2014.

[15] P. Jurczyk and L. Xiong, "Distributed anonymization: Achieving privacy for both data subjects and data providers," in *Proc. IFIP Annu. Conf. Data Appl. Secur. Privacy*. Berlin, Germany: Springer, 2009. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-03007-9_13

[16] L. Sweeney, "K-anonymity: A model for protecting privacy," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 557–570, Oct. 2002.

[17] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "*L*-diversity: Privacy beyond *K*-anonymity," *ACM Trans. Knowl. Discovery Data*, vol. 1, no. 1, p. 3, 2007.

[18] N. Li, T. Li, and S. Venkatasubramanian, "T-closeness: Privacy beyond *K*-anonymity and *L*-diversity," in *Proc. IEEE 23rd Int. Conf. Data Eng.*, Apr. 2007, pp. 106–115.

[19] A. Aminifar, Y. Lamo, K. Pun, and F. Rabbi, "A practical methodology for anonymization of structured health data," in *Proc. 17th Scandin. Conf. Health Informat.*, 2019, pp. 127–133. [Online]. Available: https://ep.liu.se/en/conference-article.aspx?series=ecp&issue=161&Article_No=22

[20] A. Aminifar, F. Rabbi, V. K. I. Pun, and Y. Lamo, "Diversity-aware anonymization for structured health data," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 2148–2154.

[21] *Health Informatics—Pseudonymization*, International Organization for Standardization, Geneva, Switzerland, Standard ISO 25237:2017, Jan. 2017. [Online]. Available: https://www.iso.org/standard/63553.html

[22] C. Dwork, "Differential privacy," in *Proc. 33rd Int. Colloq. Automata, Lang., Program. (ICALP)* (Lecture Notes in Computer Science). Berlin, Germany: Springer-Verlag, 2006. [Online]. Available: https://link.springer.com/chapter/10.1007/11787006_1

[23] M. Kantarcioglu, "A survey of privacy-preserving methods across horizontally partitioned data," in *Privacy-Preserving Data Mining*. Boston, MA, USA: Springer, 2008, pp. 313–335. [Online]. Available: https://link.springer.com/chapter/10.1007/978-0-387-70992-5_13

[24] J. Vaidya, "A survey of privacy-preserving methods across vertically partitioned data," in *Privacy-Preserving Data Mining*. Boston, MA, USA: Springer, 2008, pp. 337–358. [Online]. Available: https://link.springer.com/chapter/10.1007/978-0-387-70992-5_14

[25] W. Du and Z. Zhan, "Building decision tree classifier on private data," in *Proc. IEEE Int. Conf. Privacy, Secur. Data Mining (CRPIT)*, vol. 14. Australia: Austral. Comput. Soc., 2002, pp. 1–8. [Online]. Available: https://dl.acm.org/doi/10.5555/850782.850784
[26] J. Konečný, H. Brendan McMahan, D. Ramage, and P. Richtárik, "Federated optimization: Distributed machine learning for on-device intelligence," 2016, *arXiv:1610.02527*.
[27] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, "Communication-efficient learning of deep networks from decentralized data," 2016, *arXiv:1602.05629*.
[28] A. Aminifar, F. Rabbi, and Y. Lamo, "Scalable privacy-preserving distributed extremely randomized trees for structured data with multiple colluding parties," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 2655–2659.
[29] R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J.-J. Schmid, S. Sandhu, K. H. Guppy, S. Lee, and V. Froelicher, "International application of a new probability algorithm for the diagnosis of coronary artery disease," *Amer. J. Cardiol.*, vol. 64, no. 5, pp. 304–310, Aug. 1989.
[30] O. L. Mangasarian, W. N. Street, and W. H. Wolberg, "Breast cancer diagnosis and prognosis via linear programming," *Oper. Res.*, vol. 43, no. 4, pp. 570–577, Aug. 1995.
[31] *INTROMAT (Introducing Personalized Treatment of Mental Health Problems Using Adaptive Technology)*. Accessed: Dec. 9, 2021. [Online]. Available: https://intromat.no
[32] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th Int. Conf. Artif. Intell. Statist., Mach. Learn. Res.*, PMLR, 2017. [Online]. Available: https://proceedings.mlr.press/v54/mcmahan17a.htm
[33] V. Smith, S. Forte, C. Ma, M. Takáč, M. Jordan, and M. Jaggi, "CoCoA: A general framework for communication-efficient distributed optimization," *J. Mach. Learn. Res.*, vol. 18, p. 230, Apr. 2017.
[34] S. Baghersalimi, T. Teijeiro, D. Atienza, and A. Aminifar, "Personalized real-time federated learning for epileptic seizure detection," *IEEE J. Biomed. Health Informat.*, early access, Jul. 9, 2021, doi: 10.1109/JBHI.2021.3096127.
[35] R. Agrawal and R. Srikant, "Privacy-preserving data mining," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2000, pp. 439–450.
[36] S. Agrawal and J. R. Haritsa, "A framework for high-accuracy privacy-preserving mining," in *Proc. 21st Int. Conf. Data Eng. (ICDE)*, Apr. 2005, pp. 193–204.
[37] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacy preserving mining of association rules," *Inf. Syst.*, vol. 29, no. 4, pp. 343–364, Jun. 2004.
[38] S. Rizvi and J. Haritsa, "Maintaining data privacy in association rule mining," in *Proc. 28th Int. Conf. Very Large Databases (VLDB)*. Amsterdam, The Netherlands: Elsevier, 2002, pp. 682–693.
[39] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, "On the privacy preserving properties of random data perturbation techniques," in *Proc. 3rd IEEE Int. Conf. Data Mining*, Nov. 2003, pp. 99–106.
[40] Z. Huang, W. Du, and B. Chen, "Deriving private information from randomized data," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, 2005, pp. 37–48.
[41] M. Kantarcioglu and J. Vaidya, "An architecture for privacy-preserving mining of client information," in *Proc. IEEE Int. Conf. Privacy, Secur. Data Mining*, vol. 14, Dec. 2002, pp. 37–42.
[42] Y. Lindell and B. Pinkas, "Privacy preserving data mining," *J. Cryptol.*, vol. 15, no. 3, pp. 177–206, 2002.
[43] C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin, and M. Y. Zhu, "Tools for privacy preserving distributed data mining," *ACM SIGKDD Explor. Newslett.*, vol. 4, no. 2, pp. 28–34, Dec. 2002.
[44] L. T. Phong, Y. Aono, T. Hayashi, L. Wang, and S. Moriai, "Privacy-preserving deep learning via additively homomorphic encryption," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 5, pp. 1333–1345, May 2018.
[45] J. Vaidya and C. Clifton, "Privacy-preserving outlier detection," in *Proc. 4th IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2004, pp. 233–240.
[46] G. Jagannathan and R. N. Wright, "Privacy-preserving distributed K-means clustering over arbitrarily partitioned data," in *Proc. 11th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2005, pp. 593–599.
[47] X. Lin, C. Clifton, and M. Zhu, "Privacy-preserving clustering with distributed EM mixture modeling," *Knowl. Inf. Syst.*, vol. 8, no. 1, pp. 68–81, Jul. 2005.
[48] H. Yu, X. Jiang, and J. Vaidya, "Privacy-preserving SVM using nonlinear kernels on horizontally partitioned data," in *Proc. ACM Symp. Appl. Comput. (SAC)*, 2006, pp. 603–610.
[49] B. Pinkas, "Cryptographic techniques for privacy-preserving data mining," *ACM SIGKDD Explor. Newslett.*, vol. 4, no. 2, pp. 12–19, Dec. 2002.
[50] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," 2016, *arXiv:1610.05492*.
[51] P. Kairouz *et al.*, "Advances and open problems in federated learning," 2019, *arXiv:1912.04977*.
[52] Q. Li, Z. Wen, Z. Wu, S. Hu, N. Wang, Y. Li, X. Liu, and B. He, "A survey on federated learning systems: Vision, hype and reality for data privacy and protection," 2019, *arXiv:1907.09693*.
[53] S. Lo, Q. Lu, C. Wang, H. Paik, and L. Zhu, "A systematic literature review on federated machine learning: From a software engineering perspective," *ACM Comput. Surv.*, vol. 54, no. 5, pp. 1–39, 2021.
[54] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2019, pp. 739–753.
[55] S. Truex, L. Liu, M. E. Gursoy, L. Yu, and W. Wei, "Demystifying membership inference attacks in machine learning as a service," *IEEE Trans. Services Comput.*, vol. 14, no. 6, pp. 2073–2089, Nov. 2021.
[56] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. S. Quek, and H. V. Poor, "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 3454–3469, 2020.
[57] R. Hu, Y. Guo, H. Li, Q. Pei, and Y. Gong, "Personalized federated learning with differential privacy," *IEEE Internet Things J.*, vol. 7, no. 10, pp. 9530–9539, Oct. 2020.
[58] R. C. Geyer, T. Klein, and M. Nabi, "Differentially private federated learning: A client level perspective," 2017, *arXiv:1712.07557*.
[59] C. Zhuang, T. She, A. Andonian, M. Sobol Mark, and D. Yamins, "Unsupervised learning from video with deep neural embeddings," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9563–9572.
[60] H. Purwins, B. Li, T. Virtanen, J. Schluter, S.-Y. Chang, and T. Sainath, "Deep learning for audio signal processing," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 2, pp. 206–219, May 2019.
[61] A. M. Vartouni, M. Shokri, and M. Teshnehlab, "Auto-threshold deep SVDD for anomaly-based web application firewall," *TechRxiv*, 2021. [Online]. Available: https://www.techrxiv.org/articles/preprint/Auto-Threshold_Deep_SVDD_for_Anomaly-based_Web_Application_Firewall/15135468, doi: 10.36227/techrxiv.15135468.v1.
[62] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable AI for trees," *Nature Mach. Intell.*, vol. 2, no. 1, pp. 56–67, Jan. 2020.
[63] W. Fan, H. Wang, P. S. Yu, and S. Ma, "Is random model better? On its accuracy and efficiency," in *Proc. 3rd IEEE Int. Conf. Data Mining*, Nov. 2003, pp. 51–58.
[64] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, Mar. 1986.
[65] W. Du and Z. Zhan, "Using randomized response techniques for privacy-preserving data mining," in *Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2003, pp. 505–510.
[66] K. Cheng, T. Fan, Y. Jin, Y. Liu, T. Chen, D. Papadopoulos, and Q. Yang, "SecureBoost: A lossless federated learning framework," *IEEE Intell. Syst.*, vol. 36, no. 6, pp. 87–98, Nov. 2021.
[67] Y. Liu, Z. Ma, X. Liu, S. Ma, S. Nepal, R. H. Deng, and K. Ren, "Boosting privately: Federated extreme gradient boosting for mobile crowdsensing," in *Proc. IEEE 40th Int. Conf. Distrib. Comput. Syst. (ICDCS)*, Nov. 2020, pp. 1–11.
[68] L. Zhao, L. Ni, S. Hu, Y. Chen, P. Zhou, F. Xiao, and L. Wu, "InPrivate digging: Enabling tree-based distributed data mining with differential privacy," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Apr. 2018, pp. 2087–2095.
[69] Q. Li, Z. Wen, and B. He, "Practical federated gradient boosting decision trees," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 4642–4649.
[70] J. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, 2001.
[71] Y. Liu, Y. Liu, Z. Liu, Y. Liang, C. Meng, J. Zhang, and Y. Zheng, "Federated forest," *IEEE Trans. Big Data*, early access, May 7, 2020, doi: 10.1109/TBDATA.2020.2992755.
[72] S. Truex, N. Baracaldo, A. Anwar, T. Steinke, H. Ludwig, R. Zhang, and Y. Zhou, "A hybrid approach to privacy-preserving federated learning," in *Proc. 12th ACM Workshop Artif. Intell. Secur. (AISec)*, 2019, pp. 1–11.
[73] T. Kam Ho, "Random decision forests," in *Proc. 3rd Int. Conf. Document Anal. Recognit.*, Aug. 1995, pp. 278–282.

[74] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, Oct. 2001.
[75] A. Shamir, "How to share a secret," *Commun. ACM*, vol. 22, no. 11, pp. 612–613, Nov. 1979.
[76] A. Aminifar, F. Rabbi, K. I. Pun, and Y. Lamo, "Privacy preserving distributed extremely randomized trees," in *Proc. 36th Annu. ACM Symp. Appl. Comput.*, Mar. 2021, pp. 1102–1105.
[77] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, Apr. 2006.
[78] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning* (Springer Series in Statistics). New York, NY, USA: Springer, 2001.
[79] Z. Lipton, "The mythos of model interpretability," *Queue*, 2018.
[80] N. D. Condorcet, *Essai Sur l'Application de l'Analyse à la Probabilité des Décisions Rendues à la Pluralité des Voix*. Cambridge, U.K.: Cambridge Univ. Press, 2014.
[81] L. Rokach, *Pattern Classification Using Ensemble Methods*. Singapore: World Scientific, 2010.
[82] A. C.-C. Yao, "How to generate and exchange secrets," in *Proc. 27th Annu. Symp. Found. Comput. Sci.*, Oct. 1986, pp. 162–167.
[83] E. Bertino, D. Lin, and W. Jiang, "A survey of quantification of privacy preserving data mining algorithms," in *Privacy-Preserving Data Mining*. Boston, MA, USA: Springer, 2008, pp. 183–205. [Online]. Available: https://link.springer.com/chapter/10.1007/978-0-387-70992-5_8
[84] D. Dua and C. Graff, "UCI machine learning repository," School Inf. Comput. Sci., Univ. California, Irvine, CA, USA, 2019. [Online]. Available: http://archive.ics.uci.edu/ml and https://archive.ics.uci.edu/ml/citation_policy.html
[85] E. Garcia-Ceja, M. Riegler, P. Jakobsen, J. Tørresen, T. Nordgreen, K. Oedegaard, and O. Fasmer, "Depresjon: A motor activity database of depression episodes in unipolar and bipolar patients," in *Proc. 9th ACM Multimedia Syst. Conf.*, 2018, pp. 472–477.
[86] P. Jakobsen, E. Garcia-Ceja, L. A. Stabell, K. J. Oedegaard, J. O. Berle, V. Thambawita, S. A. Hicks, P. Halvorsen, O. B. Fasmer, and M. A. Riegler, "PSYKOSE: A motor activity database of patients with schizophrenia," in *Proc. IEEE 33rd Int. Symp. Computer-Based Med. Syst. (CBMS)*, Jul. 2020, pp. 303–308.
[87] A. Aminifar, F. Rabbi, K. Pun, and Y. Lamo, "Monitoring motor activity data for detecting patients' depression using data augmentation and privacy-preserving distributed learning," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 2163–2169.
[88] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 785–794.
[89] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, pp. 273–297, Sep. 1995.
[90] W. Stallings, *Data and Computer Communications*. Upper Saddle River, NJ, USA: Prentice-Hall, 2005.
[91] A. Pahlevan, "Multi-objective system-level management of modern green data centers," EPFL, Lausanne, Switzerland, Tech. Rep., 2019. [Online]. Available: https://infoscience.epfl.ch/record/270205?ln=en, doi: 10.5075/epfl-thesis-9457.
[92] *iPerf–The Ultimate Speed Test Tool for TCP, UDP and SCTP*. Accessed: Nov. 30, 2021. [Online]. Available: https://iperf.fr/

AMIN AMINIFAR received the M.Sc. degree in computer engineering from the K. N. Toosi University of Technology, Tehran, Iran, in 2017. He is currently pursuing the Ph.D. degree in computer science with the Western Norway University of Applied Sciences. His current research interests include artificial intelligence and machine learning and their applications, particularly in health, privacy, and security.

MATIN SHOKRI received the M.Sc. degree in computer engineering from the K. N. Toosi University of Technology, Tehran, Iran. He is currently working in machine learning algorithms, especially deep learning in image processing. His research interests include deep learning, reinforcement learning, and ensemble methods.

FAZLE RABBI is currently an Associate Professor. He has long and varied experience with information system development within a large spectrum of domain areas. His research interests include model-based software engineering, data mining, and machine learning, with emphasis on addressing the information science problems in healthcare applications. He is enthusiastic to improving the quality of living through his contribution in healthcare. Earlier, he was involved in academic research for developing reliable workflow management systems for two community-based health programs piloted at Guysborough Antigonish Strait Health Authority (GASHA), Nova Scotia, Canada. He is also interested in innovating new techniques for teaching both in classroom and online platform. Together with researchers from Kenya and Vanderbilt University Medical Center, he developed gamification approach for increasing student motivation and engagement in learning environment.

VIOLET KA I. PUN is currently an Associate Professor at the Western Norway University of Applied Sciences and the SIRIUS Centre, University of Oslo. Her research interests include using formal methods to specify, analyze, and verify the behavior of software programs, especially those running in distributed and concurrent systems. She is active in programming language theory, including language semantics, type systems, deductive verification, and formal logic. She is also interested in digitalization of healthcare domain, data privacy, and self-adaptive patient treatments. She is working on model-based business process planning with tool-supported and automated analyses in terms of formal methods.

YNGVE LAMO received the Ph.D. degree in computer science from the University of Bergen, in 2003, on formal specification of software systems with use of multialgebras. He is currently a Professor with the Department of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, Bergen, Norway. His research interests include model-based software engineering, formal methods, graph transformations, health informatics, and application of machine learning for analyzing health data.