REPUBLIC OF TURKEY
YILDIZ TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

PRIVACY-PRESERVING TECHNIQUES AND MACHINE LEARNING FOR CRITICAL SYSTEMS

Zümrüt MÜFTÜOĞLU

DOCTOR OF PHILOSOPHY THESIS
Department of Electronics and Communication Engineering
Program of Electronics

Supervisor
Prof. Dr. Tülay YILDIRIM

June, 2022

REPUBLIC OF TURKEY
YILDIZ TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

PRIVACY-PRESERVING TECHNIQUES AND MACHINE LEARNING FOR CRITICAL SYSTEMS

A thesis submitted by Zümrüt MÜFTÜOĞLU in partial fulfillment of the requirements for the degree of **DOCTOR OF PHILOSOPHY** is approved by the committee on 29.06.2022 in Department of Electronics and Communication Engineering, Program of Electronics.

Prof. Dr. Tülay YILDIRIM
Yildiz Technical University
Supervisor

Approved By the Examining Committee

Prof. Dr. Tülay YILDIRIM, Supervisor
Yildiz Technical University

Doç. Dr. Nihan KAHRAMAN, Member
Yildiz Technical University

Prof. Dr. Aydın AKAN, Member
Izmir Ekonomi Üniversitesi

Doç. Dr. Sadiye Nergis TURAL POLAT, Member
Yildiz Technical University

Prof. Dr. Fikret S. GÜRGEN, Member
Bogazici University

I hereby declare that I have obtained the required legal permissions during data collection and exploitation procedures, that I have made the in-text citations and cited the references properly, that I haven't falsified and/or fabricated research data and results of the study and that I have abided by the principles of the scientific research and ethics during my Thesis Study under the title of Privacy-Preserving Techniques and Machine Learning for Critical Systems supervised by my supervisor, Prof. Dr. Tülay YILDIRIM. In the case of a discovery of false statement, I am to acknowledge any legal consequence.

Zümrüt MÜFTÜOĞLU

Signature

Dedicated to my family

ACKNOWLEDGEMENTS

First and foremost, I'd want to express my gratitude to my advisor, Prof. Dr. Tülay YILDIRIM, for being so patient and enthusiastic in answering my many last-minute emails and phone calls. My lofty goals were given better form by her insight and experience. I owe a great debt of gratitude to her for teaching me how to think critically and approach a subject from several perspectives, as well as for inspiring me to work on the next steps when I was stuck on or confused about my thoughts.

Additionally, I want to thank the members of my thesis monitoring committee, Professor Dr. Aydn Akan and Assistant Professor Dr. Nihan KAHRAMAN, for helping me improve my work and for improving my Ph.D. adventure.

For their technical assistance during my studies, I also would like to thank Ivoline NGONG, Fatemeh MIRESHGHALLAH, Rakshit NAIDU, and Sahib SINGH for giving me opportinity to be part of a great team.I want to express my gratitude to Işıl Selen DENEMEÇ and Dr. Merve Ayyüce KIZRAK for their endless support and love. I am grateful to my pilates instructor Aslıhan TUNALI ÇİFTÇİOĞLU, who was able to keep me in top physical and mental shape throughout this really taxing procedure for me. I want to express my gratitude to my dear family last but not least. Without their love and motivation, I would not have the bravery to accept the challenge and continuously advance.

Zümrüt MÜFTÜOĞLU

TABLE OF CONTENTS

LIST OF SYMBOLS

δ	Accuracy Parameter
A	Attribute
c	Cipher Text
D	Dataset
E	Encryption Algorithm
$\mathscr{L}$	Lost Function
D'	Neighbour Dataset
ϵ	Privacy Budget
f	Query

LIST OF ABBREVIATIONS

AUROC	Area Under Receiver Operating Characteristics
AI	Artificial Intelligence
CNN	Convolutional Neural Networks
DP	Differential Privacy
DP-SGD	Differentially-Private Stochastic Gradient Descent
FL	Federated Learning
FPR	False Positive Rate
HE	Homomorphic Encryption
MR	Magnetic Resonance
mpMRI	Multiparametric MRI
PATE	Private Aggregation of Teacher Ensembles
PPML	Privacy-Preserving Machine Learning
SMPC	Secure Multi-Party Computation
SSN	Social Security Number
TPR	True Positive Rate
WHO	World Health Organization

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Privacy-Preserving Techniques and Machine Learning for Critical Systems

Zümrüt MÜFTÜOĞLU

Department of Electronics and Communication Engineering
Doctor of Philosophy Thesis

Supervisor: Prof. Dr. Tülay YILDIRIM

It is obvious that the majority of machine learning-driven applications used in daily life are fed from personal data. This data may contain personal information from people's health history to their purchase history, depending on the usage area of the applications. The increase in these applications has led to the need to increase the measures for the protection of such data. As a matter of fact, as of 2016, the data protection law for European citizens came into force. However, considering the role of data in the development of artificial intelligence algorithms, it is inevitable that legislative restrictions will create a negative situation for developments in this field. In this context, technologies that protect data privacy allow sensitive data to be protected and analyzed. The disclosure of sensitive data can be minimized with machine learning algorithms that protect privacy, developed using these technologies. Although there have been promising studies in this area recently, applications with real life data are extremely limited.

This thesis aims to examine the effects of differential privacy approaches in healthcare and biometrics applications that use critical data by conducting comprehensive research on technologies that protect privacy. In addition, the thesis study presents a comparison with the federated learning approach, which is another privacy-preserving technique. In particular, a comparison was made by applying two different differential privacy techniques for signature data, which is a biometric data type, and the effect of the privacy implementation on the similarity score is evaluated. Another aim of the thesis is to compare the behavior of these privacy-preserving approaches for two different models in each application area and data type.

Keywords: Information security, data privacy, differential privacy, biometric verification

YILDIZ TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

ÖZET

Kritik Sistemler İçin Mahremiyet Koruyucu Teknikler ve Makine Öğrenmesi

Zümrüt MÜFTÜOĞLU

Elektronik ve Haberleşme Mühendisliği Anabilim Dalı
Doktora Tezi

Danışman: Prof. Dr. Tülay YILDIRIM

Günlük hayatta kullanılan makine öğrenmesi odaklı uygulamaların büyük çoğunluğunun kişisel verilerden beslendiği aşikardır. Bu veri, uygulamaların kullanım alanına bağlı olarak, insanların sağlık geçmişinden satın alma geçmişine kadar kişisel bilgi içerebilmektedir. Bu uygulamaların artması bu tür verilerin korunmasına yönelik tedbirlerin artırılması ihtiyacını doğurmuştur. Nitekim 2016 yılı itibari ile Avrupa vatandaşları için veri koruma kanunu yürürlüğe girmiştir. Ancak yapay zeka algoritmalarının geliştirilmesinde verilerin rolü göz önüne alındığında mevzuat kısıtlamalarının bu alandaki gelişmeler için olumsuz bir durum oluşturması kaçınılmazdır. Bu bağlamda, veri mahremiyetini koruyan teknolojiler, hassas verilerin korunmasına ve analiz edilmesine olanak tanır. Bu teknolojiler kullanılarak geliştirilen mahremiyeti koruyan makine öğrenmesi algoritmaları ile hassas verilerin ifşası minimumda tutulabilmektedir. Son zamanlarda bu alanda umut verici çalışmalar yapılsa da, gerçek verilere yönelik uygulamalar son derece kısıtlıdır.

Tez kapsamında, mahremiyeti koruyan teknolojiler üzerine kapsamlı araştırmalar yapılarak, kritik veri kullanan sağlık ve biyometri uygulamalarında diferansiyel mahremiyet yaklaşımlarının etkilerinin incelenmesi amaçlanmaktadır. Ayrıca tez çalışması, bir başka mahremiyet koruyucu teknik olan federe öğrenme yaklaşımı ile bir karşılaştırma da sunmaktadır. Özellikle, bu tez çalışmasında ilk olarak biyometrik bir veri türü olan imza verileri için iki farklı diferansiyel mahremiyet tekniği uygulanarak bir karşılaştırma yapılmış ve mahremiyet tekniği eklentisinin benzerlik skoru üzerindeki etkisi değerlendirilmiştir. Tezin bir diğer amacı, mahremiyeti koruyan bu yaklaşımların davranışlarını her bir uygulama alanında ve veri tipinde iki farklı model için karşılaştırmaktır.

Anahtar Kelimeler: Bilgi Güvenliği, veri mahremiyeti, diferansiyel mahremiyet, biyometrik doğrulama

YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

1 INTRODUCTION

Artificial intelligence is an important step in human history that can be compared to the invention of writing. Especially in recent years, technological advancements have accelerated the development of artificial intelligence-supported systems. In addition to hardware competence and software capability, having sufficient data for a particular problem is indispensable to develop an artificial intelligence(AI) system. Today, AI algorithms have reached a level and speed that greatly exceeds human abilities by predicting behavior with the inferences made using the analyzed data. In addition to medicine, defense, and mathematical sciences, we witness that artificial intelligence, which already has an effect in communication, education, law, industry, and many other fields, has also a role to manage crises. The World Health Organization (WHO), which first identified Covid-19 as a novel coronavirus in February after it was discovered in Wuhan province of China, provided the most illustrative example of this function [1].

Concerns for privacy have been rising with the increasing amount of data-driven AI studies. Many transactions we do on the internet in our daily lives leave a trace in the digital world. Many transactions we do on the internet in our daily lives leave a trace in the digital world. These traces might be achieved easily through people's mobile devices or the computers they use at home, in the office, and even on the streets. The reason those concerns about data privacy have been heard frequently, especially in recent years, it is because of the analysis of those data with machine learning algorithms in many areas due to the increasing data collection capabilities. Today, the use of machine learning is becoming widespread in various applications from image recognition to suggestion systems, as it is convenient to offer multi-disciplinary solutions. Some of the developed applications need sensitive data of people. These data are usually loaded from a central location. In such a scenario, the threat is not limited to insider exploitation or an attempted attack from the outside. Studies in the literature show that neither anonymization of data nor inaccessibility of the model or data is not a sufficient countermeasure to prevent the disclosure of information about datasets.

ML, which enables one for searching, analyzing, and interpreting large amount of data, also has successful applications in the field of biometrics and health, which are challenging research areas tackled in the scope of the thesis. Considering the areas of usage of these applications and the nature of the data, it is remarkable how essential privacy-protecting approaches are in this field.

Discussions about privacy raised a new concept called "Privacy by Design (PbD)". The approach which was first proposed by Ann Cavoukian[1] was formalized by the Information and Privacy Commissioner of Ontario. This approach, which is generally used for all IT systems, presents a proactive solution by embedding privacy into the design. In this thesis, we consider privacy-preserving approaches in critical areas of study where applications are developed with machine learning algorithms.

In the thesis study, we consider privacy-preserving approaches in critical areas of study where applications are developed with machine learning algorithms.

1.1 Literature Review

In this section, machine learning methods for critical areas of study and privacy-enhancing approaches used in machine learning are presented.

1.1.1 Machine Learning and Data Privacy

Today, there are machine learning applications that we come across from many fields ranging from health to finance, from education to communication. Machine Learning (ML) algorithms perform specific tasks relying on patterns and inference [2]. Traditionally, an ML model is built using a training set that includes multiple feature vectors and related labels. The input data is usually representative of a set of samples each of which includes a set of feature values. The performance of the ML model in generalizing unseen data is demonstrated by its ability to predict the label for the test samples properly [3]. As the amount of data collected from different sources increases, so does the accuracy of these models.

Privacy gives someone to have right to use another one's information. Westin [4] defines privacy as "The claim of individuals,groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others". “Dalenius desideratum", which is a reasonable notion of privacy,

[1] Cavoukian, Ann, Ph.D., Information Privacy Commissioner, Ontario, Canada

emphasizes that the model must disclose no more about the input to which it is implemented than would have been known about the same input without implementing the model [5]. if an adversary can obtain sensitive values used as input to the model by using the model's output, it shows that a privacy breach occurs in this model [6]. In [7], Papernot et. al. described adversarial capabilities which refer to the definition and way of the available attacks and describing the potential attack vectors.

Recent researches show that machine learning models have many privacy threats. As depicted in Figure 1.1, adversaries can attack against ML systems at two phases as inference phase and the training phase. At the inference phase, they exploit model internal information (white box) or probe the system vulnerabilities (black box) during the inference time. At the training phase, the adversaries use read/write access to the training data to fake or disrupt the model. According to these points, privacy-related attacks are categorized [8] as model extraction attacks, membership inference attacks, and model inversion attacks as shown in Figure 1.2.

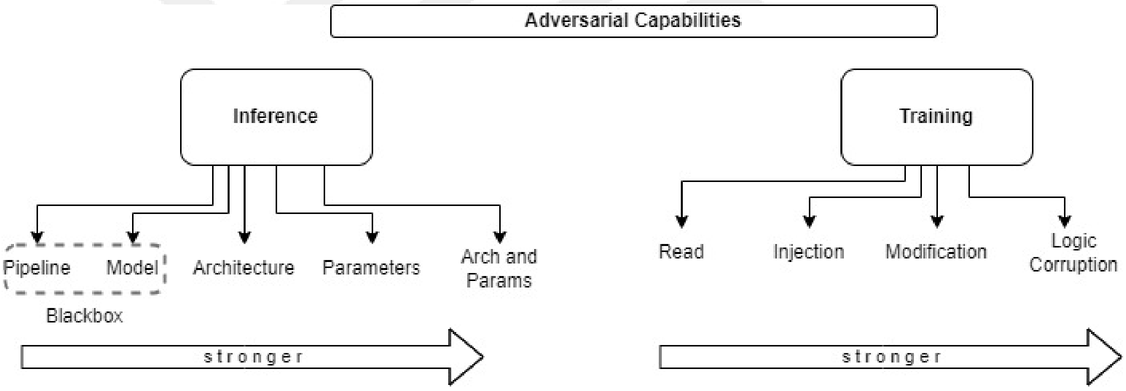


Figure 1.1 Adversarial capabilities [7]

In membership inference attack, attacker tries to determine if a certain sample is in the training set. Liuet et al. study' s [9] shows how an attacker can use publicly available services to acquire private data, reproduce a data processing model without having access to the original model, and obtain some crucial training set characteristics. In their research, Shokri et al. [10] concentrated on the fundamental membership inference attack, which determines if a record is in the model's training dataset. They tested their scenario in the most challenging environment, where the adversary has restricted model access for black-box queries that base the model's result on a specific input. The study, which is known as the first membership inference attack against ML models, presents a quantitative approach ob information leakage of ML models.

Model extraction attack is carried out by observing the output labels. The machine learning model is stolen with respect to the chosen inputs. In a model extraction attack, the attacker has black-box access without any knowledge about parameters or training data of ML model. It was firstly proposed by Tramèr et al. [11].

The model inversion attack put out by Fredrikson et al. [6] seeks to deduce details regarding the training set. They studied privacy in pharmacogenetics which uses ML models to guide medical treatments and showed that used models have privacy risks by predicting the model and some demographic information of patients.

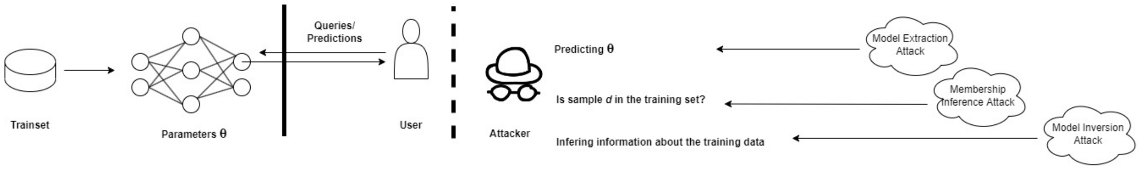


Figure 1.2 Privacy-related attacks on MLs [8]

1.1.2 Privacy-Preserving Approaches in Artificial Intelligence Algorithms

With the increase in digital services, vast amounts of information is produced and those data can be meaningful in solving some problems in daily life. Machine learning provides solutions in many areas from health to finance with data obtained from different sources. If the obtained data is sensitive, serious privacy breaches can occur.

Implementing appropriate privacy-preserving techniques is significant for defining the border of data processing and building a healthy analytics value chain[2]. Privacy-Preserving Machine Learning (PPML) [12, 13] allows analysis without changing the result and guarantees privacy by somehow changing the original data and algorithms [14]. In the literature, the taxonomy of PPML techniques is generally discussed in five main dimensions as shown in Figure 1.3 [15].

Ul Hassan et al. present a comparison of privacy preservation approaches by considering weaknesses as shown in Table 1.1 [16]. According to the priorities of algorithm, type of the privacy -preserving technique may show change.

1.1.2.1 K-Anonimity

Generally, it is not enough to anonymize the data to consider it private. Unfortunately, a common belief is that removing identifying variables such as name, phone number or age provides protection for the identities of individuals. In his study, Seweey showed how combinations of basic demographic features, such as ZIP code, date of birth, ethnicity, gender, and marital status can be identified [17].

[2] Privacy by design in big data,ENISA.December,2015.

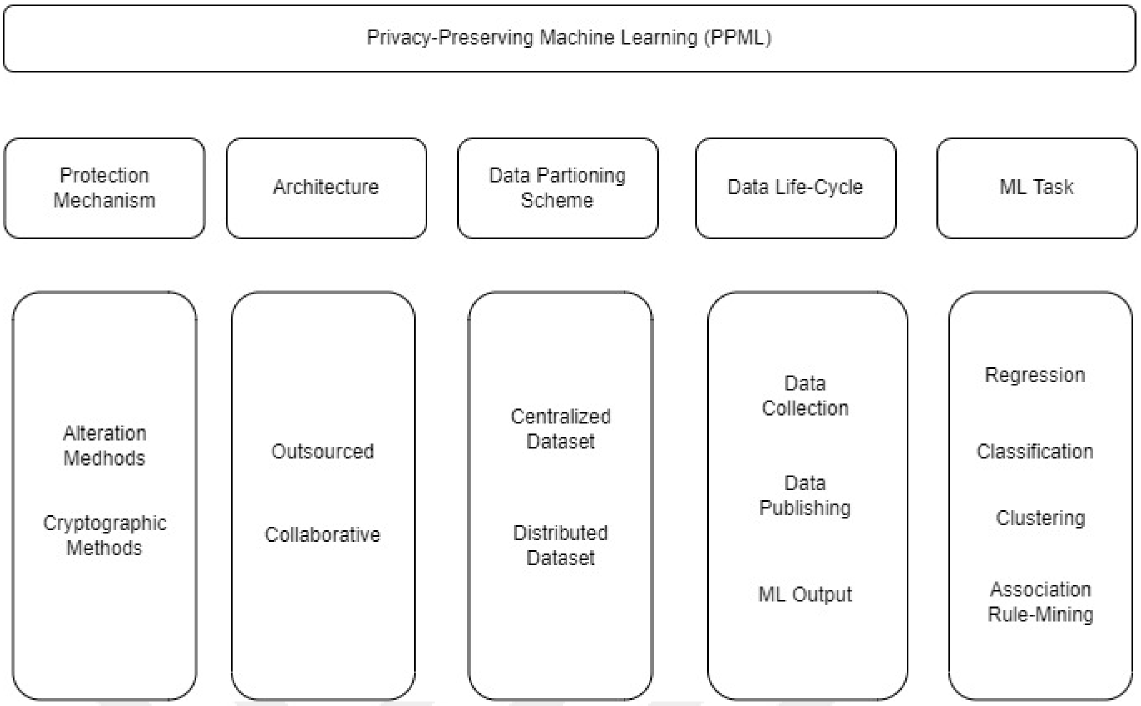


Figure 1.3 Taxonomy of PPML [15]

Table 1.1 Comparison of the privacy techniques on the basis of method and challenges

Privacy Technique	Method to Preserve Privacy	Challenges
Encryption	Public and private keys are generated to decrypt a ciphertext	Effects system speed badly
Anonymization	Personal information is altered such that the data subject cannot be recognized either directly or indirectly	Fully privacy is not guaranteed
Differential Privacy	Preserves privacy by adding random noise	Reduces data utility

Table 1.1 exemplifies this problem with a table of released medical data. The data is de-identified by removing names and Social Security Numbers (SSNs) in the table. Despite the removed attributes, values of other attributes (Date of Birth, Sex, Marital Status...etc.) can be linked through an external table and cause disclosure of privacy [18].

Samarati ve Sweeney addressed the problem of releasing person-specific data by preserving anonymity in the study which they proposed as $k-anonymity$ [18]. An original dataset containing sensitive data can be transformed in a way that makes it difficult to identify the individuals by an adversary. The property of a $k-anonymized$ dataset is that each record is similar to at least one other $k-1$ record on the potentially identifying variables [19]. The $k-anonymity$ approach

Table 1.2 Medical data sample released anonymous [18]

SSN	Name	Ethnicity	Date of Birth	Sex	ZIP	Marital of Status	Problem
*	*	white	06/04/80	Female	06001	single	obesity
*	*	asian	09/23/60	Male	33002	divorced	hypertension
*	*	asian	11/30/78	Female	16100	married	Belgium (shortness of breath)
*	*	black	02/01/55	Male	04200	married	chest pain

Table 1.3 The external table linked through Table 1.2 1.2 [24]

Name	Address	City	ZIP	DOB	Sex	Party
....	....	....	....	....	....	....
Sue K. Taylor)	1513 Wind St.	Cambridge	02150	6/4/80	female	democrat
....	....	....	....	....	....	....

requires protecting quasi-identifiers since they can turn into unique identifiers if they are combined with other quasi-identifiers. The definition of Quasi-identifers is given in Definition 1.1 [18].

Definition 1.1. (Quasi-identifier) Let $T(A_1,\ldots,A_n)$ be a table. A quasi-identifier of T is a set of attributes $\{A_i,\ldots,A_j\} \subseteq \{A_1,\ldots,A_n\}$ whose release must be controlled.

In the definition above; given a table $T(A_1,\ldots,A_n)$, a subset of attributes $\{A_i,\ldots,A_j\} \subseteq \{A_1,\ldots,A_n\}$ and a tuple $t \in T, t[A_i,\ldots,A_j]$ represents the sequence of the values of $A_i,\ldots,A_j$ in T.Furthermore, QI_T shows the set of quasi-identifiers related to T, and $|T|$ denotes cardinality which is the number of the tuples in T.

The goal of $k-anonymity$ is to provide releasing information in the T while ensuring the anonymity of the individuals. It should be noted that the released information nebulously relates to at least a given k of individuals.

Definition 1.2. (k-anonymity) Let $T(A_1,\ldots,A_n)$ be a table and QI_T be the quasi-identifiers associated with it. T is said to satisfy $k-anonymity$ iff for each quasi-identifier $QI \in QI_T$ each sequence of values in $T[QI]$ appears at least with k occurrences in $T[QI]$ [18].

So, $k-anonymity$ is a key notion that addresses the risk of re-identification of anonymized data across linkage to other datasets [26]. It is noted that each record is indistinguishable from at least k-1 other records in accordance with particular “identifying” attributes [20].

1.1.2.2 L-Diversity

$l-diversity$ approach was proposed in the study which also shows the vulnerability of $k-anonymity$ by Machanavajjhala et al. [21]. They showed how an adversary can explore the values of sensitive attributes when there is little diversity in the sensitive attributes and proposed a novel approach by using this diversity. It is not so right to call it as a threat model since It presents an improvement over the flaws of k-anonymity.

Definition 1.3. (l-diversity) A $q^* - block$ is $l - diverse$ if it contains at least l well-represented values for the sensitive attribute S. A table is l-diverse, if every $q^* - block$ is $l - diverse$.

As seen in Table 1.2, lack of diversity can cause privacy leaks in the $k - anonymized$ table. They suppose that the adversary knows some of the quasi-identifiers (such as age, zip code, sex . . . e.tc.) and also discovered the 4-anonymous table in the records published by the hospital. So if the quasi-identifiers map the 9th, 10th, 11th, and 12th records and if the adversary tries to learn the diagnosis of someone, this will not be so difficult to predict since all of the conditions are cancer.

1.1.2.3 Differential Privacy

The data holder or curator guarantees to use one's data without disclosing it through Differential Privacy proposed in 2006 [22]. For algorithms on aggregated databases, it proposes a strict privacy criteria and a probabilistic privacy method. Since the model does not learn any unique user's individual details, DP offers a strong resistance against Membership Inference Attacks, which seek to discover whether a data sample is part of training data or not [23]. In brief, in case a predicting output does not depend on a data point, a computation is differentially private [24].

Making predictions based on cross-correlation by linking sensitive data and metadata is the biggest challenge for data privacy. This may cause identifying individuals based on cross-links and raises concerns about confidentiality, responsibility, and bias. At this point, DP guarantees to strengthen privacy in data-based statistics and machine learning algorithms by presenting convenient protection for the privacy of individuals [25].

Definition 1.4. (Differential Privacy(ε, δ)) M represents a random mechanism and S denotes each of the outputs and indicates the difference between D and D' datasets when M and S are proper. If two datasets are different by a single record, they are called as adjacent [26]. In this way (ε, δ) preserves privacy [27]. The equation is defined as in Eq. (1).

$$\Pr[M(D) \in S] \leq \exp(\varepsilon). \Pr[M(D') \in S] + \delta \tag{1.1}$$

where ε represents the privacy budget and δ represents the probability of error. The ratio of the two possibilities is limited by e^{ε}. If $\delta = 0$, it provides ε-differential privacy. More powerful differential privacy is achieved through $\Delta = 0$. (ε,δ)-DP holds fracture latitude of the differential privacy [28].

The definition of privacy loss is shown in Eq. (2) :

$$\mathscr{L}_{M(D)||M(D')} = \ln \frac{\Pr[M(D) \in S]}{\Pr[M(D') \in S]}(1.2)$$

It gives the path to achieve ε-DP and (ε,δ)-DP through Laplace noise and Gaussian noise. The noise is proportional to the sensitivity of the mechanism M [29]. A smaller ε means more robust privacy [30].

Sensitivity points out the complexity caused by M. For instance, the sensitivity will modify the required volume of noise for ***f(D)*** when a specific query ***f*** of dataset D is issued.

Dwork emphasizes that DP means nothing in case it is not meaningfully implemented [31]. Recent studies present DP as a promising technique for privacy-preserving ML. There are three leading pillars :

- DP ensures a verifiable guarantee of privacy for individuals mathematically compared with other privacy-preserving approaches [32–35].
- DP maintains privacy by adding an appropriate amount of noise to the model or outputs through the concrete mechanisms in place of anonymizing data simply [36].
- Through a privacy budget, DP offers a trade-off between privacy and utility in which the lower privacy budget value ensures greater privacy. [36].

Figure 1.4 shows how to incorporate DP into an ML model. By employing fundamental techniques like the Laplace mechanism [37] and Gaussian mechanism [22], which are explained in Chapter 3, it renders it impossible for adversaries to obtain knowledge about records from the publicly disclosed ML model or results. During our study; we design our DP-SGD implementation set-up for both Covid-19 and signature datasets as shown in Figure 1.4.

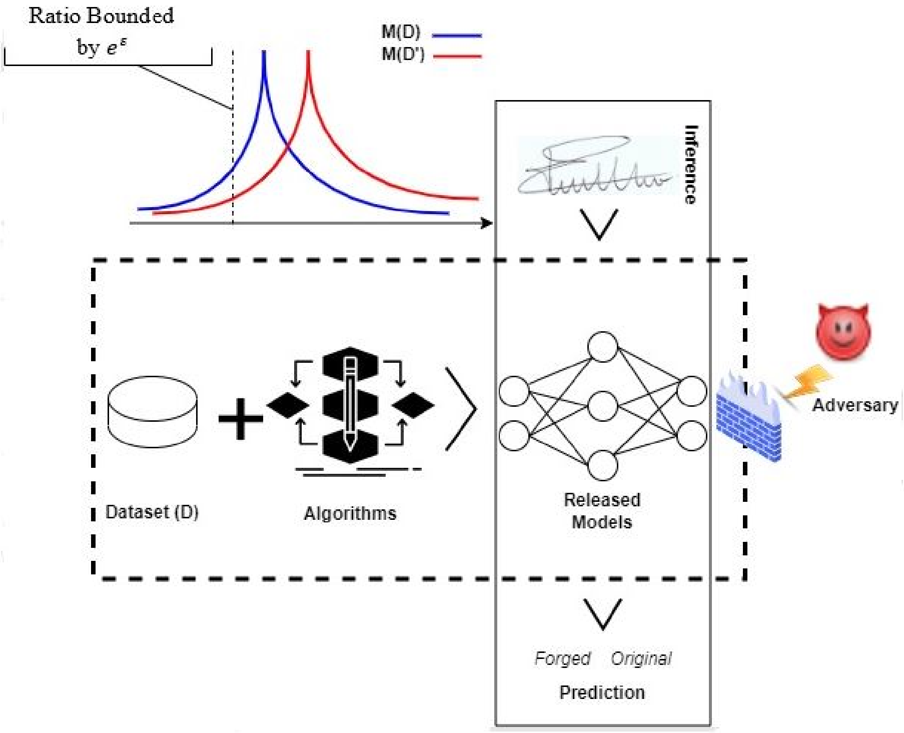


Figure 1.4 Incorporating DP in ML model, modified from [36]

1.1.2.4 Federated Learning

Federated learning provides training data without central collection. Federated learning, which offers a distributed framework, can improve the model through the model aggregation of multiple clients to ensure data privacy [38]. Vertical federated learning, horizontal federated learning, and federated transfer learning are the three categories of federated learning that have been discussed in the literature [39].

Consider N data owners, $\{F_1, \ldots, F_N\}$, each of which wanting train a machine learning model by joining their respective data $\{D_1, \ldots, D_N\}$. Conventionally, all data is brought together and used $D{=}D_1 \cup D_2 \ldots \cup D_N$ to train a model M_{SUM}. Data owners train a model M_{FED} collaboratively through Federated Learning without exposing any data D_i of owner F_i. Furthermore, V_{FED} that is denoted as the accuracy of M_{FED} should be very close to the performance of M_{SUM} (denoted as V_{SUM}). Formally, δ will be a non-negative real number, if [39]

$$[V_{FED} - V_{SUM}] < \delta$$

Here, δ demonstrates the accuracy loss of the FL algorithm.

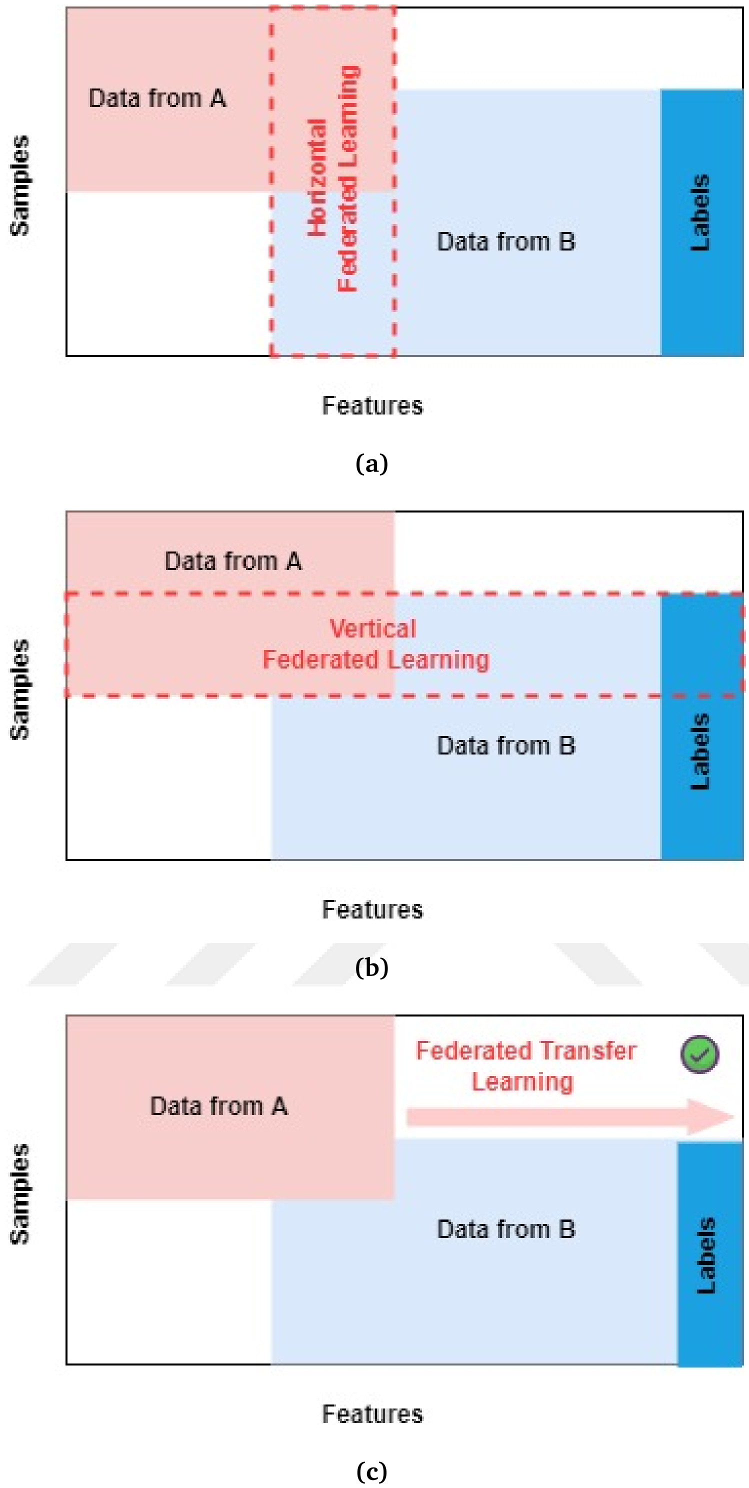


Figure 1.5 Categories of FL [39] (a) Horizontal FL, (b) Vertical FL, (c) Federated Transfer Learning

FL was proposed by Google based on the distributed data through multiple devices while providing data privacy [40–42]. Based on the features of the data, federated learning is split into two types in the literature. Suppose that D_i shows the data owned by data owner i and each row of the matrix stands for a sample, while each column means a feature. Some data sets may also include labelled data. X represents the attribute space, Y represents the label space and I denotes the sample ID space. The feature X, label Y and sample Id I create the complete training dataset (I, X, Y). The Federated Learning is classified as horizontal federated learning, vertical federated learning and federated transfer learning [39].Insight of each type of FL type is depicted in Figure 1.5

- **Horizontal Federated Learning**

  People call it sample-based FL as well. Sample-based FL is another name for it. Situations where data sets contain the same attribute space for several samples exist, as seen in 1.6

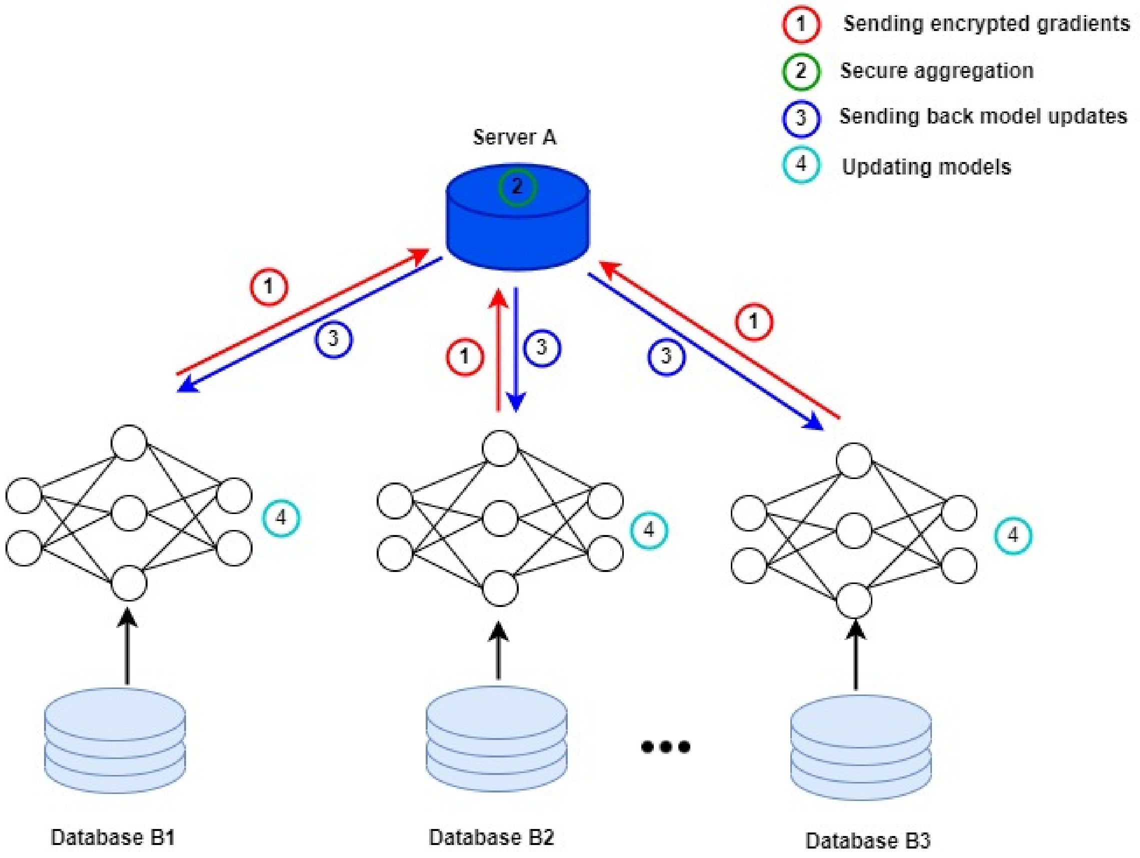


Figure 1.6 Architecture for Horizontal FL [39]

In [39], Google's proposal is presented about the implementation of a horizontal FL solution for Android devices. At this framework, By uploading the parameters to the Android cloud after locally updating the model parameters, a single user can jointly train the centralized model with other data owners. The federated horizontal learning can be summed up as [39] :

$$X_i = X_j, Y_i = Y_j, I_i \neq I_j, \forall D_i, D_j, i \neq j$$

- **Vertical Federated Learning**

  It is also known as feature-based federated learning. As illustrated in 1.7, it can be applied to situations where two datasets share the same sample ID space but differ in feature space [39]. In Vertically federated learning, different features are aggregated and training loss and gradients are computed in a privacy-preserving manner.

$$X_i \neq X_j, Y_i \neq Y_j, I_i \neq I_j, \forall D_i, D_j, i \neq j$$

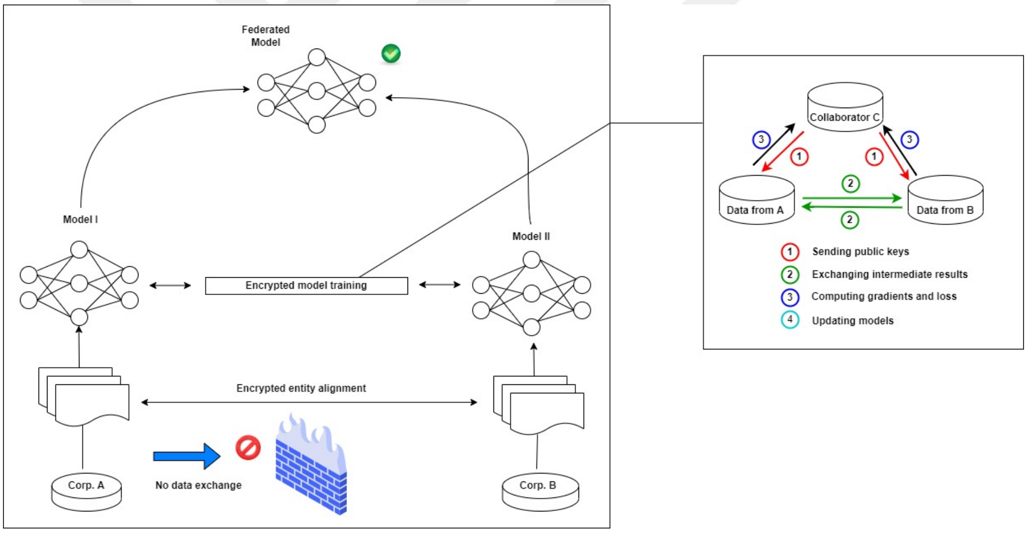


Figure 1.7 Architecture of Vertical FL [39]

In the literature, there are studies that propose privacy-preserving machine learning algorithms with vertically partitioned data, including Cooperative Statistical Analysis [43], secure linear regression [41–43], classification [44], and gradient descent [45].

- **Federated Transfer Learning**

  It is applicable for the scenarios where the two data sets differ both in sample and feature space. FTL is a considerable prolongation to the traditional federated learning systems since it handles problems overrunning the scope of existing federated learning algorithms [39] :

$$X_i \neq X_j, Y_i \neq Y_j, I_i \neq I_j, \forall D_i, D_j, i \neq j$$

1.1.2.5 Homomorphic Encryption

Encryption is a common technique to preserve the privacy of sensitive data. One of the basic challenges of this technique is that it is required to decrypt data for complicated operations. Rivest et al showed that there are encryption functions that allow encrypted data to be run without initial decryption of the operands, what they call “privacy homomorphisms”, for many sets of interesting operations [46]. In abstract algebra, homomorphism is defined as a function that transforms an element of the input domain set into an element of the output algebraic set’s range [47]. Homomorphic encryption is a technique that enables calculations to be made on encrypted material without having to first decode it. As illustrated in Figure 1.8, the result of post-decryption homomorphic operation is also the equivalent of the operation on plain text data [48]. Homomorphic encryption is a kind of encryption that provides the computations on ciphertexts without any change in the output compared with the execution on plain text.

Definition 1.5. Given an operation ‘*’ which is a homomorphic encryption scheme with encryption algorithm E , the following equation is supported :

$$E(m_1) * E(m_2) = E(m_1 * m_2), \forall m_1, m_2 \in M \tag{1.3}$$

Here, M denotes the messages spaces [49].

HE has four principal algorithms : The Key Generation Algorithm, The Encryption Algorithm, The Decryption Algorithm, The Evaluation Algorithm [47].

The Key Generation Algorithm (KeyGen), generates a pair of the secret key and public key for an asymmetric HE, and generates a single key for symmetric HE by inputting a security parameter.

The Encryption Algorithm, Enc, achieves the ciphertext $c = E(m)$ from the ciphertext space C, by inputting plaintext message m from message space M with the encryption key.

The Decryption Algorithm, Dec, gets the plaintext message $D(c) = m$ by inputting the ciphertext c with the decryption key.

The Decryption Algorithm, Dec performs the function *f(...)* over the ciphertexts to output evaluated ciphertexts $f(c1,c2) = E(f(m1,m2))$, without seeing the messages $(m1,m2)$, i.e. $D(f(c1,c2)) = f(m1,m2)$.

It is significant that the format of the ciphertexts after an evaluation process should be preserved to decrypt it accurately. The size of the ciphertext should also remain at a constant value. Otherwise, the ciphertext size will increase, and that will cause limitation of the number of performed operations.

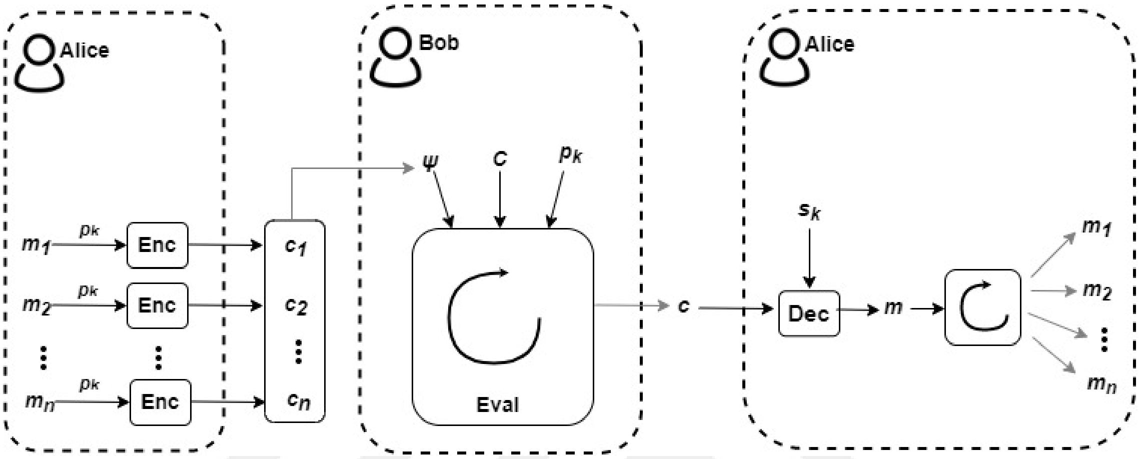


Figure 1.8 Homomorphic Encryption [50]

Homomorphic Encryption schemes are categorized according to computational abilities as Partially Homomorphic Encryption (PHE), Somewhat Homomorphic Encryption (SWHE), and Fully Homomorphic Encryption (FHE).

1.1.2.6 Secure Multi-Party Computation (SMPC)

Because of the nature of distributed machine learning, data is transparent to the system in a way that makes data privacy insufficient. In a distributed learning system, a complicated computing process is generally left to a third party [48]. The concept of secure multi-party computation was first proposed by Goldreich in 1998 [51].

1.2 Objective of the Thesis

Digitization increases the importance of identity management whose fundamental task is to establish a relationship between an individual and his personal identity. A number of applications including border crossings prefer to use biometrics because it is practical. Three methods enable people to recognize a person which is described with three w clauses: what he knows, what he possesses, and who he is. The third method addresses biometrics which bases on his inherent physical or behavioral traits [52].

The fact that biometric data is so sensitive is among the main reasons why we focus on this domain within the scope of the thesis. The health domain, which uses sensitive data at least as much as the field of biometrics, is another field chosen within the scope of the thesis.

The major objective of the thesis is to present research about privacy-preserving machine learning especially for biometrics systems and health applications. The reason we focus on these two domains is that the trained data is sensitive in those domains. Privacy-preserving is an essential term for machine learning algorithms, since protecting trained data is compulsory in case the data is individual.

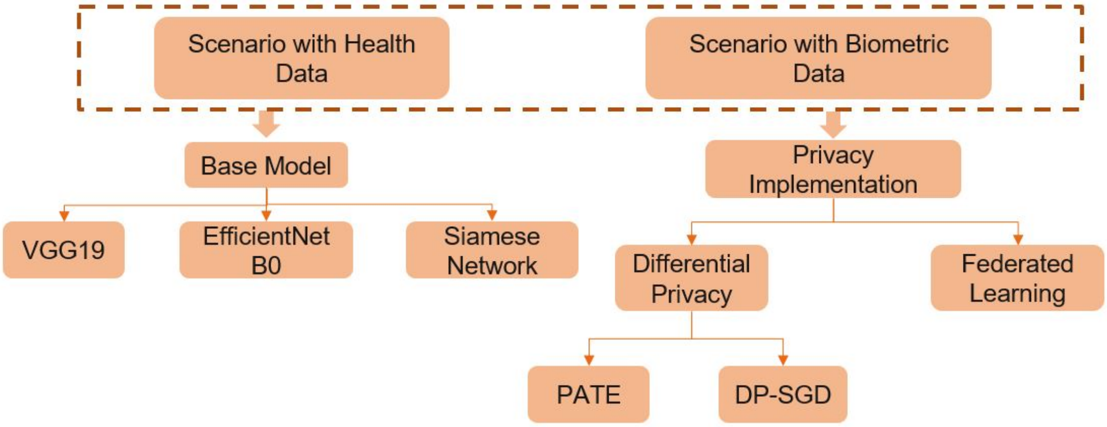


Figure 1.9 Scope of the thesis

Figure 1.9 illustrates the scope of the thesis study. We chose the differential privacy approach which enables analyzing data without compromising the privacy of the individual data samples [53]. Differential privacy adds noise to a query in order to imply plausible deniability which adds the statistical insecurity to whether the input sample has true values or not. In the scope of the thesis studies, we present a comparison between two DP approaches: Private Aggregation of Teacher Ensembles (PATE) [54] and DP-Stochastic Gradient [26].

Differentially private machine learning models can enable confidential data available for training without data usage agreements. The Fundamental Law of Information Recovery tells that heavily true answers to too many questions damage privacy. The main goal of research on differential privacy is to adjourn this inevitability as long as possible [22].

1.3 Hypothesis

The thesis will contribute to the studies in the field of privacy on the following issues:

- A comprehensive research on privacy-preserving machine learning is conducted for critical areas.

- A comparison of two different differential privacy techniques is presented for signature data for the first time.

- With the aim of showing the behavior of DP and ML in medical applications which are the most challenging domain for privacy-preserving ML, DP and FL are implemented using the Covid-19 dataset.

- Two different noise types as Laplace and Gaussian noises are implemented to observe their impact on privacy.

2
ARTIFICIAL INTELLIGENCE MODELS AND PRIVACY-PRESERVING TECHNIQUES USED IN CRITICAL APPLICATIONS

Artificial intelligence or related technologies are used in almost every field of study, and it provides advantages such as increasing efficiency and performance while enabling analysis in today's world where huge data is available. Artificial intelligence which automates tasks done only by humans previously; appears in many fields of study such as visual perception, decision making, speech recognition, and translation between languages today. Especially during the pandemic, artificial intelligence algorithms have accelerated the diagnosis and treatment [1]. Despite the increasing number of patients, the insufficient number of radiologists causes a loss of time for the treatment process, and this process often results in the death of the patient. However, accelerating the diagnosis process with the support of artificial intelligence technologies contributes to reducing these numbers. On the other hand, in addition to medical imaging, artificial intelligence is also used in assistive technologies that facilitate the adaptation of disadvantaged people to daily life [55].

Biometrics, which aims to define individuals according to their biological characteristics or behavioral attributes; is used increasingly in many applications where personal authorization is required. Such applications range from automated border control and physical access control to numerous other cases where biometrics is used to authenticate individuals [56].

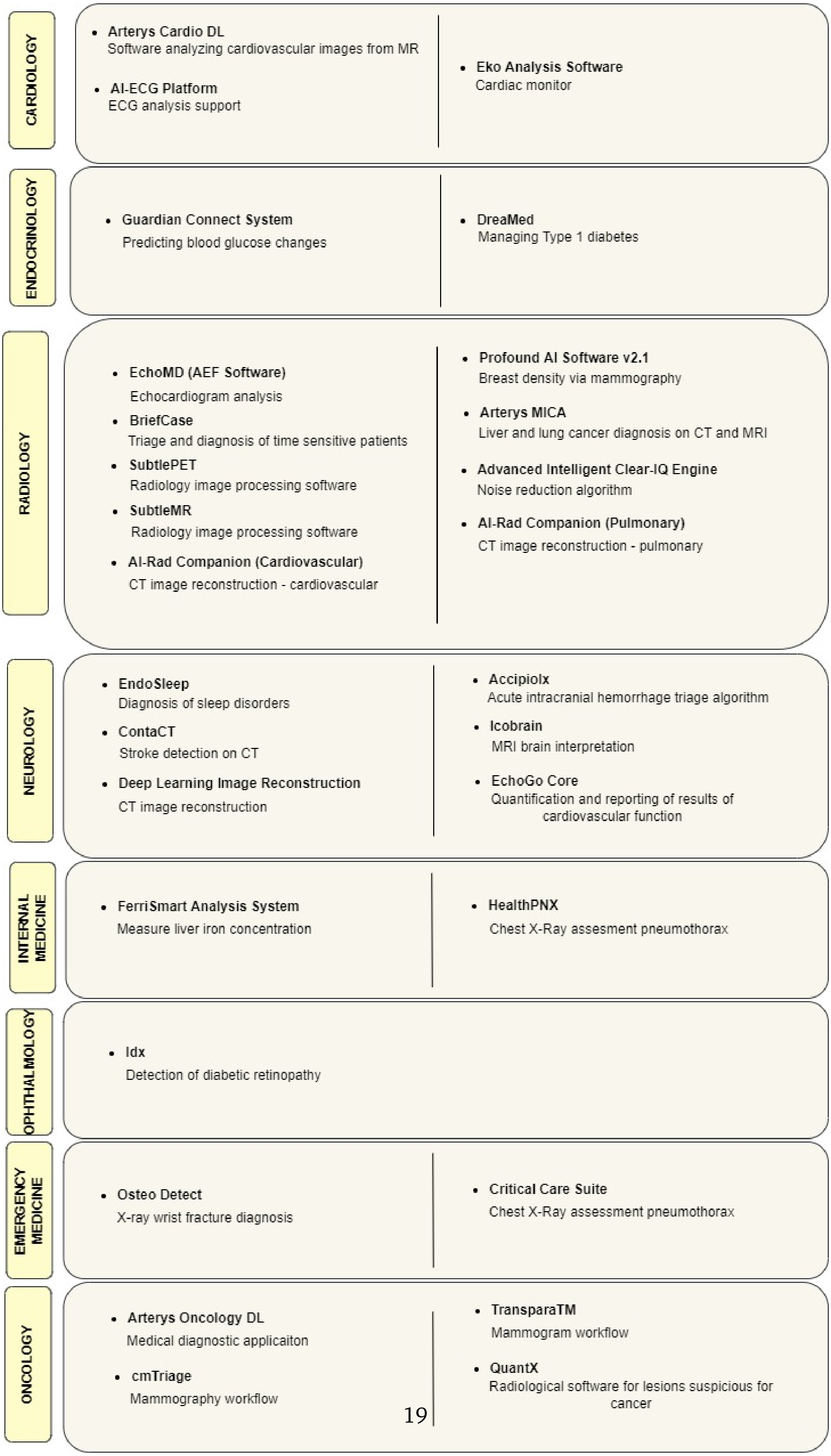
Rather than the field in which an artificial intelligence system is used, the quality of the data it uses is an important indicator for classification as critical. When we look at the applications in both fields, one of which focuses on human health and the other on human security, we see that they use individual data. Based on this indicator, we focused on machine learning algorithms in the field of health and biometrics during our thesis work.

2.1 Artificial Intelligence Algorithms Used in Medical Applications

The opportunities to analyze in the field of medicines have raised exponentially with the technological developments [57]. Machine Learning (ML) presents solutions to diagnostic and prognostic problems in the medicine. ML helps to evaluate clinical parameters and their combinations for prognosis which may be important in the course of the disease such as disease progression prediction, treatment patient management and planning. There are many studies [58] that improve the diagnostic process in medical specialties such as radiology [59], oncology [60], or general medical decision-making [59]. Analysis of continuous data utilized in the Intensive Care Unit [61]. Computer-driven medical image interpretation systems provide huge application potential by enabling an important support in medical diagnosis [62–65]. Thanks to machine learning applications, the cost of treatment in diseases where early diagnosis is important is also significantly reduced. In Figure 2.1 FDA (Food and Drug Administration (American))-approved AI software for medical usage as of June 2019 is seen.

According to a report for Cancer Research UK, late-stage cancer diagnosis requires three times more expensive treatment while causing undesirable clinical outcomes [74]. Medical diagnosis is among the most important applications of intelligent systems through the mechanisms used for generating hypotheses using patient data provided by expert systems and model-based schemes [66–68]. Recently, ML-driven approaches in medicine are making significant progress in the detection and diagnosis of diseases [69, 70]. In this section, machine learning approaches in medicine are discussed.

Chang et al. classified genetic variations of diffuse infiltrating gliomas using convolutional neural networks (CNNs). The study which was implemented through MR imaging data and molecular information shows that neural networks are competent in learning key imaging components without prior feature selection or human-directed training [71]. Roemo et al. studied whether an ML analysis using MRI-derived texture analysis (TA) features can be helpful in the assessment of the existence of placenta accreta spectrum (PAS) in patients with placenta previa (PP) [72]. Another study focuses on Sepsis which is a big health crisis in hospitals which has still no innovative and feasible tool to predict worlwide. The study shows that ML models have a better performance than potential scoring systems in predicting sepsis [73]. Wengert et al. [74] looked at early prediction of pathological full response to neoadjuvant chemotherapy and patient survival outcomes using mpMRI data. They compared the performance of different classifiers including linear support vector machine, linear discriminant analysis, logistic regression, random forests, stochastic



CARDIOLOGY	**Arterys Cardio DL** Software analyzing cardiovascular images from MR **AI-ECG Platform** ECG analysis support	**Eko Analysis Software** Cardiac monitor
ENDOCRINOLOGY	**Guardian Connect System** Predicting blood glucose changes	**DreaMed** Managing Type 1 diabetes
RADIOLOGY	**EchoMD (AEF Software)** Echocardiogram analysis **BriefCase** Triage and diagnosis of time sensitive patients **SubtlePET** Radiology image processing software **SubtleMR** Radiology image processing software **AI-Rad Companion (Cardiovascular)** CT image reconstruction - cardiovascular	**Profound AI Software v2.1** Breast density via mammography **Arterys MICA** Liver and lung cancer diagnosis on CT and MRI **Advanced Intelligent Clear-IQ Engine** Noise reduction algorithm **AI-Rad Companion (Pulmonary)** CT image reconstruction - pulmonary
NEUROLOGY	**EndoSleep** Diagnosis of sleep disorders **ContaCT** Stroke detection on CT **Deep Learning Image Reconstruction** CT image reconstruction	**Accipiolx** Acute intracranial hemorrhage triage algorithm **Icobrain** MRI brain interpretation **EchoGo Core** Quantification and reporting of results of cardiovascular function
INTERNAL MEDICINE	**FerriSmart Analysis System** Measure liver iron concentration	**HealthPNX** Chest X-Ray assesment pneumothorax
OPHTHALMOLOGY	**Idx** Detection of diabetic retinopathy	
EMERGENCY MEDICINE	**Osteo Detect** X-ray wrist fracture diagnosis	**Critical Care Suite** Chest X-Ray assessment pneumothorax
ONCOLOGY	**Arterys Oncology DL** Medical diagnostic applicaiton **cmTriage** Mammography workflow	**TransparaTM** Mammogram workflow **QuantX** Radiological software for lesions suspicious for cancer

Figure 2.1 AI/ML based algorithms in medicine [58]

gradient descent, adaptive boosting, and Extreme Gradient Boosting (XGBoost). The study which is used samples of 38 women with breast cancer concludes that XGBoost produces the most stable performance with higher accuracy. There have been other studies focusing on ophthalmic diseases such as diabetic retinopathy (DR), glaucoma, age-related macular degeneration (AMD) and retinopathy of prematurity (ROP). There have been studies which are focusing on prediction ophthalmic diseases such as diabetic retinopathy(DR) [75, 76], glaucoma [75, 77], age-related macular degeneration (AMD) [75, 78, 79] and retinopathy of prematurity (ROP) through deep learning techniques [80]. Especially, screening for DR is critical strategy for blindness prevention coupled with timely referral and treatment [81]. Deep learning is also implemented to predict error and cardiovascular risk factors (eg. age, smoking status and body mass index) [86,87]. Otha et al. focused on detecting and classifying myocardial delayed enhancement pattern through ML algorithms [82]. In another study, asymptomatic left ventricular systolic dysfunction is identified using convolutional neural network [83]. While Alaa et al. evaluate cardiovascular risk in asymptomatic people [84], Daghistani et. al [85] predicts in-hospital length of stay among cardiac patients using machine learning.

Despite all these advancements, there are still significant barriers to the widespread use of artificial intelligence applications, including the lack of standardized electronic medical records and the existence of strong legal and ethical guidelines to preserve patient privacy [86].

2.2 Artificial Intelligence Algorithms Used in Biometrics Applications

The ability to uniquely identify individuals and associate personal qualities (eg name, nationality, etc.) with an individual is becoming increasingly common. People typically use body features such as face, voice, and gait in conjunction with other contextual information (for example, location and clothing) to get to know each other. The set of attributes associated with a person creates that person's personal identity. Increasing mobility due to population growth in modern society has necessitated the development of identity management systems, which have a critical role in many applications [87]. A biometric system can measure one or more physical or behavioral characteristics. Such a system consists of two phases: recording the obtained biometric data and identification/authentication. In many applications, processed data containing distinctive features are used rather than raw biometric data. In the recognition phase, the user identity is determined by comparing the biometric data obtained from the individual with the recorded data. A biometric system consists of a sensor module, feature extraction module, database module, matching mechanism and decision module as shown in Figure 2.2.

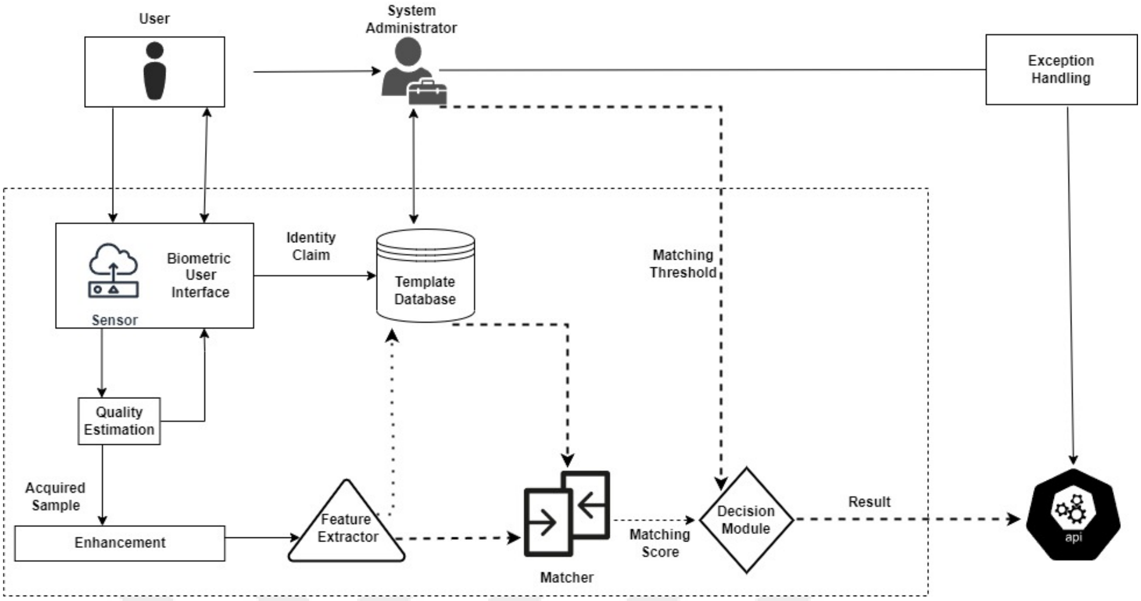


Figure 2.2 Overview of a generic biometric system [52]

Sensor module; It is the first unit to acquire data in the biometric system. The feature extraction module processes the biometric feature and extracts the features to create a biometric template. A biometric template is a collection of features that represent a biometric trait. Biometric information is saved in the database module including some personal identity information. The Matcher module generates a score representing the similarity between the input and the saved biometric template(s). The decision-making module uses this score to determine the outcome of the transaction [88].

ML is increasingly employed in biometric systems and plays big role to improve the performance of them [89]. In [90], a one-shot learning facial recognition method by using Siamese network and support vector machine (SVM) classifier is proposed. During the study they used three different datasets to train as Labeled Faces in the Wild (LFW), The Social Face Classification (SFC) and the YouTube Face (YTF). They showed that their results are close to humans' with the accuracy value of 97.35% ,97.5% and 91.4% in each data set. The authors in [91] proposes the deep CNN based on triplet inputs for one shot-learning method. They achieve better results comparing with the study in [90]. DeepIrisfor is a supervised deep learning structure for iris verification based on pair filter layer (PFL) and convolutional layers [92]. They show that they achieve better values of the accuracy and the Equal Error Rate (EER) comparing to other methods with a high genuine accept rate and very low FAR. There are also studies on application of ML in voice biometrics. In [93], It is proposed a

Voiceprint Authentication System based on Deep Learning.

Considering the areas where biometric systems are used(curtail financial fraud, secure national borders. . . etc), any breach of security may cause serious problems by bringing privacy concerns. As shown in Figure 2.3, vulnerabilities are based on modules, insiders, template databases, and interconnections between modules.

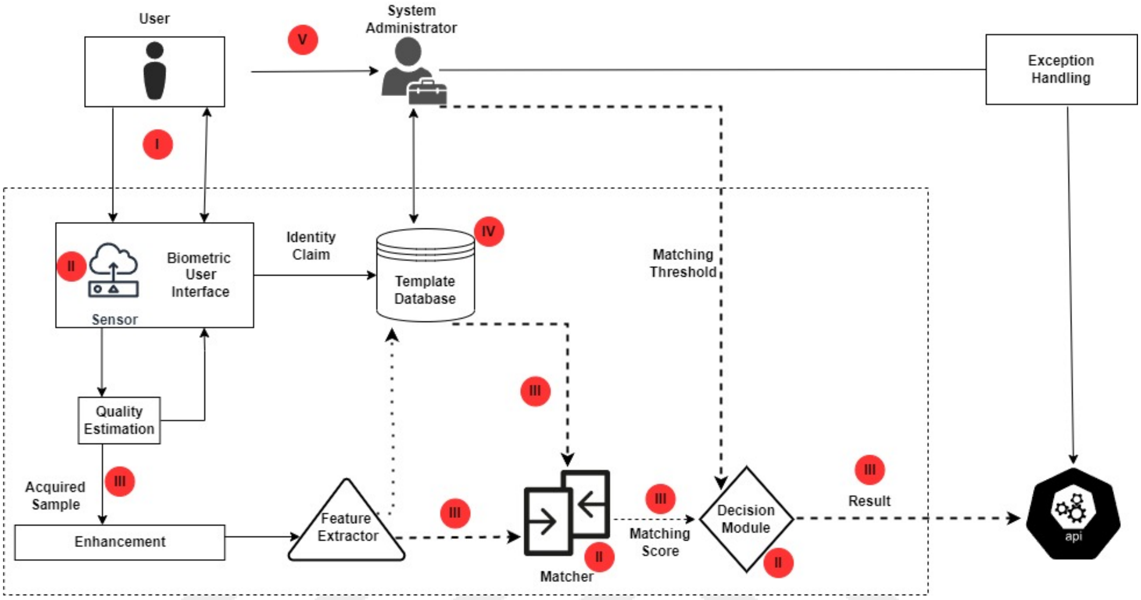


Figure 2.3 Major attack points of a generic biometric:(I) user-biometric system interface, (II) biometric system modules, (III) interconnections between biometric modules, (IV) template database, (V) attacks through insiders System [52]

2.3 Models

In the scope of the thesis, traditional VGG19 and Siamese Network were used for base model and private model.

2.3.1 VGG19

Simonyan and Zisserman propose the 16-layer VGG network (VGG16) and 19-layer VGG network (VGG19) in [94], which forms the basis of the 2014 ImageNet Challenge proposal.

The VGG19 architecture is designed to include convolutional layers and fully connected layers. Convolutional layers use kernels having 3x3 dimensions with a stride and padding to maintain that each of the maps of activation has the same dimensions as the prior one [94]. Each convolution is followed by a rectified linear unit (ReLU) activation [96], and a max pooling operation is occasionally utilized to shrink the spatial dimension. Max pooling layers utilize 2×2 kernels with a stride of 2

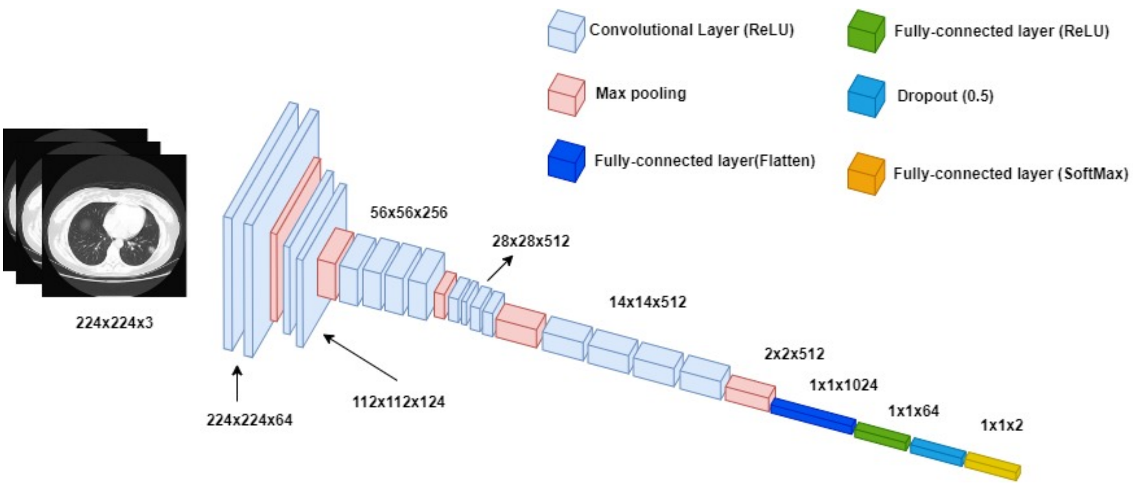

Convolutional Layer (ReLU)
Fully-connected layer (ReLU)
Max pooling
Dropout (0.5)
Fully-connected layer(Flatten)
Fully-connected layer (SoftMax)
224x224x3
56x56x256
28x28x512
14x14x512
2x2x512
1x1x1024
1x1x64
1x1x2
224x224x64
112x112x124

Figure 2.4 VGG19 model used in the thesis [95]

and no padding to provide that each spatial dimension of the activation map from the prior layer is divide into two. Two fully-connected layers with ReLU activated units are then used before the SoftMax layer. The convolutional layers may be considered as feature extraction blocks. The activation maps created by these layers are referred to as bottleneck features. VGG19 model uses three fully-connected layers to carry a classification task [94].

Figure 2.4 displays the structure of the traditional VGG19 architecture, commonly adopted for the image analysis which is also used in our study. This structure requires a test image of dimension of 224 × 224 pixels (RGB/gray scale). In the study, VGG19 is used to classify the images of the CT images.

2.3.2 Siamese Networks(SNNs)

The notion of Siamese Network was proposed by Bromley et al. for hand-written signature verification[109]. Simply, a pair of inputs are fed to sub-networks. Features of each of the inputs are extracted through sub-networks and the vectorial distance is computed. Target output is 1 for the same class input pairs while the output is 0 for different class input pairs. This operation converts classification problem into a similarity problem. Since there are two networks in SNNs, a new kind of loss function is used called Contrastive Loss. It enables to compute the similarity between the feature output vectors obtained from each of the networks [97]. This function is a method for training a similarity metric from data which consist of one or more identical sub-networks which was proposed by Chopra et al. in 2005 [98]. To elaborate on contrastive loss mathematically, let $\overrightarrow{X_1}$ and $\overrightarrow{X_2} \in I$ be a pair of inputs representing two signatures to be compared, as depicted in Figure 2.5, where Y be a binary label of the pair.

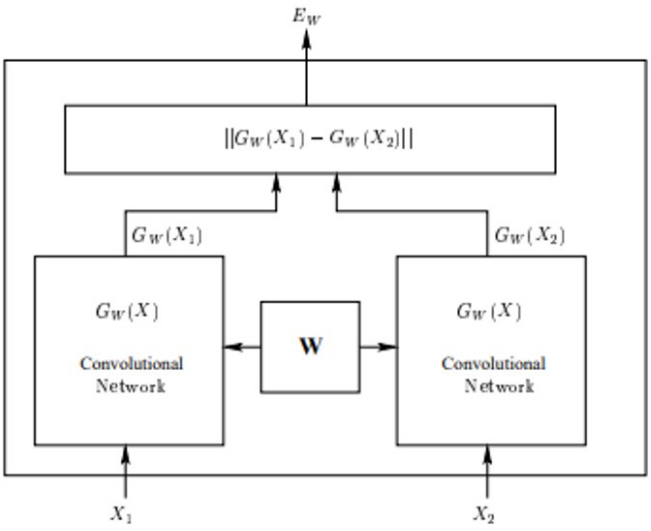


Figure 2.5 Siamese network architecture

If $\overrightarrow{X_1}$ and $\overrightarrow{X_2}$ belong to the same person, then Y will equal to 0. This pair is called as *genuine*. Otherwise, an *impostor* pair will be achieved. Let W be the shared parameter vector that depends learning while $G_W(X_1)$ and $G_W(X_2)$ be the two points in the low-dimensional space that are generated by mapping X_1 and X_2. We will describe the distance function to be learned D_W between $\overrightarrow{X_1}$, $\overrightarrow{X_2}$ as the euclidean distance between the outputs of G_W [99],

$$D_W(\overrightarrow{X}_1, \overrightarrow{X}_2) = \left\|G_w((\overrightarrow{X_1})) - G_w((\overrightarrow{X_2}))\right\|_2$$

If we shorten $D_W(\overrightarrow{X}_1, \overrightarrow{X}_2)$ by writing D_W, general form of the loss function will be,

$$\mathscr{L}(W) = \sum_{i=1}^{P} \mathscr{L}\Big(W, (Y, \overrightarrow{X}_1, \overrightarrow{X}_2)^i\Big) \tag{2.1}$$

Here $\mathscr{L}\Big(W, (Y, \overrightarrow{X}_1, \overrightarrow{X}_2)^i\Big)$ is,

$$\mathscr{L}\Big(W, (Y, \overrightarrow{X}_1, \overrightarrow{X}_2)^i\Big) = (1 - Y)L_S(D_W^i) + Y L_D(D_W^i) \tag{2.2}$$

where $(Y, \overrightarrow{X}_1, \overrightarrow{X}_2)^i$ represents i^{th} labeled sample pair, P is the number of training pairs, L_S depicts the partial loss function for a pair of similar points and L_D depicts the partial loss function for a pair of dissimilar points.

In the literature, it is seen that the use of Siamese network was extended to different computer vision tasks such as object tracking [100–104] and face verification [90, 98]. Siamese networks are preferred because they allow for the learning of broad feature representations using a similarity score from two inputs. A siamese network converts into a comparison function after training [105]. The base model used for the thesis is depicted in Figure 2.6.

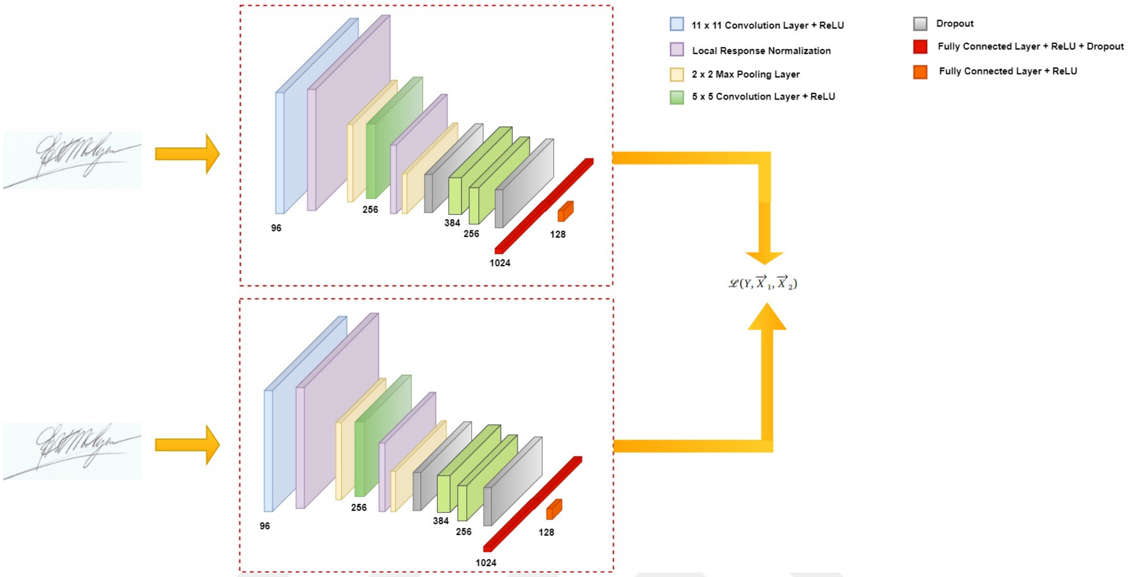

11 x 11 Convolution Layer + ReLU
Local Response Normalization
2 x 2 Max Pooling Layer
5 x 5 Convolution Layer + ReLU
Dropout
Fully Connected Layer + ReLU + Dropout
Fully Connected Layer + ReLU
96
256
384
256
1024
128
$\mathcal{L}(Y, \vec{X}_1, \vec{X}_2)$
96
256
384
256
1024
128

Figure 2.6 The SNN structure used as base model for the thesis

2.3.3 EfficientNet-B0

The Efficient which is developed by scaling with the parameters such as resolution, width, and depth has advantages in terms of cost and robustness. There are 8 models from B0 to B7. In Figure 2.7, the architecture of EfficientNet-B0 is illustrated [106].

2.4 Performance Metrics

Performance metrics are crucial to evaluate whether if an ML model make progress. In this section, the performance metrics which provide monitoring and measuring the performance of the model during the thesis study are given. These metrics are achieved through evaluation factors called as True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) [107].

2.4.1 Accuracy

Accuracy is a performance metric that gives the percentage of correct predictions [107]. It is formulated as:

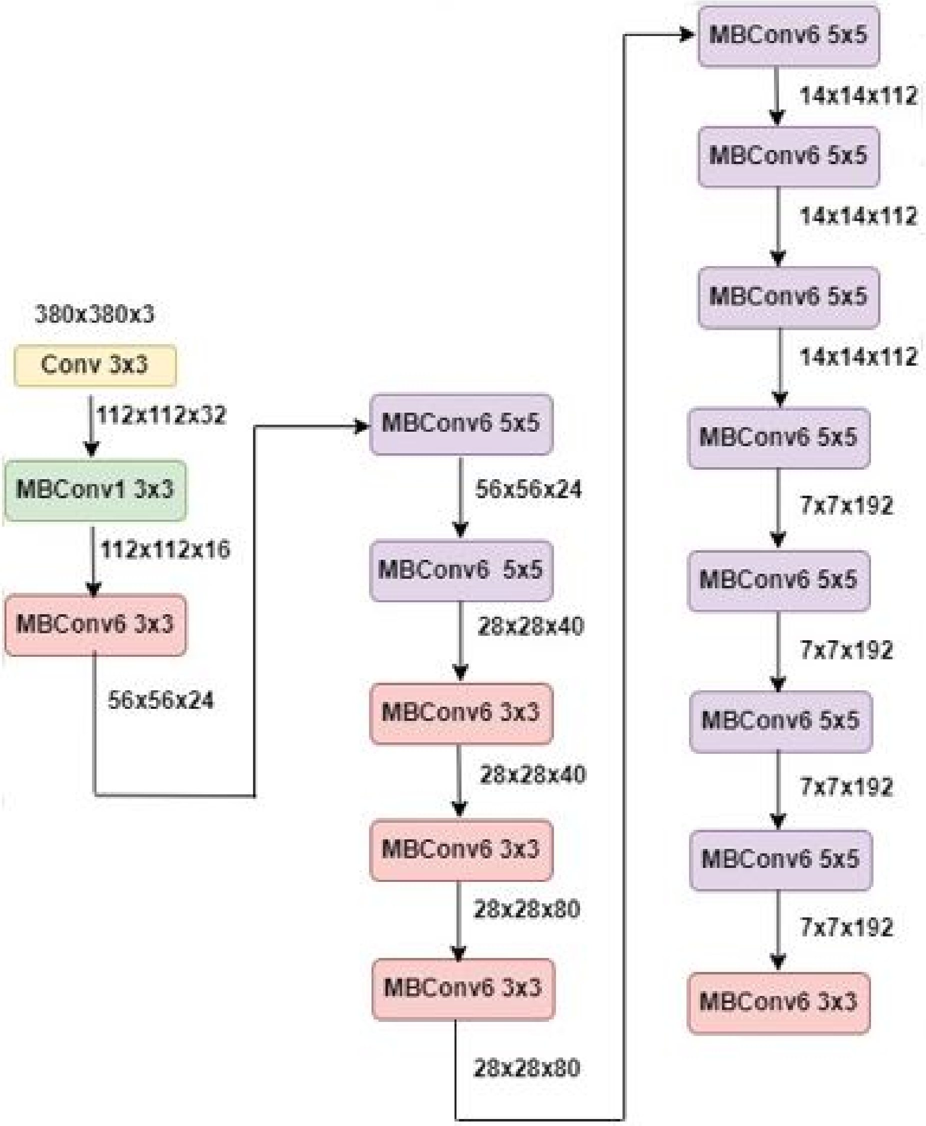


Figure 2.7 The architecture of EfficientNet B0

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.3}$$

2.4.2 Sensitivity/Recall

Sensitivity is a metric of the success of a machine learning model in detecting positive instances which shows capability of a model. It is formulated as:

$$TPR = SEN = \frac{TP}{TP + FN} \tag{2.4}$$

2.4.3 Specifity

Specifity measures the success of our model to detect the negatives. It is formulated as[136]:

$$TNR = SPE = \frac{TN}{TN + FP} \tag{2.5}$$

2.4.4 Area Under Receiver Operating Characteristics Curve (AU ROC)

AU-ROC shows the performance of the capability of the model to distinguish classes. It is plotted with TPR against the FPR.

3
DATA PRIVACY APPROACHES

It is inevitable that machine learning applications become more and more widespread. Since an ML model that performs well requires a large volume of training data, this brings inevitably potential risks of leakage of highly privacy-sensitive information with it. Imposing restrictions on data sharing as a precaution against these risks poses a danger to the progress of studies in the field. There are abundance of attacks disclose individual data by exploiting vulnerabilities in machine learning [6, 10]. One of the most widely used techniques recently investigated that guarantees privacy is differential privacy, which incorporates noise into computations. In this chapter, differential privacy approaches implemented during the thesis study are disscussed. Fredrikson et al. showed that disclosure of patients' genetic markers can jeopardize patient privacy [6].

3.1 Preliminaries

In this work, two threat points are focused on. One of them is at the data input stage and another one is at the training phase. Two different privacy approaches called Differentially-Private Stochastic Gradient Descent (DP-SGD) and Private Aggregation of Teacher Ensembles (PATE) are implemented as a countermeasure to these threats in which the noise is added to clipped gradients during training and aggregated models respectively.

3.1.1 DP-SGD (Differentially Private Stochastic Gradient Descent)

This method protects privacy using only the final parameters which are derived from the training process. The gradient is computed for a random subset of examples, the l_2 norm of each gradient is clipped, the average is computed, noise is added, and a step is taken in the direction of that average noisy gradient[26]. This method controls the influence of the training data during the training process in each iteration [26].

An important issue in DP-SGD is that it enables to follow-up privacy budget in the training process [108]. There are several studies that DP-SGD to diminish memorization [109], data generation [110] and image classification [111].

Formally, the model is updated using noisy gradient and gradient estimates gt are generated at each iteration t.

$$\tilde{g} = g_t + \eta \tag{3.1}$$

Here, η is a multivariate Gaussian distribution noise vector. The gradient of each example has l_2 norm that is assumed to be bounded by L using L/m where m means examples. In practice, the bounded $l_2 - sensitivity$ is provided by gradient clipping which is defined by Abadi et al. [26]. In practice, $l_2 - sensitivity$ is bounded through gradient clipping which reduces an individual gradient in case l_2 norm overruns threshold c [112].

$$clip(g,c) = g.min(1, \frac{c}{\|g\|}) \tag{3.2}$$

3.1.2 PATE (Private Aggregation of Teacher Ensemble)

This approach assembles plural models trained with disjoint datasets in a black-box fashion. Since this data is sensitive, the models are used as "*teachers*" for a "*student*" model instead of being published. The student model is trained to predict an output chosen by noisy voting among all of the teacher models in a way that the data or parameters are not reachable. In Figure 3.1, a generic overview of the PATE approach is depicted.

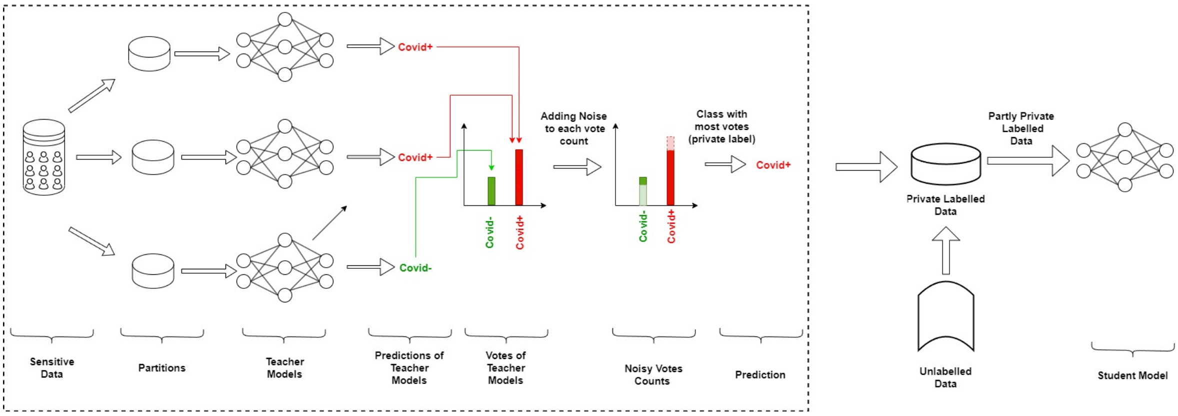


Figure 3.1 Overview of the PATE approach [21]

In formal words, the dataset (X,Y) is divided into n disjoint sets (X_n,Y_n) where X represents the set of inputs, and Y represents the set of labels. The operation provides to obtain n classifiers called teachers (f_i). Then, $f_i(x)$ prediction queries are run for each teacher model and aggregated in a single prediction to create an ensemble making predictions on unseen inputs. The aggregation guarantees privacy protection for this teacher ensemble. For m is the number of classes, j[m] and an input $\overrightarrow{x}$ is the number of teachers which assigned class j to input $\overrightarrow{x}$: $n_j(\overrightarrow{x}) = |\{i : i \in [n], f_i(\overrightarrow{x}) = j\}|$. Since the ensemble's decision shouldn't depend on a single teacher's vote, the plurality (using the label with the largest count) isn't applied and random noise is added to the vote counts n_j to introduce uncertainty.

$$f(x) = \underset{j}{\operatorname{argmax}} \Big\{ n_j(\overrightarrow{x}) + \text{Lap}(\frac{1}{\gamma}) \Big\} \tag{3.3}$$

where γ denotes a privacy parameter and Lap(b) the Laplacian Distribution with location 0 and scale b. A large Y offers a high guarantee of privacy. However, because the noisy maximum f can diverge from the real plurality, it may lessen the labels' accuracy [54].

3.1.3 Privacy Mechanisms

There are three basic privacy mechanisms to provide differential privacy in the literature as the Laplace Mechanism [37], the Gaussian Mechanism [22] and the Exponential Mechanism [113]. We experienced the Laplace Mechanism and the Gaussian Mechanism in our study.

- **The Laplace Mechanism**

  The $l_1 - sensitivity$

  Numeric queries such as, functions $f : \mathbb{N}^{|x|} \rightarrow \mathbb{R}^{|k|}$ rank among the most basic kinds of database queries. These queries convert database data into k real numbers. One of the important parameters that determines just how accurately we can answer such queries is their $l_1 - sensitivity$. The $l_1 - sensitivity$ of a function f is described as:

  $$\Delta f = \max_{\substack{x,y \in N^{|x|} \\ \|f(x)-f(y)\|_1}} \|f(x) - f(y)\|_1 \tag{3.4}$$

  The $l_1 {-} sensitivity$ of a function f captures how much a single individual's data might change the function f in the worst case, and thus intuitively captures the uncertainty in the response we must give to hide the involvement of a single individual. Here, this intuition is expressed mathematically by saying that the sensitivity gives an upper bound on how much we can distort the output of the f function to maintain privacy[26].

The distribution with probability density function defined as follows is known as the Laplace Distribution (centered at 0) with scale b:

$$Lap(x/b) = \frac{1}{2b}\exp(-\frac{|\mathrm{x}|}{\mathrm{b}}) \tag{3.5}$$

The Laplace distribution is a symmetric version of the exponential distribution.

The Laplace mechanism computes f and adds noise derived from the Laplace distribution to each coordinate. The noise scale is adjusted according to the sensitivity of f (divided by ε) [22]

- **The Gaussian Mechanism**

  Let $f : \mathbb{N}^{|x|} \rightarrow \mathbb{R}^{|k|}$ be an arbitrary d-dimensional function, and let us define its $l_2 - sensitivity$ to be

$$\Delta_2 f = \max_{\mathrm{adjacentx,y}} \|f(x) - f(y)\|_2 \tag{3.6}$$

  The Gaussian Mechanism with parameter σ adds noise scaled to $\mathcal{N}(0, \sigma^2)$ to each of the components of the output.

  Let $\varepsilon \in (0,1)$ be arbitrary. For $c^2 > \ln(1.25/\delta)$, the Gaussian Mechanism with parameter $\sigma \geq \Delta_2 f/\varepsilon$ is (ε,δ) - differentially private [26].

3.2 Privacy Approaches

In this section, the approaches which are tackled in the scope thesis are discussed. As it is known, machine learning applications in the field of health are becoming more and more widespread. Especially during the Covid-19 pandemic, many applications have been developed for the rapid diagnosis of diseases and the course of the disease. On the other hand, biometric systems, whose popularity has increased recently, attract attention, especially with applications that require high-level security. The criticality of both the usage fields and the data they use has been a factor in the evaluation of the health and biometrics fields within the scope of this thesis.

3.2.1 Privacy Approach in the AI-Driven Diagnosis of COVID-19

The pandemic has quickly become the most important threat to people's lives and artificial intelligence-driven applications have played a life-saving part in the rapid diagnosis and treatment of the disease. These AI driven applications, although met with enthusiasm initially, have led to type of data to privacy concerns over time. Computed tomography (CT) is an advantageous data type to diagnose COVID-19 during the outbreak. Since publicly available COVID-19 CT datasets are hard to acquire due to privacy issue, it is extremely difficult to develop AI-powered diagnosis methods using CTs [114]. Figure 3.2 illustrates the privacy implementation points which are used in the scope of thesis study. Here, it is seen that an adversary can target the model or data. These are critical attack points for ML algorithms as already mentioned in detail in Section 1.1.1.

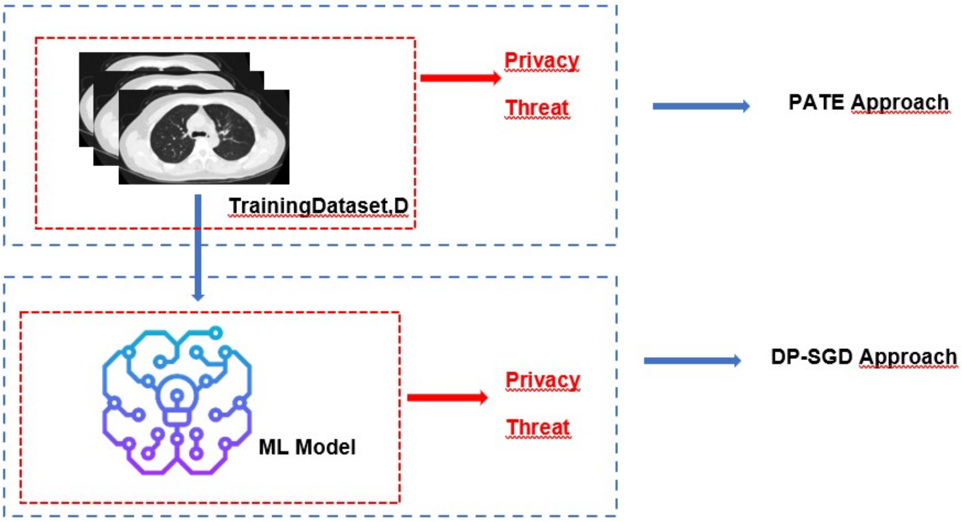


Figure 3.2 Overview of the privacy points of privacy implementation

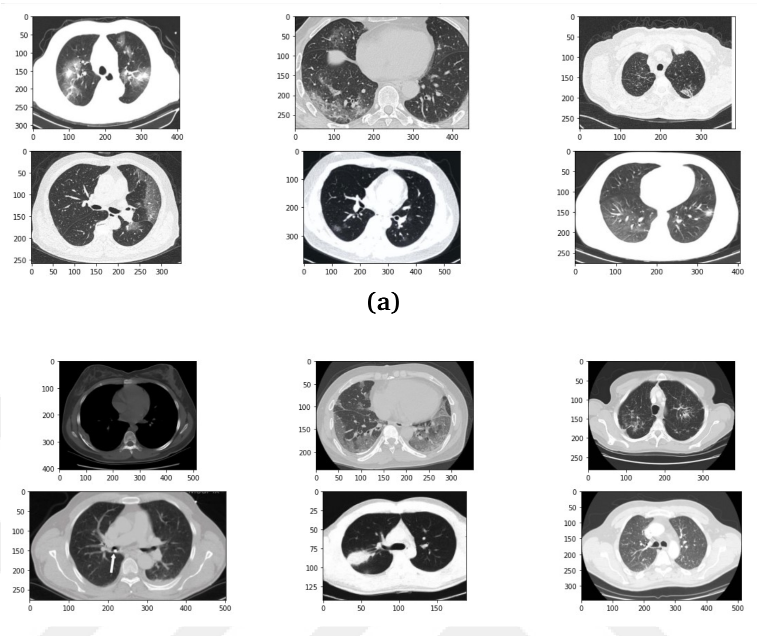
3.2.1.1 COVID-19 Dataset

During our studies we used the open-source dataset which was built to address this issue. The dataset contains 349 CT images with positive label and 397 CT images with negative label of 216 patients which are collected from medRxiv[1] and bioRxiv[2] for COVID-19 through PyMuPDF[3] to extract the low-level structure information of the PDF files of preprints. Figure 3.3 shows samples from the dataset which are given detailed in Table 3.1.

[1] https://www.medrxiv.org/
[2] https://www.biorxiv.org/
[3] https://github.com/pymupdf/PyMuPDF

Table 3.1 Number of samples in Covid-19 dataset

	Train Set	Test Set	Validation Set
Covid-19	279	35	35
Non Covid-19	318	40	39



(a)

(b)

Figure 3.3 Samples from Covid-19 dataset : (a) COVID-19 positive, (b) COVID-19 negative

3.2.2 Privacy Approach in ML-Driven Signature Verification Application

Another issue that gained momentum in our lives due to hygiene reasons during the pandemic is the use of our biometric data. Today, biometric applications can be encountered in a variety of fields from airport crossing points to personnel crossing points. In our study, signature data, which is accepted as a valid characteristic for identity verification in public, legal and commercial transactions, is used. Especially in banking transactions, machine learning is used to prevent forgery in valuable documents. Considering that the signatures cannot be changed like ordinary passwords, accessing data in such an authentication system causes a serious privacy vulnerability.

3.2.2.1 Signature Dataset

During the thesis study BHSig260 dataset is used for the both models' implementation. It includes 24 genuine and 30 forged signatures collected from 260 signaturers. 100 of them were signed in Bengali and 160 of them signed in Hindi. It means that there are 100 × 24=2400 genuine and 100× 30=3000 forged signatures in Bengali, and

$160 \times 24 = 3840$ genuine and $160 \times 30 = 4800$ forged signatures in Hindi.

3.3 Simulation and Results

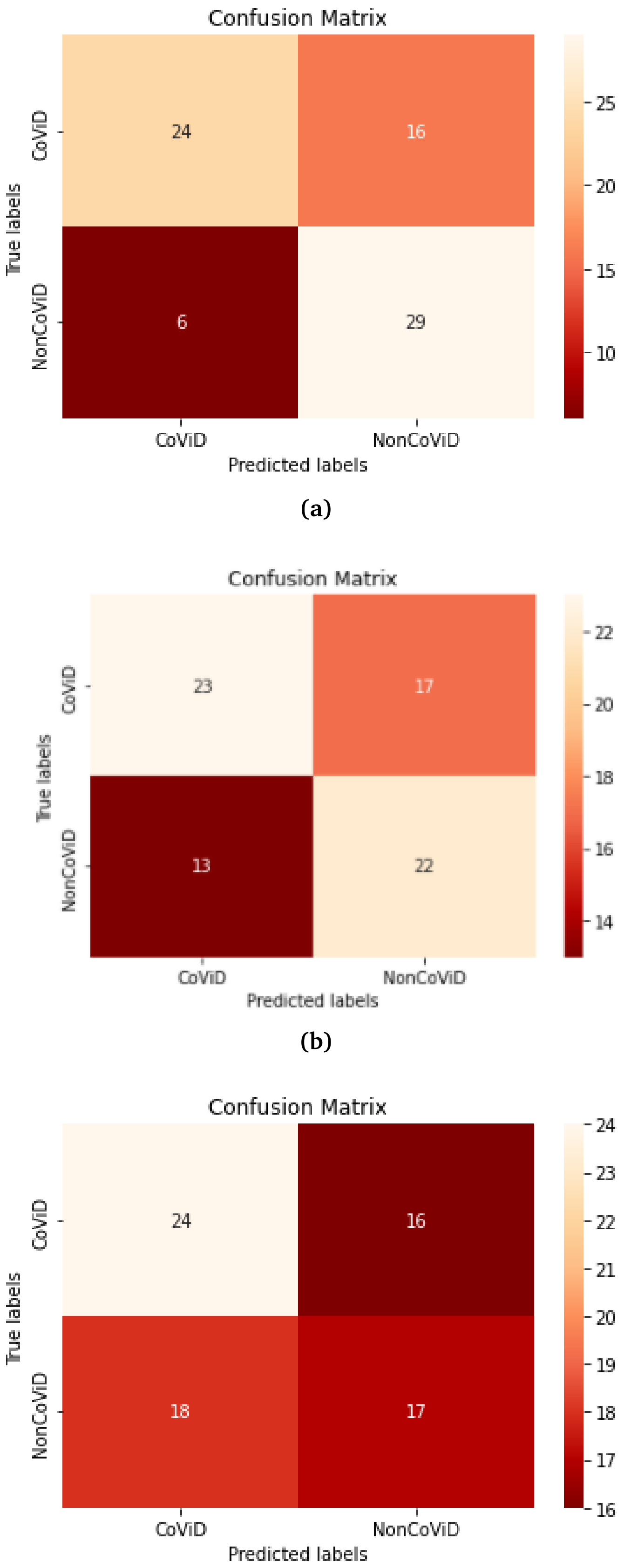
During the study, it is seen that differential privacy implementation deteriorates the accuracy. Compared with DP-SGD method, PATE implementation provides better accuracy. It is known that the confusion matrix is a good indicator to describe the performance of the ML classifier. Figure 3.4 shows the confusion matrices for both the base and the privacy implementations training with COVID-19 dataset. It is seen that the rate of true prediction for PATE implementation is higher than DP-SGD.

For the PATE approach, multiple models which are from different subsets with disjoint datasets are trained [21]. These subsets which are called “teacher” and “student” datasets are used to train "teacher" and "student" models, respectively. The training set is reserved for creating "the teacher" dataset, while the test set is reserved for creating "the student" dataset. After aggregating the outputs of the teacher models, each of which is trained independently, we add the noise to implement privacy. Now, we have private labels. The PATE process is completed by training the student model with private labels. We have used Convolutional Neural Network to achieve private labels. After getting private labels, we have trained Efficient-B0 model, which is our student model, by employing the obtained private labels. In the thesis study, we focus on the effect of the type of the noise on test accuracy by adding Laplace Noise and Gaussian Noise to obtain private labels.

To perform PATE Analysis we use Pysyft, a Python library used to develop privacy-preserving ML [115]. PATE Analysis returns a value called data-dependent epsilon. This value is important as it gives an idea of the success of the quorum among the teachers. The higher the quorum among the teachers, the smaller the privacy cost is. Convolutional Neural Networks algorithm has been used to design teacher ensemble models and the EfficientNet-B0 model is used for the student model. We also implement Laplace Noise and Gaussian Noise respectively.

As seen in Figure 3.5, AUC values show that the performance of seperating the positive class values from the negative class values is better in PATE implementation than in DP-SGD.

One of the privacy parameter which controls the privacy is the privacy budget(ε) [30]. The privacy budget (ε) defines the quantity of the privacy guarantee provided by DP. The smaller value of the ε presents a qualitatively higher privacy guarantee [116]. In Table 3.2 and Table 3.3 show the impact of ε on the differentially private model

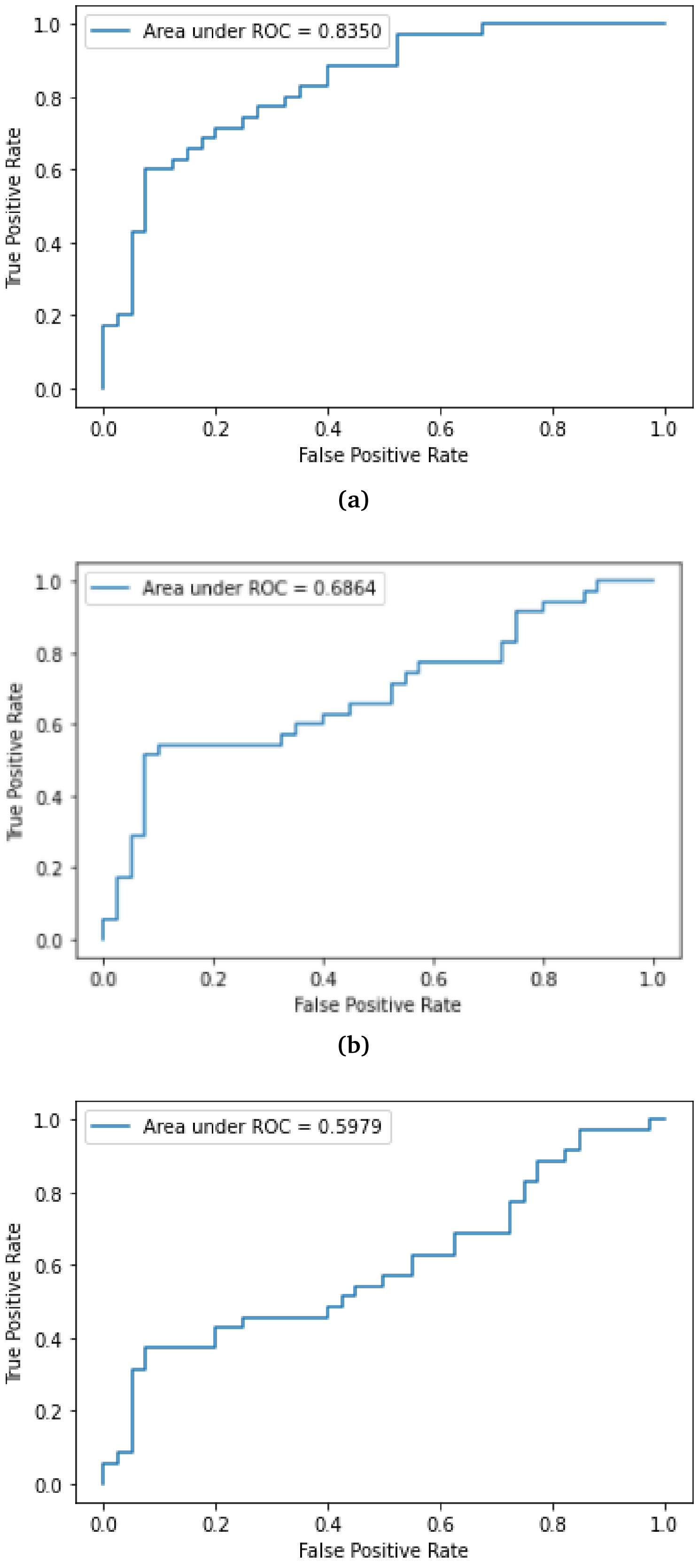


(a)

(b)

(c)

Figure 3.4 Confusion matrices (a) base model, (b) DP-SGD implementation, (c) PATE implementation



(a)

(b)

(c)

Figure 3.5 AU ROC curves (a) base model, (b) DP-SGD implementation, (c) PATE Implementation

using the Laplace Mechanism and Gaussian Mechanism respectively. As mentioned in Section 3.1.1.3, The Laplace Mechanism uses l_1 *sensitivity* while the Gaussian Mechanism uses l_1 *or* l_2 *sensitivity*. Here, sensitivity enables the calibration of the amount of noise. Since the Laplace Mechanism and the Gaussian Mechanism can be extended to vector-valued functions, the length of this vector carries meaningful information about privacy amount. l_2 *sensitivity* will be much lower than the l_1 *sensitivity* for applications with long vectors such as machine learning. In other words, the Gaussian mechanism lets adding much less noise and the accuracy is affected much less compared with the base model.

Table 3.2 DP with the Laplace Mechanism for different ε values (PATE)

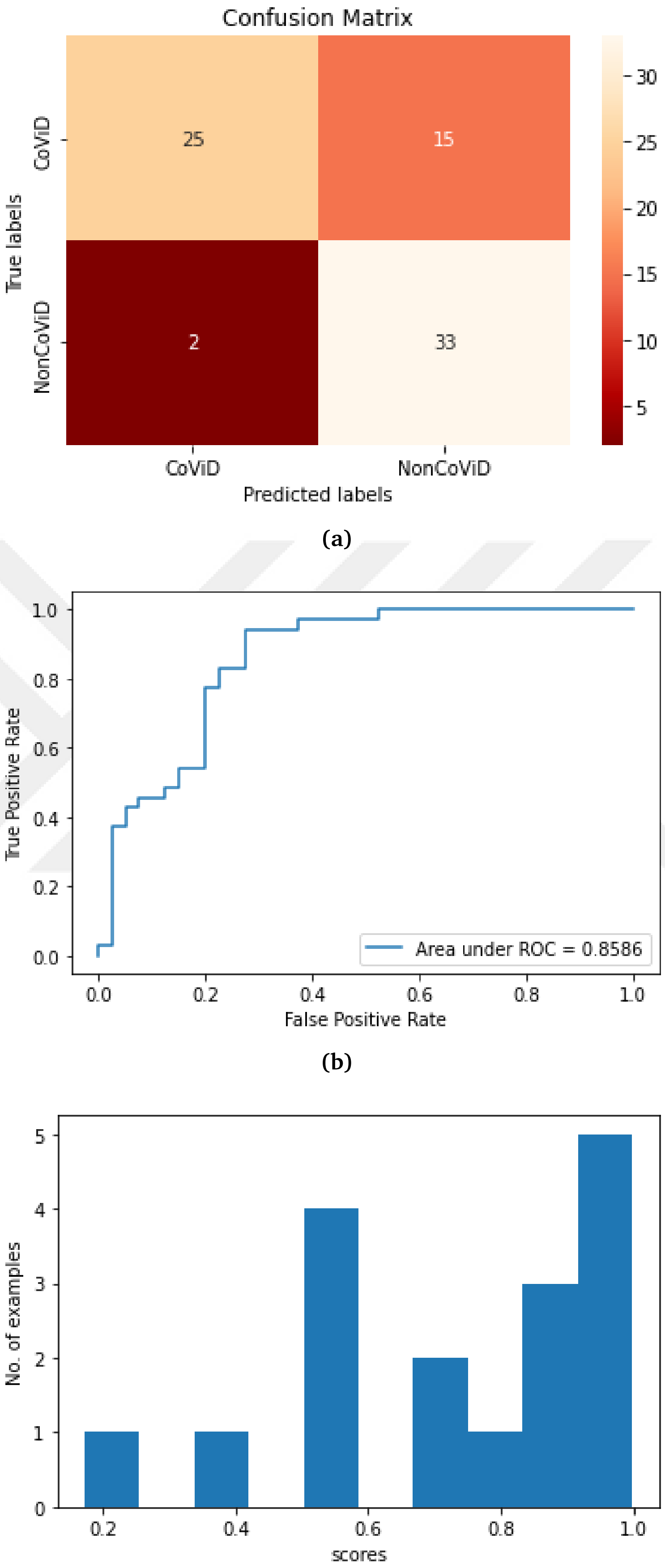
	$\varepsilon = 0.001$	$\varepsilon = 0.05$	$\varepsilon = 0.5$	$\varepsilon = 1.0$
Acc	0.486	0.627	0.640	0.773
Sensitivity	0.371	0.429	0.600	0.943
Specificity	0.590	0.800	0.675	0.625
Area Under ROC	0.493	0.686	0.742	0.859

Table 3.3 DP with the Gaussian Mechanism for different ε values (PATE)

	$\varepsilon = 0.001$	$\varepsilon = 0.05$	$\varepsilon = 0.5$	$\varepsilon = 1.0$
Acc	0.627	0.653	0.507	0.720
Sensitivity	0.914	0.514	0.571	0.914
Specificity	0.375	0.775	0.450	0.550
Area Under ROC	0.746	0.657	0.496	0.686

We also apply DP using Laplace Noise and Gaussian Noise separately for Covid-19 dataset. The results of both implementations are given in Table 3.2 and Table 3.3. We use EfficientNet-B0 model for this study which is used for the diagnosis of COVID-19 from CXR images. During the study, to ensure privacy, we employ Private Aggregation of Teacher Ensembles (PATE) approach as the differential privacy method which was proposed by Papernot et al. [21]. As already mentioned in Figure 3.1, the PATE framework includes the aggregation of teacher models, an aggregation mechanism, and a student model.

We have approximately same data-dependent epsilon value ($\varepsilon \approxeq$ 6.0782). The data-dependent epsilon gives an idea about the consensus among the teachers [28]. The results show that privacy implementation affects the accuracy of the base model negatively while painting a promising picture for future works. Comparing both noise models, we have achieved lower test accuracy rate with DP implementation using Gauss Noise. The Gaussian noise is used to achieve $(\varepsilon, \delta) - DP$. Gauss mechanism uses $l_2 - sensitivity$, while Laplace Noise uses $l_1 - sensitivity$ as mentioned. Even if a negligible delta value is chosen, it is observed that this affects the amount of added noise.

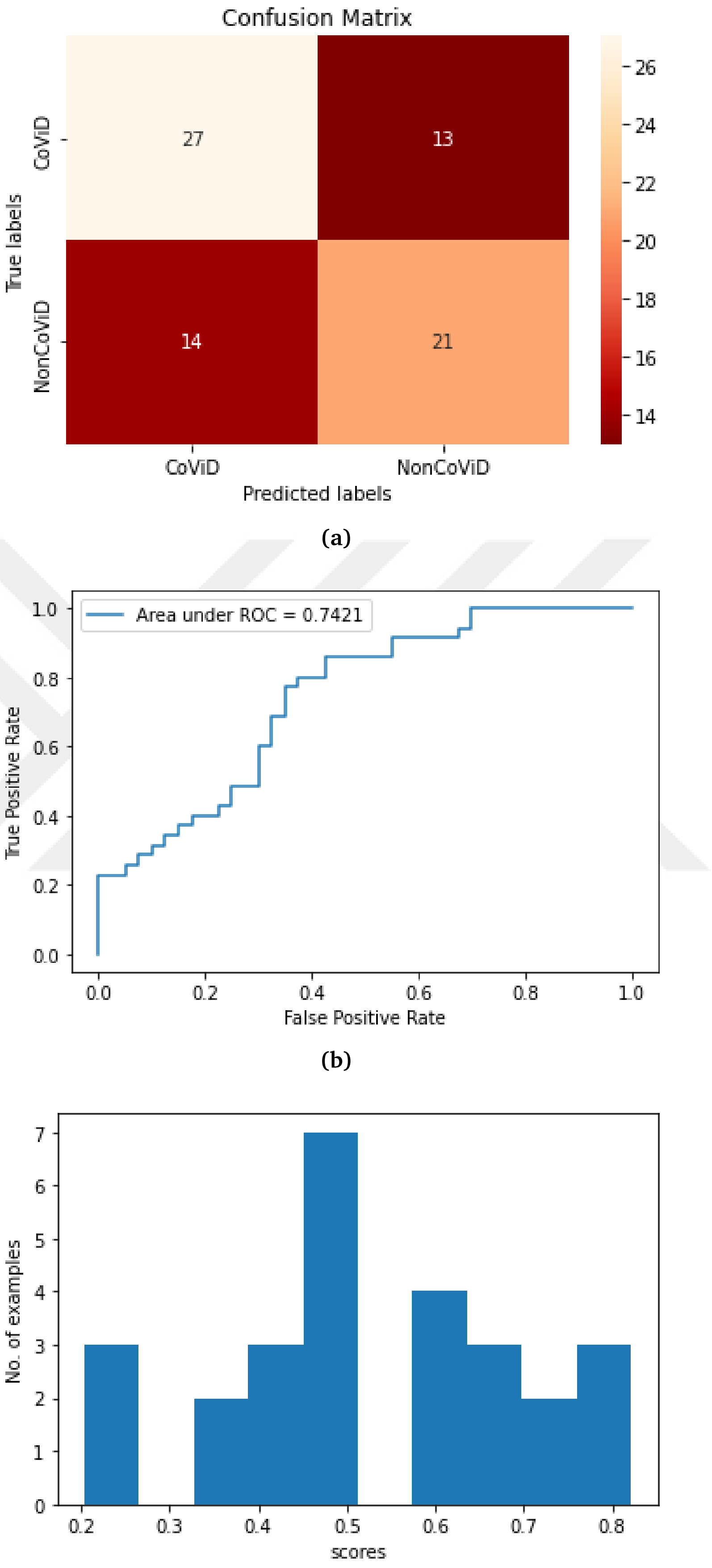


(a)

(b)

(c)

Figure 3.6 PATE implementation with the Laplace Mechanism for ε =0.001 (a) confusion matrix, (b) AU-ROC, (c) misclassified scores

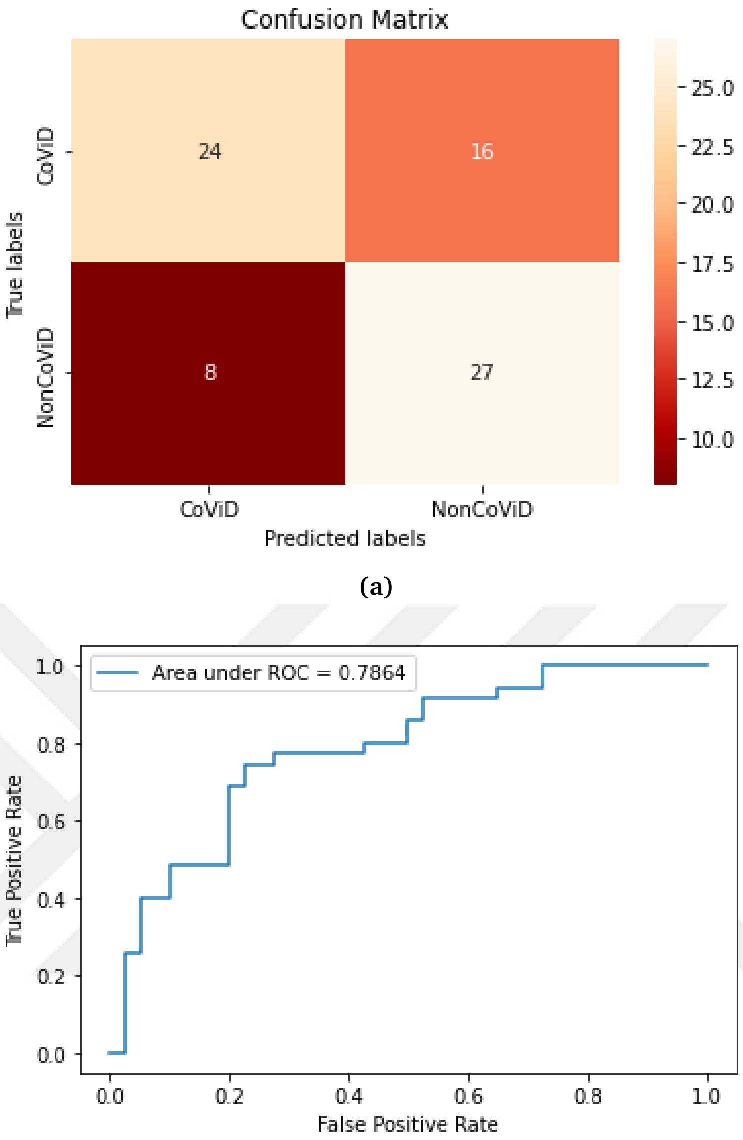


(a)

(b)

(c)

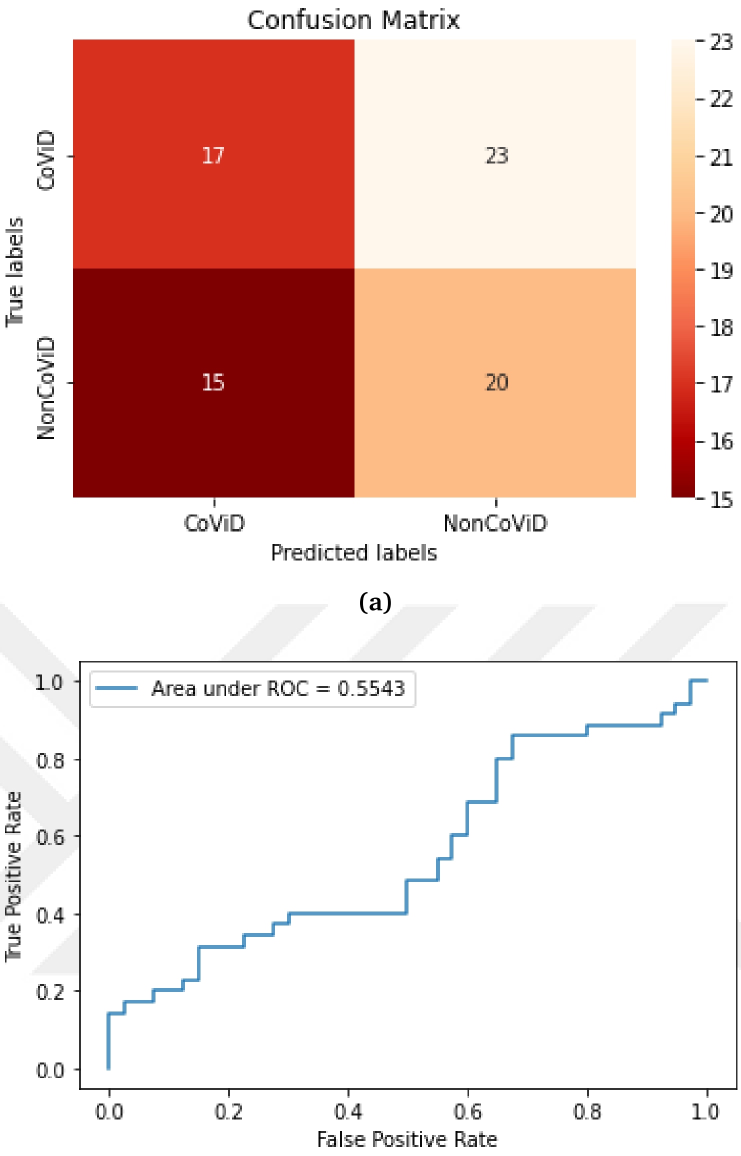
Figure 3.7 PATE implementation with the Laplace Mechanism for ε =0.05 (a) confusion matrix, (b) AU-ROC, (c) misclassified scores



(a)

(b)

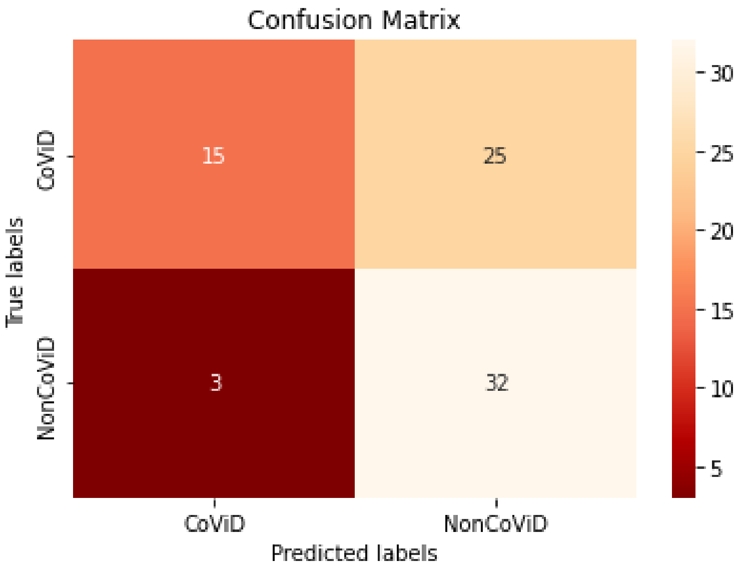
Figure 3.8 PATE implementation with the Laplace Mechanism for ε =0.5 (a) confusion matrix, (b) AU-ROC



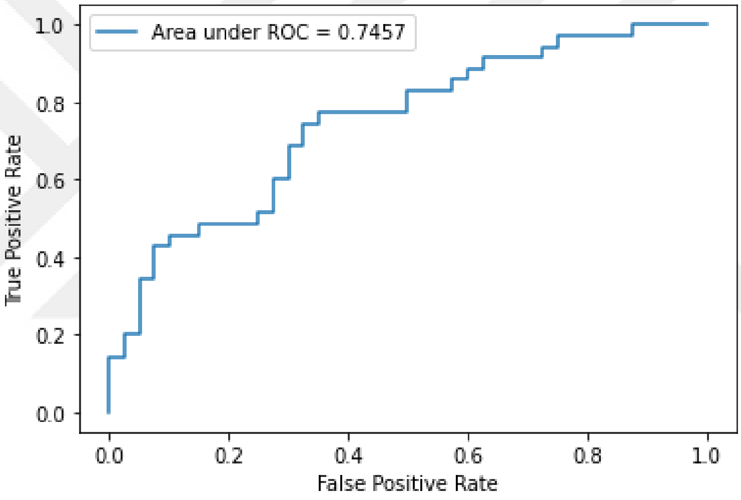
(a)

(b)

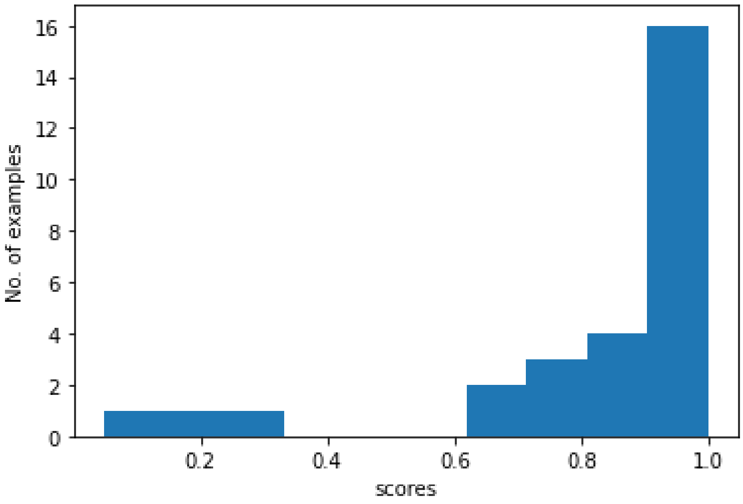
Figure 3.9 PATE implementation with the Laplace Mechanism for ε =1.0 (a) confusion matrix, (b) AU-ROC



(a)

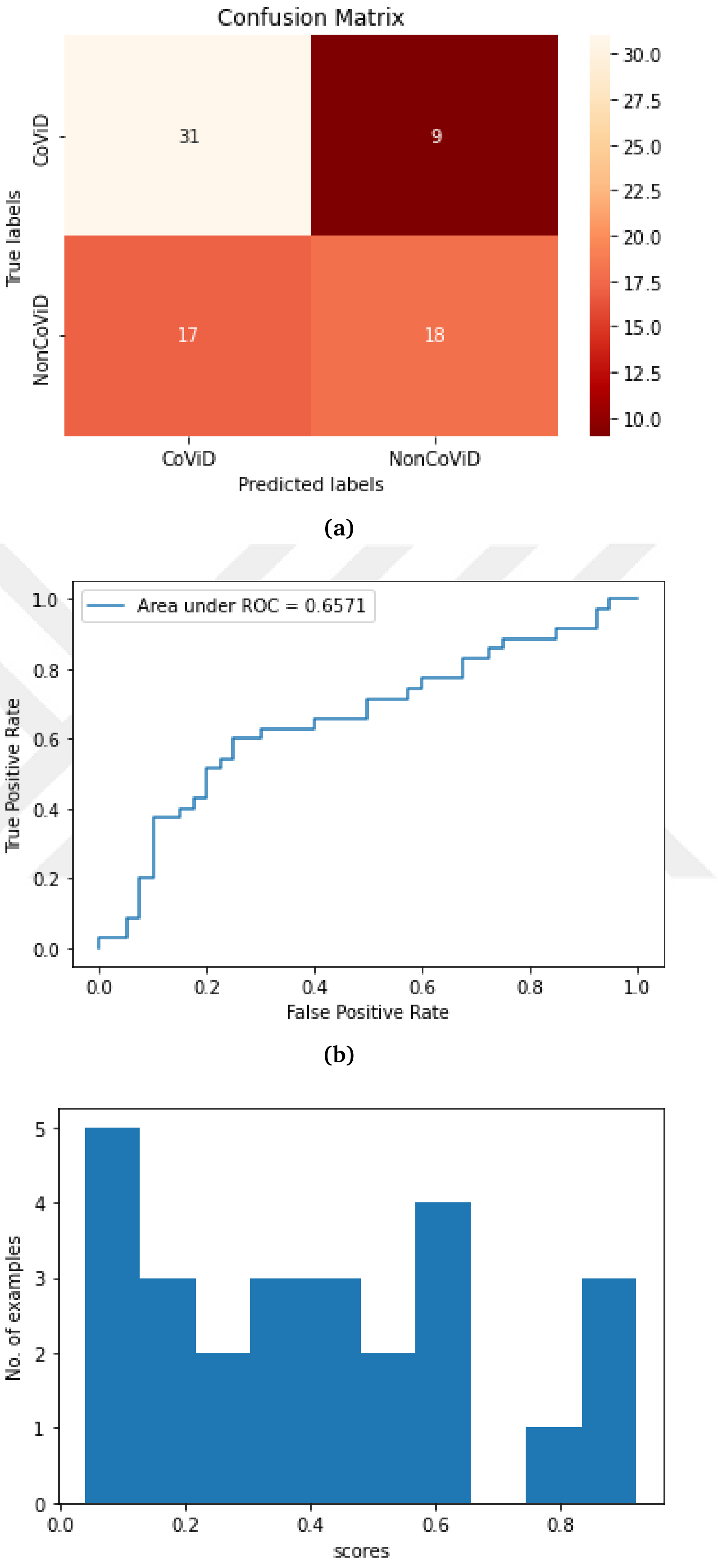


(b)



(c)

Figure 3.10 PATE implementation with the Gaussian Mechanism for ε =0.001 (a) confusion matrix, (b) AU-ROC, (c) misclassified scores

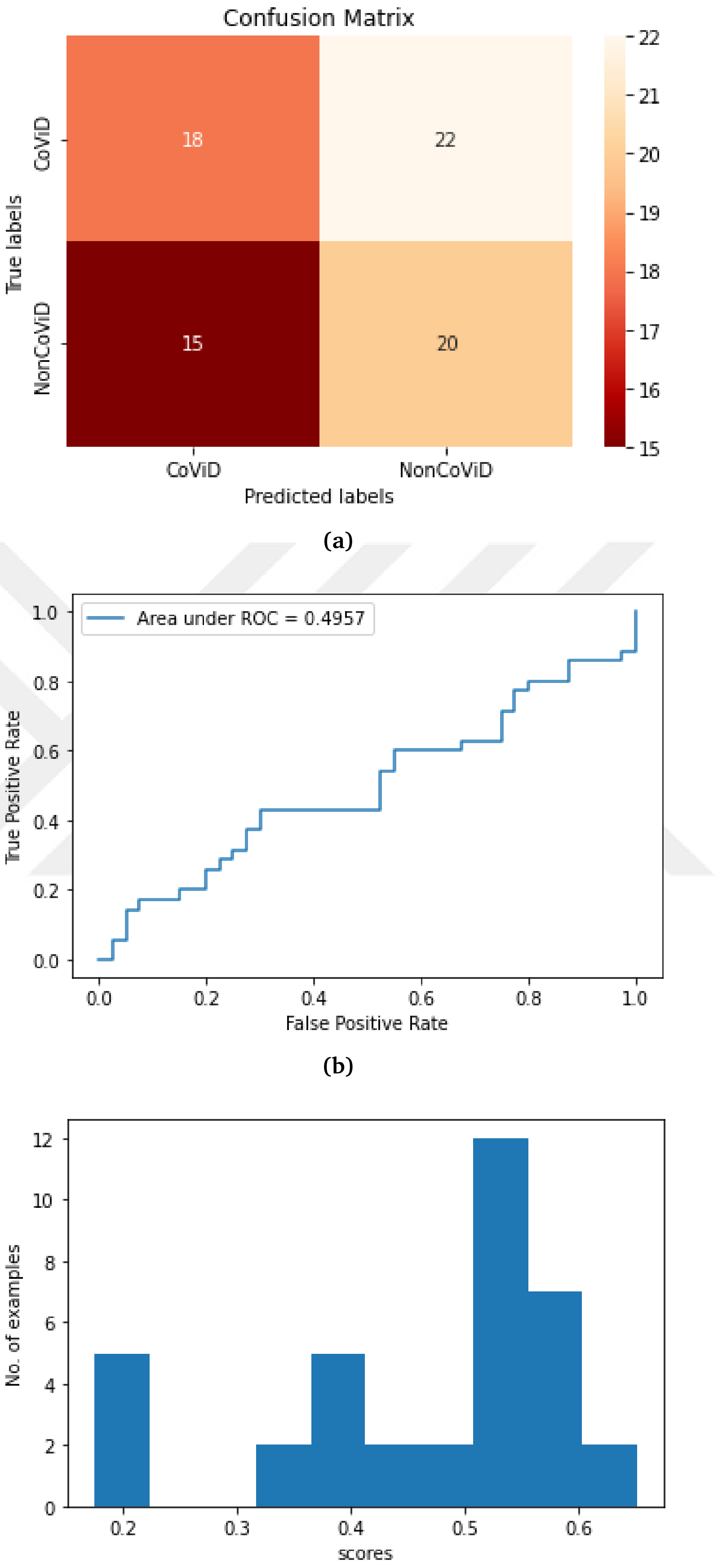


(a)

(b)

(c)

Figure 3.11 PATE implementation with the Gaussian Mechanism for ε =0.05 (a) confusion matrix, (b) AU-ROC, (c) misclassified scores

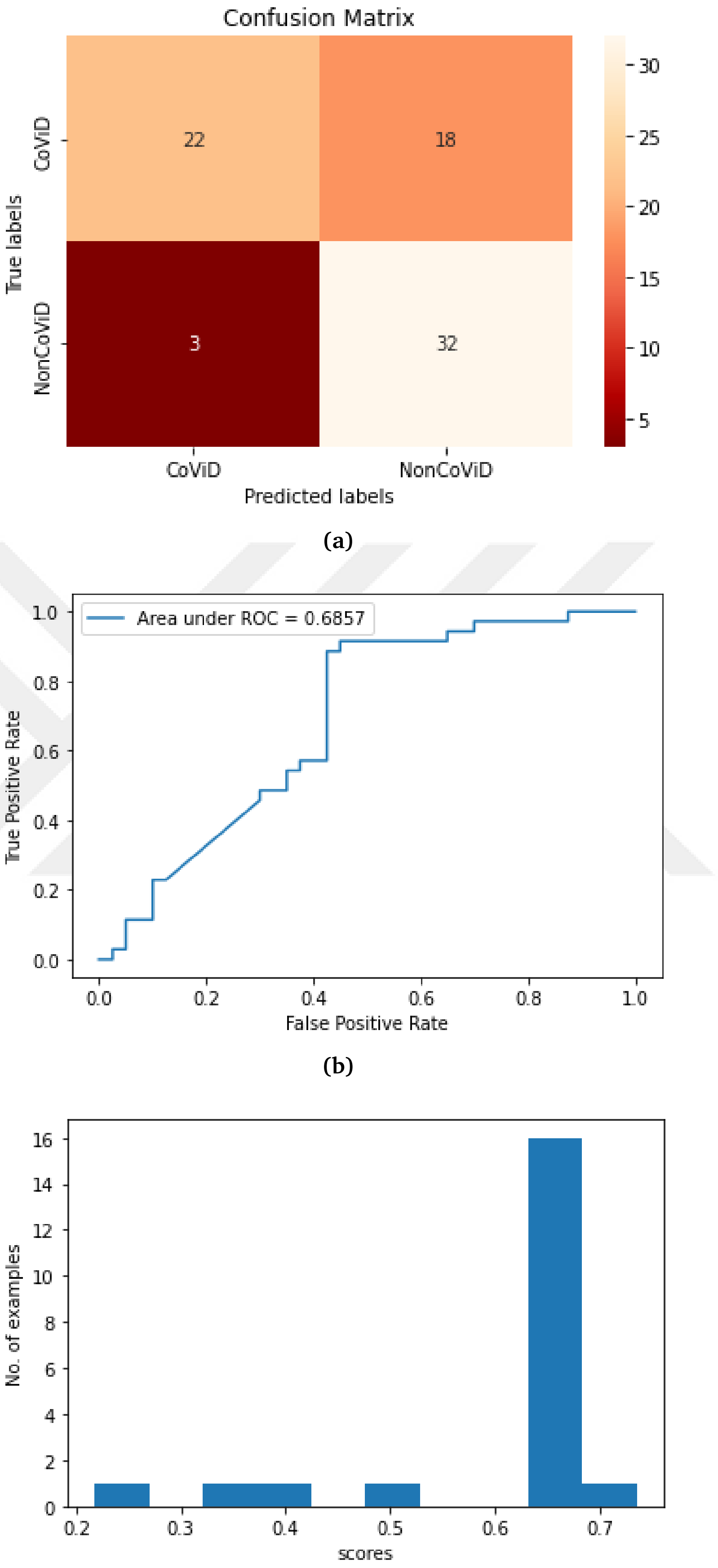


(a)

(b)

(c)

Figure 3.12 PATE implementation with the Gaussian Mechanism for ε =0.5 (a) confusion matrix, (b) AU-ROC, (c) misclassified scores

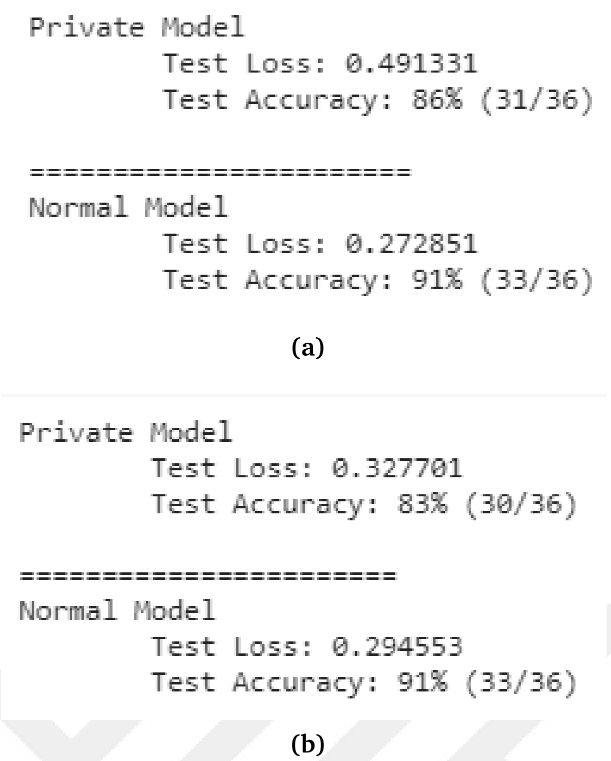


(a)

(b)

(c)

Figure 3.13 PATE implementation with the Gaussian Mechanism for ε =1.0 (a) confusion matrix, (b) AU-ROC



```
Private Model
        Test Loss: 0.491331
        Test Accuracy: 86% (31/36)

=======================
Normal Model
        Test Loss: 0.272851
        Test Accuracy: 91% (33/36)
```

(a)

```
Private Model
        Test Loss: 0.327701
        Test Accuracy: 83% (30/36)

=======================
Normal Model
        Test Loss: 0.294553
        Test Accuracy: 91% (33/36)
```

(b)

Figure 3.14 The effect of noise type on test accuracy of private model: (a) Laplace noise, (b) Gauss noise

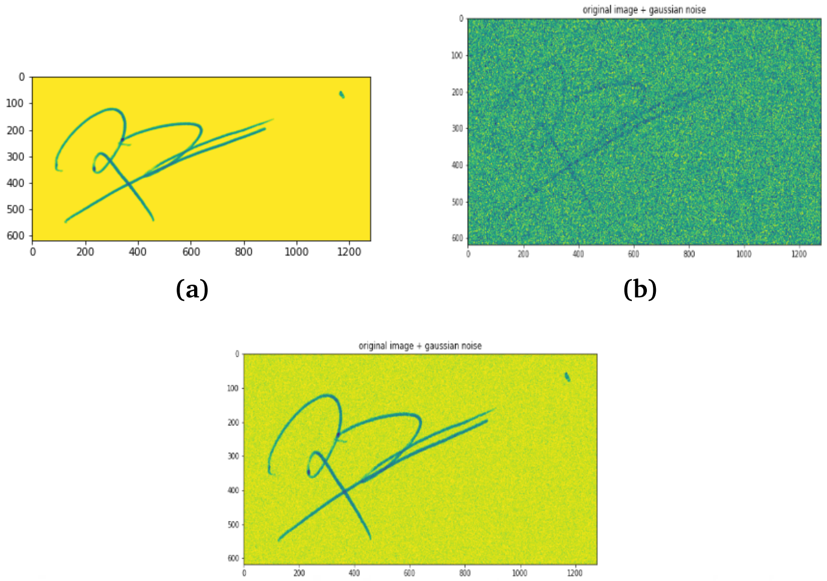
The VGG19 model used for the Covid-19 study is experienced for the signature dataset as well. We achieve an accuracy of 76% for private implementation while it is 81% for the base model. We also observe a decrease in dissimilarity scores between the base and private models, when we implement Siamese Network.

In Figure 3.15, it is shown the effect of the privacy budget on the signature sample from the dataset which is trained. As it is seen, the smaller ε causes more deterioration which means more privacy while decreasing the accuracy of the model.

Since SNN classifies the signatures according to the dissimilarity vector of the images, we use the dissimilarity score to evaluate the privacy effect. Figure 3.16, It is observed a decrease in dissimilarity scores between the base and private models when we implement Siamese Network.

Figure 3.17 demonstrates the misclassified examples. It is obtained using assigned scores after getting the indices of the misclassified examples. The histogram enables to view the nature of the errors committed by the model. In the figure, it is seen that there are spikes. These spikes show our model is confident in misclassifying the examples.

In Figure 3.18 is compared area under ROC curves for the SNN training with the



(a)

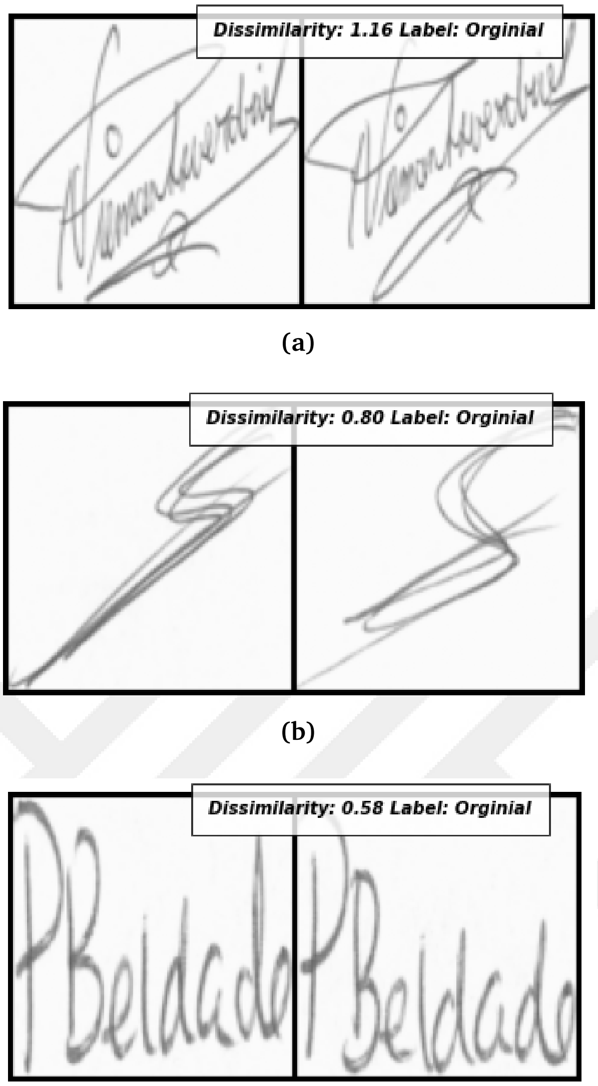
(b)

(c)

Figure 3.15 Samples with Gaussian noisy data (a) original signature, (b) noisy signature for $\varepsilon = 0.001$, (b) noisy signature for $\varepsilon = 0.01$

signature dataset, It is seen that PATE implementation is more advantageous than DP-SGD implementation. It is evident that the AUC for the PATE implementation is higher than that for DP-SGD implementation. Therefore, we conclude that the model with PATE implementation has a better overall performance of classifying the positive class in the dataset.

Lastly, we present a Federated Learning implementation to present a comparison with DP using the signature dataset. The scenario works as depicted in 3.19. A server is created and it initialize the model. In our study, we create three local datasets which are trained with a local copy of the model. Then, the updated model transfers the training results from the local copy to the server. The server also updates the model from the aggregated training results. When we compare the results achieved by using the signature dataset, we achieve 78% accuracy for the VGG19 model and 75% for the SNN model in FL implementation. And we achieve 76% accuracy for the VGG19 model and 71% for the SNN model in DP implementation. While the accuracies of base models are 81% and 88% respectively. We also implement the same setup for the Covid-19 dataset. We achieve 83% accuracy for the Efficient-B0 model and 81% for the VGG19model in FL implementation. And we achieve 50% accuracy for the Efficient-B0 model and 73% for the VGG19 model in DP implementation. While the accuracies of base models are 91% and 87% respectively.

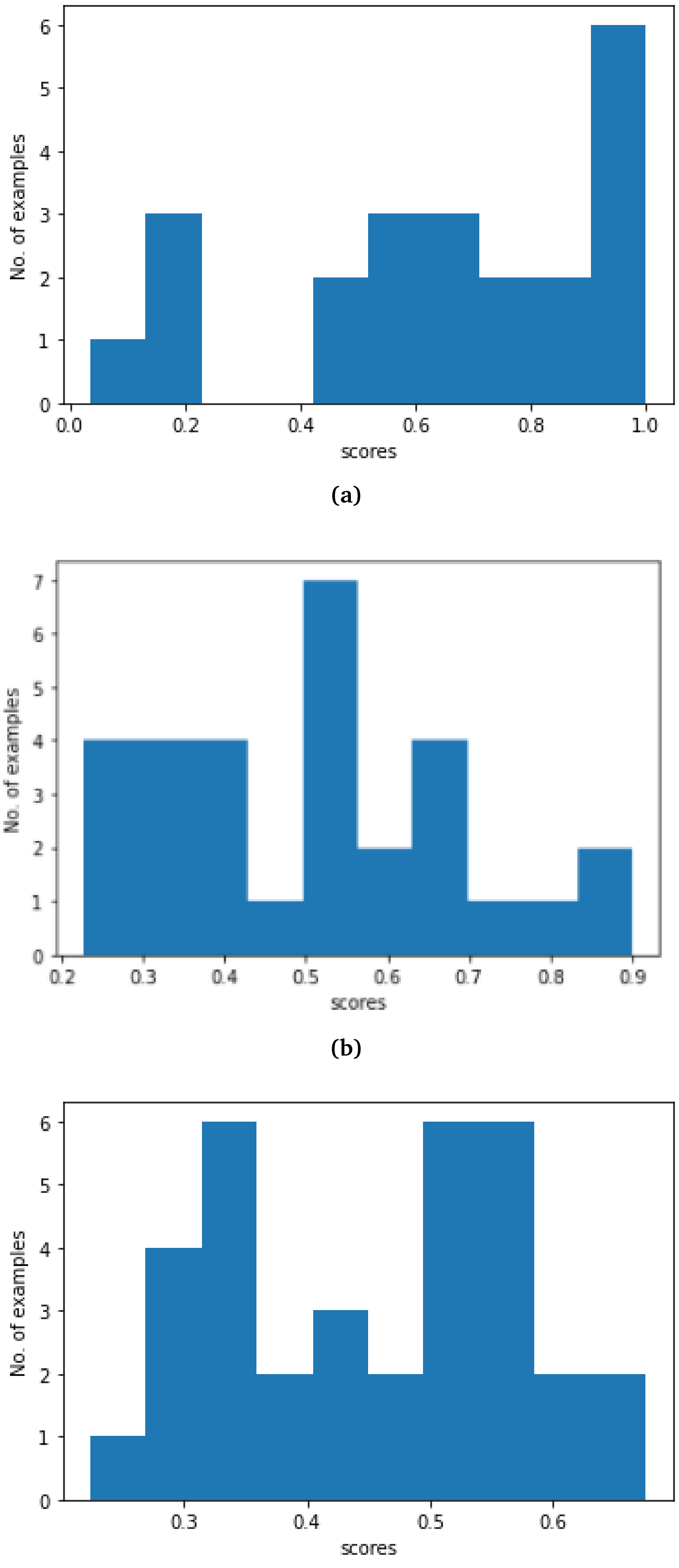

Dissimilarity: 1.16 Label: Orginial
Dissimilarity: 0.80 Label: Orginial
Dissimilarity: 0.58 Label: Orginial

(a)

(b)

(c)

Figure 3.16 Privacy impact on the dissimilarity score (a) base model, (b) private model (PATE), (c) private model (DP-SGD)

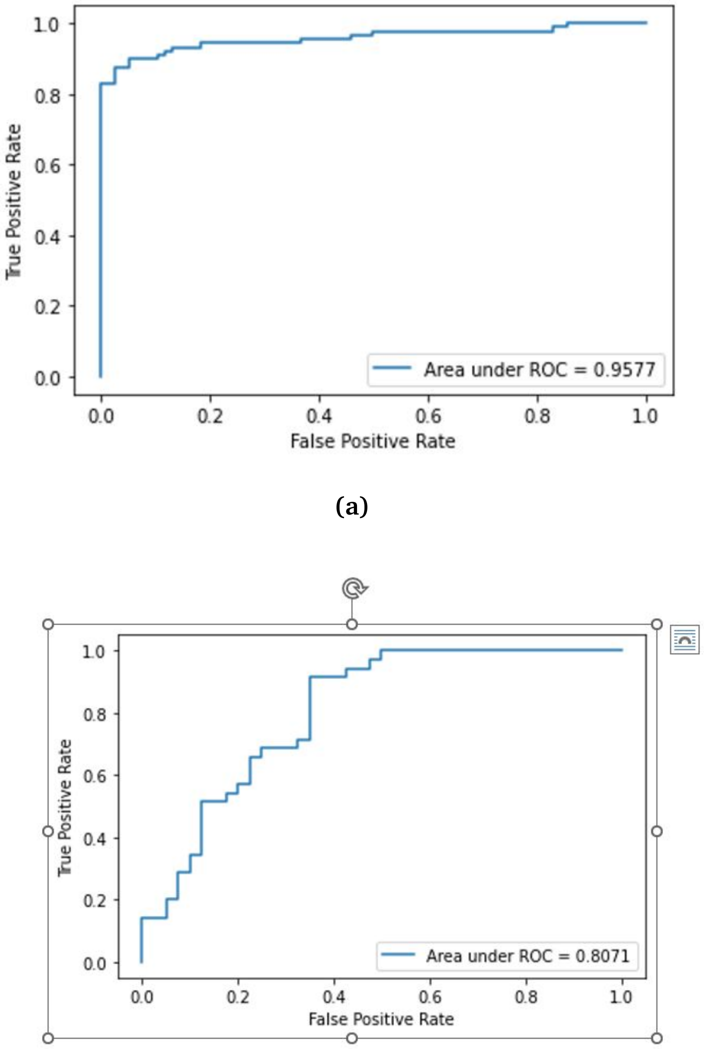


(a)

(b)

(c)

Figure 3.17 Missclassified samples (a) base model, (b) DP-SGD implementation, (c) PATE implementation



(a)

(b)

Figure 3.18 AU-ROC Curves (a) private model (PATE),(b) private model (DP-SGD)

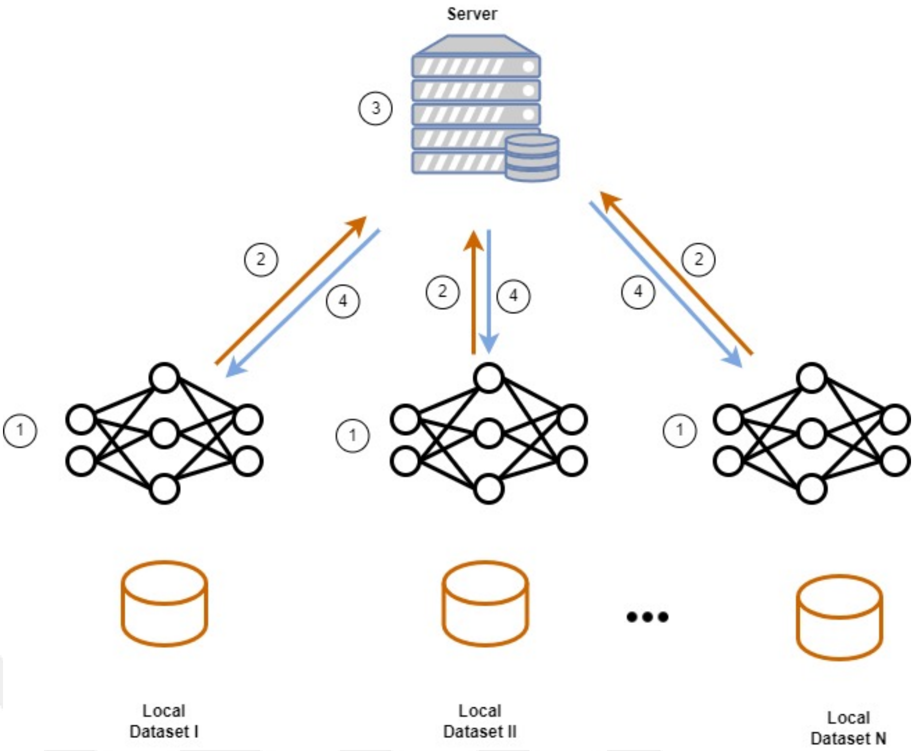


Figure 3.19 Overview of the FL architecture

4
RESULTS AND DISCUSSION

4.1 Conclusions

Although machine learning applications facilitate the quality of our everyday life, they also have a very serious vulnerability in terms of data privacy. Applications developed using personal data directly, especially in the fields of health and biometrics, carry a data privacy risk. Privacy protection technologies are promising tools to address these issues. There are many studies that propose cryptographic techniques to preserve privacy[117, 118]. Since users require to keep the set of encryption keys, these techniques are computationally costly. When the anonymization techniques are considered, it is seen that they are insufficient to protect privacy in applications where the data needs to be analyzed public[119]. Within the scope of the thesis study, different types of differential privacy approaches are investigated to test the effects of privacy implementations by conducting comprehensive research on privacy protection technologies. The privacy implementation at the data point gives better results, whereas implementing privacy at the optimizer level negatively affects the accuracy of the base model. Although it seems that there are studies on this aspect conducted to date, it may be useful to say that there is still some time for real data applications.

The thes presents a comparison that was made for the biometric data and medical data by applying two different differential privacy techniques and federated learning. The aim of the thesis is to compare the behavior of these privacy-preserving approaches for two different models in each field of application. In the future, applications to increase performance for differential privacy will provide good motivation.

We generally use accuracy for the model evaluation. Table 4.1 presents the results of the FL Model by comparing it with the Base Model and DP Model. It is shown that the FL model performs better than the DP model. Here, the noise has a negative effect on the accuracy of the model. And it also takes attraction that the model of EfficientNet-B0 gives worse results.

Table 4.1 Comparison of the accuracies of the private models

	Base Model	FL Model	DP Model	Model Type
Signature	81%	78%	76%	VGG19
Signature	88%	75%	71%	SNN
COVID-19	91%	83%	50%	EfficientNet B0
COVID-19	87%	81%	73%	VGG19

This thesis study clearly demonstrates that it is beneficial to use ML algorithms developed with techniques enhancing privacy, especially in medical data and biometric systems. It is also expected that this study motivates future studies on the subject and paves the way for those future propects. The study also shows that:

- Privacy implementation approach may show differences according to the type of the data used in the system.
- It is more feasible to apply privacy at data points rather than model parameters, especially in critical areas.
- Even though it is still challenging, privacy-preserving mechanisms enable the usage of sensitive data securely.
- While the damage that federated learning creates on model performance paints a more advantageous picture than differential privacy, the comparison of the durability of privacy assurances they provide is open to debate. Therefore, how much accuracy can be compromised when planning privacy will be an important criterion in determining the appropriate privacy protection method.

4.2 Main Contribution

Employing differential privacy in machine learning algorithms using biometric data is the principal novelty of the thesis. Privacy effect on dissimilarity score is achieved in this study for the first time. The study also deepens differential privacy at the local and global levels by adding DP on data and optimizer levels to evaluate performance.

To sum up, one of the privacy-preserving techniques called as DP is applied to biometric and health data in the thesis. The reason for the selection of those areas is that the data used in both domains is sensitive. We especially focuse on two different DP approaches: PATE and DP-SGD. PATE implements privacy to inputs, while DP-SGD implements it to optimizer of the model. Because differential privacy still has many open questions, the study will hopefully have an impact that could change the course of studies on this subject. It should be emphasized that PATE is a supervised approach.

It means that we need to have access to such data since PATE requires doing training on unlabeled data partially.

4.3 Future Works

The point that technology has reached shows that the use of AI in critical areas will become increasingly widespread. While applications in the field of health and biometrics especially come to the fore due to the areas in which they are used, the sensitivity of the data used makes it essential to ensure data and model robustness. At this point, it would be useful to compare different data types and different ML models using different privacy approaches. Thus, privacy mechanisms that may be more suitable for specific data types can be identified.

To decrease the negative effect of DP which presents a robust privacy guarantee on the model, further implementations may be performed by optimizing the ML models with different DP metrics. The results of this thesis study are promising and can be improved by optimizing observed trade-off points and by employing other DP metrics. Studies should be deepened on fair and secure ML models without sacrificing the performance of the model.

REFERENCES

[1] M. A. Kızrak, Z. Müftüoğlu, T. Yıldırım, "Limitations and challenges on the diagnosis of covid-19 using radiology images and deep learning," in *Data Science for COVID-19*, Elsevier, 2021, pp. 91–115.

[2] C. M. Bishop, N. M. Nasrabadi, *Pattern recognition and machine learning*, 4. Springer, 2006, vol. 4.

[3] M. Al-Rubaie, J. M. Chang, "Privacy-preserving machine learning: Threats and solutions," *IEEE Security & Privacy*, vol. 17, no. 2, pp. 49–58, 2019.

[4] A. F. Westin, "Privacy and freedom," *Washington and Lee Law Review*, vol. 25, no. 1, p. 166, 1968.

[5] C. Dwork, M. Naor, "On the difficulties of disclosure prevention in statistical databases or the case for differential privacy," *Journal of Privacy and Confidentiality*, vol. 2, no. 1, 2010.

[6] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, T. Ristenpart, "Privacy in pharmacogenetics: An {end-to-end} case study of personalized warfarin dosing," in *23rd USENIX Security Symposium (USENIX Security 14)*, 2014, pp. 17–32.

[7] N. Papernot, P. McDaniel, A. Sinha, M. P. Wellman, "Sok: Security and privacy in machine learning," in *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*, IEEE, 2018, pp. 399–414.

[8] M. Xue, C. Yuan, H. Wu, Y. Zhang, W. Liu, "Machine learning security: Threats, countermeasures, and evaluations," *IEEE Access*, vol. 8, pp. 74 720–74 742, 2020.

[9] B. Liu *et al.*, "Cloning your mind: Security challenges in cognitive system designs and their solutions," in *Proceedings of the 52nd Annual Design Automation Conference*, 2015, pp. 1–5.

[10] R. Shokri, M. Stronati, C. Song, V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE symposium on security and privacy (SP)*, IEEE, 2017, pp. 3–18.

[11] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, T. Ristenpart, "Stealing machine learning models via prediction {apis}," in *25th USENIX security symposium (USENIX Security 16)*, 2016, pp. 601–618.

[12] R. Agrawal, R. Srikant, "Privacy-preserving data mining," in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 2000, pp. 439–450.

[13] Y. Lindell, B. Pinkas, "Privacy preserving data mining," in *Annual International Cryptology Conference*, Springer, 2000, pp. 36–54.

[14] V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, Y. Theodoridis, "State-of-the-art in privacy preserving data mining," *ACM Sigmod Record*, vol. 33, no. 1, pp. 50–57, 2004.

[15] R. Talbi, "Robust and privacy preserving distributed machine learning," Ph.D. dissertation, Université de Lyon, 2021.

[16] M. U. Hassan, M. H. Rehmani, J. Chen, "Differential privacy techniques for cyber physical systems: A survey," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 1, pp. 746–789, 2019.

[17] L. Sweeney, "Weaving technology and policy together to maintain confidentiality," *The Journal of Law, Medicine & Ethics*, vol. 25, no. 2-3, pp. 98–110, 1997.

[18] P. Samarati, L. Sweeney, "Protecting privacy when disclosing information: K-anonymity and its enforcement through generalization and suppression," 1998.

[19] K. El Emam, F. K. Dankar, "Protecting privacy using k-anonymity," *Journal of the American Medical Informatics Association*, vol. 15, no. 5, pp. 627–637, 2008.

[20] A. Machanavajjhala, J. Gehrke, D. Kifer, M. Venkitasubramaniam, "L-diversity: Privacy beyond k-anonymity," in *22nd International Conference on Data Engineering (ICDE'06)*, 2006, pp. 24–24. DOI: `10.1109/ICDE.2006.1`.

[21] A. Machanavajjhala, D. Kifer, J. Gehrke, M. Venkitasubramaniam, "L-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, 3–es, 2007.

[22] C. Dwork, A. Roth, *et al.*, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[23] H. Hu, Z. Salcic, L. Sun, G. Dobbie, P. S. Yu, X. Zhang, "Membership inference attacks on machine learning: A survey," *ACM Computing Surveys (CSUR)*, 2021.

[24] C. Dwork, "Differential privacy," in *Encyclopedia of Cryptography and Security*, H. C. A. van Tilborg, S. Jajodia, Eds. Boston, MA: Springer US, 2011, pp. 338–340, ISBN: 978-1-4419-5906-5. DOI: `10.1007/978-1-4419-5906-5_752`. [Online]. Available: `https://doi.org/10.1007/978-1-4419-5906-5_752`.

[25] ——, "Differential privacy: A survey of results," in *International conference on theory and applications of models of computation*, Springer, 2008, pp. 1–19.

[26] M. Abadi *et al.*, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 2016, pp. 308–318.

[27] C. Dwork, "A firm foundation for private data analysis," *Communications of the ACM*, vol. 54, no. 1, pp. 86–95, 2011.

[28] A. Beimel, K. Nissim, U. Stemmer, “Private learning and sanitization: Pure vs. approximate differential privacy,” in *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, Springer, 2013, pp. 363–378.

[29] B. Jayaraman, D. Evans, “Evaluating differentially private machine learning in practice,” in *28th USENIX Security Symposium (USENIX Security 19)*, 2019, pp. 1895–1912.

[30] A. Haeberlen, B. C. Pierce, A. Narayan, “Differential privacy under fire,” in *20th USENIX Security Symposium (USENIX Security 11)*, 2011.

[31] C. Dwork, N. Kohli, D. Mulligan, “Differential privacy in practice: Expose your epsilons!” *Journal of Privacy and Confidentiality*, vol. 9, no. 2, 2019.

[32] L. Sweeney, “K-anonymity: A model for protecting privacy,” *International journal of uncertainty, fuzziness and knowledge-based systems*, vol. 10, no. 05, pp. 557–570, 2002.

[33] N. Li, T. Li, S. Venkatasubramanian, “T-closeness: Privacy beyond k-anonymity and l-diversity,” in *2007 IEEE 23rd international conference on data engineering*, IEEE, 2006, pp. 106–115.

[34] P. Samarati, L. Sweeney, “Generalizing data to provide anonymity when disclosing information,” in *PODS*, vol. 98, 1998, pp. 10–1145.

[35] M. E. Nergiz, M. Atzori, C. Clifton, “Hiding the presence of individuals from shared databases,” in *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, 2007, pp. 665–676.

[36] M. Gong, Y. Xie, K. Pan, K. Feng, A. K. Qin, “A survey on differentially private machine learning,” *IEEE computational intelligence magazine*, vol. 15, no. 2, pp. 49–64, 2020.

[37] C. Dwork, F. McSherry, K. Nissim, A. Smith, “Calibrating noise to sensitivity in private data analysis,” in *Theory of cryptography conference*, Springer, 2006, pp. 265–284.

[38] K. Zhang, X. Song, C. Zhang, S. Yu, “Challenges and future directions of secure federated learning: A survey,” *Frontiers of computer science*, vol. 16, no. 5, pp. 1–8, 2022.

[39] Q. Yang, Y. Liu, T. Chen, Y. Tong, “Federated machine learning: Concept and applications,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–19, 2019.

[40] J. Konecn, H. B. McMahan, D. Ramage, P. Richtárik, “Federated optimization: Distributed machine learning for on-device intelligence (2016),” *arXiv preprint arXiv:1610.02527*, 2016.

[41] J. Konečn, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, D. Bacon, “Federated learning: Strategies for improving communication efficiency,” *arXiv preprint arXiv:1610.05492*, 2016.

[42] H. McMahan, E. Moore, D. Ramage, B. Arcas, “Federated learning of deep networks using model averaging. corr abs/1602.05629,” *arXiv preprint arXiv:1602.05629*, 2016.

[43] W. Du, M. J. Atallah, "Privacy-preserving cooperative statistical analysis," in *Seventeenth Annual Computer Security Applications Conference*, IEEE, 2001, pp. 102–110.

[44] W. Du, Y. S. Han, S. Chen, "Privacy-preserving multivariate statistical analysis: Linear regression and classification," in *Proceedings of the 2004 SIAM international conference on data mining*, SIAM, 2004, pp. 222–233.

[45] L. Wan, W. K. Ng, S. Han, V. C. Lee, "Privacy-preservation for gradient descent methods," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2007, pp. 775–783.

[46] R. L. Rivest, L. Adleman, M. L. Dertouzos, *et al.*, "On data banks and privacy homomorphisms," *Foundations of secure computation*, vol. 4, no. 11, pp. 169–180, 1978.

[47] A. Wainakh, "Homomorphic encryption for data security in cloud computing," M.S. thesis, Middle East Technical University, 2018.

[48] H. Fang, Q. Qian, "Privacy preserving machine learning with homomorphic encryption and federated learning," *Future Internet*, vol. 13, no. 4, p. 94, 2021.

[49] X. Yi, R. Paulet, E. Bertino, "Homomorphic encryption," in *Homomorphic encryption and applications*, Springer, 2014, pp. 27–46.

[50] K. Benzekki, A. El Fergougui, A. E. B. El Alaoui, "A secure cloud computing architecture using homomorphic encryption," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 2, 2016.

[51] O. Goldreich, "Secure multi-party computation," *Manuscript. Preliminary version*, vol. 78, p. 110, 1998.

[52] A. K. Jain, A. A. Ross, K. Nandakumar, "Multibiometrics," in *Introduction to Biometrics*, Springer, 2011, pp. 209–258.

[53] C. Dwork, "Differential privacy," in *33rd International Colloquium on Automata, Languages and Programming, part II (ICALP 2006)*, ser. Lecture Notes in Computer Science, vol. 4052, Springer Verlag, Jul. 2006, pp. 1–12, ISBN: 3-540-35907-9. [Online]. Available: `https://www.microsoft.com/en-us/research/publication/differential-privacy/`.

[54] N. Papernot, M. Abadi, U. Erlingsson, I. Goodfellow, K. Talwar, "Semi-supervised knowledge transfer for deep learning from private training data," *arXiv preprint arXiv:1610.05755*, 2016.

[55] Z. Müftüoğlu, M. Kızrak, T. Yıldırım, "Privacy-preserving mechanisms with explainability in assistive ai technologies," in *Advances in Assistive Technologies*, Springer, 2022, pp. 287–309.

[56] C. Berghoff, M. Neu, A. von Twickel, "The interplay of ai and biometrics: Challenges and opportunities," *Computer*, vol. 54, no. 09, pp. 80–85, 2021.

[57] S. B. Scruggs *et al.*, "Harnessing the heart of big data," *Circulation research*, vol. 116, no. 7, pp. 1115–1119, 2015.

[58] S. Benjamens, P. Dhunnoo, B. Meskó, "The state of artificial intelligence-based fda-approved medical devices and algorithms: An online database," *NPJ digital medicine*, vol. 3, no. 1, pp. 1–8, 2020.

[59] S. Horng, D. A. Sontag, Y. Halpern, Y. Jernite, N. I. Shapiro, L. A. Nathanson, “Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning,” *PloS one*, vol. 12, no. 4, e0174708, 2017.

[60] N. M. Patel *et al.*, “Enhancing next-generation sequencing-guided cancer care through cognitive computing,” *The oncologist*, vol. 23, no. 2, pp. 179–185, 2018.

[61] G. D. Magoulas, A. Prentza, “Machine learning in medical applications,” in *Advanced course on artificial intelligence*, Springer, 1999, pp. 300–307.

[62] R. Hanka, T. Harte, A. Dixon, D. Lomas, P. Britton, “Neural networks in the interpretation of contrast-enhanced magnetic resonance images of the breast,” *CURRENT PERSPECTIVES IN HEALTHCARE COMPUTING*, pp. 275–283, 1996.

[63] S. Andreassen, O. K. Hejlesen, “The role of model-based systems in medical decision support,” in *Proceedings of the International Conference on Neural Networks and Expert Systems in Medicine and Healthcare: NNESMED 94*, University of Plymouth Press, 1994, pp. 310–318.

[64] P. R. Innocent, M. Barnes, R. John, “Application of the fuzzy art/map and minmax/map neural network models to radiographic image classification,” *Artificial Intelligence in Medicine*, vol. 11, no. 3, pp. 241–263, 1997.

[65] S. Phee, W. Ng, I. Chen, F. Seow-Choen, B. Davies, “Automation of colonoscopy part ii: Visual-control aspects. interpreting images with a computer to automatically maneuver the colonoscope,” *IEEE Engineering in Medicine and Biology*, pp. 81–88,

[66] K. Kralj, M. Kukar, *Using machine learning to analyse attributes in the diagnosis of coronary artery disease*. 1998.

[67] J. Stausberg, M. Person, “A process model of diagnostic reasoning in medicine,” *International Journal of Medical Informatics*, vol. 54, no. 1, pp. 9–23, 1999.

[68] B. Zupan, J. A. Halter, M. Bohanec, “Qualitative model approach to computer assisted reasoning in physiology,” *Proceedings of Intelligent Data Analysis in Medicine and Pharmacology-IDAMAP98*, 1998.

[69] P. Sajda, “Machine learning for detection and diagnosis of disease,” *Annual review of biomedical engineering*, vol. 8, no. 1, pp. 537–565, 2006.

[70] R. Cuocolo, T. Perillo, E. De Rosa, L. Ugga, M. Petretta, “Current applications of big data and machine learning in cardiology,” *Journal of geriatric cardiology: JGC*, vol. 16, no. 8, p. 601, 2019.

[71] P. Chang *et al.*, “Deep-learning convolutional neural networks accurately classify genetic mutations in gliomas,” *American Journal of Neuroradiology*, vol. 39, no. 7, pp. 1201–1207, 2018.

[72] V. Romeo *et al.*, “Machine learning analysis of mri-derived texture features to predict placenta accreta spectrum in patients with placenta previa,” *Magnetic resonance imaging*, vol. 64, pp. 71–76, 2019.

[73] M. M. Islam, T. Nasrin, B. A. Walther, C.-C. Wu, H.-C. Yang, Y.-C. Li, "Prediction of sepsis patients using machine learning approach: A meta-analysis," *Computer methods and programs in biomedicine*, vol. 170, pp. 1–9, 2019.

[74] A. Tahmassebi *et al.*, "Impact of machine learning with multiparametric magnetic resonance imaging of the breast for early prediction of response to neoadjuvant chemotherapy and survival outcomes in breast cancer patients," *Investigative radiology*, vol. 54, no. 2, p. 110, 2019.

[75] D. S. W. Ting *et al.*, "Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes," *Jama*, vol. 318, no. 22, pp. 2211–2223, 2017.

[76] R. Gargeya, T. Leng, "Automated identification of diabetic retinopathy using deep learning," *Ophthalmology*, vol. 124, no. 7, pp. 962–969, 2017.

[77] Z. Li, Y. He, S. Keel, W. Meng, R. T. Chang, M. He, "Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs," *Ophthalmology*, vol. 125, no. 8, pp. 1199–1206, 2018.

[78] P. M. Burlina, N. Joshi, M. Pekala, K. D. Pacheco, D. E. Freund, N. M. Bressler, "Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks," *JAMA ophthalmology*, vol. 135, no. 11, pp. 1170–1176, 2017.

[79] F. Grassmann *et al.*, "A deep learning algorithm for prediction of age-related eye disease study severity scale for age-related macular degeneration from color fundus photography," *Ophthalmology*, vol. 125, no. 9, pp. 1410–1420, 2018.

[80] J. M. Brown *et al.*, "Automated diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks," *JAMA ophthalmology*, vol. 136, no. 7, pp. 803–810, 2018.

[81] D. S. W. Ting *et al.*, "Artificial intelligence and deep learning in ophthalmology," *British Journal of Ophthalmology*, vol. 103, no. 2, pp. 167–175, 2019.

[82] Y. Ohta, H. Yunaga, S. Kitao, T. Fukuda, T. Ogawa, "Detection and classification of myocardial delayed enhancement patterns on mr images with deep neural networks: A feasibility study," *Radiology. Artificial intelligence*, vol. 1, no. 3, 2019.

[83] Z. I. Attia *et al.*, "Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram," *Nature medicine*, vol. 25, no. 1, pp. 70–74, 2019.

[84] A. M. Alaa, T. Bolton, E. Di Angelantonio, J. H. Rudd, M. Van der Schaar, "Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 uk biobank participants," *PloS one*, vol. 14, no. 5, e0213653, 2019.

[85] T. A. Daghistani, R. Elshawi, S. Sakr, A. M. Ahmed, A. Al-Thwayee, M. H. Al-Mallah, "Predictors of in-hospital length of stay among cardiac patients: A machine learning approach," *International journal of cardiology*, vol. 288, pp. 140–147, 2019.

[86] G. A. Kaissis, M. R. Makowski, D. Rückert, R. F. Braren, “Secure, privacy-preserving and federated machine learning in medical imaging,” *Nature Machine Intelligence*, vol. 2, no. 6, pp. 305–311, 2020.

[87] H. Mohan, *Textbook of pathology*. Jaypee Brothers, Medical Publishers Pvt. Limited, 2018.

[88] J. Van Hulse, T. M. Khoshgoftaar, A. Napolitano, “Experimental perspectives on learning from imbalanced data,” in *Proceedings of the 24th international conference on Machine learning*, 2007, pp. 935–942.

[89] L. Chato, S. Latifi, “Application of machine learning to biometric systems-a survey,” in *Journal of Physics: Conference Series*, IOP Publishing, vol. 1098, 2018, p. 012 017.

[90] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, “Deepface: Closing the gap to human-level performance in face verification,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1701–1708.

[91] F. Schroff, D. Kalenichenko, J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[92] N. Liu, M. Zhang, H. Li, Z. Sun, T. Tan, “Deepiris: Learning pairwise filter bank for heterogeneous iris verification,” *Pattern Recognition Letters*, vol. 82, pp. 154–161, 2016.

[93] A. Boles, P. Rad, “Voice biometrics: Deep learning-based voiceprint authentication system,” in *2017 12th System of Systems Engineering Conference (SoSE)*, IEEE, 2017, pp. 1–6.

[94] T. Carvalho, E. R. De Rezende, M. T. Alves, F. K. Balieiro, R. B. Sovat, “Exposing computer generated images by eye’s region classification via transfer learning of vgg19 cnn,” in *2017 16th IEEE international conference on machine learning and applications (ICMLA)*, IEEE, 2017, pp. 866–870.

[95] M. Y. Kamil, “A deep learning framework to detect covid-19 disease via chest x-ray and ct scan images.,” *International Journal of Electrical & Computer Engineering (2088-8708)*, vol. 11, no. 1, 2021.

[96] V. Nair, G. E. Hinton, “Rectified linear units improve restricted boltzmann machines,” in *Icml*, 2010.

[97] A. B. López, “Deep learning in biometrics: A survey,” *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*, vol. 8, no. 4, pp. 19–32, 2019.

[98] S. Chopra, R. Hadsell, Y. LeCun, “Learning a similarity metric discriminatively, with application to face verification,” in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, IEEE, vol. 1, 2005, pp. 539–546.

[99] R. Hadsell, S. Chopra, Y. LeCun, “Dimensionality reduction by learning an invariant mapping,” in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06)*, IEEE, vol. 2, 2006, pp. 1735–1742.

[100] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, P. H. Torr, “Fully-convolutional siamese networks for object tracking,” in *European conference on computer vision*, Springer, 2016, pp. 850–865.

[101] S. Sun, N. Akhtar, H. Song, A. Mian, M. Shah, “Deep affinity network for multiple object tracking,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 1, pp. 104–119, 2019.

[102] D. Guo, J. Wang, Y. Cui, Z. Wang, S. Chen, “Siamcar: Siamese fully convolutional classification and regression for visual tracking,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 6269–6277.

[103] P. Voigtlaender, J. Luiten, P. H. Torr, B. Leibe, “Siam r-cnn: Visual tracking by re-detection,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 6578–6588.

[104] Z. Chen, B. Zhong, G. Li, S. Zhang, R. Ji, “Siamese box adaptive network for visual tracking,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 6668–6677.

[105] K. Fu, D.-P. Fan, G.-P. Ji, Q. Zhao, J. Shen, C. Zhu, “Siamese network for rgb-d salient object detection and beyond,” *IEEE transactions on pattern analysis and machine intelligence*, 2021.

[106] M. Tan, Q. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in *International conference on machine learning*, PMLR, 2019, pp. 6105–6114.

[107] R. R. Sanni, H. Guruprasad, “Analysis of performance metrics of heart failured patients using python and machine learning algorithms,” *Global transitions proceedings*, vol. 2, no. 2, pp. 233–237, 2021.

[108] Z. Shen, T. Zhong, “Analysis of application examples of differential privacy in deep learning,” *Computational Intelligence and Neuroscience*, vol. 2021, 2021.

[109] H. B. McMahan, D. Ramage, K. Talwar, L. Zhang, “Learning differentially private recurrent language models,” *arXiv preprint arXiv:1710.06963*, 2017.

[110] L. Xie, K. Lin, S. Wang, F. Wang, J. Zhou, “Differentially private generative adversarial network,” *arXiv preprint arXiv:1802.06739*, 2018.

[111] M. A. Rahman, T. Rahman, R. Laganière, N. Mohammed, Y. Wang, “Membership inference attack against differentially private deep learning model.,” *Trans. Data Priv.*, vol. 11, no. 1, pp. 61–79, 2018.

[112] X. Chen, S. Z. Wu, M. Hong, “Understanding gradient clipping in private sgd: A geometric perspective,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 13 773–13 782, 2020.

[113] F. McSherry, K. Talwar, “Mechanism design via differential privacy,” in *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS’07)*, IEEE, 2007, pp. 94–103.

[114] Z. Müftüoğlu, M. A. Kizrak, T. Yildirim, “Differential privacy practice on diagnosis of covid-19 radiology imaging using efficientnet,” in *2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, IEEE, 2020, pp. 1–6.

[115] A. Ziller *et al.*, “Pysyft: A library for easy federated learning,” in *Federated Learning Systems*, Springer, 2021, pp. 111–139.

[116] A. Dandekar, D. Basu, S. Bressan, “Differential privacy at risk: Bridging randomness and privacy budget,” *arXiv preprint arXiv:2003.00973*, 2020.

[117] L. Chen *et al.*, “Robustness, security and privacy in location-based services for future iot: A survey,” *IEEE Access*, vol. 5, pp. 8956–8977, 2017.

[118] W. Meng, E. W. Tischhauser, Q. Wang, Y. Wang, J. Han, “When intrusion detection meets blockchain technology: A review,” *Ieee Access*, vol. 6, pp. 10 179–10 188, 2018.

[119] T. Wang, Z. Zheng, M. H. Rehmani, S. Yao, Z. Huo, “Privacy preservation in big data from the communication perspective—a survey,” *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 753–778, 2018.

PUBLICATIONS FROM THE THESIS

Papers

1. Müftüoğlu, Z., Yildirim, T. (2019). Comparative Analysis of Crypto Systems Using Biometric Key. Procedia Computer Science, 154, 327-331.

Conference Papers

1. Müftüoğlu, Z., Kizrak, M. A., Yildirim, T. (2020, August). Differential privacy practice on diagnosis of covid-19 radiology imaging using efficientnet. In 2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA) (pp. 1-6). IEEE.

Books

1. Müftüoğlu, Z., Kızrak, M. A., Yıldırım, T. (2022). Data sharing and privacy issues arising with COVID-19 data and applications. In Data Science for COVID-19 (pp. 61-75). Academic Press.

2. Müftüoğlu, Z., Kızrak, M. A., Yıldırım, T. (2022). Privacy-Preserving Mechanisms with Explainability in Assistive AI Technologies. In Advances in Assistive Technologies (pp. 287-309). Springer, Cham.

3. Kızrak, M. A., Müftüoğlu, Z., Yıldırım, T. (2021). Limitations and challenges on the diagnosis of COVID-19 using radiology images and deep learning. In Data Science for COVID-19 (pp. 91-115). Academic Press.