UNIVERSITÀ degli STUDI di CATANIA

DIPARTIMENTO DI INGEGNERIA ELETTRICA, ELETTRONICA E INFORMATICA

PH.D. PROGRAM IN SYSTEMS, ENERGETICS, COMPUTER AND TELECOMMUNICATIONS ENGINEERING
XXXVI CYCLE

Ph.D. Thesis

FROM CENTRALIZATION TO COLLABORATION: HARNESSING GENERATIVE MODELS IN FEDERATED LEARNING FOR MEDICAL IMAGE ANALYSIS

FEDERICA PROIETTO SALANITRI

PhD Coordinator
Prof. P. ARENA

Supervisor
Prof.ssa D. GIORDANO
Co-Supervisor
Prof. C. SPAMPINATO
Prof. U. BAGCI

To Giovanni, my safe harbor in life and work.
For enjoying every milestone with me and even more
for sharing and overcoming every challenge with me.
For always pushing me towards improvement, achievement and happiness.
To my family, my pillar.
For the unconditional love, strength and support
you have given me over the years.
To Concetto, an invaluable leader and constant supporter.
For being a precious guide,
for the enthusiasm with which you drive us to grow professionally,
for believing in me from the beginning.
To Simone, a point of reference.
For your expertise and your contribution to each of our works.
To the PeRCeiVe Lab team, companions in successes and challenges.
For the coffee breaks that lightened the difficult moments
and for those that were the turning point of our works.
These years would not have been the same without you.

From my heart
Thank you

ABSTRACT

The advent of Artificial Intelligence (AI) in healthcare has marked a new era of medical diagnostics and treatment. Particularly in the field of medical imaging, Deep Learning (DL), a subset of AI, has demonstrated unprecedented success. Complex neural network architectures have been developed, capable of detecting, classifying, and segmenting diseases from medical images with remarkable accuracy, often rivaling or surpassing human experts. However, the effectiveness of deep learning is contingent upon access to vast, diverse, and high-quality datasets, the acquisition of which is often hindered by privacy concerns, data sharing restrictions, and the inherent variability in medical data across different institutions.

Federated Learning (FL), an innovative machine learning paradigm, offers a compelling solution to these challenges. FL enables the training of AI models across multiple decentralized devices or servers holding local data samples, without the need to exchange the data itself. This approach not only preserves data privacy but also allows for the utilization of diverse datasets from different institutions, thereby enhancing the robustness and generalizability of the

AI models.

In addition to the core principles of FL, the use of generative models, such as Generative Adversarial Networks (GANs), has emerged as a powerful tool for maintaining privacy in the FL scenario. These models can generate synthetic data that mimic the statistical properties of the original data, allowing for the training of robust models without exposing sensitive patient information.

This thesis explores the transition from the classic *centralized* deep learning approach to federated learning in medical imaging, with a specific focus on the use of generative models for privacy preservation. We delve into the technical aspects of implementing FL and generative models, discuss the challenges and potential solutions, and present case studies where these methods have been successfully applied.

By investigating the shift from deep learning to federated learning and the role of generative models, this research aims to contribute to the ongoing efforts to integrate AI into healthcare more effectively, responsibly, and inclusively. The advent of federated learning and generative models marks a new era in medical imaging analysis, paving the way for a future of effective, collaborative and privacy-preserving healthcare.

CONTENTS

Part I

INTRODUCTION

At the beginning of this journey, it is essential to establish a clear context for upcoming exploration. The initial chapters are critical in setting this context. Chapter 1 reveals the inspirations behind our investigation and outlines the specific goals we seek to achieve. Next, Chapter 2 presents the key concepts and state-of-the-art approaches that have shaped the field. This chapter not only provides a comprehensive background, but also highlights the key works and methodologies that will be used in our thesis. Together, these sections ensure a comprehensive introduction to the research narration and the central role of our contributions.

CHAPTER

ONE

MOTIVATIONS AND OBJECTIVES

In clinical practice, huge volumes of data are routinely produced, spanning a myriad of formats such as images (MRI, CT scans, sonography, radiography), time series data (like EEG, EMG, ECG), audio (speech, Doppler sounds), and text documents (structured and unstructured reports, annotations, and comments). Comprehensive digitalization of medical data has paved the way for extensive processing and training of artificial intelligence (AI) systems, aiming to assist in various tasks within clinical practice. However, despite AI's inception in the 1950s, it's only in the past two decades that digital medical data has become widespread.

The "AI-winter" from the 1960s to the 1980s, marked by a lack of data and computational power, saw AI facing widespread criticism, leading to a decline in its research and development. In particular, the medical community was reluctant to adopt these new technological methods. The reliance on technology for patient care and healthcare

optimization was viewed with skepticism.

However, the landscape has evolved radically. With advancements in data processing capabilities, deep learning has found its way in clinical medicine: it is now pivotal in generating diagnoses through medical imaging and signal analysis, planning treatments via chatbots and questionnaires, and even assisting in surgeries with AI-guided robotic systems.

As we delve into the core of this thesis, the initial focus is on traditional centralized approaches in medical imaging. We specifically explore and propose solutions for COVID-19 assessment from CT scans, pancreas segmentation, and pancreatic cyst classification. These centralized methodologies, while effective, often grapple with issues related to generalization, scalability and data privacy.

The dissertation then shifts to Federated Learning, a paradigm that attempts to address the limitations of centralized systems. Here, the emphasis is on the innovative use of generative models that offer data-driven approaches for federated learning. We finally conclude this dissertation with the introduction of a new methodology: by combining experience replay from continual learning with federated learning principles and by enforcing privacy-preserving capabilities to GAN, we present a distributed federated learning approach that achieves remarkable results in real-world medical scenarios.

The motivation behind this exploration is to harness the transformative potential of AI in addressing the pressing challenges of medical imaging, offering solutions that are not only effective but also prioritize the sanctity of patient data.

The advantages of integrating AI into medicine are multiple: indeed, AI can play a pivotal role in helping clinicians, especially those

in the early stages of their careers, and by providing decision support, reducing the likelihood of oversight and enhancing overall quality of care.

As we navigate through the chapters, this journey will not only highlight the challenges faced in this domain but also present groundbreaking solutions that leverage the myriad benefits of AI in enhancing medical practice.

CHAPTER

TWO

BACKGROUND

In this chapter, we will provide an introduction to the key concepts and state-of-the-art approaches that form the foundation of this thesis. Each concept introduced here will be explored in greater depth in the subsequent chapters, ensuring that the reader has a comprehensive understanding of the specifics and nuances of our work. This foundational knowledge will serve as a roadmap, guiding the reader through the intricate details and methodologies presented in each chapter.

2.1 Machine Learning in Healthcare

Machine learning, specifically deep learning, has been instrumental in the evolution of medical imaging analysis. The application of these techniques has significantly improved the detection, classification, and segmentation of diseases from medical images [162].

Segmentation and classification stand as pivotal processes in the

medical imaging analysis domain. In segmentation, each pixel within an image is labeled. Pixels belonging to the same type of object receive identical labels, as seen in *semantic segmentation* [118]. Conversely, in *instance segmentation* [55], distinct objects of the same category are identified as separate entities. The primary objective of image segmentation through deep learning is to equip machines with a more human-centric perception and understanding of images. Classification, in contrast, is about labeling an entire image or a specific segment based on its content. Numerous deep learning architectures, notably Convolutional Neural Networks (CNNs) [92], have been extensively employed for both these segmentation and classification tasks [2, 111].

CNN-based models have been crucial in the field of medical imaging research. However, a new type of architecture, known as the Transformer, has recently made a significant impact. Transformer models, introduced by [170], have shown exceptional performance across various artificial intelligence domains, including natural language processing [169], audio processing [42], and more recently, computer vision tasks [85]. Transformers are capable of effectively learning arbitrary functions. They are composed of two primary operational blocks: an attention-based block that models relationships between elements, and a multi-layer perceptron (MLP) that models relationships within elements. A sequence of attention and MLP blocks, combined with residual connections, has proven to generalize well across multiple tasks.

Transformer-based architectures offer several key advantages in medical imaging analysis. They can capture long-range dependencies, provide an inherent method for explaining model decisions, and can achieve comparable performance to traditional deep learning with simpler models.

Vision Transformers (ViTs) [43] have become increasingly popular. Following this, numerous ViT-based approaches have been developed and continuously adapted for medical practice. Many of these approaches, such as TransUNet [28] and TransFuse [191], focus on image segmentation by integrating transformer modules with traditional deep learning ones. Others, like ViT-V-Net [29], explore the use of ViTs for volumetric image registration using a hybrid ConvNet-Transformer architecture. Some models use transformers for detection [44] or classification [36] tasks.

Despite the significant advancements in these areas, several challenges persist in medical imaging analysis. One of the primary issues is the lack of large, diverse, and high-quality datasets. This problem is further exacerbated by distribution shifts, where the data distribution changes between the training and testing phases, leading to a decrease in model performance.

Another challenge is the evaluation of model performance; several metrics are commonly used to evaluate the performance of segmentation and classification models. One of the most straightforward metrics is *accuracy*, which provides a general measure of how often the model's predictions align with the actual outcomes.

However, in many scenarios, like the medical one, the stakes are high, accuracy alone might not provide a comprehensive view of the model's performance and the consequences of incorrect predictions can be significant. Therefore, certain metrics become especially important to ensure that models are both sensitive to actual cases and specific in their predictions, especially when dealing with imbalanced datasets. Some metrics particularly crucial in medical contexts are the following:

- **Sensitivity** (True Positive Rate): measures the proportion of actual positive cases that the model correctly identifies. In medical scenarios, this metric is vital because it indicates how well the model detects true cases of a disease or condition. High sensitivity is crucial for conditions where missing a positive case (false negative) could have severe consequences, such as failing to diagnose a malignant tumor.

- **Precision**: evaluates how many of the positive identifications made by the model are actually correct. In medical tests, high precision ensures that patients aren't falsely diagnosed with a condition they don't have, which could lead to unnecessary stress, further testing, or even unwarranted treatment.

- Area Under the Receiver Operating Characteristic Curve (**AUC-ROC**): Given that medical datasets can sometimes be imbalanced (e.g., few positive cases of a rare disease), the AUC-ROC becomes essential. It provides a comprehensive view of the model's performance across all possible classification thresholds, ensuring that the model can distinguish between positive and negative cases effectively. An AUC of 1 indicates perfect classification, while an AUC of 0.5 suggests that the model's performance is no better than random guessing.

- **F-measure** (F1-score): The F1-score is the harmonic mean of precision and recall (sensitivity). In medical contexts, where both false positives and false negatives can have significant implications, the F1-score provides a balanced metric that considers both types of errors.

While these metrics are crucial, it's essential to consider the specific medical scenario when evaluating a model. For instance: in a preliminary screening test for a severe disease, a high sensitivity might be prioritized to ensure no positive cases are missed, even if it means some false positives that can be ruled out with further testing. In a confirmatory diagnostic test, precision might be more critical to ensure that only true positive cases are identified to avoid unnecessary treatments. In all cases, the choice of metrics should align with the clinical objectives and the potential consequences of model errors.

In summary, while building and training models is a complex task, evaluating their performance is equally crucial. Leveraging a combination of these metrics ensures a more holistic understanding of a model's strengths and areas of improvement.

In the following chapters, we will delve deeper into the techniques used for segmentation and classification in medical imaging analysis, discuss the challenges in more detail, and explore potential solutions.

2.2 Interpretabilty in Deep Models

The complexity of deep networks often makes them difficult to understand, which hinders their broader use in critical decision-making fields like healthcare. It's crucial to make these models more transparent so users can grasp the reasoning behind their decisions. Explainable AI (XAI) offers insights into how these models work. In fact, regulations like the European Union's General Data Protection Regulation (GDPR, Article 15) emphasize the importance of understanding how medical decisions are derived.

Many AI experts are turning to XAI to demystify their algorithms. XAI techniques can be grouped based on whether they're built into the model or applied afterward, if they're specific to a certain model or can be used universally, and if they provide a broad or narrow explanation.

Methods built directly into the model, i.e. *model-based* methods, aim to make the model inherently explainable. But, these aren't suitable for large-scale models with numerous parameters. On the other hand, *posthoc* techniques first train a model and then analyze its features to explain its behavior. For instance, they might visually highlight which parts of an input image were most influential in a decision.

Some XAI methods, focus solely on the relationship between a model's input and output, without needing details about the model itself; these are also considered post-hoc techniques. Others, instead, are designed specifically for certain types of models, and have limited versatility.

These techniques can be used to explain a model's reasoning either broadly or in specific cases. For instance, in medicine, they might pinpoint which parts of an MRI were crucial in identifying a tumor.

In medical imaging, XAI has been used to offer visual or written explanations for a model's decisions. Visual methods, like Grad-CAM [152], highlight areas of an image that influenced a decision, while textual explanations, like image captioning, describe what happens in an image. Some approaches even combine both visual and textual explanations [94].

However, not all XAI methods are universally lauded. Some post-hoc techniques have been criticized for not offering deep enough insights into the inner workings of models [147]. Some experts recom-

mend built-in explainability methods over post-hoc ones.

Transformers, naturally offer more transparency due to their attention mechanisms, which can be visualized to elucidate the model's reasoning process [3, 23]. By seamlessly blending performance with explainability, Transformers obviate the need for major modifications, effectively bridging the advantages of both built-in and post-hoc XAI approaches.

2.3 Federated Learning

Federated Learning (FL) [113] is a groundbreaking machine learning paradigm that enables multiple entities to collaboratively develop a robust machine learning model without sharing raw data. This approach addresses crucial issues such as data privacy, data security, data access privileges, and access to heterogeneous data.

FL can be broadly categorized into two types: centralized and decentralized. In centralized FL, a central server is responsible for aggregating model updates from all participating devices and distributing the updated global model back to the devices [98]. In contrast, decentralized FL allows devices to exchange model updates directly with each other without relying on a central server.

One of the most popular algorithms used in FL is Federated Averaging (FedAvg) [113], which averages the model updates from all devices to update the global model. However, FedAvg and similar algorithms can face challenges when dealing with non-independently and identically distributed (non-i.i.d.) data, where the data distribution varies across devices. Several strategies have been proposed

to mitigate the negative effects caused by non-i.i.d. datasets. For instance, FedProx [98] addresses this limitation by adding a penalty term to the loss, driving the local models to a shared optimum. On the other hand, FedCurv [157] incorporates curvature information to improve the aggregation of local model updates. FedMA [171] builds a shared global model in a layer-wise manner by matching and averaging hidden elements with similar feature extraction signatures. Meanwhile, FedBN [100] keeps batch normalization layers private, while other model parameters are aggregated by the central node.

FL has found applications in various domains, including healthcare [183, 9], Internet of Things (IoT) [84], and edge systems [173].

Despite its advantages, FL also presents several challenges. One of the primary challenges is system efficiency, which pertains to the computational and communication overheads associated with training models across distributed nodes [98]. Data heterogeneity and statistical heterogeneity arise due to the diverse nature of data sources and the varying data distributions across different nodes, respectively. System heterogeneity, on the other hand, refers to the differences in computational capabilities and resources of participating devices or nodes. Data imbalance is another challenge where some nodes might have abundant data while others have scarce data. Resource allocation pertains to the optimal distribution of computational resources for efficient training. Lastly, privacy concerns revolve around ensuring that the shared model updates do not leak sensitive information. To address these challenges, various solutions have been proposed. Blockchain technology, for instance, has been introduced to secure data sharing and enhance the training process in FL [107]. For handling data heterogeneity and ensuring privacy, methods based

on secret sharing and generative adversarial networks have been suggested [196]. Additionally, to tackle the challenges of data imbalance and resource allocation, advanced optimization techniques and adaptive algorithms have been developed [41].

In terms of communication efficiency, FL algorithms have been evaluated and compared, highlighting the need for solutions that address both communication and privacy perspectives. Furthermore, new efficient FL algorithms have been proposed to optimize client and cost selections, addressing the major problems of communication, system heterogeneity, and data heterogeneity faced by FL [108, 104].

Federated learning can be implemented using either a centralized or decentralized approach. In the centralized approach [113, 98, 100], a central server coordinates the learning process, receiving model updates from each device and sending back the updated global model. This approach is simpler and easier to manage, but it can be a single point of failure and may not scale well to large networks [101].

In contrast, the decentralized approach [90, 164] involves peer-to-peer communication between devices, without the need for a central server. This approach can be more resilient and scalable, but it also requires more sophisticated coordination mechanisms to ensure consistent learning.

2.4 Federated Learning in Healthcare

Federated learning (FL) has emerged as a transformative approach in the domain of healthcare, addressing the intricate challenges associated with the management and utilization of medical data. The

essence of FL lies in its commitment to safeguarding patient privacy. Unlike traditional centralized learning methods, FL ensures that sensitive health data remains localized, never leaving its original location. This decentralized approach is paramount in the healthcare sector, where data spans a spectrum from electronic health records and diagnostic images to real-time measurements from wearable devices [16].

One of the standout features of healthcare data is its inherent heterogeneity. Given the vast diversity in patient demographics, health conditions, and data collection methodologies, healthcare often grapples with non-i.i.d. scenarios. Here, data distribution can vary dramatically across different devices or institutions. FL's architecture is uniquely suited to navigate this heterogeneity, fostering models that are not only robust but also exhibit enhanced generalizability across a myriad of patient profiles and conditions [154].

In the fields of medical imaging, for example, Sheller et al. [154] utilized FL for brain tumor segmentation in MRI scans across multiple institutions, highlighting the value of diverse data for model generalization. Similarly, Li et al. [98] demonstrated the efficacy of FL in chest X-ray classification, achieving top-tier performance across decentralized datasets. Beyond imaging, Brisimi et al [16]. showcased FL's potential in predicting patient hospitalizations using electronic health records, achieving enhanced accuracy without direct data sharing.

However, the integration of FL into healthcare is not without its challenges. The technical intricacies of FL, combined with the diverse nature of healthcare data and the ever-present need for stringent privacy measures, necessitate continuous research and innovation. Addressing these challenges requires a multidisciplinary approach, combining expertise from medical professionals, data scientists, and pri-

vacy experts. As solutions to these challenges evolve, FL stands poised to redefine healthcare, offering a harmonious blend of data-driven insights and uncompromised patient privacy. Its potential to revolutionize healthcare practices, from diagnosis to treatment, positions FL as a cornerstone of future medical advancements.

2.5 Privacy in Federated Learning

The significance of privacy in FL is crucial, especially in sectors like healthcare or finance where data sensitivity is of the highest importance [97, 149].

To ensure that privacy is maintained, various techniques and protocols are employed. Differential privacy [1, 56, 93], for instance, introduces noise to the data or results, making it challenging to reverse-engineer individual data points. Homomorphic encryption is another technique that allows computations on encrypted data without the need for decryption, ensuring data remains secure during processing [188]. Secure aggregation protocols are also in place to ensure that data shared during the model aggregation phase remains private [47, 75].

However, ensuring privacy is not without its challenges. Some studies have highlighted potential vulnerabilities in FL [121], emphasizing the need for continuous research and improvement in privacy-preserving mechanisms. One of the emerging concerns is the ability of adversaries to reconstruct the original input data from shared model weights or gradients [51, 197]. Such reconstruction attacks can compromise the very essence of privacy that FL aims to uphold.

In conclusion, as the digital landscape becomes increasingly data-driven, the emphasis on privacy in frameworks like FL underscores the evolving priorities in the fields of artificial intelligence and machine learning.

Part II

CENTRALIZED DEEP LEARNING METHODS FOR MEDICAL IMAGE ANALYSIS

In the intricate scenario of medical diagnostics, the fusion of artificial intelligence (AI) with medical imaging has emerged as a source of innovation, heralding a new era of enhanced precision and patient-centric care. Central to this transformative journey are the two pillars of classification and segmentation, each playing a pivotal role in shaping the future of medical interventions.

Classification, beyond its foundational role of categorizing medical images into diagnostic groups, has profound implications for patients' perspectives. By accurately identifying and categorizing pathologies, from the early stages of diseases like cancer to the nuanced variations of conditions like pneumonia, classification ensures that patients receive timely and appropriate care. In an era where early detection can significantly alter disease outcomes, the importance of precise classification cannot be overstated. It's not merely about categorizing images;

it's about outlining the way of patient care, ensuring interventions are timely, relevant, and effective.

Segmentation, while technical in its delineation of regions of interest, is crucial in its applications. Accurate segmentation forms the backbone of many therapeutic procedures. In radiation therapy, for instance, precise segmentation ensures that radiation is delivered solely to the tumor, sparing healthy tissue and mitigating side effects. In surgical planning, understanding the exact boundaries and relationships of anatomical structures can mean the difference between a successful surgery and post-operative complications. Through segmentation, we're not just viewing images; we're visualizing the roadmap for interventions, ensuring they're both precise and safe.

In this dynamic landscape, the following works represent efforts to harness the potential of AI-driven medical image analysis techniques. Each proposed work faces unique challenges as it seeks to provide innovative solutions.

The Chapter 3 delves into the pressing global challenge of the COVID-19 pandemic. In the face of the COVID-19 pandemic, rapid and accurate diagnosis became the pivot of effective patient care. This research not only introduces a state-of-the-art AI-powered pipeline for detecting COVID-19 from CT scans but emphasizes the crucial aspect of explainability. By bridging the gap between AI predictions and clinical understanding, it ensures that radiologists are not just passive recipients of AI decisions but active participants in a collaborative diagnostic process. Chapter 4 ventures into the intricate realm of pancreatic pathologies. IPMN, a precursor to one of the most lethal cancers, requires nuanced and precise identification. This work showcases the potential of neural transformers in medical diagnostics, high-

lighting their capability to capture complex patterns and offer accurate classifications, which can significantly influence therapeutic decisions.

Lastly, Chapter 5 delves into the challenges of accurate organ segmentation, a foundational step in many diagnostic and therapeutic procedures. By focusing on the hierarchical learning of 3D features, this work emphasizes the significance of understanding spatial relationships in three-dimensional medical images, leading to enhanced precision in pancreas segmentation.

Together, these works not only provide a comprehensive exploration of the current state and future potential of centralized deep learning methods in medical image analysis but also highlight the continuous evolution of methodologies, ensuring that the medical community is equipped with the best tools for patient care.

CHAPTER

THREE

AN EXPLAINABLE AI SYSTEM FOR AUTOMATED COVID-19 ASSESSMENT AND LESION CATEGORIZATION FROM CT-SCANS

We begin our journey with an emergency that has impacted the entire global population: the COVID-19 pandemic. Recognizing the critical need for efficient diagnosis, we venture into addressing two fundamental tasks in medical image analysis: lung segmentation and the subsequent categorization of lesions within it.We will delve into the intricacies of these challenges and present a comprehensive pipeline that excels at initially extracting the lung parenchyma and lobes, enabling the accurate detection and categorization of COVID-19 lesions within CT scans.

3.1 Motivation

At the end of 2019 in Wuhan (China) several cases of an atypical pneumonia, particularly resistant to the traditional pharmacological treatments, were observed. In early 2020, the COVID-19 virus [198] has been identified as the responsible pathogen for the unusual pneumonia. From that time, COVID-19 has spread all around the world hitting, to date about 155 million of people (with about 3.5M deaths), stressing significantly healthcare systems in several countries. Since the beginning, it has been noted that 20% of infected subjects appear to progress to severe disease, including pneumonia and respiratory failure and in around 2% of cases death [129]. Currently, the standard diagnosis of COVID-19 is de facto based on a biomolecular test through Real-Time Polimerase Chain Reaction (RT-PCR) test [67, 126]. However, although widely used, this biomolecular method is time-consuming requiring up to several hours for being processed.

Recent studies have outlined the effectiveness of radiology imaging through chest X-ray and mainly Computed Tomography (CT) given the pulmonary involvement in subjects affected by the infection [105, 30]. Given the extension of the infection and the number of cases that daily emerge worldwide and that call for fast, robust and medically sustainable diagnosis, CT scan appears to be suitable for a robust-scale screening, given the higher resolution w.r.t. X-Ray. In this scenario, artificial intelligence may play a fundamental role to make the whole diagnosis process automatic, reducing, at the same time, the efforts required by radiologists for visual inspection [148].

In this paper, thus, we present an AI-based system to achieve both *COVID-19 identification* and *lesion categorization* (ground glass,

crazy paving and consolidation) that are instrumental to evaluate lung damages and the prognosis assessment. Our method relies only on radiological image data avoiding the use of additional clinical data in order to create AI models that are useful for large-scale and fast screening with all the subsequent benefits for a favorable outcome. More specifically, we propose an innovative automated pipeline consisting of 1) lung/lobe segmentation, 2) COVID-19 identification and interpretation and 3) lesion categorization. We tested the AI-empowered software pipeline on multiple CT scans, both publicly released and collected at the Spallanzani Institute in Italy, and showed that: 1) our segmentation networks is able to effectively extract lung parenchyma and lobes from CT scans, outperforming state of the art models; 2) the COVID-19 identification module yields better accuracy (as well as specificity and sensitivity) than expert radiologists. Furthermore, when attempting to interpret the decisions made by the proposed AI model, we found that it learned automatically, and without any supervision, the CT scan features corresponding to the three most common lesions spotted in the COVID-19 pneumonia, i.e., consolidation, ground glass and crazy paving, demonstrating its reliability in supporting the diagnosis by using only radiological images. Finally, we integrate the tested AI models into a user-friendly GUI to support further AI explainability for radiologists. The GUI processes entire CT scans and reports if the patient is likely to be affected by COVID-19, showing, at the same time, the scan slices that supported the decision.

To sum up, the main contributions of the work are the following:

- We propose a novel lung-lobe segmentation network outperforming state-of-the-art models;

- We employ the segmentation network to drive a classification network that first identifies CT scans of COVID-19 patients, and, afterwards, automatically categorizes specific lesions;

- We then provide interpretation of the decisions made by the employed models and discover that, indeed, the proposed approach focuses on specific COVID-19 lesions for distinguishing whether a CT scan is related to positive patients or not;

- We finally integrate the whole AI pipeline into a web platform to ease use for radiologists, supporting them in their investigation on COVID-19 disease. To the best of our knowledge, this is the first publicly available platform that offers COVID-19 diagnosis services based on CT scans with explainability capabilities. The free availability to the general public for such an important task, while the pandemic is still in full effect, is, in our opinion, an invaluable aid to the medical community.

3.2 Related Work

The COVID-19 epidemic caught the scientific community unprepared and in response a high volume of research has been dedicated to the related compelling issues to address, at all possible levels. In particular, since the beginning of the epidemic, AI models have been employed for disease spread monitoring [8, 103, 193], for disease progression [13] and prognosis [102], for predicting mental health ailments inflicted upon healthcare workers [35] and for drug repurposing [119, 82] and discovery [139].

However, the lion's share in employing AI models for the fight against COVID-19 belongs to the processing of X-rays and CT scans with the purpose of detecting the presence of COVID-19 or not. In fact, recent scientific literature has demonstrated the high discriminative and predictive capability of deep learning methods in the analysis of COVID-19 related radiological images[18, 66]. The key radiological techniques for COVID-19 induced pneumonia diagnosis and progression estimation are based on the analysis of CT and X-ray images of the chest, on which deep learning methodologies have been widely used with good results for segmentation, predictive analysis, and discrimination of patterns [122, 124, 115]. If, on one hand, X-Ray represents a cheaper and most effective solution for large scale screening of COVID-19 disease, on the other hand, its low resolution has led AI models to achieve lower accuracy compared to those obtained with CT data.

For the above reasons, CT scan has become the gold standard for investigation on lung diseases. In particular, deep learning, mainly in the form of Deep Convolutional Neural Networks (DCNN), has been largely applied to lung disease analysis from CT scans images, for evaluating progression in response to specific treatment (for instance immunotherapy, chemotherapy, radiotherapy) [153, 26], but also for interstitial lung pattern analysis [14, 48] and on segmentation and discrimination of lung pleural tissues and lymph-nodes [120, 156]. This latter aspect is particularly relevant for COVID-19 features and makes artificial intelligence an extremely powerful tool for supporting early diagnosis of COVID-19 and disease progression quantification. As a consequence, several recent works have reported using AI models for automated categorization of CT scans [115] and also on COVID-19 [96, 155, 12] but without being able to distinguish between the

various types of COVID-19 lesions.

3.3 Explainable AI for COVID-19 data understanding

The proposed AI system aims at 1) extracting lung and lobes from chest CT data, 2) categorizing CT scans as either COVID-19 positive or COVID-19 negative; 3) identifying and localizing typical COVID-19 lung lesions (consolidation, crazy paving and ground glass); and 4) explaining eventually what CT slices it based its own decisions.

3.3.1 AI Model for Lung Segmentation

Our lung-lobe segmentation model is based on the *Tiramisu* network [72], a fully-convolutional DenseNet [65] in a U-Net architecture [142]. The model consists in two data paths: the downsampling one, that aims at extracting features and the upsampling one that aims at generating the output images (masks). Skip connections (i.e., connections starting from a preceding layer in the network's pipeline to another one found later bypassing intermediate layers) aim at propagating high-resolution details by sharing feature maps between the two paths.

In this work, our segmentation model follows the Tiramisu architecture, but with two main differences:

- Instead of processing each single scan individually, convolutional LSTMs [181] are employed at the network's bottleneck layer to exploit the spatial axial correlation of consecutive scan slices.

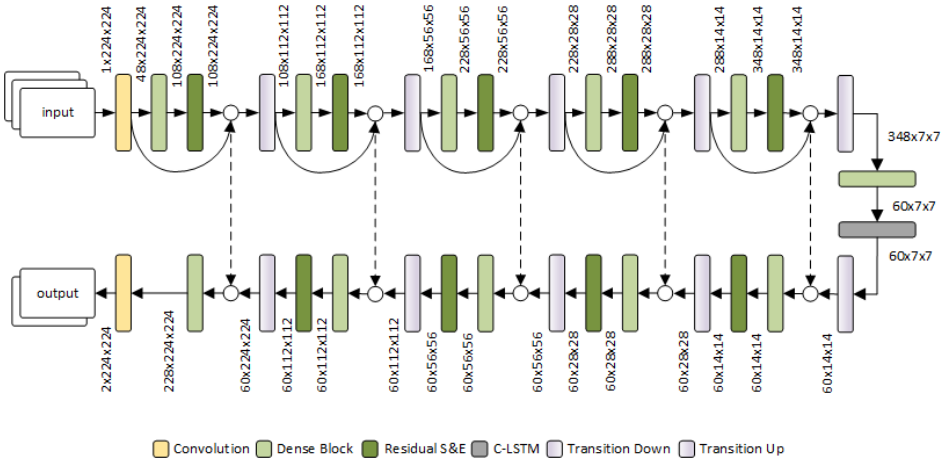


Figure 3.1: *The proposed segmentation architecture, consisting of a downsampling path (top) and an upsampling path (bottom), interconnected by skip connections and by the bottleneck layer.*

- In the downsampling and upsampling paths, we add residual squeeze-and-excitation layers [64], in order to emphasize relevant features and improve the representational power of the model.

Before discussing the properties and advantages of the above modifications, we first introduce the overall architecture, shown in Fig. 3.1.

The input to the model is a sequence of 3 consecutive slices – suitably resized to 224×224 – of a CT scan, which are processed individually and combined through a convolutional LSTM layer. Each slice is initially processed with a standard convolutional layer to expand the feature dimensions. The resulting feature maps then go through the downsampling path of the model (the encoder) consisting of five sequences of dense blocks, residual squeeze-and-excitation layers and

transition-down layers based on max-pooling. In the encoder, the feature maps at the output of each residual squeeze-and-excitation layer are concatenated with the input features of the preceding dense block, in order to encourage feature reuse and improve their generalizability. At the end of the downsampling path, the *bottleneck* of the model consists of a dense block followed by a convolutional LSTM. The following upsampling path is symmetric to the downsampling one, but it features: 1) skip connections from the downsampling path for concatenating feature maps at the corresponding layers of the upsampling path; 2) transition-up layers implemented through transposed convolutions. Finally, a convolutional layer provides a 6-channel segmentation map, representing, respectively, the log-likelihoods of the lobes (5 channels, one for each lobe) and non-lung (1 channel) pixels.

In the following, we review the novel characteristics of the proposed architecture.

Residual squeeze-and-excitation layers. Explicitly modeling interdependencies between feature channels has demonstrated to enhance performance of deep architectures; squeeze-and-excitation layers [64] instead aim to select informative features and to suppress the less useful ones. In particular, a set of input features of size $C \times H \times W$ is squeezed through average-pooling to a $C \times 1 \times 1$ vector, representing global feature statistics. The "excitation" operator is a fully-connected non-linear layer that translates the squeezed vector into channel-specific weights that are applied to the corresponding input feature maps.

Convolutional LSTM. We adopt a recurrent architecture to pro-

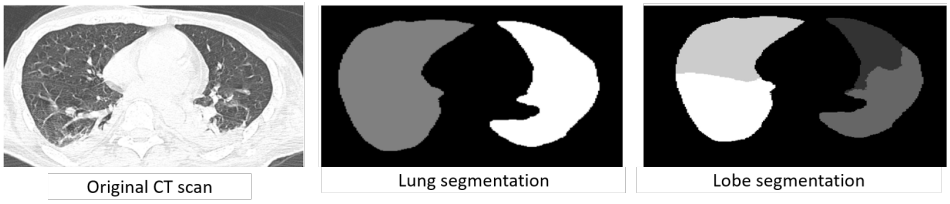


Figure 3.2: *Example of lung and lobes segmentation.*

cess the output of the bottleneck layer, in order to exploit the spatial axial correlation between subsequent slices and enhance the final segmentation by integrating 3D information in the model. Convolutional LSTMs [181] are commonly used to capture spatio-temporal correlations in visual data (for example, in videos), by extending traditional LSTMs using convolutions in both the *input-to-state* and the *state-to-state* transitions. Employing recurrent convolutional layers allows the model to take into account the context of the currently-processed slice, while keeping the sequentiality and without the need to process the entire set of slices in a single step through channel-wise concatenation, which increases feature sizes and loses information on axial distance.

Fig. 3.2 shows an example of automated lung and lobe segmentation from a CT scan by employing the proposed segmentation network. The proposed segmentation network is first executed on the whole CT scan for segmenting only lung (and lobes); the segmented CT scan is then passed to the downstream classification modules for COVID-19 identification and lesion categorization.

3.3.2 Automated COVID-19 Diagnosis: CT classification

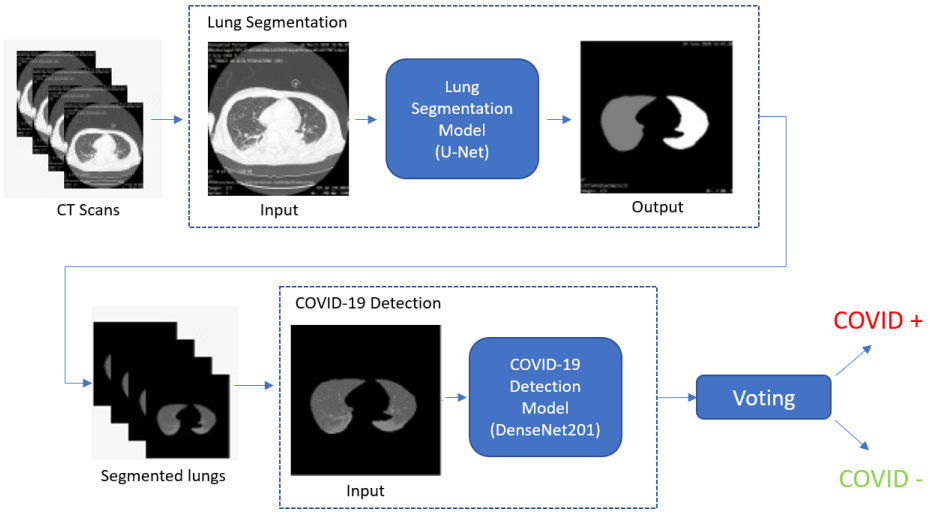


Figure 3.3: *Overview of the* ***COVID-19 detection approach*** *for CT scan classification as either COVID-19 positive or negative.*

After parenchima lung segmentation (through the segmentation model presented in Sect. 3.3.1) a deep classification model analyzes slice by slice each segmented CT scan, and decides whether a single slice contains evidence of the COVID-19 disease. Note that slice-based COVID-19 classification is only the initial step towards the final prediction, which takes into account *all* per-slice predictions, and assigns the "positive" label in presence of a certain number of slices (10% of the total) that the model has identified as COVID-19 positive. Hence, COVID-19 assessment is actually carried out per patient, by combin-

ing per-slice predictions.

At this stage, the system does not carry out any identification and localization of COVID-19 lesions, but it just identifies all slices where patterns of interest may be found and according to them, makes a guess on the presence or not of COVID-19 induced infection. An overview of this model is shown in Fig. 3.3: first the segmentation network, described in the previous section, identifies lung areas from CT scan, then a deep classifier (a DenseNet model in the 201 configuration [65]) processes the segmented lung areas to identify if the slice shows signs of COVID-19 virus.

Once the COVID-19 identification model is trained, we attempt to understand what features it employs to discriminate between positive and negative cases. Thus, to interpret the decisions made by the trained model we compute class-discriminative localization maps that attempt to provide visual explanations of the most significant input features for each class. To accomplish this we employ GradCAM [151] combined with VarGrad [5]. More specifically, GradCAM is a technique to produce such interpretability maps by investigating output gradient with respect to feature map activations. More specifically, GradCAM generates class-discriminative localization map for any class c by first computing the gradient of the score for class c, s^c, w.r.t feature activation maps A_k of a given convolutional layer. Such gradients are then global-average-pooled to obtain the activation importance weights w, i.e.:

$$w_k^c = \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{3.1}$$

Afterwards, the saliency map S^c, that provides an overview of the

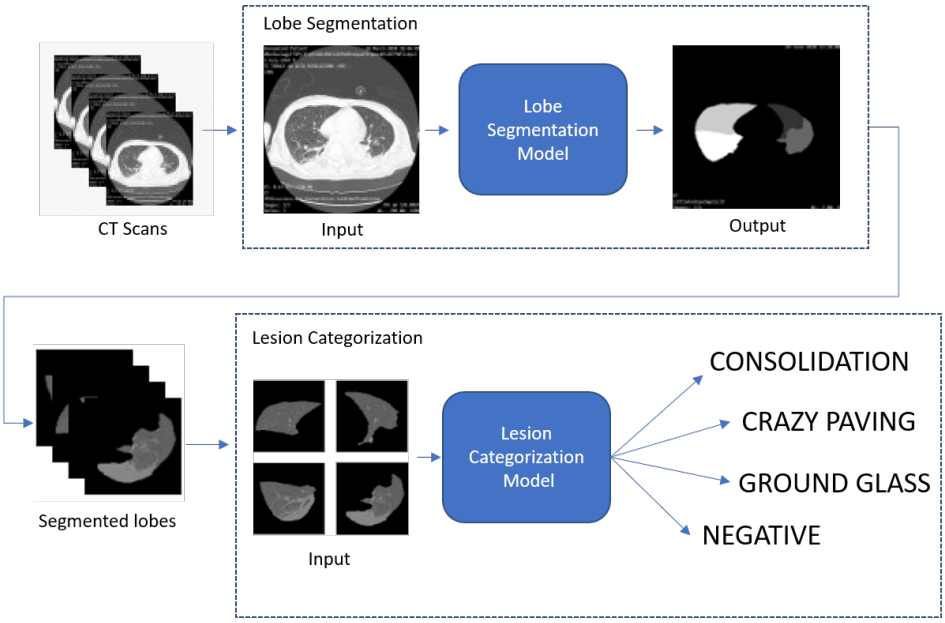


Figure 3.4: *Overview of* ***COVID-19 lesion categorization*** *approach.*

activation importance for the class c, is computed through a weighted combination of activation maps, i.e.:

$$S^c = ReLU\left(\sum_k w_k^c A^k\right) \tag{3.2}$$

VarGrad is a technique used in combination to GradGAM and consists in performing multiple activation map estimates by adding, each time, Gaussian noise to the input data and then aggregating the estimates by computing the variance of the set.

3.3.3 COVID-19 lesion identification and categorization

An additional deep network activates only if the previous system identifies a COVID-19 positive CT scan. In that case, it works on the subset of slices identified as COVID-19 positives by the first AI system with the goal to localize and identify specific lesions (consolidation, crazy paving and ground glass). More specifically, the lesion identification system works on segmented lobes to seek COVID-19 specific patterns. The subsystem for lesion categorization employs the knowledge already learned by the COVID-19 detection module (shown in Fig. 3.3) and refines it for specific lesion categorization. An overview of the whole system is given in Fig. 3.4.

3.3.4 A Web-based Interface for Explaining AI decisions to Radiologists

In order to explain to radiologists, the decisions made by a "black-box" AI system, we integrated the inference pipeline for COVID-19 detection into a web-based application. The application was designed to streamline the whole inference process with just a few clicks and visualize the results with a variable grade of detail (Fig. 3.5). If the radiologists desire to see which CT slices were classified as positive or negative, they can click on "Show slices" where a detailed list of slices and their categorization is showed (Fig. 3.6).

Because the models may not achieve perfect accuracy, a single slice inspection screen is provided, where radiologists can inspect more closely the result of the classification. It also features a restricted set

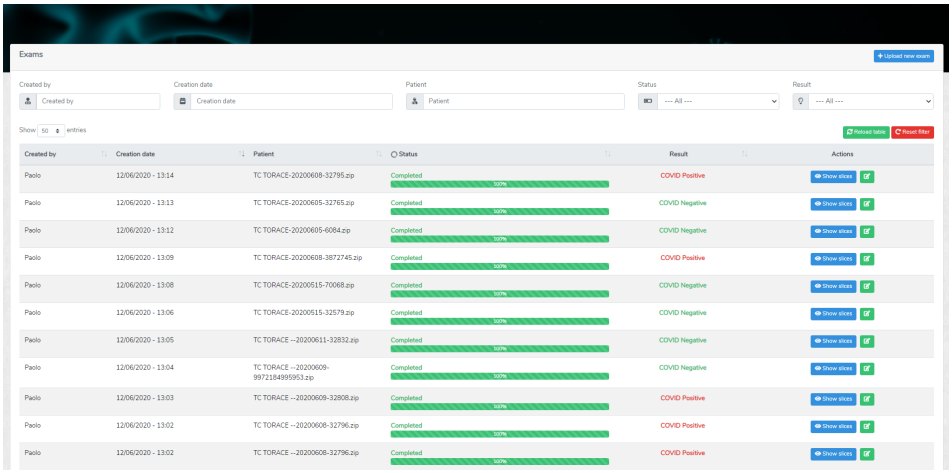


Figure 3.5: *The main page of the AI-empowered web GUI for explainable AI. The user is presented with a list of the CT scan classifications reporting the models' prediction.*

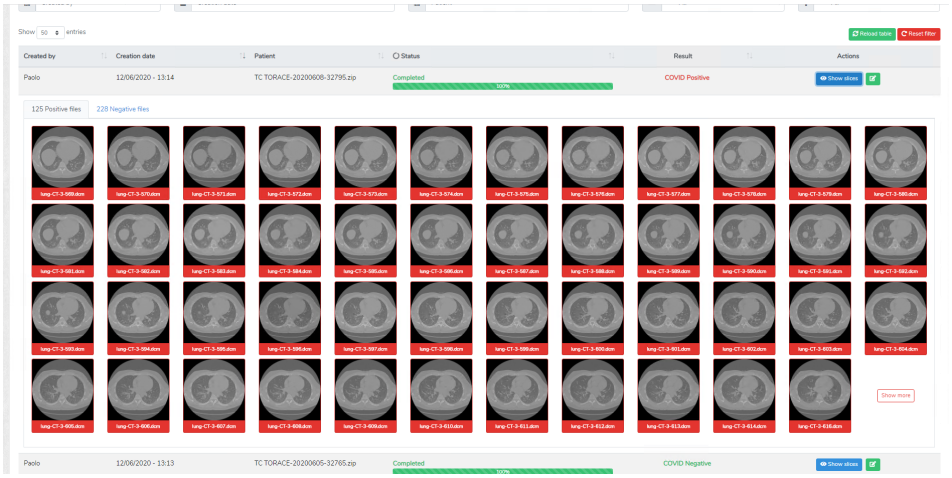


Figure 3.6: *The summarized classification result showing the CT slices that the neural network classified as positive or negative.*

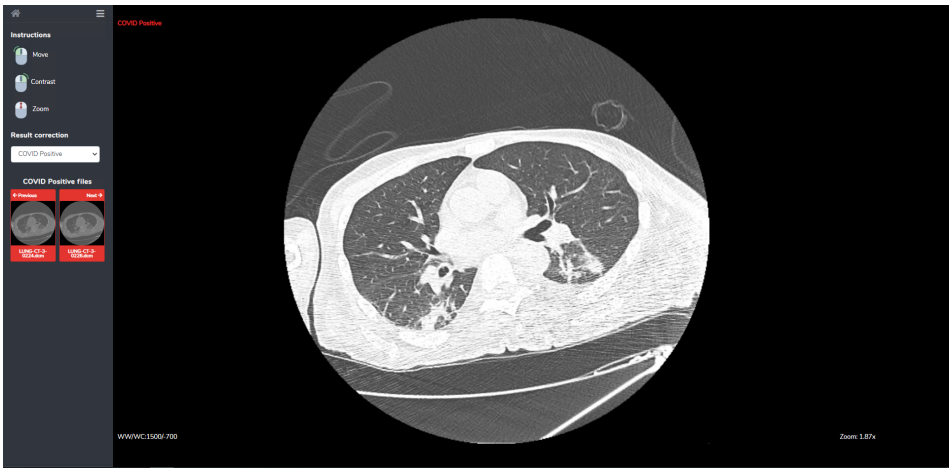


Figure 3.7: *The slice inspection screen. In this screen the user can inspect each single slice and the AI models' decisions.*

of image manipulation tools (move, contrast, zoom) for aiding the user to make a correct diagnosis (Fig. 3.7).

The AI-empowered web system integrates also a relevance feedback mechanism where radiologists can correct the predicted outputs, and the AI module exploits such a feedback to improve its future assessments. Indeed, both at the CT scan level and at the CT slice level, radiologists can correct models' prediction. The AI methods will then use the correct labels to enhance their future assessments.

3.4 Results and Discussion

3.4.1 Dataset and annotations

Data. Our dataset contains overall 166 CT scans: 72 of COVID-19 positive patients (positivity confirmed both by a molecular — reverse transcriptase–polymerase chain reaction for SARS-coronavirus RNA from nasopharyngeal aspirates — and an IgG or IgM antibody test) and 94 of COVID-19 negative subjects (35 patients with interstitial pneumonia but tested negative to COVID-19 and 59 controls).
CT scans were performed on a multi-detector row helical CT system scanner [1] using 120 kV pp, 250 mA, pitch of 1.375, gantry rotation time of 0,6 s and time of scan 5,7 s. The non-contrast scans were reconstructed with slice thicknesses of 0.625 mm and spacing of 0.625 mm with high-resolution lung algorithm. The images obtained on lung (window width, 1,000–1,500 H; level, –700 H) and mediastinal (window width, 350 H; level, 35–40 H) settings were reviewed on a picture archiving and communication system workstation[2]. For training the lung/lobe segmentation model we adopted a combination of the LIDC [10], LTRC[3] and [61] datasets, for a total of 300 CT scans.
Annotations. We perform both COVID-19 identification and lesion categorization, thus the annotations are different according to the task. For COVID-19 identification, ground truth consists of the results of the molecular and an IgG/IgM antibody test. Among the set of 166 CT scans, we used 95 scans (36 positives and 59 negatives) for train-

[1] Bright Speed, General Electric Medical Systems, Milwaukee, WI
[2] Impax ver. 6.6.0.145, AGFA Gevaert SpA, Mortsel, Belgium
[3] https://ltrcpublic.com/

ing, 9 scans for validation (5 positives and 4 negatives) and 62 scans (31 positives and 31 negatives) for test. To compare the AI performance to the human one, the test set of 62 CT scans was provided to three expert radiologists for blind evaluation.
For lesion categorization, instead, CT scans of positive patients were also annotated by three expert radiologists (through consensus) who selected a subset of slices and annotated them with the type (Consolidation, Ground Glass and Crazy Paving) and the location (left-/right/central and posterior/anterior) of the lesion. In total, about 2,400 slices were annotated with COVID-19 lesions and about 3,000 slices of negative patients with no lesion. Tab. 3.1 provides an overview of all the CT scans and lesion annotations in our dataset.
As for lung segmentation, annotations on lung/lobe areas were done manually by the same three expert radiologists who carried out lesion categorization.

CT scans		**Annotated slices**			
		Ground glass	Crazy paving	Consolidation	Total
Positive	72	1,035	757	598	2,390
Negative	94	–	–	–	2,988

Table 3.1: *CT Dataset for training and testing the deep models.*

3.4.2 Training Procedure

COVID-19 Identification Model. The COVID-19 detection network is a DenseNet201, which was used pretrained on the ImageNet dataset [39]. The original classification layers in DenseNet201 were re-

placed by a 2-output linear layer for the COVID-19 positive/negative classification. Given the class imbalance in the training set, we used the weighted binary cross-entropy (defined in 3.3) as training loss and RT-PCR virology test as training/test labels. The weighted binary cross-entropy loss for a sample classified as x with target label y is then calculated as:

$$WBCE = -w\left[y \cdot \log x + (1-y) \cdot \log(1-x)\right] \tag{3.3}$$

where w is defined as the ratio of the number negative samples to the total number of samples if the label is positive and vice versa. This way the loss results higher when misclassifying a sample that belongs to the less frequent class. It is important to highlight that splitting refers to the entire CT scan and not to the single slices: we made sure that full CT scans were not assigned to different splits to avoid any bias in the performance analysis. This is to avoid the deep models overfit the data by learning spurious information from each CT scan, thus invalidating the training procedure, thus enforcing robustness to the whole approach. Moreover, for the COVID-19 detection task, we operate at the CT level by processing and categorizing each single slice. To make a decision for the whole scan, we perform voting: if 10% of total slices is marked as positive then the whole exam is considered as a COVID-19 positive, otherwise as COVID-19 negative. The choice of the voting threshold was selected according to the best operating point in the ROC curve.

COVID-19 lesion categorization model. The lesion categorization deep network is also a DenseNet201 model where classification layers were replaced by a 4-output linear layer (*ground glass, con-*

solidation, crazy paving, negative). The lesion categorization model processes lobe segments (extracted by our segmentation model) with the goal to identify specific lesions. Our dataset contains 2,488 annotated slices; in each slice multiple lesion annotations with relative location (in lobes) are available. Thus, after segmenting lobes from these images we obtained 5,264 lobe images. We did the same on CT slices of negative patients (among the 2,950 available as shown in Tab. 3.1) and selected 5,264 lobe images without lesions. Thus, in total, the entire set consisted of 10,528 images. We also discarded the images for which lobe segmentation produced small regions indicating a failure in the segmentation process. We used a fixed test split consisting of 195 images with consolidation, 354 with crazy paving, 314 with ground glass and 800 images with no lesion. The remaining images were split into training and validation sets with the ratio 80/20. Given the class imbalance in the training set, we employed weighted cross-entropy as training loss. The weighted cross-entropy loss for a sample classified as x with target label y is calculated as:

$$WCE = -w \sum^{C} y \cdot log(x) \tag{3.4}$$

where C is the set of all classes. The weight w for each class c is defined as:

$$w_c = \frac{N - Nc}{N} \tag{3.5}$$

where N is the total number of samples and N_c is the number of samples that have label c.

Since the model is the same as the COVID identification network, i.e., DenseNet201, we started from the network trained on the COVID-

identification task and fine-tune it on the categorization task to limit overfitting given the small scale of our dataset.

For both the detection network and the lesion categorization network, we used the following hyperparameters: batch-size = 12, learning rate = 1e-04, ADAM back-propagation optimizer with beta values 0.9 and 0.999, eps = 1e-08 and weight decay = 0 and the back-propagation method was used to update the models' parameters during training. Detection and categorization networks were trained for 20 epochs. In both cases, performance are reported at the highest validation accuracy.

Lung/lobe segmentation model. For lung/lobe segmentation, input images were normalized to zero mean and unitary standard deviation, with statistics computed on the employed dataset. In all the experiments for our segmentation model, input size was set to 224×224, initial learning rate to 0.0001, weight decay to 0.0001 and batch size to 2, with RMSProp as optimizer. When C-LSTMs were employed, recurrent states were initialized to zero and the size of the input sequences to the C-LSTM layers was set to 3. Each training was carried out for 50 epochs. All experiments have been executed using the HPC4AI infrastructure [7].

3.4.3 Performance Evaluation

In this section we report the performance of the proposed model for lung/lobe segmentation, COVID-19 identification and lesion categorization.

Lobe segmentation

Our segmentation model is based on the Tiramisu model [72] with the introduction of *squeeze-and-excitation* blocks and of a convolutional LSTM (either unidirectional or bidirectional) after the bottleneck layer. In order to understand the contribution of each module, we first performed ablation studies by testing the segmentation performance of our model using different architecture configurations:

- Baseline: the vanilla Tiramisu model described in [72];
- Res-SE: residual *squeeze-and-Excitation* module are integrated in each dense block of the Tiramisu architecture;
- C-LSTM: a unidirectional convolutional LSTM is added after the bottleneck layer of the Tiramisu architecture;
- Res-SE + C-LSTM: variant of the Tiramisu architecture that includes both residual *squeeze-and-Excitation* at each dense layer and a unidirectional convolutional LSTM after the bottleneck layer.

We also compared the performance against the U-Net architecture proposed in [61] that is largely adopted for lung/lobe segmentation.

All architectures were trained for 50 epochs by splitting the employed lung datasets into a training, validation and test splits using the 70/10/20 rule. Results in terms of Dice score coefficient (DSC) are given in Tab. 3.2. It has to be noted that unlike [61], we computed DSC on all frames, not only on the lung slices.
The highest performance is obtained with the Res-SE + C-LSTM

configuration, i.e., when adding *squeeze-and-excitation* and the unidirectional C-LSTM at the bottleneck layer of the Tiramisu architecture. This results in an accuracy improvement of over 4 percent points over the baseline. In particular, adding *squeeze-and-excitation* leads to a 2 percent point improvement over the baseline. Segmentation results are computed using data augmentation obtained by applying random affine transformations (rotation, translation, scaling and shearing) to input images. The segmentation network is then applied to our COVID-19 dataset for prior segmentation without any additional fine-tuning to demonstrate also its generalization capabilities.

Model	Lung segmentation	Lobe segmentation
Baseline Tiramisu [72]	89.41 ± 0.45	77.97 ± 0.31
Baseline + Res-SE	91.78 ± 0.52	80.12 ± 0.28
Baseline + C-LSTM	91.49 ± 0.57	79.47 ± 0.38
Baseline + Res-SE + C-LSTM	**94.01 ± 0.52**	**83.05 ± 0.27**

Table 3.2: *Ablation studies of our segmentation network in terms of dice score. Best results are shown in bold. Note: we did not compute confidence intervals on these scores as they are obtained from a very large set of CT voxels.*

COVID-19 diagnosis

We here report the results for COVID-19 diagnosis, i.e., classification between positive and negative cases. In this analysis, we compare model results to those yielded by three experts with different degree of expertise:

1. Radiologist 1: a physician expert in thoracic radiology (∼30 years of experience) with over 30,000 examined CT scans;

2. Radiologist 2: a physician expert in thoracic radiology (∼10 years of experience) with over 9,000 examined CT scans;

3. Radiologist 3: a resident student in thoracic radiology (∼3 years of experience) with about 2,000 examined CT scans.

	Sensitivity	C.I. (95%)
Radiologist 1	83.9	[71.8 – 91.9]
Radiologist 2	87.1	[75.6 – 94.3]
Radiologist 3	80.6	[68.2 – 89.5]
AI Model without lung segmentation	83.9	[71.8 – 91.9]
AI Model with lung segmentation	**90.3**	**[79.5 – 96.5]**

Table 3.3: *Sensitivity (in percentage together with 95% confidence interval) comparison between manual readings of expert radiologists and the AI model for COVID-19 detection without lung segmentation and AI model with segmentation.*

It should be noted that the gold standard employed in the evaluation is provided by molecular and antibody tests, hence radiologists' assessments are not the reference for performance comparison.

We also assess the role of prior segmentation on the performance. This means that in the pipelines showed in Figures 3.3 and 3.4 we removed the segmentation modules and performed classification using the whole CT slices using also information outside the lung areas.

Results for COVID-19 detection are measured in terms of sensitivity, specificity and AUC, and are given in Tables 3.3, 3.4 and 3.5. Note that the AUC is a reliable metric in our scenario, since we explicitly defined the test set to be balanced among classes. More recent techniques [25] may be suitable when this assumption does not hold, as is often the case for new or rare diseases.

	Specificity	C.I. (95%)
Radiologist 1	87.1	[75.6 – 94.3]
Radiologist 2	87.1	[75.6 – 94.3]
Radiologist 3	90.3	[79.5 – 96.5]
AI Model without lung segmentation	87.1	[75.6 – 94.3]
AI Model with lung segmentation	**93.5**	**[83.5 – 98.5]**

Table 3.4: *Specificity (in percentage together with 95% confidence interval) comparison between manual readings of expert radiologists and the AI model for COVID-19 detection without lung segmentation and AI model with segmentation.*

Our results show that the AI model with lung segmentation achieves higher performance than expert radiologists. However, given the relatively small scale of our dataset, statistical analysis carried out with the Chi-squared test does not show any significant difference between AI models and radiologists.
Furthermore, performing lung segmentation improves by about 6 percent points both the sensitivity and the specificity, demonstrating its effectiveness.

	AUC	C.I. (95%)
Radiologist 1	0.83	[0.72 – 0.93]
Radiologist 2	0.87	[0.78 – 0.96]
Radiologist 3	0.80	[0.69 – 0.91]
AI Model without lung segmentation	0.94	[0.87 – 1.00]
AI Model with lung segmentation	**0.95**	**[0.89 – 1.00]**

Table 3.5: *AUC (together with 95% confidence interval) comparison between manual readings of expert radiologists and the AI model for COVID-19 detection without lung segmentation and AI model with segmentation.*

In addition, we also measure how the sensitivity of the COVID-19 identification changes w.r.t. the level of disease severity. In particular, we categorize the 31 positive cases into three classes according to the percentage of the affected lung area: low severity (11 cases), medium severity (11 cases), high severity (9 cases). Results are reported in Table 3.6 that shows how our AI-based method seems to be yielding better assessment than the domain experts, especially at the beginning of the disease (low severity). This is important as an earlier disease detection may lead to a more favourable outcome. In case of high severity, two out of three radiologists showed difficulties in correctly identifying the COVID-19, mainly because when the affected lung area is significant, the typical COVID patterns are less visible. However, even in this case, our deep learning model was able to discriminate

robustly COVID cases.

	Low Severity	Medium severity	High severity
Radiologist 1	72.7(50.6 – 88.5)	100.0(90.9 – 70.6)	77.8(54.7 – 92.6)
Radiologist 2	72.7(50.6 – 88.5)	90.9(70.6 – 100.0)	100.0(81.5 – 100.0)
Radiologist 3	63.6(42.3 – 81.3)	100.0(90.9 – 70.6)	77.8(54.7 – 92.6)
Model_{wo}_segmentation	72.7(50.6 – 88.5)	90.9(70.6 – 100.0)	88.9(67.0 – 99.2)
Model_{w}_segmentation	81.8(59.6 – 94.9)	90.9(70.6 – 100.0)	100.0(81.5 – 100.0)

Table 3.6: *Sensitivity (in percentage) changes w.r.t. disease severity. From the 31 test CTs for positive patients: 11 are with low severity, 11 with medium severity, and 9 with high severity. Values in parentheses indicate 95% confidence intervals (CI).*

As a backbone model for COVID-19 identification, we employ DenseNet201 since it yields the best performance when compared to other state of the art models, as shown in Table 3.7. In all tested cases, we use upstream segmentation through the model described in Sect. 3.3.1. Voting threshold was set to 10% on all cases.

In order to enhance trust in the devised AI models, we analyzed what features these methods employ for making the COVID-19 diagnosis decision. This is done by investigating which artificial neurons fire the most, and then projecting this information to the input images. To accomplish this we combined GradCAM [151] with VarGrad [5][4] and Fig. 3.8 shows some examples of the saliency maps generated by interpreting the proposed AI COVID-19 classification network. It is interesting to note that the most significant activation areas correspond to the three most common lesion types, i.e., ground glass,

[4]https://captum.ai/

Model	Variant	Sensitivity (CI)	Specificity (CI)	Accuracy (CI)
AlexNet	-	71.0(57.9 – 81.6)	90.3(79.5 – 96.5)	80.7(68.3 – 89.5)
ResNet	18	71.0(57.9 – 81.6)	93.5(83.5 – 98.5)	82.3(70.1 – 90.7)
	34	80.7(68.3 – 89.5)	90.3(79.5 – 96.5)	85.5(73.7 – 93.1)
	50	83.9(71.9 – 91.9)	90.3(79.5 – 96.5)	87.1(75.6 – 94.3)
	101	77.4(64.7 – 89.9)	87.1(75.6 – 94.3)	82.3(70.1 – 90.7)
	152	77.4(64.7 – 89.9)	90.3(79.5 – 96.5)	83.9(71.9 – 91.9)
DenseNet	121	77.4(64.7 – 89.9)	93.5(83.5 – 98.5)	85.5(73.7 – 93.1)
	169	67.9(83.5 – 98.5)	93.5(83.5 – 98.5)	81.4(68.7 – 90.2)
	201	90.3(79.5-96.5)	93.5(83.5 – 98.5)	91.9(81.5 – 97.5)
SqueezeNet	-	66.7(54.5 – 78.9)	93.5(83.5 – 98.5)	81.4(68.7 – 90.2)
ResNeXt	-	77.4(64.7 – 86.9)	90.3(79.5 – 96.5)	83.9(71.9 – 91.9)

Table 3.7: *COVID-19 classification accuracy (in percentage) by several state of the art models. Values in parentheses indicate 95% confidence intervals (CI).*

consolidation and crazy paving. This is remarkable as the model has indeed learned the COVID-19 peculiar patterns without any information on the type of lesions (to this end, we recall that for COVID-19 identification we only provide, at training times, the labels "positive" or "negative", while no information on the type of lesions is given).

COVID-19 lesion categorization

For COVID-19 lesion categorization we used mean (and per-class) classification accuracy over all lesion types and per lesion that are provided, respectively, in Table 3.8. Note that no comparison with

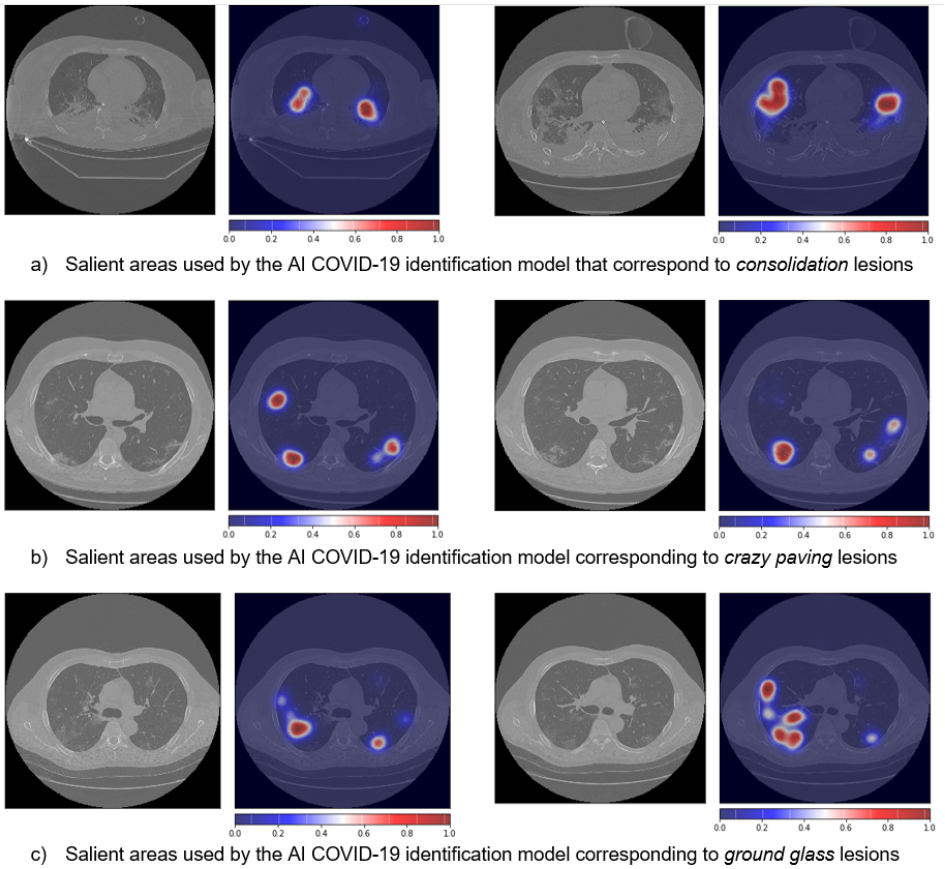


Figure 3.8: *Lung salient areas identified automatically by the AI model for CT COVID-19 identification.*

radiologists is carried out in this case, since ground-truth labels on lesion types are provided by radiologists themselves, hence they are the reference used to evaluate model accuracy.

	Model no_segm	Model w_segm
Consolidation	77.8%(69.9 – 84.1)	97.9%(93.6 – 99.8)
Ground glass	18.6%(14.1 – 24.1)	41.3%(35.1 – 47.7)
Crazy Paving	57.1%(49.4 – 64.4)	98.3%(94.8 – 99.8)
Negative	99.3%(98.6 – 99.7)	99.9%(99.5 – 100)
Average	63.2%	84.4%

Table 3.8: *Per-class accuracy for lesion categorization between AI model without lung segmentation and AI model with segmentation. Values in parentheses indicate 95% confidence intervals (CI).*

Mean lesion categorization accuracy reaches, when operating at the lobe level, about 84% of performance. The lowest performance is obtained on ground glass, because ground glass opacities are specific CT findings that can appear also in normal patients with respiratory artifact. Operating at the level of single lobes yields a performance enhancement of over 21 percent points, and, also in this case, radiologists did not have to perform any lobe segmentation annotation, reducing significantly their efforts to build AI models. The most significant improvement when using lobe segmentation w.r.t. no segmentation is obtained on the *Crazy Paving* class, i.e., 98.3% against 57.1%.

3.4.4 Discussion

Although COVID-19 diagnosis from CT scans may seem an easy task for experienced radiologists, our results show that this is not always the case: in this scenario, the approach we propose has demonstrated its capability to carry out the same task with an accuracy that is at least on par with, or even higher than, human experts, thus showing the potential impact that these techniques may have in supporting physicians in decision making. Artificial intelligence, in particular, is able to accurately identify not only if a CT scan belongs to a positive patient, but also the type of lung lesions, in particular the smaller and less defined ones (as those highlighted in Fig. 3.8). As shown, the combination of segmentation and classification techniques provides a significant improvement in the sensitivity and specificity of the proposed method.

Of course, although the results presented in this work are very promising in the direction of establishing a clinical practice that is supported by artificial intelligence models, there is still room for improvement. One of the limitations of our work is represented by the relatively low number of samples available for the experiments. In order to mitigate the impact of this issue, we carried out confidence level analysis to demonstrate the statistical significance of our results. Moreover, the employed dataset consists of images taken by the same CT scanner, not tested in multiple scanning settings. This could affect the generalization of the method on images taken by other CT scanner models; however, this issue can be tackled by domain adaptation techniques for the medical imaging domain, which is an active research topic [134, 110, 109].

Finally, one of the key features of our approach is the integration of explainability functionalities that may help physicians in understanding the reasons underlying a model's decision, increasing in turn, the trust that experts have in AI–enabled methods. Future developments in this regard should explore, in addition to model explainability, also *causability* features in order to evaluate the quality of the explanations provided [62, 63].

3.5 Conclusion

In this work we have presented an AI-based pipeline for automated lung segmentation, COVID-19 detection and COVID-19 lesion categorization from CT scans. Results showed a sensitivity of 90.3% and a specificity of 93.5% for COVID-19 detection and average lesion categorization accuracy of about 84%. Results also show that a significant role is played by prior lung and lobe segmentation, that allowed us to enhance diagnosis performance of about 6 percent points.

The AI models are then integrated into a user-friendly GUI to support AI explainability for radiologists. To the best of our knowledge, this is the first AI-based software, publicly available, that attempts to explain radiologists what information is used by AI methods for making decisions and that proactively involves in the loop to further improve the COVID-19 understanding.

The results obtained both for COVID-19 identification and lesion categorization pave the way to further improvements, driven towards the implementation of an advanced COVID-19 CT/RX diagnostic pipeline, that is interpretable, robust and able to provide not only

disease identification and differential diagnosis, but also the risk of disease progression.

3.6 Publications

Pennisi, M., Kavasidis, I., Spampinato, C., Schinina, V., Palazzo, S., Proietto Salanitri, F., ... & Conoci, S. (2021). An explainable AI system for automated COVID-19 assessment and lesion categorization from CT-scans. Artificial intelligence in medicine, 118, 102114.

CHAPTER

FOUR

NEURAL TRANSFORMERS FOR INTRADUCTAL PAPILLARY MUCOSAL NEOPLASMS (IPMN) CLASSIFICATION IN MRI IMAGES

We then focused on the classification of precancerous cysts, particularly intraductal papillary mucinous neoplasms (IPMNs). Specifically, we propose an AI-based classifier of IPMNs that leverages the capabilities of neural transformers, with the goal of increasing accuracy and while targeting interpretability issues.

4.1 Motivation

Pancreatic cancer, also known as pancreatic ductal adenocarcinoma (PDAC), is a growing public health issue around the world. In the United States in 2021, an estimated 60,430 new cases of pancreatic cancer will be diagnosed, with 48,220 people dying from the disease [160]. Pre-cancerous cysts or neoplasms in the pancreatic ducts are known as Intraductal Papillary Mucosal Neoplasms (IPMN) and can develop anywhere in the pancreas' ductal zone. Grading the severity of IPMNs is an important diagnosis step: most IPMNs are low-grade, and should be monitored over time; high-grade IPMNs, however, may turn into invasive cancer if left untreated. In these cases, surgery is the first choice to prevent them from expanding into malignant pancreatic tumors. Therefore, there is an unmet need for early detection techniques of IPMNs, in order to identify which IPMNs may lead to cancer. Automated image analysis in radiology imaging plays a key role in diagnosis, treatment and intervention of pancreas diseases; thus there is a strong potential for machine learning tools to support IPMN grade prediction that can serve better than the current radiographic standards. The most popular imaging modalities for the pancreas are computed tomography (CT) and magnetic resonance imaging (MRI). In the last few years, transformer architectures [170, 43] have proven to be a valid alternative to standard convolutional networks on a variety of different tasks. More specifically, transformers enable learning arbitrary functions and consists of two main operation blocks: first, an attention-based block for modeling inter-element relations; second, a multi-layer perceptron (MLP) modeling relations intra-element. A sequence of attention and MLP blocks intertwined with residual con-

nections and normalization has showed to allow for generalization over multiple tasks. Following this trend, herein we propose an automated IPMN classifier based on transformer architecture. We, in particular, show how transformers generalize better than standard and state-of-the-art CNNs (namely, DenseNet, AlexNet, etc.) also for extremely complex tasks, as IPMN classification, while providing similar accuracy to the state of the art IPMN classification study with deep learning [91].
The major contributions of this study are the following:

1. Our work on IPMN classification is an important application contribution, which is not widely done due to the difficult nature of the problem, and hence there is a very limited published research on this task using MRI data with deep learning. Our method provides a significant state-of-the-art baseline to be compared with for further MRI pancreas research just before critical surgery decision or surveillance.

2. Our study contributes to the recent AI research in the strive to demonstrate architectural universalism of Transformers that can be used in a wide variety of tasks using little inductive bias, beside validating their better interpretability than CNN counterparts. To the best of our knowledge, transformers have never been tried on high-risk medical diagnoses tasks before, particularly for pancreas imaging research.

4.2 Related Work

While significant progress has been made for automated approaches to segment the pancreas and its cysts [194], the use of advanced machine learning algorithms to perform fully automatic risk-stratification of IPMNs is still limited.

Some recent works, employing machine learning techniques for predicting the risk of malignancy in IPMN, have used endoscopic ultrasound (EUS) images [89, 54] yielding high accuracy of 94.0%, outperforming both human diagnosis (56%) and conventional guidelines (40–68%). CT imaging has been also adopted for investigating IPMN as in [57, 50] where low-level imaging features, such as texture, strength, and shape, have been extracted from segmented cysts or pancreas for IPMN classification. Recently, deep learning methods based on standard convolutional neural networks have been proposed to diagnose IPMN from MRI scans [68, 34, 91]. Sarfaraz et. al. [68] proposed an architecture for automated IPMN classification based on feature extraction with canonical correlation using a pre-trained 3D CNN, while [34] propose a novel CNN for recognizing high grade dysplasia or cancer on MR-images, yielding promising results. Finally, Rodney et al. [91] constructed two novel "inflated" CNN architectures, InceptINN and DenseINN, for the task of diagnosing IPMN from multisequence (T1 and T2) MRI obtaining an accuracy of about 73% in grading IPNM into three classes (no risk, low and high-risk). In this work, we employ transformers that are specific neural architectures originally proposed for machine translation tasks [170]. Transformer-based models in NLP are generally pre-trained on large corpora and then fine-tuned for the task at hand [40, 135]. The increasing interest in their application to

vision tasks starts with Vision Transformers [43] and Detection Transformer [23]. Recently, several methods have explored transformer-based architectures for medical image analysis mainly for segmentation tasks [28, 180, 58]. However, these method employ an hybrid architecture combining both convolutions and transformers. Our approach builds upon pure vision transformers and employs a strategy similar to that one employed in NLP (as in [40, 135]), i.e., pre-training transformers on natural images and then fine-tuning them to MRI IPNM images. Experimental results show that our pre-trained transformers perform significantly better than state-of-the-art CNN classifiers.

4.3 Method

In our study, we follow the recently emerging approach of *Transformers* [170] for vision tasks. In particular, we use the ViT [43] setting, in which the encoder of the original transformer model is used on a sequence of image "patches". However, since [43] is trained on natural images, it is necessary to adapt the input representation to be able to process MRI scans, which are instead composed by an aggregation of multiple slices, providing anatomical volumetric information.

Fig. 4.1 describes the proposed procedure in detail. We use T1- and T2-weighted MRI scans of the same patients in an early fusion fashion to enrich diagnostic and anatomical (localization) information. For each modality, we first sample k=9 consecutive slices and use them to create a single image, rearranging the selected slices in a $\sqrt{k} \times \sqrt{k}$ grid. k can be set differently depending on the memory availability and z-direction resolution of the MRI scan. In our experiments, we optimize

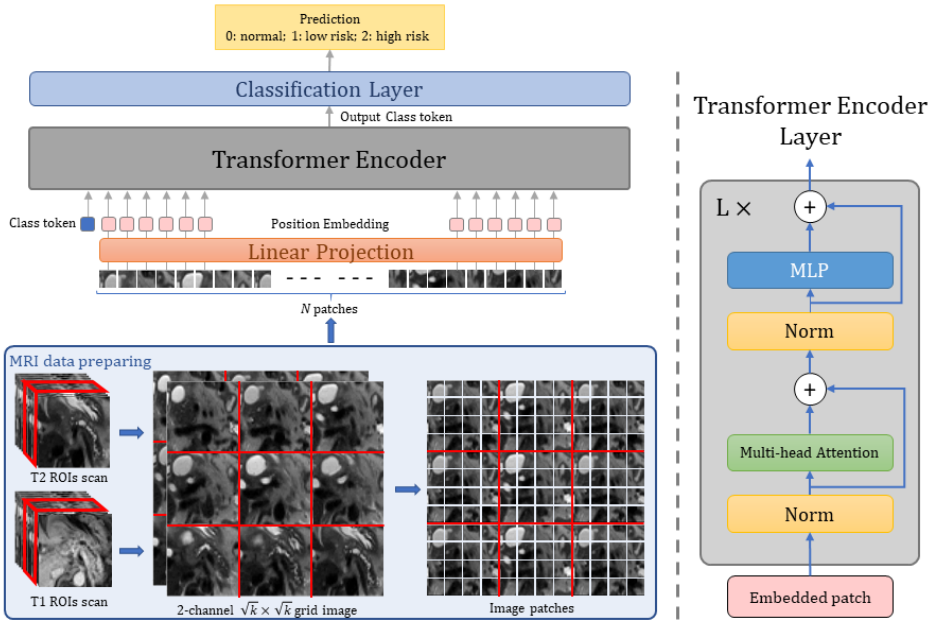


Figure 4.1: *The proposed transformer-based architecture. T1 and T2 slices are concatenated along the channel dimension and sequences of 9 consecutive slices are arranged in a 3×3 grid. Patches are then extracted from the resulting image, and are used as input to the transformer architecture. After encoding the pach set through transformer layers (consisting of a cascade of multihead attention block and MLP layers), a special* classification token *encodes global image representation, and is used for final classification into three IPMN classes: normal, low risk and high risk.*

this number to appreciate full anatomical information of the pancreas. The two images (one for each modality) are then concatenated along the channel dimension: the resulting tensor, of size $\sqrt{k}H \times \sqrt{k}W \times 2$,

with $H \times W$ being the original size of each slice, is provided as input to the transformer. Without loss of generality, let us assume that $H = W$. As in [43], the input image is then divided into N patches of size $P \times P$, where $N = \frac{kHW}{P^2}$. As a result of this procedure, an input image $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ becomes a sequence of 2D patches $\mathbf{X}_p \in \mathbb{R}^{N \times (P \times P \times C)}$ with C being the channel dimension. The 2D patches are then flattened into vectors of size P^2C and projected to an *embedding space* of size D, obtaining a sequence of *token embeddings*. As a last pre-processing step, learnable positional encodings are summed to token embeddings, producing the actual input data sequence to the transformer. We extend the token sequence with a special *class* token, whose state at the output of the transformer describes the overall input image representation for classification purposes [43, 167, 40].

Formally, the input $\mathbf{z}_0$ to the transformer is defined as:

$$\mathbf{z}_0 = \left[\mathbf{x}_{\texttt{class}}, \mathbf{x}_p^1\mathbf{E}, \ldots, \mathbf{x}_p^N\mathbf{E}\right] + \mathbf{E}_{\text{pos}}, \tag{4.1}$$

where each $\mathbf{x}_p^i \in \mathbb{R}^{P^2C}$ is a flattened patch vector, $\mathbf{E} \in \mathbb{R}^{(P^2C) \times D}$ is the embedding matrix and $\mathbf{E}_{\text{pos}} \in \mathbb{R}^{(N+1) \times D}$ is the positional encoding matrix.

The transformer encoder [170] alternates multi-head self-attention and MLP (multilayer perceptron) blocks. These blocks are then intertwined with layer normalization and residual connections (see Fig. 4.1), as follows:

$$\mathbf{z}'_l = \text{MSA}(\text{LN}(\mathbf{z}_{l-1})) + \mathbf{z}_{l-1}, \tag{4.2}$$

$$\mathbf{z}_l = \text{MLP}(\text{LN}(\mathbf{z}'_l)) + \mathbf{z}'_l, \tag{4.3}$$

where $l = 1 \cdots L$ identifies the transformer layer, $\text{LN}(\cdot)$ performs layer

normalization, MLP represents a multilayer perceptor, and MSA(·) computes the standard *query-key-value* multi-head self-attention [170].

At the last transformer layer, the output embedding corresponding to *class* token is finally used for classification into 3 classes, since the MRI dataset includes normal scans, low-grade and high-grade IPMN lesions:

$$\mathbf{y} = \mathrm{LN}(\mathbf{z}_L^0), \tag{4.4}$$

with $\mathbf{y}$ being the vector of output class scores.

4.4 Experimental Results

4.4.1 Dataset

We evaluate the accuracy of our proposed IPMN risk assessment method in MRI (with both T1 and T2 modalities). We use a total of 139 scans from distinct patients, retrospectively collected at Mayo Clinic [91]. Patients have either IPMN cysts detected in their pancreases or they are normal control cases selected to match the IPMN patients. Out of 139 cases, 58 (42%) were male; mean (standard deviation) age was 65.3 (11.9) years. 22% had normal pancreas; 34%, low-grade dysplasia; 14%, high-grade dysplasia; and 29%, adenocarcinoma [34]. Two expert radiologists graded the cases in a pathology report after surgery: 0) normal, 1) low-grade IPMN, and 2) high-grade IPMN. We did not consider invasive carcinoma in our analysis as they are outside the scope of IPMN risk stratification.

MRI images were resized (in the transverse plane) to 256×256 pixels. Voxel spacing of MRI scans were varying from 0.468 mm to

1.406 mm. We applied a set of pre-processing steps: N4 bias field correction followed by an edge-preserving Gaussian smoothing, and intensity standardization procedure to normalize MRI scans across patients, scanners, and time. All MRIs were performed using Siemens scanners 1.5 or 3 T (Siemens, Berlin, Germany).

The experimental procedures involving human subjects described in this paper were approved by the Institutional Review Board.

4.4.2 Training Procedure

We use the Vision-Transformer pre-trained on 300 million images [163] and released in [43]. During training, we fine-tune all transformers layers with the training data from the MRI dataset. MRI slices are cropped around the pancreas areas by expert physicians for all scans, and each set of 9 consecutive slices, extracted in a sliding window fashion, is arranged in a 3×3 grid (from top-left to bottom-right), where each cell of the grid is filled by a 64×64 MRI slice (see Fig. 4.1). Input MRI scans are re-oriented using the RAS axes convention and normalized, individually, between 0 and 1. Data augmentation is performed through random horizontal flipping and random 90-degrees rotation (identically applied to all slices within a grid). We minimize the cross-entropy loss with gradient descent using the Adam optimizer (learning rate: 0.003) and batch size of 8, for a total of 3000 epochs. At inference time, we classify each input MRI by feeding the sequence of 9 central slices to the model.

We employ the same training and evaluation procedure for CNN models used as baselines, i.e., DenseNet-121 [65], AlexNet [88] ResNet18 [59], EfficientNet_b5 [166] and MobileNet_v2 [150]. Exper-

iments are performed on a NVIDIA RTX 3090 GPU. The proposed approach was implemented in PyTorch and MONAI.

Method	Accuracy	Precision	Recall
AlexNet	$0.42_{\pm 0.17}$	0.37 ± 0.15	0.39 ± 0.11
DenseNet	0.51 ± 0.12	0.54 ± 0.14	0.50 ± 0.14
ResNet18	0.53 ± 0.11	0.55 ± 0.23	0.32 ± 0.08
MobileNet_v2	0.43 ± 0.11	0.54 ± 0.26	0.35 ± 0.11
EfficientNet_b5	0.55 ± 0.10	0.60 ± 0.14	0.36 ± 0.08
Ours	**0.70** ± 0.11	**0.67** ± 0.19	**0.64** ± 0.12

Table 4.1: *Performance of tested models with 10-fold nested cross-validation. We report results in term of mean ± standard deviation of metrics computed over all validation folds.*

4.4.3 Performance

We perform 10-fold nested cross-validation in order to estimate the accuracy of the proposed approach and the methods under comparison. Results are reported in Table 4.1, where the proposed model largely outperforms the CNN models, confirming the better generalization capabilities of transformer-based architectures compared to standard convolutional models.

We also evaluate the role of early fusion and of combining the T1 and T2 modalities, by assessing classification performance when the model receives only one modality at a time (either T1 or T2) and when performing late fusion. In this case, we train two transformer models, one for each modality, and we then concatenate the two class to-

Method	Accuracy	Precision	Recall
T1	0.53 ± 0.08	0.60 ± 0.11	0.58 ± 0.14
T2	0.64 ± 0.12	0.64 ± 0.13	0.63 ± 0.11
T1+T2 modalities			
Late fusion	0.60 ±0.16	0.61 ± 0.13	0.59 ± 0.11
Early fusion	**0.70** ± 0.11	**0.67** ± 0.19	**0.64** ± 0.12

Table 4.2: *Performance of our model using different input data modality with 10-fold nested cross-validation. We report results in terms of mean ± standard deviation of metrics computed over all validation folds.*

kens before classification. Performance is reported in Table 4.2 which demonstrates how using T1 and T2 in an early fusion setting yields the highest performance.

It has to be noted that, comparing on the same dataset, the performance achieved by our transformer-based approach is slightly lower than those obtained in [91], i.e., about 73%. However, the architecture in [91] was specifically designed and tuned for solving the IPMN classification problem, while our transformer architecture is general, designed for natural image classification and applied directly without significant architectural changes to IPMN classification problem. This is remarkable, as we demonstrate that a general architecture performs similarly to an ad-hoc one for a complex task with limited training data.

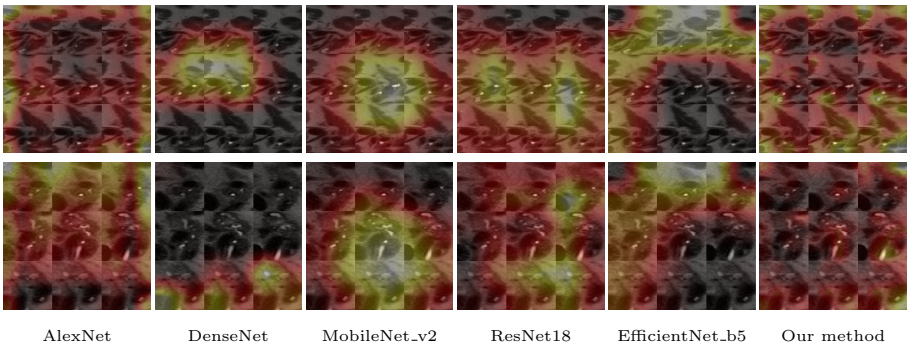


Figure 4.2: *Comparison between the attention maps of several state-of-the-art models and our model in case of correct (top row) and erroneous (bottom row) predictions on a 3×3 grid of MRI images.*

4.4.4 Interpretability of results

Transformers allow for a more direct interpretation of their internal representations through visualizing the attention weights [43], thus supporting the sought interpretability necessary in safety-critical contexts as the medical domain one. We apply Attention Rollout [3] to track down the information propagated from the input layer to the embeddings in the higher layers. Thus, we average attention weights of all heads of each transformer layer and then multiply these averages across all layers. Fig. 4.3 shows some examples of interpretabilty maps in cases of correct cyst classification.

It can be observed how our transformer-based model focuses its attention mainly on cysts, thus it provides robust predictions. Conversely, CNN-based models lead to weak decisions, as their attention maps (estimated using GradCam [152]) reveal that features not strictly related to cysts are used for classification (see Fig. 4.2, top row). Fi-

nally, although our model fails in some cases, as demonstrated by the classification accuracy in Tab. 4.1, its attention maps often point to the correct cyst regions (see Fig. 4.2, bottom row); thus, the wrong prediction is due to either using directly raw data, rather than a more powerful representation, or lack of enough training data.

4.5 Discussion

In this work, our overall goal was to classify pancreas (IPMN) cysts automatically. We utilized *transformers* for the first time for pancreas risk predictions and obtained promising results that can be used for MRI-based IPMN risk stratification routinely. Compared to the (few) existing methods, transformers showed higher performance overall. We found that training transformer for IPMN risk stratification is easier than conventional CNN based systems and generalizes better. Furthermore, the proposed transformer-based classifier allows for better interpretation of results than standard CNNs, revealing how

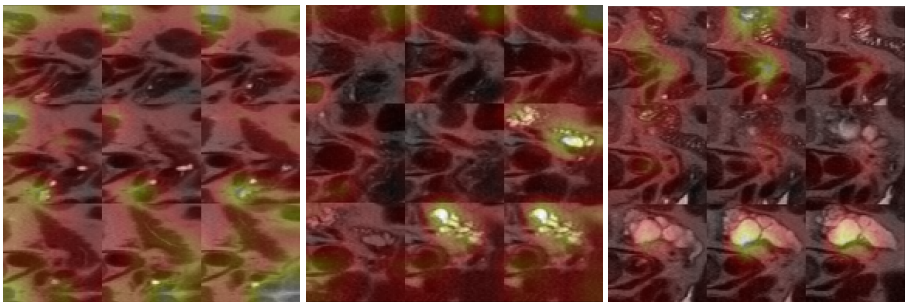


Figure 4.3: *Attention maps of our transformer-based classifier on 3×3 grid of MRI images for correct IPMN classification.*

it employs cues exclusively related to cysts, providing more robustness to the automated diagnosis than the comparing methods. These findings highlight the contribution that transformers can give to the future research in medical image understanding, in general, and IPMN classification, in particular, beside contributing the recent AI research efforts towards universal architectures.

4.6 Publications

Proietto Salanitri, F., Bellitto, G., Palazzo, S., Irmakci, I., Wallace, M., Bolan, C., ... & Spampinato, C. (2022, July). Neural Transformers for Intraductal Papillary Mucosal Neoplasms (IPMN) Classification in MRI images. In 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) (pp. 475-479). IEEE.

CHAPTER

FIVE

HIERARCHICAL 3D FEATURE LEARNING FOR PANCREAS SEGMENTATION

In this chapter, we continue our exploration of the pancreas, shifting our focus to the foundational task of segmentation. However, when it comes to the pancreas, segmentation is no a straightforward endeavor. The organ's shapes and sizes exhibit vast variability across patients; further challenges arise from the pancreas's intensity similarities to surrounding tissues and its often indistinct boundaries, a consequence of the resolution limitations inherent to medical scanners. The presence of cysts, tumors, or other abnormalities only exacerbates these challenges, often confounding segmentation algorithms and leading to inaccurate boundary delineations. Yet, in the face of these complexities, we propose an effective automated segmentation method that works both MRI and CT imaging modality.

5.1 Motivation

Pancreatic cancer is a growing public health concern worldwide. In 2021, an estimated 60,430 new cases of pancreatic cancer will be diagnosed in the US and 48,220 people will die from this disease [160]. Early detection of pancreas cancer [127] is very hard and options in treatment are very limited. Radiology imaging and automated image analysis play key roles in diagnosis, prognosis, treatment, and intervention of pancreatic diseases; thus, there is a strong, unmet, need for computer aided analysis tools supporting these tasks. The first step in such analysis is to automate the medical image segmentation procedures, since manual segmentation (current standard) is tedious, prone to error, and it is not practical in routine clinical evaluation of the diseases [128]. Beyond the known challenges of medical image segmentation problems, pancreas is one of the most difficult organs to segment despite the recent advances in deep segmentation models.

Computed tomography (CT) and magnetic resonance imaging (MRI) are the two most common modalities for pancreas imaging. CT is the modality of choice for pancreatic cancer at the moment, while MRI is mostly used for finding other pancreatic diseases including cysts and diabetes. Compared to CT, MRI has advantages such as the lack of ionizing radiation, better resolution and soft tissue contrast. However, MRI has other unique difficulties, including field inhomogeneity, non-standard intensity distributions due to variations in scanners, patients, field strengths, and high similarity in pancreas and non-pancreas tissue densities.

Image-based pancreas analysis is by itself a challenging task. Shapes and sizes greatly vary across different patients, making it difficult to

use robust priors for improving the delineation procedures. Intensity similarities to non-pancreatic tissues, and smooth or invisible boundaries (due to resolution limitations of medical scanners) are other challenges that need to be addressed in a successful segmentation method. Moreover, in presence of a cyst, tumor, or other abnormalities in pancreases, segmentation algorithms may easily fail to delineate correct boundaries.

To address these challenges, in this work we propose a novel 3D fully convolutional encoder-decoder network with hierarchical multi-scale feature learning, for general, fully-automated pancreas segmentation applicable to CT and MRI scans. Major contributions of this study are the following:

- Our segmentation network is unique in the sense that it is volumetric, learns to extract 3D volume features at different scales, and decodes features hierarchically, leading to improved segmentation results;

- We show the efficacy of our work both on CT and MRI scans. Our architecture successfully extracts pancreases from CT and MRI with high accuracy, obtaining new state-of-the-art results on a publicly-available CT benchmark and first-ever volumetric pancreas segmentation from MRI in the literature.

- Our work on MRI pancreas segmentation is an important application contribution, due to the very limited published research on this task using MRI data with deep learning. It is our belief that our method provides a significant state-of-the-art baseline to be compared with for further MRI pancreas research.

5.2 Related Work

Following the success of deep learning methods applied in medical image segmentation, researchers have recently shown an increasing interest in pancreas segmentation, in order to support physicians in early stage diagnosis for pancreas cancer. Although this application field is still in its infancy — also due to variabilities in texture, size and imaging contrast — a line of promising approaches has been proposed in the literature, mainly on CT scans [20, 86, 95, 106, 112, 143, 144, 145, 172, 175, 186, 195]. We here describe the most significant ones which relate to our proposed model.

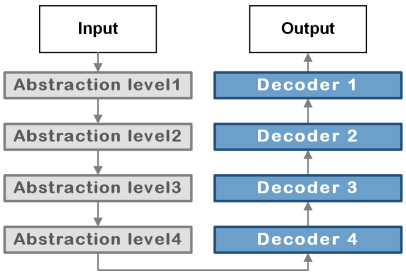
In [144], a two-stage cascaded approach for pancreas localization and pancreas segmentation is proposed. In the first stage, the method localizes the pancreas in the entire 3D CT scan, providing a reliable bounding box for a more refined segmentation step, based on an efficient application of holistically-nested convolutional networks (HNNs) on the three views of pancreas CT image. Per-pixel probability maps are then fused to produce a 3D bounding box of the pancreas. Projective adversarial networks [86] incorporate high-level 3D information through 2D projections and introduce an attention module that supports a selective integration of global information from the segmentation module to an adversarial network. More recently, [175] proposes a dual-input v-mesh fully-convolutional network, which receives original CT scans and images processed by contrast-specific graph-based visual saliency, in order to enhance the soft tissue contrast and highlight differences among local regions in abdominal CT scans.

All of the above works tackle the problem of pancreas segmentation on CT scans. However, as already mentioned, MRI acquisitions have

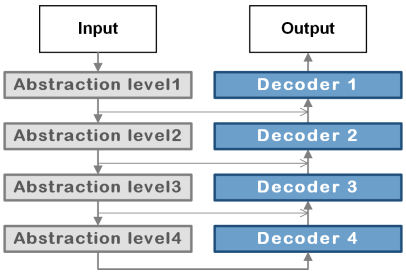
several advantages over CT — most importantly, fewer risks to the patients. On the other hand, MRI pancreas segmentation presents additional challenges to automated visual analysis. For this reason and others (e.g., the lack of public benchmarks), very few works have addressed pancreas segmentation on MRI data: to the best of our knowledge, the major attempts are [11, 20, 21]. In [21], two CNN models are combined to perform, respectively, tissue detection and boundary detection; the results are provided as input to a conditional random field (CRF) for final segmentation. In [11], an algorithmic approach based on hand-crafted features is proposed, employing an ad-hoc multi-stage pipeline: contrast enhancement within coarsely detected pancreas regions is applied to differentiate between pancreatic and surrounding tissue; 3D segmentation and edge detection through max-flow and min-cuts approach and structured forest are performed; finally, non-pancreatic contours are removed via morphological operations on area, structure and connectivity.

5.3 Method

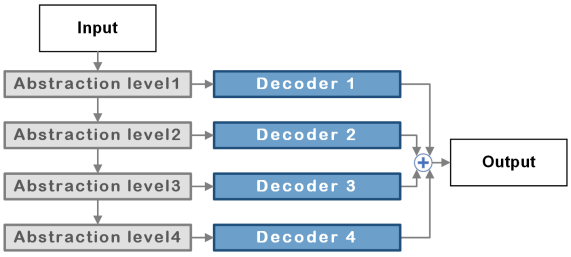
Our 3D fully-convolutional pancreas segmentation model — *PankNet* — is based on an encoder-decoder architecture; however, unlike standard encoder-decoder schemes with a single decoding path (see Fig. 5.1a), we have parallel decoders at different abstraction levels, generating multiple intermediate segmentation maps (Fig. 5.1c). Hierarchical decoding is also fundamentally different from using skip connections (Fig. 5.1b), since these have the purpose to ease gradient flow and forward low-level features for output reconstruction, while



(a) *encoder-decoder*



(b) *skip-connection*



(c) *hierarchical decoding*

Figure 5.1: *A comparison between our proposed architecture and other types of networks used for segmentation: (a) standard encoder–decoder architecture; (b) encoder–decoder architecture with skip connections; (c) encoder–hierarchical decoder architecture (ours).*

our multiple decoders aim to extract local and global dependencies. The detailed architecture is shown in Fig. 5.2: the input data (either CT or MRI volume) is first processed by the encoder stream of the model which aggregates volumetric features at different abstraction levels. These features are then given as input to different decoder streams, each generating a segmentation mask volume.

All intermediate masks are concatenated along the channel dimension and finally merged through a convolutional layer in order to predict the final segmentation mask for all input slices.

5.3.1 Volume feature encoding

The model's encoder performs aggregation of volumetric features from the input data. It is based on S3D [179], a network originally proposed for action recognition using 3D spatial and temporal separable 3D convolution layers, pretrained on the Kinetics Dataset [81]. We use the pretrained network, similarly to other works [21, 86, 91], to ease convergence given the limited training data we have from both CT and MRI datasets. Our encoder processes $D = 48$ slices from an input scan by progressively aggregating volumetric cues down to a more compact representation of size $1024 \times \frac{W}{8} \times \frac{H}{32} \times \frac{D}{32}$ (channels $\times$ width $\times$ height $\times$ depth). Features at the bottleneck and at the outputs of the second, third and fourth pooling layers are fed to separate decoders, described in the following section, to implement our hierarchical decoding strategy.

The proposed approach can be easily adapted to different encoder architectures. Thus, we additionally design a lightweight variant of our *PanKNet* network by replacing the S3D-based encoder with an

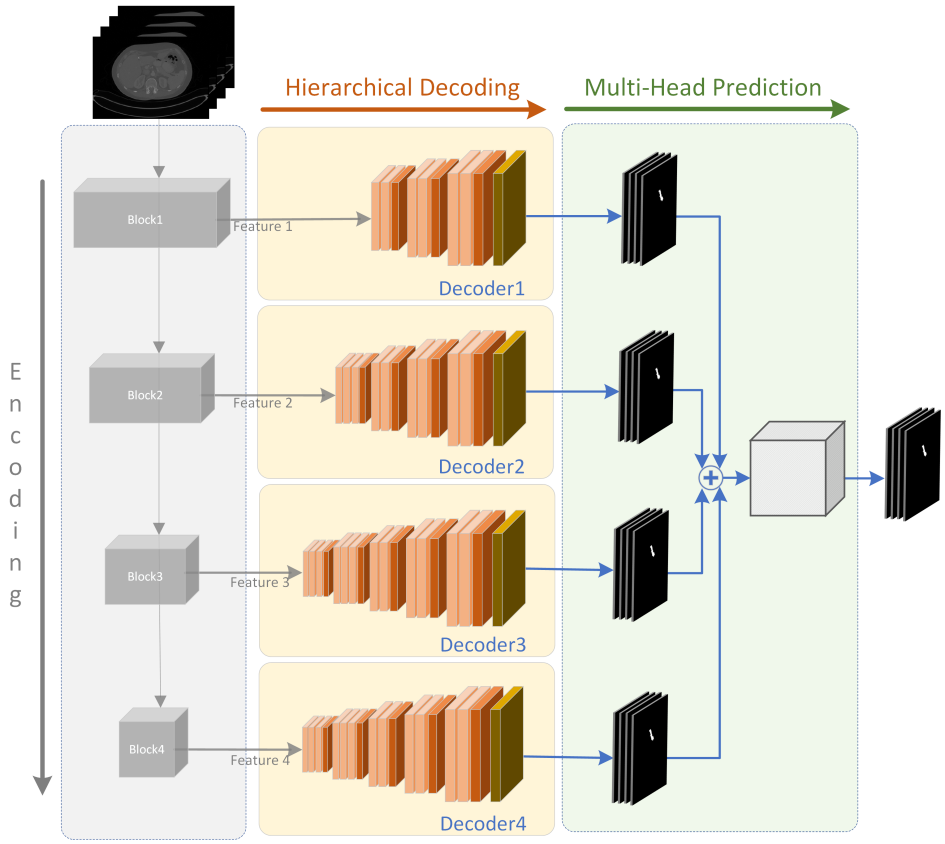


Figure 5.2: *PanKNet architecture: the encoding path extracts aggregated volumetric features, while the decoding path predicts four different intermediate segmentation masks (coarse to fine). Finally, intermediate segmentations are integrated into a detailed output mask.*

encoder based on MobileNetV2 [150], where 2D convolutions are replaced with 3D ones through inflation. In particular, the 2D kernels are replicated along the third dimension, and the values of the weights are divided by the number of replications as proposed in [24]. In this

case, as input to the decoders, we select the output of the second, third, fourth and sixth *bottleneck* blocks of MobileNetV2, providing a more compact feature map of size $160 \times \frac{W}{16} \times \frac{H}{32} \times \frac{D}{32}$. This lightweight variant has 10 times fewer parameters (2.5 millions of parameters, 9.33 MB) and than the S3D counterpart (25.6 millions of parameters, 97.88 MB).

5.3.2 Hierarchical Decoding

Our hierarchical decoding strategy employs features at different points of the encoder stream to generate intermediate segmentation masks that aim to capture and combine fine segmentation (derived from decoders of deeper features) to coarse segmentation (derived from decoders of initial features). We include four decoders: each one processes a set of volumetric features taken from the corresponding level in the encoder stack and performs segmentation on the input volume (see Fig. 5.2, yellow blocks). Each decoder consists of a cascade of upsampling blocks, depending on the size of the input feature map: decoders operating on deeper features require less blocks to recover the original input size. Each upsampling block contains a 3D convolutional block (convolutional layer + batch normalization + ReLU), one or two 3D separable convolutional blocks, and a trilinear upsample layer. As last layer, a pointwise 3D convolution outputs a volume with size $2 \times W \times H \times D$, where W, H and D are the same as the input volume.

5.3.3 Pancreas Segmentation

Intermediate segmentation maps predicted by each of the model's decoders are combined into a global mask. In particular, the four intermediate maps are concatenated into a $8{\times}W \times H \times D$ tensor, which then goes through a last layer performing a voxel-wise convolution to generate a single segmentation map of size $2 \times W \times H \times D$.

The whole model (encoder, hierarchical decoders and output layer) is trained end-to-end using a hierarchical Dice loss [116] between ground-truth mask, intermediate generated masks and the output segmentation mask. Formally, given the predicted output segmentation masks $\mathbf{S}_v$ for the input volume, the four maps $\hat{\mathbf{S}}_{v^i}$ estimated by the decoders, and the ground-truth segmentation maps $\mathbf{G}_v$ for the input data, the *segmentation loss* $\mathcal{L}_s$ is:

$$\mathcal{L}_s\left(\mathbf{S}_v, \hat{\mathbf{S}}_{v^i}, \mathbf{G}_v\right) = \sum_{i=1}^{4} \frac{2\sum_j \hat{S}_{v^{i,j}} G_{v^j}}{\sum_j \hat{S}^2_{v^{i,j}} + \sum_j G^2_{v^j}} + \frac{2\sum_j S_{v^j} G_{v^j}}{\sum_j S^2_{v^j} + \sum_j G^2_{v^j}} \quad (5.1)$$

where index i iterates over the four intermediate maps and index j iterates over voxels.

5.4 Experiments and Results

5.4.1 Dataset

We evaluate the accuracy of our proposed deep segmentation method in both CT and MRI modalities. For the former, we use the publicly available NIH Pancreas-CT dataset, which is the most used pancreas

segmentation dataset for benchmarking [143]. This dataset includes 82 abdominal contrast-enhanced 3D CT scans. The resolution of the CT scans is $512 \times 512 \times Z$, with Z (between 181 and 466) indicating the number of slices along the transverse axis. Voxel spacing ranges from 0.5 mm to 1 mm. More details on this dataset are available in [143].

In our experiments with MRI data, we use 40 in-house collected T2-weighted MRI scans from 40 patients, who have either IPMN (intraductal papillary mucinous neoplasm) cysts detected in their pancreases or invasive pancreatic ductal carcinoma. Two expert radiologists annotated pancreases manually and consensus segmentation masks were generated at the end of the ground-truth labeling procedure with agreement. MRI images were resized (in the transverse plane) to 256 x 256 pixels, with voxel spacing of varying from 0.468 mm to 1.406 mm. To minimize uncertainties in MRI scans, we applied a set of pre-processing steps: N4 bias field correction followed by an edge-preserving Gaussian smoothing, and intensity standardization procedure to standardize MRI scans across patients, scanners, and time.

5.4.2 Training and evaluation procedure

We apply the same training procedure for the two datasets, with the only difference regarding how model backbones are pre-trained. On the NIH Pancreas-CT dataset, we pre-train S3D on Kinetics [81] and MobileNetV2 on ImageNet [39] with weight inflation; on our MRI data, *Pancreas-MRI*, we employ the backbones pre-trained on the CT task.

Input CT and MRI scans are re-oriented using the RAS axes convention for consistency. We then perform voxel resampling through trilinear interpolation in order to have isotropic (1 mm) voxel spacing, and normalize the values of each scan between 0 and 1. During training, data augmentation is performed with random horizontal flipping, random 90-degrees rotation and random crops of size 128×128×48 (in RAS coordinates). We minimize our multi-part Dice loss with mini-batch gradient descent using the Adam optimizer (learning rate: 0.001) and batch size 8, for a total of 3000 epochs.

At inference time, we compute output segmentation masks by running a sliding window routine over an entire input scan, using 256×256×48 windows overlapping by 25%. Voxel labels from overlapping segmentations are obtained by averaging the set of predictions. For evaluation, we carry out 4-fold cross-validation. At each iteration, the set of training folds is further split into the actual training set and a validation set, that is used to select the epoch at which Dice score on the test fold is reported. As metrics for quantitative evaluation, we employ: *Dice score coefficient* (DSC), *Positive Predictive Value* (PPV) and *Sensitivity*.

Experiments are performed on an NVIDIA Quadro P6000 GPU. The proposed approach was implemented in PyTorch and MONAI.

5.4.3 Results

We first test our model (as well as its lightweight variant) on the NIH Pancreas-CT dataset and compare it to existing methods (which share our evaluation strategy with 4-fold cross-validation), namely, [20, 86, 95, 106, 112, 143, 144, 145, 172, 175, 186, 195]. Summarized

Method	DSC			PPV	SENS
	Avg	Max	Min		
Roth et al. [143]	71.42±10.11	86.29	23.99	–	–
Roth et al. [144]	78.01±8.20	88.65	34.11	–	–
Roth et al. [145]	81.27±6.27	88.96	50.69	–	–
Zhou et al. [195]	82.37±5.68	90.85	62.43	–	–
Cai et al. [20]	82.40±6.70	90.10	60.00	–	–
Li et al.(2019) [95]	83.50±6.20	–	–	84.50±6.90	83.70±10.40
Liu et al.(2020) [106]	84.10±4.90	–	–	83.60±5.90	85.30±8.20
You et al. [186]	84.50±4.97	91.02	62.81	–	–
Khosravan et al. [86]	85.53±1.23	88.71	83.20	–	–
Wang et al.(2020) [172]	85.90±3.40	–	–	–	–
Man et al.(2019) [112]	86.90±4.90	–	–	–	–
Wang et al. [175]	87.04±6.80	–	–	**89.50±5.80**	87.70±7.90
$\text{PanKNet}_{\text{Light}}$	*87.13±4.58*	93.49	72.77	86.85±6.52	*88.48±5.12*
PanKNet	**88.01±4.74**	93.84	70.62	*88.25±5.45*	**88.69±5.99**

Table 5.1: *Comparison of PanKNet against multiple state-of-the-art models for pancreas segmentation on NIH Pancreas-CT dataset using 4-fold cross-validation.* ***Best performance in bold****, second best in italic.*

in Table 5.1, our results indicate that *PanKNet* outperforms existing methods over different metrics. Note that PanKNet does not require any auxiliary regularization networks [86], nor additional inputs [175], nor upstream pancreas localization module [112]. Remarkably, even the lightweight variant of PanKNet yields accuracy comparable to the full model, while outperforming existing models, showing that the choice of the backbone is not as important as the overall employed

Method	**DSC**			**PPV**	**SENS**
	Avg	Max	Min		
3D-UNet [83]	65.05±9.17	84.58	49.80	61.55±7.55	74.42±13.99
$\text{Baseline}_{\text{Light}}$	69.17±8.10	83.86	49.92	64.64±7.49	84.19±11.72
Baseline	65.16±9.11	84.00	49.49	61.92±8.22	75.22±12.46
$\text{PanKNet}_{\text{Light}}$	72.96±10.33	88.54	49.90	71.39±11.21	79.76±11.53
PanKNet	**77.46±08.62**	**89.07**	**52.30**	**76.63±8.66**	**80.91±10.51**

Table 5.2: *Segmentation performance on Pancreas-MRI dataset (4-fold cross-validation).*

hierarchical architecture. The best trade-off between accuracy and computational resources for CT pancreas segmentation is represented by $\text{PanKNet}_{\text{Light}}$, whose memory occupation is about 10 MB compared to about 100 MB of PanKNet, but with very similar performance.
We then test our model on pancreas segmentation from MRI data. In this case, we compare the 3D-UNet, proposed in [83], pre-trained on the NIH Pancreas-CT dataset and fine-tuned on our MRI dataset. Furthermore, we add to this evaluation some control experiments to show the effectiveness of the designed architecture. Consequently, we define as baseline our encoder-decoder architecture without hierarchical decoding strategy, decoding only the features at the model's bottleneck. Results in Table 5.2 indicate that both PanKNet variants outperform the state-of-the-art 3D U-Net model [83]. The baseline (with either backbones) also performs better than 3D U-Net model [83] demonstrating that even our 3D fully convolutional network, ablated from the hierarchical decoding, is effective for MRI pancreas segmentation. Adding hierarchical decoding leads to enhanced segmentation

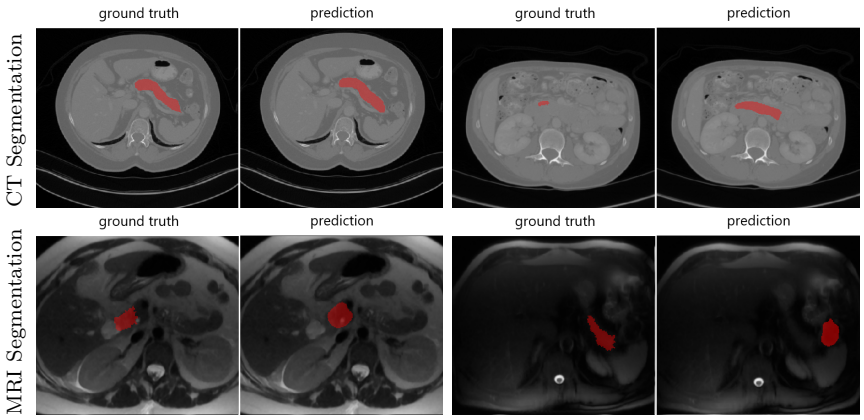


Figure 5.3: *Segmentation masks at the highest (left column) and lowest (right column) Dice score on NIH Pancreas-CT (first row) and Pancreas-MRI dataset (second row).*

performance, especially on DSC and PPV. Different from CT segmentation and from baseline models, PanKNet largely outperforms its lightweight counterpart, demonstrating that MRI pancreas segmentation is far more complex and challenging than CT segmentation and calls for high-capacity networks to be solved.

Example segmentation masks, corresponding to the highest and lowest Dice scores reported in Tables 5.1 and 5.2 for CT and MRI pancreas segmentation, are illustrated in Fig. 5.3.

5.5 Discussion

In this study, we propose a novel 3D fully-convolutional network for pancreas segmentation from MRI and CT scans. Our proposed deep

network aims at learning and combining multi-scale features, namely a hierarchical decoding strategy, to generate intermediate segmentation masks for a coarse-to-fine segmentation process. The intermediate masks, capturing fine details, are derived from decoders of deeper features while coarse segmentation details are derived from decoders of initial features. We evaluated the efficacy of our method (a) on CT scans from the publicly available NIH CT-Pancreas benchmark, and obtained a new state of the art Dice score **88.01%**, outperforming all previous methods; and (b) on MRI scans, obtaining a Dice score of **77.46%**, which can be used as a baseline for future works on MRI pancreas segmentation. Noting that MRI pancreas segmentation methods are extremely limited due to the challenging nature of the problem, our study offers a fresh insight into MRI analysis of pancreas from a fully automated volumetric segmentation strategy. PanKNet is tested for pancreas segmentation, but its architecture is general and can be applied to any 3D object segmentation problem in medical domain.

5.6 Publications

Proietto Salanitri, F., Bellitto, G., Irmakci, I., Palazzo, S., Bagci, U., & Spampinato, C. (2021). Hierarchical 3d feature learning for pancreas segmentation. In Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 12 (pp. 238-247). Springer International Publishing.

Part III

GENERATIVE-BASED FEDERATED LEARNING STRATEGIES

In the preceding section of this thesis, we explored some important centralized approaches in medical imaging. While these centralized methodologies have showcased remarkable efficacy, they inherently grapple with challenges, especially those tied to data centralization. Aggregating data from diverse sources at a single point not only introduces logistical complexities but also amplifies concerns related to data privacy and security. In the sensitive realm of medical imaging data, where the protection of patient data is of paramount importance, these concerns cannot be overlooked.

At its core, federated learning is a decentralized strategy where the model is trained across multiple devices or servers without the need to centralize the data. This paradigm ensures that data remains localized, thereby addressing the primary concerns of data transfer and central storage. More critically, federated learning emerges as a beacon for data privacy. Instead of sharing raw data, which could potentially expose sensitive information, federated learning shares model updates,

ensuring that the individual data points never leave their original location. This decentralized approach not only preserves the integrity and confidentiality of the data but also harnesses the collective intelligence of diverse datasets in a privacy-centric manner.

However, the shift towards federated learning is not trivial. A primary concern is ensuring that decentralized data meaningfully contributes to the global model without breaching privacy constraints. Recognizing this challenge, we introduced an innovative approach leveraging the capabilities of Generative Adversarial Networks (GANs). GANs, renowned for their ability to generate synthetic yet realistic data, are the cornerstone of our solution. By synthesizing data that mirrors the essential characteristics of the original without directly replicating it, GANs offer a pathway to share and enhance federated learning models without compromising data privacy. In essence, instead of sharing actual data, we propose sharing its synthetic counterpart that maintains privacy while still keeping the significant visual features. Building on this foundational idea of using GANs for privacy-preserving federated learning, we embarked on a series of research endeavors, each addressing unique challenges and offering innovative solutions within this framework. These works not only underscore our commitment to advancing the field but also highlight the practical applications of our proposed methods.

In Chapter 6 we delve deeper into the latent space of GANs, focusing on its manipulation and aggregation. We propose innovative techniques that generate privacy-preserving synthetic images and introduce novel methods for aggregating this information in a federated context. The emphasis here is on how the latent space can be harnessed and manipulated to enhance federated learning, ensuring that

medical insights can be shared and aggregated without compromising the integrity of the original data. Building upon the foundational concepts of the latent space, Chapter 7 introduces a complementary approach to the previous research. While many methods, including the prior work and techniques like k-same, focus on aggregating information in the latent space, and generating samples from this aggregation in a strainghtforward way, in this research we propose an intelligent method to navigate the latent space, ensuring equidistance, privacy, and class consistency during data generation. Moving to the last work presented in this thesis, Chapter 8 seamlessly integrates the concepts of continual learning and federated learning to address the challenges posed by distribution shifts. Continual learning predominantly addresses the issue of distribution shift over time, as models continuously evolve and adapt to incoming data. In contrast, federated learning confronts distribution shifts in space due to the decentralized nature of its data sources. By combining these concepts with generative adversarial models, our FedER framework effectively integrates features from local nodes, crafting models that can generalize across diverse datasets while preserving privacy. This strategy sidesteps the constraints of traditional federated learning solutions, presenting a robust and privacy-focused methodology. Furthermore, the real-world relevance of this framework is emphasized by its application in actual medical scenarios, showcasing its resilience and effectiveness in practical contexts.

CHAPTER

SIX

GAN LATENT SPACE MANIPULATION AND AGGREGATION FOR FEDERATED LEARNING IN MEDICAL IMAGING

As we embark on the next phase of our exploration, we move into the world of federated learning, approaching it from a distinctly data-driven perspective. The overarching goal is clear: to generate and disseminate synthetic samples that maintain the integrity of the original data while ensuring utmost privacy. The synthesis of data is achieved through a multistage strategy: exploiting the power of Generative Adversarial Networks (GANs) to grasp the distribution of original data, projecting these private data samples into the GAN's latent space, and then clustering these projections. By interpolating the cluster centroids, we generate synthetic images, meticulously ensuring a minimized risk of sensitive information leakage.

6.1 Motivation

The recent success of deep learning in the medical domain has shown it to be a promising tool to support medical diagnosis and treatment, but large amounts of training data are still needed to build models able to achieve good accuracy and generalization. However, medical institutions generally curate their own datasets and keep them private for privacy concerns. Due to their small size, models trained on private datasets tend to overfit, introduce biases and generalize badly on other data sources that address the same task [187].

A viable solution for increasing the size and diversity of data is to employ a collaborative learning strategy, where multiple distributed nodes support the training of a model for a shared task [184]. Federated Learning [113, 154], in particular, has emerged as a training paradigm where each node trains a copy of a shared model on its private data and sends the local updates to a central server, where model parameters are tuned based on aggregated local updates. However, aggregating gradients or weights from multiple nodes does not deal with the non-i.i.d. nature of distributed data. Furthermore, gradient integration raises privacy issues as training data might be reconstructed, to a certain degree, starting from the shared gradients as demonstrated in [51, 192, 197].

In this work, we propose a generative approach where each distributed node generates, and shares, a synthetic version of its own data through manipulation and aggregation of latent spaces learned by a Generative Adversarial Network (GAN). In particular, our synthetic samples are drawn from the same distribution as the original ones, but are designed to prevent the inclusion of patient-specific visual

patterns. Sharing the manipulated images, rather than the generation model, prevents the reconstructions of real data through attacks to the model and circumvents the gradient/weight aggregation problem.

We tested our approach on the task of tuberculosis classification from X-ray images of two different datasets, namely, the Montgomery County X-ray Set and Shenzhen Hospital X-ray Set [22, 71, 70]. Our experiments simulate a multi-node multimodal data scenario, where each dataset is located on a different node. It achieves 75% and 60% in classification accuracy on the Shenzhen and the Montgomery datasets, respectively, whereas standard centralized training on the dataset union (i.e., not in a federated learning setting) yields 78% and 43%. The capabilities of our approach to synthesize images visually distant from the real ones are measured quantitatively by evaluating LPIPS (Learned Perceptual Image Patch Similarity) distance [190] between real images and samples generated through latent space optimization on a standard (non-privacy-preserving) GAN and by the proposed approach. Qualitatively, we also show several examples of generated images with corresponding closest match in the real dataset, demonstrating significant differences that prevent tracing back to the original real distribution.

6.2 Related Work

Federated learning (FL) embraces a family of privacy-preserving distributed learning strategies that allow nodes to keep training data private, while supporting the creation of a shared model. Typically, a central server sends a model to a set of client nodes; local model

updates are aggregated by the server, which sends the new model to the clients in an iterative process. In FedAvg [113], the server computes model averaging combining local stochastic gradient descent updates of each client. FedProx [98] is a generalization and re-parametrization of FedAvg proving theoretically convergence guarantee when training over non-identical distributed data (statistical heterogeneity). FedMA [171] builds a shared global in a layer-wise manner by matching and averaging hidden elements with similar feature extraction signatures. All these methods attempt to train a central model using the gradients gathered from multiple models trained on local private data.

FL particularly suits medical field applications, where data privacy is a critical concern. Li et al. [99] present the first FL system for medical image analysis, employing FedAvg and differential privacy [1] for brain tumor segmentation. Roy et al. [146] also apply FL for whole-brain segmentation in MRI. Recently, several other collaborative learning methods [37, 46, 130] have been proposed, especially because of the emergency need raised by the COVID-19 pandemic, in order to harness multiple data sources to promptly react to emergency scenarios. However, gradient aggregation does not seem to guarantee the required level of data privacy, as it has been demonstrated that network inputs can be recovered from gradient updates [51, 176, 197]. Differential privacy [1, 56, 93] attempts to reduce this issue by obfuscating gradients through noise. Zhu et al. [197], for instance, add Gaussian/Laplacian noise to gradients and compress the model with gradient pruning. However, adding noise to the gradients significantly compromises model's performance.

In this work, we tackle the problem of federated learning from a data-

perspective: rather than sharing weights/updates, which can be attacked, we share a synthetic version of private data — generated through a GAN — that retains visual content to support distributed training, but improves privacy by hiding specific visual patterns of patients. GANs have been also employed in federated learning regime, but always in the view of aggregating parameters to create a general model. In GS-WGAN [27], a gradient-sanitized Wasserstein GAN improves differential privacy, by carefully distorting gradient information in a way that reduces loss of information and generates more informative samples. Federated CycleGAN [161] is designed to perform unsupervised image translation; however, they still share local gradients, which may introduce the above privacy concerns. FedDPGAN [189] designs a distributed DPGAN [178] trained in a FL framework, to train models for COVID-19 diagnosis from chest X-ray images, without data sharing. In [136], the authors propose a framework to extend a large family of GANs to a FL setting utilizing a centralized adversary.

6.3 Method

6.3.1 Overview

In our approach, shown in Fig. 6.1, a set of distributed nodes create synthetic images and share them with a central node, where a model is trained using the received data. Specifically, each node trains a GAN to transform its own private dataset into a privacy-preserved one where patient information leak is minimized. The visual features of the privacy-preserved dataset still come from the same distribution

of the real private one (as per GAN training) in order to support the training of the centralized model. Although we do not perform a

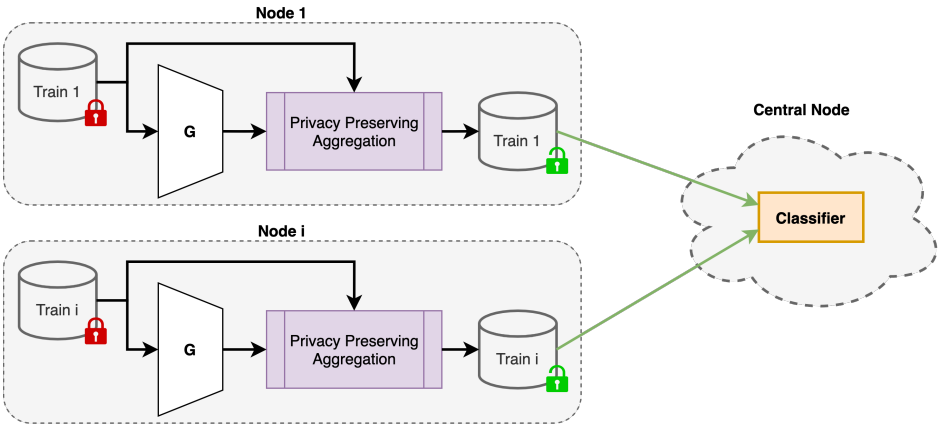


Figure 6.1: *The proposed federated learning framework*

formal security analysis of our approach, for the sake of readability we will refer to it as "privacy-preserving", to distinguish it from the cases where no precaution is taken to prevent patient information leak in the sharing and learning process (referred to as "non privacy-preserving").

6.3.2 Generative Adversarial Network

Generative Adversarial Networks (GANs) [52] consist of two networks, a generator model and a discriminator model: the former is trained to generate realistic images, while the latter is trained to distinguish between real and synthetic samples. In the conditional settings, where the generation process is controlled by a label to synthesize samples for a specific class, the two models are alternately trained to minimize

the following losses, respectively:

$$L_D = \mathbb{E}_{x,y}[\log(D(x,y))] + \mathbb{E}_{z,y}[\log(1 - D(G(z,y),y))] \tag{6.1}$$

$$L_G = \mathbb{E}_{z,y}[\log(D(G(z,y),y))] \tag{6.2}$$

where (x, y) is sampled from the real data distribution $\mathcal{D}$, z is sampled from a latent distribution $\mathcal{Z}$ (mapped by generator G to the real distribution for class y) and D is the discriminator model that predicts the likelihood of the input being real, given the target label. During training, the better D becomes at recognizing fake samples, the more G has to improve its generation capabilities, thus increasing the realism of synthetic data.

In this work, our GAN architecture is based on StyleGAN2 [78], where an auxiliary network maps a class-conditioned latent vector z to an intermediate latent vector $w \in \mathcal{W}$, which helps to improve generation quality and simplifies the projection of real images in $\mathcal{D}$ to the latent space $\mathcal{W}$. Indeed, given a real image x of class y, it is then possible to find an intermediate latent point $\hat{w}$ such that $G(\hat{w}) \approx x$, by optimizing the LPIPS distance loss [190] between x and $G(\hat{w})$ with respect to $\hat{w}$, which measures the similarity of activations by a pre-trained model. Of course, this projection property negatively affects the sought privacy in FL, as the generated synthetic distribution may contain visual patterns highly similar to those of the original samples.

6.3.3 Privacy-Preserving Aggregation

To address the privacy limitation of existing GAN methods, we propose a *Privacy-Preserving Aggregation* strategy (shown in Fig. 6.2) injected in the GAN training during data generation to encourage

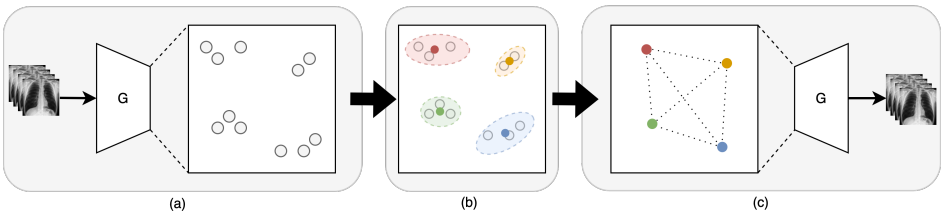


Figure 6.2: *Privacy Preserving Aggregation: a) a generator G is trained for each node using its own private dataset. Training images are then projected in the generator latent space; b) projected latent vectors are clustered through spectral clustering, based on pairwise LPIPS distance between corresponding images; c) linear interpolation among cluster centroids produces new latent vectors, which are used to generate synthetic samples that are sent to the central node.*

privacy. Let $\hat{W} = \{\hat{w_1}, \hat{w_2}, ..., \hat{w_N}\}$ be a set of points obtained by projecting N images onto the GAN latent space, for a given dataset class. We carry out spectral clustering [125] based on LPIPS distances between the images corresponding to $\hat{W}$ projections. Cluster centroids $\hat{W^c} = \{\hat{w_1^c}, \hat{w_2^c}, ..., \hat{w_M^c}\}$, representing latent aggregations with similar visual features in terms of LPIPS distance, are then employed as a starting point for data synthesis. Working with centroids allows us to capture shared patterns between dataset samples while improving privacy, since the resulting latent vectors cannot be traced back to specific patients. To create enough synthetic samples to allow model's training, we then carry out an augmentation procedure based on linearly interpolating the $\hat{W^c}$ centroids in the latent space and generating training samples using points along the trajectories between them. This is also beneficial for increasing dataset variability,

as it allows to produce samples that combine patterns of groups of patients (e.g., interpolating clusters with lesions on left/right lung may produce synthetic images with lesions on both lungs), leading to better generalization capabilities. Note that clustering and interpolation are carried out independently for each dataset class, by exploiting the conditional generation capabilities of the generator. This ensures that sampled latent vectors are assigned a well-defined label, making the corresponding synthetic images suitable for training the central node classifier. Clusters with only one sample are discarded in the process.

6.4 Experiments and Results

We test the proposed approach on the task of tuberculosis classification from X-ray images in a non-i.i.d. federated learning setting, where different datasets are used for each node, to simulate a more realistic training scenario. Each node generates synthetic X-ray images by applying our aggregation approach on its private dataset; images generated by each node are shared with a central node and used to train a classification model.

6.4.1 Datasets and training procedure

We employ the Montgomery County X-ray Set and the Shenzhen Hospital X-ray set[1] [22, 71, 70]. The Montgomery Set contains 138 frontal chest X-ray images (80 negatives and 58 positives), captured with a Eureka stationary machine (CR) at 4020×4892 or 4892×4020 pixel

[1]This dataset was released by National Library of Medicine, National Institute Of health, Bethesda, USA

resolution. The Shenzhen dataset was collected using a Philips DR Digital Diagnostic system. It includes 662 frontal chest X-ray images (326 negatives and 336 positives), with a variable resolution of approximately 3000×3000 pixels. In our federated learning setting, each dataset is associated to a node. We employ 80% of each dataset to train a GAN and generate synthetic images using the proposed approach. The remaining 20% of each dataset is used for testing the model trained on the central node. Test labels are balanced: 65 positives and 65 negatives on the Shenzhen dataset, and 15 positives and 15 negatives on the Montgomery dataset.

We use StyleGAN2-ADA [77] for image generation on each node, because of its suitability in low-data regimes and its intrinsic latent projection mechanism. GANs are trained in a label-conditioned setting and yield a Fréchet inception distance (FID) of 21.36 and 55.38 on the Shenzhen and Montgomery datasets, respectively. Latent space projection is carried out as in [78] for 500 iterations. Spectral clustering is carried out using 20 clusters on the Shenzhen Dataset and 10 on Montgomery one, due to the difference in sizes. Centroid interpolation computes 9 intermediate points for each pair of centroids. The resulting synthetic datasets include 1,730 samples per class on Shenzen and 415 samples per class on Montgomery. On the central node, we use a ResNet-50 classifier, trained by minimizing a cross-entropy loss with mini-batch gradient descent using the Adam optimizer for a total of 1,000 epochs; mini-batch size is set to 64 and the learning rate is 10^{-6}. All images are resized to 256×256, and data augmentation is carried out with random horizontal flip and random 90-degree rotations. Experiments are performed on an NVIDIA GeForce RTX 3090, using PyTorch.

6.4.2 Experimental Results

We evaluate the performance of our approach by considering three different data usage scenarios:

1. **Real data:** the central server trains a classifier on the original joint dataset using images of all nodes (this is the standard supervised centralized setting).

2. **Synthetic (non privacy-preserving) data:** each node generates a synthetic training set by sampling from a GAN trained on the real data; synthetic samples are then used to train on the central server. No privacy-preserving mechanism is enforced: sampled images are drawn from the original distribution as learned by the GAN.

3. **Synthetic privacy-preserving data:** the training set for the central server is created by employing our privacy-preserving generation procedure (see Sect. 6.3.3).

Tab. 6.1 reports the test accuracy on each dataset under the above three scenarios. On the Shenzhen dataset, our approach is close to centralized training using all data, respectively 0.75 and 0.78 classification accuracy. Interestingly, the non-privacy-preserving synthetic setting achieves even higher performance, which is explained by the larger number of training samples (662 real samples in Shenzhen, compared to 3,460 synthetic samples), confirming that sample synthesis helps making up for data scarcity – although in this case no precautions are taken to improve privacy. This phenomenon is even more evident on the smaller Montgomery dataset (138 samples), where the

usage of synthetic data yields significantly improved accuracy (0.43 on the original dataset vs 0.60 on the synthetic one).

Dataset	Training data	Accuracy
Shenzhen	Real	0.78
	Synthetic (non privacy-preserving)	0.82
	Synthetic (privacy-preserving)	0.75
Montgomery	Real	0.43
	Synthetic (non privacy-preserving)	0.60
	Synthetic (privacy-preserving)	0.60

Table 6.1: *Classification accuracy on the test set of each dataset, in different training scenarios.*

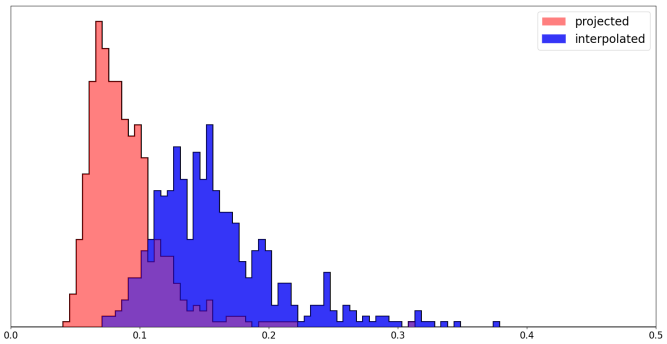


Figure 6.3: *In red, LPIPS distance histogram between real images and the corresponding images obtained through latent space projection. In blue, LPIPS distance histogram between real images and the closest images generated with the proposed approach.*

Privacy-preserving capabilities of the proposed approach are measured quantitatively by computing the LPIPS distance between real

training images and a) their projected counterparts using StyleGAN2, and b) the most similar samples from the pool of images generated by our strategy. Ideally, we would expect that, when using a standard StyleGAN2 network, the latent projection procedure should be able to recover an image that the model has used at training time — which is undesirable, since knowledge of the model would allow an attacker to reconstruct original samples; we also expect that images synthesized through generative aggregation should be significantly dissimilar to any real sample. Indeed, LPIPS distance histograms in Fig. 6.3 show that a distribution shift can be observed between the two sets of measured distances: latent space projection of real images tends to produce samples with significantly smaller distances than those obtained with most similar synthetic images generated by our approach. This

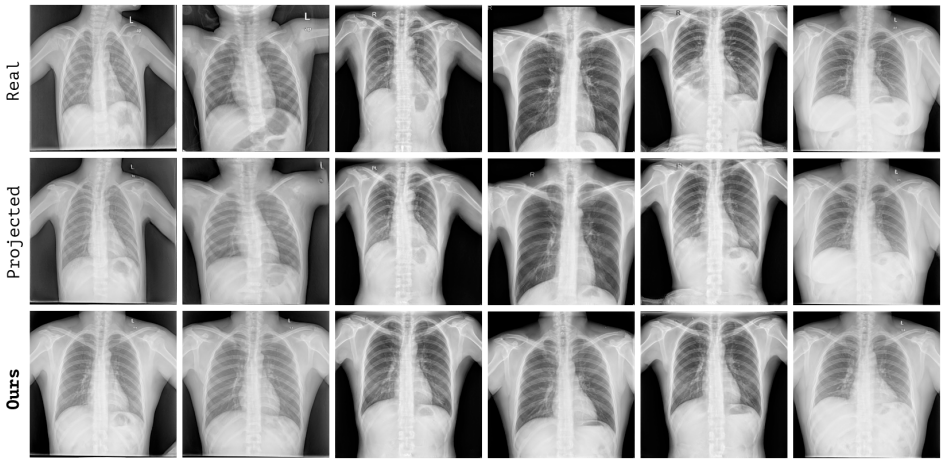


Figure 6.4: *Top: real images from Shenzhen Dataset; middle: images generated by latent projection; bottom: most similar synthetic images obtained with the proposed method.*

effect can be also appreciated qualitatively in the samples reported in Fig. 6.4, showing six images randomly sampled from the Shenzhen Dataset (top row) compared to their projection in the generator latent space (middle row) and the closest image in the aggregated dataset (bottom row).

6.5 Discussion

In this study we propose a synthetic data aggregation approach as an alternative to classic federated learning with gradient aggregation, which is subject to privacy concerns due to the risk of reconstructing the original inputs. Rather than training a central model by aggregating gradients from individual nodes, we propose to generate a synthetic dataset for each node and use the union of these datasets to train the central model. We tested our approach in a realistic scenario, using two X-Rays datasets for Tuberculosis classification, simulating a system with two nodes and non-i.i.d. data. The results demonstrated the validity of our approach, which obtains comparable performance to those obtained when training on the union of all datasets. Moreover, we showed, both quantitatively and qualitatively, that the generated images exhibit visual features typical of the original data, while being significantly different from any actual real image, thus preventing to trace them back to individual patients. Still, this is a preliminary work: future developments will investigate its validity in the presence of more nodes or in the presence of i.i.d. distributions.

6.6 Publications

Pennisi, M., Proietto Salanitri, F., Palazzo, S., Pino, C., Rundo, F., Giordano, D., & Spampinato, C. (2022, September). GAN Latent Space Manipulation and Aggregation for Federated Learning in Medical Imaging. In International Workshop on Distributed, Collaborative, and Federated Learning (pp. 68-78). Cham: Springer Nature Switzerland.

CHAPTER

SEVEN

A PRIVACY-PRESERVING WALK IN THE LATENT SPACE OF GENERATIVE MODELS FOR MEDICAL APPLICATIONS

Building on the foundation laid in the previous chapter, we now take one step deeper into the intricacies of navigating latent spaces. Although linear interpolation offers a simple method for generating synthetic samples, it poses the risk of producing data points that are very similar to real ones, potentially compromising privacy. To solve this problem, in this chapter we introduce a new strategy to navigate latent spaces, ensuring the generation of diverse and privacy-preserving synthetic samples. Using an auxiliary identity classifier, we learn a non-linear pathway between points in latent spaces, reducing the chances of inadvertently replicating real samples. Through rigorous empirical evaluations, we demonstrate the superiority of our method over linear interpolation in terms of security. Moreover, its applicability goes be-

yond GANs. Indeed, any network architecture with a navigable latent space can benefit from our method, underscoring its versatility and wide relevance.

7.1 Motivation

The success of deep learning for medical data analysis has demonstrated its potential to become a core component of future diagnosis and treatment methodologies. However, in spite of the efforts devoted to improve data efficiency [87], the most effective models still rely on large datasets to achieve high accuracy and generalizability. An effective strategy for obtaining large and diverse datasets is to leverage collaborative efforts based on data sharing principles; however, current privacy regulations often hinder this possibility. As a consequence, small private datasets are still used for training models that tend to overfit, introduce biases and generalize badly on other data sources addressing the same task [187]. As a mitigation measure, generative adversarial networks (GANs) have been proposed to synthesize highly-realistic images, extending existing datasets to include more (and more diverse) examples [131], but they pose privacy concerns as real samples may be encoded in the latent space. *K-same* techniques [73, 114] attempt to reduce this risk by following the *k-anonymity* principle [165] and replacing real samples with synthetic aggregations of groups of k samples. As a downside, these methods reduce the dataset size by a factor of k, which greatly limits their applicability.

To address this issue, we propose an approach, complementing k-

same techniques, for generating an extended variant of a dataset by sampling a privacy-preserving walk in the GAN latent space. Our method directly optimizes latent points, through the use of an *auxiliary identity classifier*, which informs on the similarity between training samples and synthetic images corresponding to candidate latent points. This optimized navigation meets three key properties of data synthesis for medical applications: 1) *equidistance*, encouraging the generation of diverse realistic samples suitable for model training; 2) *privacy preservation*, limiting the possibility of recovering original samples, and, 3) *class-consistency*, ensuring that synthesized samples contain meaningful clinical information. To demonstrate the generalization capabilities of our approach, we experimentally evaluate its performance on two medical image tasks, namely, tuberculosis classification using the Shenzhen Hospital X-ray dataseet [22, 71, 70] and diabetic retinopathy classification on the APTOS dataset [80]. On both tasks, our approach yields classification performance comparable to training with real samples and significantly better than existing *k-same* techniques such as k-SALSA [73], while keeping the same robustness to membership inference attacks.

Contributions: 1) We present a latent space navigation approach that provides a large amount of diverse and meaningful images for model training; 2) We devise an optimization strategy of latent walks that enforces privacy; 3) We carry out several experiments on two medical tasks, demonstrating the effectiveness of our generative approach on model's training and its guarantees to privacy preservation.

7.2 Related Work

Conventional methods to protect identity in private images have involved modifying pixels through techniques like masking, blurring, and pixelation [138, 15]. However, these methods have been found to be insufficient for providing adequate privacy protection [4]. As an alternative, GANs have been increasingly explored to synthesize high-quality images that preserve information from the original distribution, while disentangling and removing privacy-sensitive components [182, 185]. However, these methods have been mainly devised for face images and cannot be directly applicable to medical images, since there is no clear distinction between identity and non-identity features [73].

Recent approaches, based on the *k-same* framework [114], employ GANs to synthesize clinically-valid medical images principle by aggregating groups of real samples into synthetic privacy-preserving examples [73, 132]. In particular, k-SALSA [73] uses GANs for generating retinal fundus images by proposing a local style alignment strategy to retain visual patterns of the original data. The main downside of these methods is that, in the strive to ensure privacy preservation following the *k-anonymity* [165] principle, they significantly reduce the size of the original dataset.

Our latent navigation strategy complements these approaches by synthesizing large and diverse samples, suitable for downstream tasks. In general, latent space navigation in GANs manipulates the latent vectors to create new images with specific characteristics. While many works have explored this concept to control semantic attributes of generated samples [78, 17], to the best of our knowledge, no method has tackled the problem from a privacy-preservation standpoint, especially

on a critical domain such as medical image analysis.

7.3 Method

The proposed **P**rivacy-preserving **LA**tent **N**avigation (**PLAN**) strategy envisages three separate stages: 1) GAN training using real samples; 2) latent privacy-preserving trajectory optimization in the GAN latent space; 3) privacy-preserving dataset synthesis for downstream applications. Fig. 7.1 illustrates the overall framework and provides a conceptual interpretation of the optimization objectives.

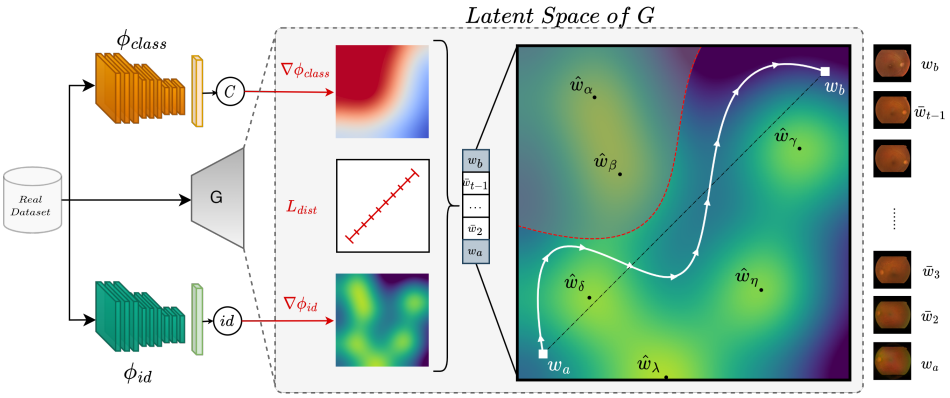


Figure 7.1: Overview of the PLAN approach. *Using real samples, we train a GAN, an* identity classifier ϕ_{id} *and an* auxiliary classifier ϕ_{class}. *Given two arbitrary latent points,* $\mathbf{w}_a$ *and* $\mathbf{w}_b$, *PLAN employs* ϕ_{id} *and* ϕ_{class} *to gain information on latent space structure and generate a privacy-preserving navigation path (right image), from which synthetic samples can be sampled (far right images, zoom-in for details).*

Formally, given a GAN generator $G : \mathcal{W} \to \mathcal{X}$, we aim to navigate its latent space $\mathcal{W}$ to generate samples in image space $\mathcal{X}$ in a privacy-preserving way, i.e., avoiding latent regions where real images might be embedded. The expected result is a synthetic dataset that is safe to share, while still including consistent clinical features to be used by downstream tasks (e.g., classification).

Our objective is to find a set of latent points $\bar{\mathcal{W}} \subset \mathcal{W}$

from which it is safe to synthesize samples that are significantly different from training points: given the training set $\hat{\mathcal{X}} \subset \mathcal{X}$ and a metric d on $\mathcal{X}$, we want to find $\bar{\mathcal{W}}$ such that $\min_{\mathbf{x} \in \hat{\mathcal{X}}} d\left(G\left(\bar{\mathbf{w}}\right), \mathbf{x}\right) > \delta$, $\forall \bar{\mathbf{w}} \in \bar{\mathcal{W}}$, for a sufficiently large δ. Manually searching for $\bar{\mathcal{W}}$, however, may be unfeasible: generating a large $\bar{\mathcal{W}}$ is computationally expensive, as it requires at least $|\bar{\mathcal{W}}|$ forward passes through G, and each synthesized image should be compared to all training images;

moreover, randomly sampled latent points might not satisfy the above condition.

To account for latent structure, one could explicitly sample away from latent vectors corresponding to real data. Let $\hat{\mathcal{W}}_i \subset \mathcal{W}$ be the set of latent vectors that produce near-duplicates of a training sample $\mathbf{x}_i \in \mathcal{X}$, such that $G(\hat{\mathbf{w}}_i) \approx \mathbf{x}_i$, $\forall \hat{\mathbf{w}}_i \in \hat{\mathcal{W}}_i$. We can thus define $\hat{\mathcal{W}} = \bigcup_{i=1}^{N} \hat{\mathcal{W}}_i$ as the set of latent points corresponding to all N samples of the training set: knowledge of $\hat{\mathcal{W}}$ can be used to move the above constraint from $\mathcal{X}$ to $\mathcal{W}$, by finding $\bar{\mathcal{W}}$ such that $\min_{\hat{\mathbf{w}} \in \hat{\mathcal{W}}} d\left(\bar{\mathbf{w}}, \hat{\mathbf{w}}\right) > \delta$, $\forall \bar{\mathbf{w}} \in \bar{\mathcal{W}}$. In practice, although $\hat{\mathcal{W}}_i$ can be approximated through latent space projection [78, 6] from multiple initialization points, its cardinality $|\hat{\mathcal{W}}_i|$ cannot be determined *a priori* as it is potentially unbounded.

From these limitations, we pose the search of seeking privacy-

preserving latent points as a trajectory optimization problem, constrained by a set of objectives that mitigate privacy risks and enforce sample variability and class consistency. Given two arbitrary latent points (e.g., provided by a *k-same* aggregation method), $\mathbf{w}_a, \mathbf{w}_b \in \mathcal{W}$, we aim at finding a latent trajectory $\bar{\mathbf{W}}_T = [\mathbf{w}_a = \bar{\mathbf{w}}_1, \bar{\mathbf{w}}_2, \ldots, \bar{\mathbf{w}}_{T-1}, \mathbf{w}_b = \bar{\mathbf{w}}_T]$ that traverses the latent space from $\mathbf{w}_a$ to $\mathbf{w}_b$ in T steps, such that none of its points can be mapped to any training sample. We design our navigation strategy to satisfy three requirements, which are then translated into optimization objectives:

1. **Equidistance.** The distance between consecutive points in the latent trajectory should be approximately constant, to ensure sample diversity and mitigate mode collapse. We define the equidistance loss, $\mathcal{L}_{\text{dist}}$, as follows:

$$\mathcal{L}_{\text{dist}} = \sum_{i=1}^{T-1} \|\bar{\mathbf{w}}_i, \bar{\mathbf{w}}_{i+1}\|_2^2 \tag{7.1}$$

   where $\|\cdot\|_2$ is the L_2 norm. Note that without any additional constraint, $\mathcal{L}_{\text{dist}}$ converges to the trivial solution of linear interpolation, which

   gives no guarantee that the path will not contain points belonging to $\hat{\mathcal{W}}$.

2. **Privacy preservation.** To navigate away from latent regions corresponding to real samples, we employ an auxiliary network ϕ_{id}, trained on $\hat{\mathcal{X}}$ to perform *identity classification*. We then set the privacy preservation constraint by imposing that a sampled

trajectory must maximize the uncertainty of ϕ_{id}, thus avoiding samples that could be recognizable from the training set. Assuming ϕ_{id} to be a neural network with as many outputs as the number of identities in the original dataset, this constraint can be mapped to a privacy-preserving loss, $\mathcal{L}_{\text{id}}$, defined as the Kullback-Leibler divergence between the softmax probabilities of ϕ_{id} and the uniform distribution $\mathcal{U}$:

$$\mathcal{L}_{\text{id}} = \sum_{i=1}^{T} \text{KL}[\phi_{id}(G(\bar{\mathbf{w}}_i)) \parallel \mathcal{U}(1/n_{\text{id}})] \tag{7.2}$$

where n_{id} is the number of identities.

This loss converges towards points with enhanced privacy, on which a trained classifier is maximally uncertain.

3. **Class consistency.** The latent navigation strategy, besides being privacy-preserving, needs to retain discriminative features to support training of downstream tasks on the synthetic dataset. In the case of a downstream classification task, given $\mathbf{w}_a$ and $\mathbf{w}_b$ belonging to the same class, all points along a trajectory between $\mathbf{w}_a$ and $\mathbf{w}_b$ should exhibit the visual features of that specific class. Moreover, optimizing the constraints in Eq. 7.1 and Eq. 7.2 does not guarantee good visual quality, leading to privacy-preserving but useless synthetic samples. Thus, we add a third objective that enforces class-consistency on trajectory points. We employ an additional *auxiliary classification network* ϕ_{class}, trained to perform classification on the original dataset, to ensure that sampled latent points share the same visual properties (i.e., the same class) of $\mathbf{w}_a$ and $\mathbf{w}_b$. The corresponding

loss $\mathcal{L}_{\text{class}}$ is as follows:

$$\mathcal{L}_{\text{class}} = \sum_{i=1}^{T} \text{CE}\left[\phi_{\text{class}}(G(\bar{\mathbf{w}}_i)), y\right] \tag{7.3}$$

where CE is the cross-entropy between the predicted label for each sample and the target class label y.

Overall, the total loss for privacy-preserving latent navigation is obtained as:

$$\mathcal{L}_{\text{PLAN}} = \mathcal{L}_{dist} + \lambda_1 \mathcal{L}_{id} + \lambda_2 \mathcal{L}_{label} \tag{7.4}$$

where λ_1 and λ_2 weigh the three contributions.

In a practical application, we employ PLAN in conjunction with a privacy-preserving method that produces synthetic samples (e.g., a *k-same* approach). We then navigate the latent space between random pairs of such samples, and increase the size of the dataset while retaining privacy preservation. The resulting extended set is then used to train a *downstream classifier* ϕ_{down} on synthetic samples only. Overall, from an input set of N samples, we apply PLAN to $N/2$ random pairs, thus sampling $TN/2$ new points.

7.4 Experiments and Results

We demonstrate the effectiveness and privacy-preserving properties of our PLAN approach on two classification tasks, namely, tuberculosis classification and diabetic retinopathy (DR) classification.

7.4.1 Training and evaluation procedure

Data preparation. For tuberculosis classification, we employ the Shenzhen Hospital X-ray set[1] [22, 71, 70] that includes 662 frontal chest X-ray images (326 negatives and 336 positives). For diabetic retinopathy classification, we use the APTOS fundus image dataset [80] of retina images labeled by ophthalmologists with five grades of severity. We downsample it by randomly selecting 950 images, equally distributed among classes, to simulate a typical scenario with low data availability (as in medical applications), where GAN-based synthetic sampling, as a form of augmentation, is more needed. All images are resized to 256×256 and split into train, validation and test set with 70%, 10%, 20% proportions.

Baseline methods. We evaluate our approach from a privacy-preserving perspective and by its capability to support downstream classification tasks. For the former, given the lack of existing methods for privacy-preserving GAN latent navigation, we compare PLAN to standard linear interpolation. After assessing privacy-preserving performance, we measure the impact of our PLAN sampling strategy when combined to k-SALSA [73] and the latent cluster interpolation approach from [132] (LCI in the following) on the two considered tasks.

Implementation details. We employ StyleGAN2-ADA [77] as GAN model for all baselines, trained in a label-conditioned setting on the original training sets. For all classifiers (ϕ_{id}, ϕ_{class} and ϕ_{down}) we employ a ResNet-18 network [59]. Classifiers ϕ_{id} and ϕ_{class} are trained on the original training set, while ϕ_{down} (i.e., the task classifier, one

[1] This dataset was released by the National Library of Medicine, NIH, Bethesda, USA.

for each task) is trained on synthetic samples only. For ϕ_{id}, we apply standard data augmentation (e.g., horizontal flip, rotation) and add five GAN projections for each identity, to mitigate the domain shift between real and synthetic images. ϕ_{down} is trained with a learning rate of 0.001, a batch size of 32, for 200 (Shenzhen) and 500 (APTOS) epochs. Model selection is carried out at the best validation accuracy, and results are averaged over 5 runs. When applying PLAN on a pair of latent points, we initialize a trajectory of $T = 50$ points through linear interpolation, and optimize Eq. 7.4 for 100 steps using Adam with a learning rate of 0.1; λ_1 and λ_2 are set to 0.1 and 1, respectively. Experiments are performed on an NVIDIA RTX 3090.

7.4.2 Results

To measure the privacy-preserving properties of our approach, we employ the *membership inference attack* (MIA) [158], which attempts to predict if a sample was used in a classifier's training set. We use attacker model and settings defined in [123, 74], training the attacker on 30% of the training set (seen by PLAN through ϕ_{id} and ϕ_{class}) and 30% of the test set (unseen by PLAN); as a test set for MIA, we reserve 60% of the original test set, leaving 10% as a validation set to select the best attacker. Ideally, if the model preserves privacy, the attacker achieves chance performance (50%), showing inability to identify samples used for training. We also report the FID of the generated dataset, to measure its level of realism, and the mean of the minimum LPIPS [190] ("mmL" for short) distances between each generated sample and its closest real image, to measure how generated samples differ from real ones. We compare PLAN to a linear interpolation between arbitrary

pairs of start and end latent points, and compute the above measures on the images corresponding to the latent trajectories obtained by two approaches. We also report the results of the classifier trained on real data to provide additional bounds for both classification accuracy and privacy-preserving performance.

	Shenzhen				**Aptos**			
	Acc. (%)(↑)	MIA (↓)	FID (↓)	mmL (↑)	Acc. (%)(↑)	MIA (↓)	FID (↓)	mmL (↑)
Real	81.23±1.03	71.41±3.59	–	–	50.74±2.85	73.30±4.04	–	–
Linear	82.14±1.40	56.28±1.60	63.85	0.125	41.58±2.11	50.53±3.06	**85.17**	0.118
PLAN	**83.85**±1.33	**50.13**±3.99	**63.22**	**0.159**	**46.95**±3.06	**48.51**±2.85	90.81	**0.131**

Table 7.1: *Comparison between the downstream classifier (ϕ_{down}) model trained with real samples and those trained with synthetic samples generated from the linear path and privacy path, respectively.*

Results in Table 7.1 demonstrate that our approach performs similarly to training with real data, but with higher accuracy with respect to the linear baseline. Privacy-preserving results, measured through MIA and mmL, demonstrate the reliability of our PLAN strategy in removing sensitive information, reaching the ideal lower bound of MIA accuracy.

Fig. 7.2 shows how, for given start and end points, PLAN-generated samples keep high quality but differ significantly from real samples, while latent linear interpolation may lead to near-duplicates. This is confirmed by the higher LPIPS distance between generated samples and the most similar real samples for PLAN.
After verifying the generative and privacy-preserving capabilities of

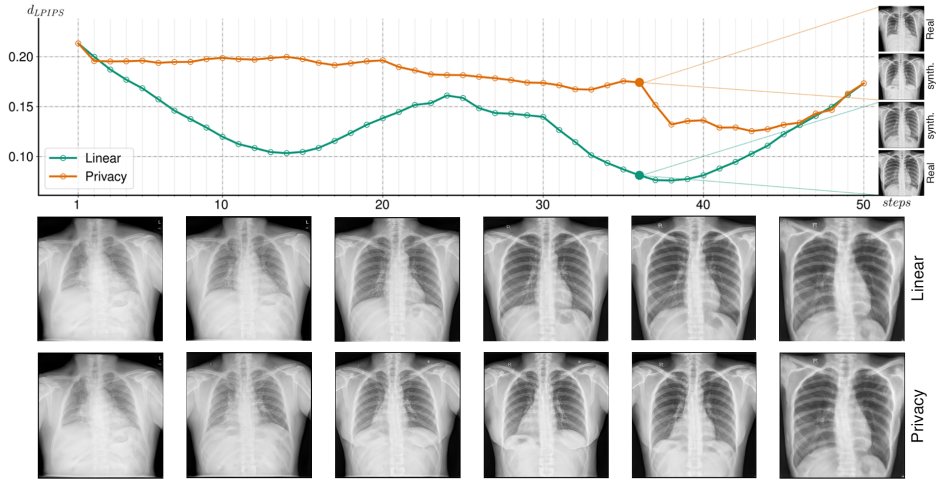


Figure 7.2: Linear vs PLAN navigation between two arbitrary points. *For each step of the latent trajectory, we compute the LPIPS distance between each synthetic sample and its closest real image. On the right, a qualitative comparison of images at step 35 and their closest real samples: the synthetic image obtained with PLAN differs significantly from its closest real sample; in linear interpolation, synthetic and real samples look similar. Bottom images show synthetic samples generated by linear interpolation and PLAN at the same steps (zoom-in for details).*

our approach, we evaluate its contribution to classification accuracy when combined with existing *k-same* methods, namely k-SALSA [73] and LCI [132]. Both methods apply latent clustering to synthesize a privacy-preserving dataset, but exhibit low performance transferability to classification tasks, due to the reduced size of the resulting synthetic dataset. We carry out these experiments on APTOS, using

$k = 5$ and $k = 10$, for comparison with [73][2]. Results are given in Table 7.2 and show how our PLAN strategy enhances performance of the two baseline methods, reaching performance similar to training the retinopathy classifier with real samples (i.e., 50.74 on real data vs 44.95 when LCI [132] is combined with PLAN) and much higher than the variants without PLAN. We also measured MIA accuracy between the variants with and without PLAN, and we did not observe significant change among the different configurations: accuracy was at the chance level in all cases, suggesting their privacy-preserving capability.

	k-SALSA [73]	**k-SALSA +PLAN**	LCI [132]	**LCI +PLAN**
k = 5	25.58±6.32	**36.59**±3.48	38.74±4.51	**43.16**±2.71
k = 10	27.47±3.42	**34.21**±1.62	36.42±3.77	**44.95**±1.61

Table 7.2: *Impact of our navigation strategy on k-same methods on the APTOS dataset. Performance are reported in terms of accuracy.*

7.5 Discussion

We presented PLAN, a latent space navigation strategy designed to reduce privacy risks when using GANs for training models on synthetic data. Experimental results, on two medical image analysis tasks, demonstrate how PLAN is robust to membership inference attacks

[2]Values of k smaller than 5 led to vulnerabilities to MIA on APTOS, as shown in [73].

while effectively supporting model training with performance comparable to training on real data. Furthermore, when PLAN is combined with state-of-the-art *k-anonymity* methods, we observe a mitigation of performance drop while maintaining privacy-preservation properties. Future research directions will address the scalability of the method to large datasets with a high number of identities, as well as learning latent trajectories with arbitrary length to maximize privacy-preserving and augmentation properties of the synthetic datasets.

7.6 Publications

Pennisi, M., Proietto Salanitri, F. , Bellitto, G., Palazzo, S., Bagci, U., & Spampinato, C. (2023). A Privacy-Preserving Walk in the Latent Space of Generative Models for Medical Applications. In Medical Image Computing and Computer-Assisted Intervention - MICCAI 2023: 26th International Conference, Vancouver, Canada, October 8-12, 2023, Proceedings. Springer International Publishing.

CHAPTER

EIGHT

FEDER: FEDERATED LEARNING THROUGH EXPERIENCE REPLAY AND PRIVACY-PRESERVING DATA SYNTHESIS

Approaching the end of our journey, this chapter attempts to bring together all the concepts we have explored through our dissertation. We propose a comprehensive framework for distributed federated learning. The core of this framework is a privacy-preserving GAN, trained to generate data that can be shared safely, facilitating the exchange of knowledge among federation participants. Inspired by the principles of Continual Learning, we incorporate the concept of Experience Replay to manage the distribution shifts present among the federation nodes. This strategy, not only ensures patterns that can generalize across different datasets, but also supports privacy. The real proof of the effectiveness of our approach, is its impressive performance in real medical scenarios, which underscores its practical importance.

8.1 Motivation

Recent advances of deep learning in the medical imaging domain have shown that, while data-driven approaches represent a powerful and promising tool for supporting physicians' decisions, the availability of large-scale datasets plays a key role in the effectiveness and reliability of the resulting models [69, 174, 32]. However, the curation of large medical imaging datasets is a complex task: data collection at single institutions is relatively slow and the integration of historical data may require significant efforts to deal with different data formats, storage modalities and acquisition devices; moreover, medical institutions are often reluctant to share their own data, due to privacy concerns. As a consequence, this affects the quality, reliability and generalizability of models trained on local datasets, which unavoidably suffer from bias and overfitting issues, reducing the ability to address future data distribution shifts [187]. In order to overcome the lack of large-scale datasets, methodological solutions can be adopted: in particular, federated learning [184] encompasses a family of strategies for distributed training over multiple nodes, each with its own private dataset, which typically communicate with a central node by sending local model updates, used to train the main model. In this scenario, no data is explicitly shared between nodes, thus addressing the required privacy issues. However, this family of techniques generally performs well when dataset distributions are approximately *i.i.d.* and local gradients/models contribute to learning shared features: unfortunately, in practice this hypothesis rarely holds, due to differences in the acquisition and in the clinical nature of data collected by multiple institutions. Moreover, the presence of a central node, besides representing

a single point of failure, requires that all nodes trust it to correctly and fairly treat updates from all sources: indeed, privacy issues arise when transferring local updates to the "semi-honest" central node [45], which might attempt to reconstruct original inputs from gradients or parameter variations [197, 51, 192]. To address the above limitations, we present *FedER*, a federated learning approach that, leveraging experience replay from continual learning [137, 140, 141, 19] and generative models [52, 117, 78], proposes a principled way for training local models that approximately converge to the same decisions, without the need of a shared model architecture and of central coordination. *FedER* also enforces privacy preservation through the transmission of synthetic data generated in a way to obfuscate real data patterns.
Specifically, FedER's learning strategy envisages multiple nodes that initially train their local models and a GAN on their own datasets. The GAN will be used in order to generate a privacy-preserving synthesized version of the dataset (buffer). Once local training is completed in a node, its model and the "buffer" of generated synthetic data are sent to a random node of the network. The receiving node then adapts the incoming model using its own data and the received buffer data, in order to limit model's forgetting. Data privacy is ensured through a privacy-preserving generative adversarial network (GAN) that employs a specific loss designed to maximize the distance from real data, while keeping a high level of realism and — as importantly — clinically-consistent features, in order to allow models to be trained effectively.

FedER is tested on two tasks, simulating a *non-i.i.d.* medical scenario: 1) classification of tuberculosis from X-ray data, using Montgomery County and Shenzhen Hospital datasets [22, 71, 70],

and 2) melanoma classification using skin images of the ISIC 2019 dataset [33, 168, 31]. The experimental setting is specifically designed to emulate a realistic medical *non-i.i.d.* scenario, where each node in the federation uses its own dataset. This is in stark contrast with common procedures where non-*i.i.d.* distributions are simulated by splitting a single source dataset. Results show how our approach is able to reach performance similar to using centralized training on all real data together in a single node, while outperforming current state-of-the-art methods, such as FedAvg [113], FedProx [98] and FedBN [100]. Privacy-preserving capabilities are measured quantitatively by evaluating LPIPS distance [190] between real images and samples generated, respectively, through latent space optimization on a standard GAN and by the proposed approach. Qualitatively, we also show several examples of generated images with corresponding closest match in the real dataset, demonstrating significant differences that prevent tracing back to the original real distribution.

In summary, the overall contributions of the proposed work are the following:

- We propose a decentralized federated learning strategy, based on continual learning principles, designed for medical imaging data, which outperforms server-based federated learning approaches and yields performance similar to standard (non-federated) training settings. Furthermore, experience replay allows local node models to converge to the same decisions, thus making the whole approach behave similarly to server-based aggregation models.

- We propose a GAN-based privacy-preserving mechanism that

supports synthetic data sharing through a GAN-based technique designed to minimize patient information leak. This is different from most privacy-preserving techniques based on differential privacy, which degrades performance due to added noise.

- Most approaches for model aggregation in federated learning employ gradient/parameter averaging. These solutions completely neglect any similarity or dissimilarity between merged features, possibly resulting in interference that harm convergence. FedER, instead, takes feature semantics into account when merging models: if a node receives a model that extracts useful features for the local dataset, these can be readily employed and re-used, without the risk of randomly averaging them with other less important features. *FedER*, thus, surpasses the common and straightforward weight/gradient averaging paradigm, replacing it with a principled way for knowledge transfer, which relaxes two of the constraints of the leading federated learning approaches: the presence of a central node and model homogeneity.

8.2 Related Work

In a typical FL setting, a central server sends a model to a set of client nodes; each node fine-tunes the model on its own data, then sends local model updates back to the server; the server aggregates the updates by all nodes into the global model, which is sent back to nodes iteratively until convergence. Given the constraints existing in the medical domain, especially in terms of data sharing, it represents an

appropriate test-bench for federated learning methods [99, 146, 37, 46]. The most straightforward way to aggregate information from multiple nodes is through averaging local models of each client, as proposed in FedAvg [113] and FedProx [98]. However, statistical data heterogeneity is an issue as it may lead to catastrophic forgetting [76, 53]. FedCurv [157] addresses this limitation by adding a penalty term to the loss driving the local models to a shared optimum. FedMA [171] builds a shared global model in a layer-wise manner by matching and averaging hidden elements with similar feature extraction signatures. Our method differs from existing feature integration approaches in that, instead of averaging model updates or gradients, which can be subject to input reconstruction attacks [51, 178, 197], each node attempts to learn features that perform well on its own dataset while retaining knowledge from other nodes, in a more principled way than parameter averaging. The strategy of fitting the global model to local data is also sought by the recent federated *personalized methods*. FedBN [100], for instance, keeps batch normalization layers private, while other model parameters are aggregated by the central node.

However, the presence of a central node that aggregates local updates simplifies the communication protocol when the number of clients is very large (thousands or millions), but introduces several downsides: it represents a single point of failure; it can become a bottleneck when the number of clients increases [101]; in general, it may not always be available or desirable in collaborative learning scenarios [76]. In this work, we deal with *decentralized federated learning*, in which the central node is replaced by peer-to-peer communication between clients: there is no longer a global shared model as in standard FL, but the communication protocol is designed so that all local mod-

els approximately converge to the same solution. Decentralized learning is particularly suitable to application in the medical domain, where the number of nodes (i.e., institutions) is relatively low; however, research is still ongoing, and no effective solutions have been established. In [90], a Bayesian approach is proposed to learn a shared model over a graph of nodes, by aggregating information from local data with the model of each node's one-hop neighbors. A secure weight averaging algorithm is proposed in [177], where model parameters are not shared between nodes, but all converge to the same numerical values (with the disadvantages associated to parameter averaging with *non-i.i.d.* data distributions). Other approaches implement different communication strategies based on parameter sharing (e.g., decentralized variants on FedAvg [164, 113]). In general, many of the existing solutions do not target, nor are they tested on, the medical domains — most employ toy datasets, such as MNIST and CIFAR10. Two works, similar in the decentralized learning spirit to ours, are proposed in [146, 49], where use cases of decentralized and swarm learning for medical image segmentation are presented. However, like other approaches, they adopt simple parameter averaging to integrate features or predictions from multiple nodes.

8.3 Method

8.3.1 Overview

An overview of FedER is shown in Fig. 8.1. In this scenario, a *federation* consists of a set of N peer nodes, each owning a private dataset.

Before the decentralized training algorithm is started, each node

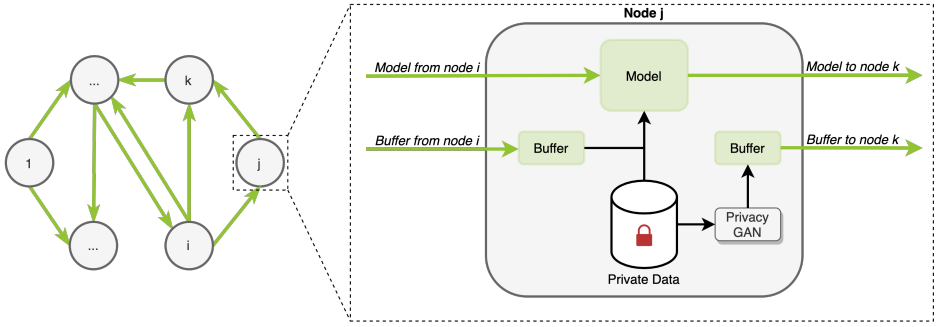


Figure 8.1: Overview of FedER learning strategy. *Each node initially trains a* privacy-preserving GAN*, that is used to sample synthetic data from the local distribution, without retaining features that may be used to identify patients. Then, each node iteratively receives the local model and a buffer of synthetic samples from a random node, and fine-tunes the received model on its own private data, using the buffer to prevent forgetting of previously-learned features.*

internally trains a *privacy-preserving generative adversarial network*, which is used to generate synthetic samples from its private data distribution. The training objective of the GAN is designed to enforce the constraint that sampled data do not include privacy-sensitive information, while maintaining the clinical features required for successful training.

At each round of decentralized training, each node receives a model and a set of synthetic samples — "buffer" — from a random node in the federation. The input model to the node is fine-tuned on both the private dataset and the buffer, in a way that is reminiscent of experience replay techniques in continual learning (e.g., [19]), in order to learn features that transfer between nodes and that can handle

non-*i.i.d.* distributions. At the end of each round (i.e., after performing several training iterations), the locally-trained model is sent to a randomly-chosen successor node together with a buffer of local synthetic samples, and the whole procedure is repeated.

In this work we specifically address the problem of federated learning for medical image classification; thus, the method is presented by considering this task, but the whole strategy can be applied to any other task without losing generalization.

8.3.2 Privacy-preserving GAN

In the proposed method, nodes exchange both models and data, implementing a knowledge transfer procedure based on experience replay (see Sect. 8.3.3 below). Of course, sharing real samples would go against federated learning policies; hence, exchanged samples are generated so that they are representative of the local data, while taking precautions against privacy violations — which may happen, for instance, if the generative model overfits the source dataset.

Formally, we assume that each node n_i, from a set of N nodes, owns a private dataset $\mathcal{D}_i = \{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \ldots, (\mathbf{x}_M, \mathbf{y}_M)\}$, where each $\mathbf{x}_j \in \mathcal{X}$ represents a sample in the dataset, and each $\mathbf{y}_j \in \mathcal{Y}$ represents the corresponding target[1]. The local dataset can then be used to train a conditional GAN [117], consisting of a generator G, that synthesizes samples for a given label by modeling $P(\mathbf{x}|\mathbf{y}, \mathbf{z}))$, where $\mathbf{z} \in \mathcal{Z}$ is a random vector sampled from the generation latent space, and a

[1]The proposed approach is task-agnostic, as long as it is possible to sample from the $\mathcal{Y}$ distribution. For simplicity, within the scope of this work, we will focus on classification tasks, and we will assume that targets are class labels.

discriminator D, which outputs the probability of an input sample being real, modeling $P\left(\text{real}|\mathbf{x},\mathbf{y}\right)$. The standard GAN formulation introduces a discrimination loss, which trains D to distinguish between real and synthetic samples:

$$\mathcal{L}_D = -\mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\log\left(D\left(\mathbf{x},\mathbf{y}\right)\right)\right] - \mathbb{E}_{\mathbf{z},\mathbf{y}}\left[\log\left(1 - D\left(G\left(\mathbf{z},\mathbf{y}\right),\mathbf{y}\right)\right)\right], \quad (8.1)$$

and a generation loss, which trains G to synthesize samples that appear realistic to the discriminator:

$$\mathcal{L}_G = -\mathbb{E}_{\mathbf{z},\mathbf{y}}\left[D\left(G\left(\mathbf{z},\mathbf{y}\right),\mathbf{y}\right)\right]. \quad (8.2)$$

While it has been theoretically proven that, at convergence, the distribution learned by the generator matches and generalizes from the original data distribution [53], unfortunately GAN architectures may be subject to training anomalies, including mode collapse and overfitting: as a consequence, the basic GAN formulation may lead to the generation of samples that are near duplicates of the original samples, which would be unacceptable in a federated learning scenario.

In order to mitigate this risk, we introduce a *privacy-preserving loss*, enforcing the generation of samples that do not retain potentially sensitive information, but still include features that are clinically relevant to the target $\mathbf{y}$ of the synthetic sample. In other words, if $\mathbf{y}$ encodes generic features for the diagnosis of a certain disease, we want the generator to learn how to synthesize samples conditioned by $\mathbf{y}$, that exhibit evidence of that disease but cannot be traced back to any of the dataset's samples of the same disease.

To do so, our privacy-preserving loss aims at penalizing the model proportionally to the similarity between pairs of real and synthetic samples. We measure "similarity" by means of the LPIPS metric [190],

which has been shown to capture perceptual similarity by calibrating the distance between feature vectors extracted from a pre-trained VGG model [159].

In practice, given a batch of real samples $\left\{\mathbf{x}_1^{(r)}, \mathbf{x}_2^{(r)}, \ldots, \mathbf{x}_b^{(r)}\right\}$ and a batch of synthetic samples $\left\{\mathbf{x}_1^{(s)}, \mathbf{x}_2^{(s)}, \ldots, \mathbf{x}_b^{(s)}\right\}$, the privacy-preserving loss term is computed as:

$$\mathcal{L}_{\text{PP}} = \frac{1}{b} \sum_{\mathbf{x}^{(r)}} \sum_{\mathbf{x}^{(s)}} d_L \left(\mathbf{x}^{(r)}, \mathbf{x}^{(s)}\right), \tag{8.3}$$

where d_L is the LPIPS distance defined as:

$$d_{\text{L}}(\mathbf{x}^{(r)}, \mathbf{x}^{(s)}) = \sum_i w_i \cdot \|\phi_i(\mathbf{x}^{(r)}) - \phi_i(\mathbf{x}^{(s)})\|_2 \tag{8.4}$$

where ϕ_i represents the feature maps extracted from the i^{th} layer of a deep neural network and w_i is a weight learned to reflect the perceptual importance of that layer. Note that, in this formulation, we ignore the $\mathbf{y}$ targets associated to each $\mathbf{x}$: we want to prevent the model from generating near-duplicates of real samples in general, regardless of class correspondence. Also, we intentionally employ a pairwise metric on samples, rather than an aggregated metric such as Frêchet Inception Distance [60], since we want to prevent similarity between samples, not between distributions, which would conflict with the GAN objective.

The resulting new loss for the Generator is a combination of Eq. 8.2 and Eq. 8.3:

$$\mathcal{L}_{G\text{-PP}} = \mathcal{L}_G - \alpha \mathcal{L}_{\text{PP}} \tag{8.5}$$

where $\mathcal{L}_{\text{PP}}$ is sign reversed as we want to maximize Eq. 8.3, while α is a hyperparameter used to balance the two terms.

The combined effect of the three loss terms — $\mathcal{L}_D$, $\mathcal{L}_G$, $\mathcal{L}_{\text{PP}}$ — pushes the generator to explore the sample space to match the dataset distribution, while "avoiding" latent space mappings that would project to actual real samples.

8.3.3 Federated learning with experience replay

Current approaches for federated learning are mostly based on parameter averaging (e.g., FedAvg), which is, however, a straightforward way to combine knowledge from multiple sources: feature locations are not aligned over different models and may be disrupted by updates, before slowly converging to consensus: hypothetically, two models could learn the same set of features at different locations of the same layer, to only have them cancel each other when averaging. In a decentralized scenario, this issue is even exacerbated, due to the lack of an entity that enforces global agreement on node features.

In our approach, we address this problem by taking inspiration from continual learning strategies [38] that learn how to perform a task with a non-*i.i.d* data stream without forgetting previously-learned knowledge: as a consequence, models are encouraged to reuse and adapt features so that they can equally serve the current and previous tasks. Analogously, in the federated learning setting, the objective is to train a global model trained on disjoint non-*i.i.d.* data distributions coming from different nodes.

Given these premises, we define a federated learning strategy where a node receives another node's model and surrogate data (generated through our privacy-preserving GAN) — the *"previous task"* — and fine-tunes that model on its own private date — the *"current task"* —

while using received synthetic data as a reference to what is necessary to retain/adapt from the knowledge learned by the previous node. The idea is to build for each node a model able to tackle its internal data while not forgetting about the data seen in previous nodes/iterations.

We first introduce the terminology used in the method's description. In our approach, we define a *set of N tasks* $\mathcal{T} = (T_1, T_2, \ldots, T_N)$, where T_i is the task to be solved within node n_i.

Definition 1. Task T_i aims at optimizing a model M_i, parameterized by $\boldsymbol{\theta}_i$, on dataset $\mathcal{D}_i$ residing on node n_i and that cannot be shared to other nodes.

Definition 2. A buffer $\mathcal{B}_i$ is a set of synthetic images, drawn from a latent space learned through a generative model $\mathcal{G}_i$ using data $\mathcal{D}_i$ available on node n_i.

Definition 3. Training is organized in parallel *rounds*. At the end of round r, each node n_i produces a model M_i^r trained on dataset $\mathcal{D}_i$ and on a buffer $\mathcal{B}_j$, received from another node n_j, to optimize an objective $\mathcal{L}$, i.e., to find $\arg\min_{\boldsymbol{\theta}_i^r} = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}_i\cup\mathcal{B}_j}[\mathcal{L}(M_i^r(\mathbf{x}, \boldsymbol{\theta}_i^r), \mathbf{y})]$. For each training round, all nodes in parallel share to/receive from other nodes, buffer of synthetic images and trained models.

In the following, we describe our method (whose graphical representation is given in Fig. 8.1) from the point of view of a single node n_j. At a given round r, training for node n_j can be seen as learning a new task T_j, from dataset $\mathcal{D}_j$, in a continual learning setting by finetuning the incoming model $\mathcal{M}_i^{r-1}$ (with parameters $\boldsymbol{\theta}_i^{r-1}$) on $\mathcal{D}_j$

and on the incoming buffer $\mathcal{B}_i$ in order to learn T_j while mitigating the forgetting of T_i. Thus, unlike other federated learning approaches, each node does not have its own local model: as the decentralized learning strategy proceeds, a node iteratively receives a model from another node and updates it with local information, while preserving previously-learned knowledge, before sending it to the next node. Formally, the loss function for model $\mathcal{M}_j^r$ in node n_j at round r is given as:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}_j^r) = \lambda\, \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}_j}[\mathcal{L}(M_j^r(\mathbf{x},\boldsymbol{\theta}_j^r),\mathbf{y})]+ \\ + (1-\lambda)\, \mathbb{E}_{(\mathbf{x}',\mathbf{y}')\sim\mathcal{B}_i}[\mathcal{L}(M_j^r(\mathbf{x}',\boldsymbol{\theta}_j^r),\mathbf{y}')]\end{aligned} \tag{8.6}$$

where λ controls the importance between real samples from the local dataset D_i and replayed synthetic samples from node n_i. Note that, for a given n_j, the predecessor node n_i is not fixed: in a practical asynchronous implementation, a node may receive a model and buffer from any random node in the federation at any time, using queues to handle incoming data.

After optimizing the $\mathcal{L}(\boldsymbol{\theta}_j^r)$ objective through mini-batch gradient descent for a certain number of training iterations, the resulting model $M_j^r(\boldsymbol{\theta}_j^r)$, with updated parameters $\boldsymbol{\theta}_j^r$, is sent to a random node n_k of the federation, along with a buffer $\mathcal{B}_j$ of locally-generated synthetic samples. The number of training rounds/iterations and the size of the buffer is discussed in the next section.

Then, the general federated model $\mathcal{M}$, after all training rounds, is given by the union of all the N node models, i.e., $\mathcal{M} = M_1 \cup M_2 \cup \cdots \cup M_N$. However, experimental results, reported below in Sect. 8.4, demonstrate that all models converge to similar decisions,

thus each node model can be considered as a general model for the entire network.

To ease the understanding of the whole training strategy we also report the algorithm pseudo-code in Alg. 1.

8.4 Experimental Results

We test FedER on two applications simulating real case scenarios with multiple centers holding, and not sharing, their own data: 1) tuberculosis classification from X-ray images using two different datasets, and 2) skin lesion classification with three different datasets. In this section we present the employed benchmarks, the training procedure and report the obtained results to demonstrate the advantages of the proposed approach w.r.t. the state-of-the-art.

8.4.1 Datasets

X-ray image datasets for tuberculosis classification. We assume two separate nodes in the federation: one with the Montgomery County X-ray set and another one with the Shenzhen Hospital X-ray set [22, 71, 70]. The Montgomery Set consists of 138 frontal chest X-ray images (80 negatives and 58 positives), captured with a Eureka stationary machine (CR) at 4020×4892 or 4892×4020 pixel resolution. The Shenzhen dataset was collected using a Philips DR Digital Diagnostic system. It includes 662 frontal chest X-ray images (326 negatives and 336 positives), with a variable resolution of approximately 3000×3000 pixels.

Skin lesion classification. We employ the ISIC 2019 challenge

Algorithm 1: FedER Learning Procedure

Notations *The N nodes are indexed by n_i; E is the number of local epochs for each round. R the total round of communications between nodes.*

Each node n_i contains:

$\mathcal{D}_i$ Private Dataset

$\mathcal{G}_i$ Generator (privacy-preserving) trained on D_i

$\mathcal{M}_i^r$ Model for node n_i at round r

$\mathcal{B}_i$ Synthetic data buffer sampled using G_i

`// Before Federated Training`

for *each node $n_i \in N$* **do**

  Train $\mathcal{G}_i$ on $\mathcal{D}_i$

  Generate Buffer $\mathcal{B}_i$ using $\mathcal{G}_i$

  Train $\mathcal{M}_i^0$ on $\mathcal{D}_i$

end

`// Federated Training`

for *each round $r = 1, 2, ..., R$* **do**

  for *each node $n_j \in N$ in parallel* **do**

    Send $\mathcal{M}_j^{r-1}$, $\mathcal{B}_j$ to a node $n_k \in \{N \setminus n_j\}$

    Receive $\mathcal{M}_i^{r-1}$, $\mathcal{B}_i$ from a node $n_i \in \{N \setminus n_j\}$

    $\mathcal{M}_j^r \leftarrow \mathcal{M}_i^{r-1}$

    Train $\mathcal{M}_j^r$ on $\{\mathcal{D}_j \cup \mathcal{B}_i\}$ for E epochs

  end

end

Rounds	Epochs	Tuberculosis		Melanoma		
		Shenzhen	**Montgomery**	**BCN**	**HAM**	**MSK4**
		Accuracy	Accuracy	Accuracy	Accuracy	Accuracy
10	1	82.39±6.91	56.13±3.03	76.73±2.07	82.24±4.01	67.93±4.84
	10	82.86±2.44	86.73±4.22	83.83±1.96	84.72±2.29	73.67±2.59
	100	83.56±1.72	90.79±3.92	85.51±1.85	88.65±1.12	71.81±2.04
100	1	83.31±2.59	88.71±3.82	78.94±2.55	87.34±1.62	72.07±3.45
	10	85.22±2.42	89.72±3.46	84.62±1.40	85.05±1.62	73.72±2.41
	100	87.10±2.31	91.50±2.60	86.06±0.96	89.26±1.11	72.41±1.53

Table 8.1: Rounds and epochs in FedER. *Results (mean ± standard deviation) obtained with 5-fold cross-validation. Buffer size = 512.*

dataset, which contains 25,331 skin images belonging to nine different diagnostic categories. In this case, we assume a federation with three nodes as data provided belongs to three different sources: 1) the BCN20000 [33] dataset, consisting of 19,424 images of skin lesions captured from 2010 to 2016 in the Hospital Clínic in Barcelona; 2) the HAM10000 dataset [168], which contains 10,015 skin images collected over a period of 20 years from two different sites, the Department of Dermatology at the Medical University of Vienna, Austria, and the skin cancer practice of Cliff Rosendahl in Queensland, Australia; 3) the MSK4 [31] dataset, which is anonymous and includes 819 samples. Among all skin lesion classes, we only consider the melanoma class, posing the problem as a binary classification task.

In all tasks and datasets we adopt 80% of the available data to train both the privacy-preserving GAN and the classification model, while

the remaining 20% of each dataset is used as test set. Test sets are also balanced w.r.t. the label to avoid performance biases due to class imbalance. For all tested federated methods (including state-of-the-art ones), model selection is carried out through with 5-fold cross-validation on the training set, as a grid search on number of training rounds, number of rounds per epoch and learning rate. For FedProx [98], we also include the μ hyperparameter.

8.4.2 Training procedure and metrics

Federated training

In all settings, we employ ResNet-18 as classification model, trained by minimizing the cross-entropy loss with mini-batch gradient descent using the Adam optimizer. Mini-batch size is set to 32 and 8 for the Shenzhen and Montgomery datasets, respectively, and to 64 for skin lesion datasets. The learning rate was found, through cross-validation, to be 10^{-4}. Data augmentation is carried out with random horizontal flip; for skin images we additionally apply random 90-degree rotations. All images are resized to 256×256. The ratio between real and synthetic samples controlled by λ in Eq. 8.6 is set to 0.5 for all experiments, i.e., each mini-batch is composed by the same quantity of real and synthetic images. This also ensures that our method performs the same number of optimization steps as other approaches that do not use any synthetic data.

The node federation is trained for R rounds. In our implementation, at each round nodes are randomly ordered to establish each node's predecessor and successor: given our focus on medical applica-

tions, we can assume that the number of nodes is low enough that synchronization is not an issue. However, asynchronicity can be achieved by assuming that nodes can store incoming data in a queue: if the distribution of successor nodes is uniform and computation times are similar for all nodes, this is on average equivalent to the synchronous case. The number of rounds R and epochs E for FedER on the tuberculosis and melanoma classification tasks are set both to 100, according the 5-fold cross-validation results shown in Table 8.1. Buffer size is set for all experiments to 512.

Method	**Tuberculosis**		
	Shenzhen	**Montgomery**	**Mean**
Standalone	82.31	90.00	86.16
Centralized training	82.77	77.67	80.22
Centralized training with synthetic data only	76.92	79.33	78.13
Centralized training with synthetic and real data	85.38	86.67	86.03
FedER (ours)	80.15	86.67	83.41

Table 8.2: Comparison between FedER and centralized baselines for Tuberculosis. *Results for FedER are obtained with a buffer size of 512, 100 rounds and 100 epochs per round.*

GAN training

We recall that GAN training is carried out **before** federated learning using training data only, while leaving out test samples, as mentioned in Sect. 8.4.1. Our privacy-preserving GAN employs StyleGAN2-ADA [79] as a backbone, because of its suitability in low-data regimes and its generation capabilities. Training is carried out in two steps:

Method	Melanoma			
	BCN	**HAM**	**MSK4**	**Mean**
Standalone	82.90	82.55	69.75	78.40
Centralized training	78.80	82.90	71.23	77.64
Centralized training with synthetic data only	60.71	61.09	61.23	61.01
Centralized training with synthetic and real data	81.53	80.44	73.46	78.48
FedER (ours)	82.11	84.58	68.40	78.36

Table 8.3: Comparison between FedER and centralized baselines for Melanoma. *Results for FedER are obtained with a buffer size of 512, 100 rounds and 100 epochs per round.*

1) the GAN is initially trained without any privacy-preserving loss to support learning of high-quality visual features; 2) afterwards, we enable privacy-preserving loss and fine-tune the model in order to limit the embedding of patient-specific patterns in the GAN latent space. For classification purposes, GANs are trained in a label-conditioned fashion with a mini-batch size of 32 and learning rate of 0.0025 for both the generator and the discriminator. Early-stopping criteria are based on the Frêchet Inception Distance (FID) [60] between real and synthetic distributions: in the first training step, we stop training if FID does not improve for 10,000 iterations; in the second training step, we employ a criterion which stops training if FID increases by a factor of 2.5 w.r.t. the value obtained in the first step. As for the α parameter in Eq. 8.3, we tested multiple values of α (0, 0.5, 1, 1.5, 2 and 3) and found that the value of 1 yields the best compromise between image generation quality and pairwise LPIPS distance [190] over all tested datasets. In order to quantitatively evaluate privacy

preservation, we also compute the average LPIPS distance between each real image and its closest synthetic sample by means of latent space projection (described in Sect. 8.4.4): the higher value of LPIPS, the lower the possibility to reconstruct real images from the generator.

	Dataset	FedER	Standalone
Tuberculosis	**Shenzhen**	80.54 ± 1.20	66.15 ± 22.84
	Montgomery	85.67 ± 2.36	70.00 ± 28.28
Melanoma	**BCN**	82.87 ± 1.22	65.06 ± 19.68
	HAM	84.45 ± 0.75	59.94 ± 20.47
	MSK4	67.78 ± 1.28	65.43 ± 5.05

***Table 8.4: Accuracy convergence among distributed node models**. Each local model is evaluated on all test sets of the federation in order to measure convergence and generalization (lower standard deviation corresponds to higher convergence).*

8.4.3 Federated learning performance

We first evaluate the performance (in terms of classification accuracy) of FedER in the non-*i.i.d.* setting, and compare it to several centralized baselines, namely:

- **Centralized training**: all datasets are merged in a single node where all training happens. In this setting, no federated learning constraints are applied.
- **Centralized training with synthetic data only**. In this setting, each node trains a privacy-preserving GAN model and

shares a synthetic version of its own data with the central node, where global training is performed. In this case, we aim to assess how much information is retained by synthetic data to support classification.

- **Centralized training with synthetic and real data.** This setting is a combination of the previous two: real and synthetic samples are centrally merged and used for training a global classifier. This scenario measures the contribution of synthetic data as a data augmentation approach.

We also compare FedER against standard training of the local node models, referred to as "Standalone" . Classification accuracy is computed using local node models on their own data. The results, reported in Table 8.2 and Table 8.3, show that standalone training appears to be the most favourable scenario. Centralized strategies perform generally worse than standalone training, because of the non-*i.i.d.* nature of the data. However, when the centralized approach is trained with original data augmented with synthetic samples, its classification accuracy is on par with the standalone training, possibly due to the learned generative latent spaces that likely tend to smooth different modes of non-*i.i.d.* data. FedER, instead, outperforms its centralized counterpart and yields slightly worse performance (1.5 percent points less) than standalone training. Although this may appear, at a first glance, as a shortcoming of FedER, we recall that in a federated learning scenario, we aim at building a model that, leveraging multiple data distributions present in the federation, may generalize better, thus addressing possible future data drifts. In order to assess the capabilities of the trained models to achieve such a generaliza-

tion, we measure the decision convergence by evaluating how a local node model performs on other node datasets. Results are in Table 8.4 and show a good average accuracy, with a low standard deviation, by FedER, indicating that each node model performs equally well on its own dataset and on the others (i.e., all node models converge to similar decisions). Conversely, standalone training yields significantly lower accuracy and higher standard deviation than ours, demonstrating to be an unsuitable strategy for the sought generalization properties.

	Shenzhen	**Montgomery**	**Mean**
FedAvg [113]	72.31	83.33	77.82
FedProx [98]	78.46	76.67	77.56
FedBN [100]	63.08	70.00	66.54
FedER (ours)	**80.15**	**86.67**	**83.41**

***Table 8.5: Comparison with state-of-the-art methods for Tuberculosis**. In bold, best accuracy values.*

	BCN	**HAM**	**MSK4**	**Mean**
FedAvg [113]	77.55	75.15	67.28	73.33
FedProx [98]	78.80	81.87	64.81	75.16
FedBN [100]	**82.19**	81.12	59.26	74.19
FedER (ours)	82.11	**84.58**	**68.40**	**78.36**

***Table 8.6: Comparison with state-of-the-art methods for Melanoma**. In bold, best accuracy values.*

Thus, Tables 8.2 and 8.3 show the performance obtained by each node model on its internal test data, while Table 8.4 shows, instead,

the performance obtained when each node model is tested again all other nodes' data. The latter results indicate that in FedER, any arbitrary node model can be used for the final evaluation, as all federation models converge to the same decisions. However, we further investigate whether building an ensemble of all node models yields better performance than using one arbitrary model. Results are given in Table 8.7 indeed showing higher accuracy by the ensemble. However, the models' ensemble leads to increased communication overhead (after training, all models have to be shared across the federation) and inference costs (each node needs to make a forward pass for all its available models to make the prediction). For this reason, the following experiments are carried out without using ensemble.

Method	**Tuberculosis**	**Melanoma**
No ensemble	83.41 ± 4.61	78.36 ± 8.72
Ensemble	$\mathbf{84.77 \pm 4.57}$	$\mathbf{80.35 \pm 9.42}$

Table 8.7: Accuracy performance with and without models' ensemble. *Results are computed by testing (first line) each node model with its own data and (second line) creating an ensemble and testing it on all nodes' data.*

We then compare our approach (without ensemble) to state-of-the-art federated learning approaches, namely: a) server-based federated methods, FedAvg [113] and FedProx [98], which have shown to perform generally better than decentralized methods [164, 90], and b) a personalized method, FedBN [100]. As already mentioned, to avoid

biased assessment, we use the official code repository[2] of FedBN [100] and hyper-parameter selection on the tested datasets was carried out through grid search on training rounds/epochs, learning rate and μ for FedProx [98] using 5-fold cross validation as for our approach. Results, for the tuberculosis and the melanoma tasks, are reported in Table 8.5 and Table 8.6 respectively, and show that FedER outperforms all methods under comparison. Interestingly, FedER learning strategy does better than: a) *server-based methods*, FedAvg [113] and FedProx [98], suggesting that experience replay is a more effective feature aggregation approach than naive parameter averaging; b) personalized methods, such as FedBN [100], which affects a limited aspect of feature representation (i.e., input layer distributions), while our approach adapts the entire model to local and remote tasks.

	Node Convergence	
Buffer	**Shenzhen**	**Montgomery**
0	70.62 ± 11.97	80.33 ± 10.84
256	80.46 ± 2.96	81.67 ± 4.24
512	80.54 ± 1.20	85.67 ± 2.36
1024	82.23 ± 1.31	86.00 ± 3.01
2048	82.08 ± 1.39	88.67 ± 2.97

Table 8.8: FedER classification accuracy w.r.t. buffer size. *Each local model is evaluated on all test sets of the federation in order to measure convergence and generalization (lower standard deviation corresponds to higher convergence).*

[2] https://github.com/med-air/FedBN

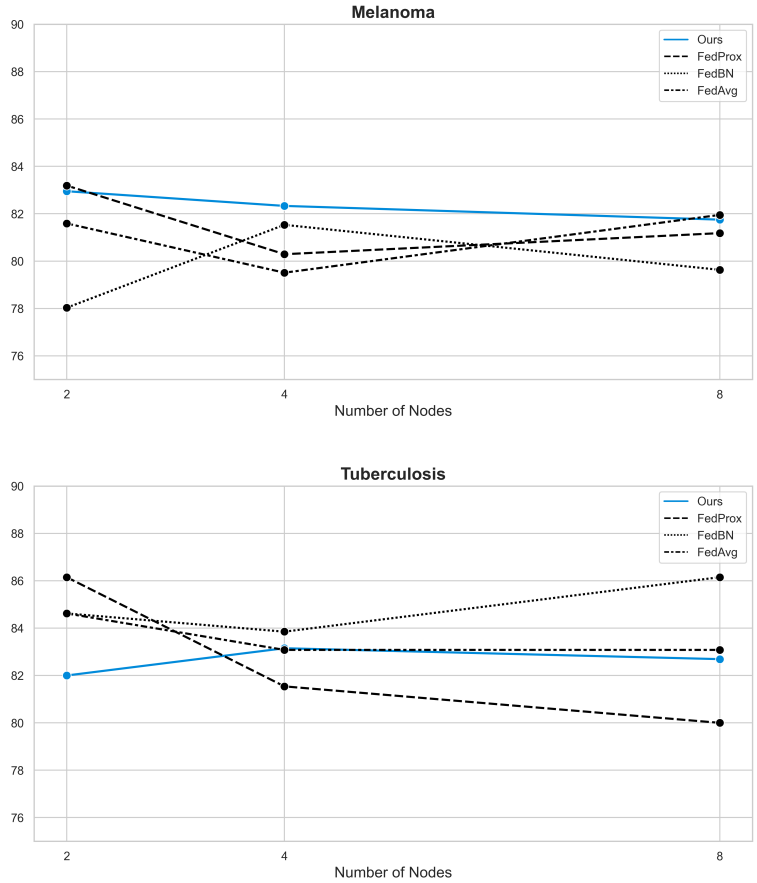


Figure 8.2: Scalability performance in the* i.i.d. *setting *w.r.t. number of nodes for the proposed approach and state-of-the-art methods.*

These above results suggest that experience replay plays a key role in federated models as a principled way to integrate features coming from different data distributions. To further assess its contribution, we evaluate FedER performance when using buffer at different sizes.

Results on the tuberculosis task, measured as mean and standard deviation of the local node models over a given dataset, are shown in Table 8.8 and indicate a clear contribution of the buffer in terms of overall performance and models' agreement. Indeed, with no buffer we obtain the lowest average performance and the highest standard deviation. As the buffer is enabled, we can observe a performance gain (mainly for the Shenzhen dataset) and a significant drop in standard deviation. Performance improves as buffer size increases, although gain becomes negligible above 512. Since higher buffer sizes result in more data to be shared among nodes, we use a buffer size of 512, as the best trade-off between accuracy and communication costs. We finally evaluate the capability of FedER to scale with the size of the federated network. Accordingly, we quantify this property using an *i.i.d.* setting on both tuberculosis (Shenzhen dataset) and skin lesion classification (BCN dataset) tasks, by equally splitting the available data on multiple nodes. Fig 8.2 shows how the proposed approach is able to keep classification accuracy high and performs on par with state-of-the-art approaches (namely, FedAvg, FedProx and FedBN).

8.4.4 Privacy-preserving performance

In this section we quantify how much information of real samples is retained by our privacy-preserving method, and in particular in the mapping between latent space and synthetic images. To do so, we employ the projection method proposed in [78]: given a real image $\mathbf{x}$, we find an intermediate latent point $\mathbf{w}$ such that the generated image $G(\mathbf{w})$ is most similar to $\mathbf{x}$, by optimizing $\mathbf{w}$ to minimize the LPIPS distance [190] between $\mathbf{x}$ and $G(\mathbf{w})$.

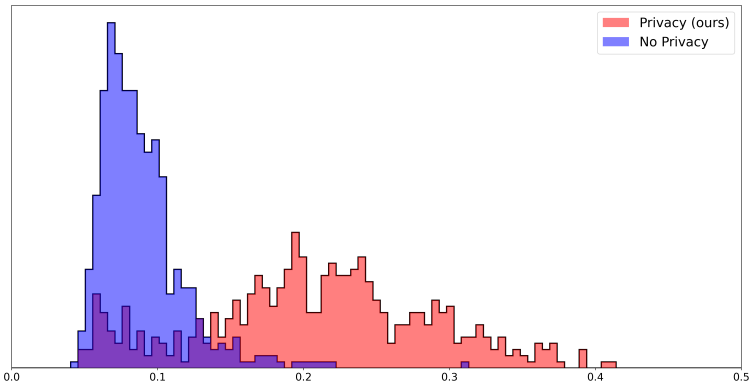


Figure 8.3: Quantitative analysis of privacy-preserving generation. *In blue, LPIPS distance histogram between real images and the corresponding images obtained through latent space projection using a GAN trained without the proposed privacy-preserving loss. In red, LPIPS distance histogram between real images and the closest images generated with the proposed approach.*

In practice, for each image of the dataset used for GAN training, we perform backprojection to find its most similar synthetic sample, and measure the LPIPS distance between the original and projected images. Fig. 8.3 shows the histograms of the resulting distances on the Shenzhen dataset, using GAN models trained with and without the proposed privacy-preserving loss (both models start from the same $\mathbf{w}$, for fairness). The histograms show that standard GAN training, with no privacy-preserving loss, tends to yield distances closer to 0, demonstrating that real images are indeed included into the generator latent space; while our model significantly mitigates this issue, by synthesizing samples that are substantially different than the original ones. In order to qualitatively substantiate these findings, Fig. 8.4 compares

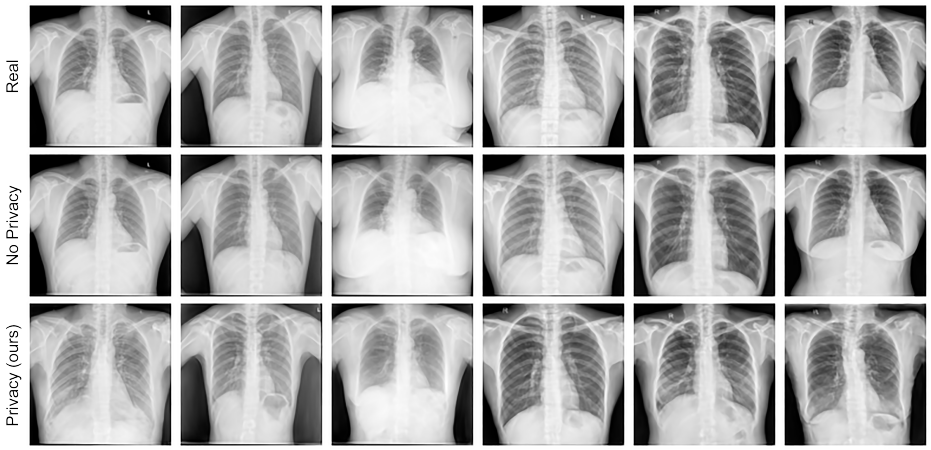


Figure 8.4: Qualitative samples of our privacy-preserving generation. *Top row: real images from the Shenzhen dataset. Middle row: projection with a standard GAN. Bottom row: projection with our privacy-preserving GAN.*

original samples from the Shenzhen dataset with the corresponding projections, generated with and without our privacy-preserving loss[3]. It is easy to notice that generated samples with a traditional GAN highly resemble real data, making it impossible to share such samples, albeit synthetic, in a privacy-safe manner, as they clearly contain patient information. Instead, comparing real images with the projections obtained from privacy-preserving GAN confirms the inability of the generator to find latent representations that recover real images used during training.

Given the high realism of generated samples, we run additional

[3]We show only X-Ray synthesized samples, as the effect is of our privacy-preserving strategy, is more appreciable than in skin lesion data.

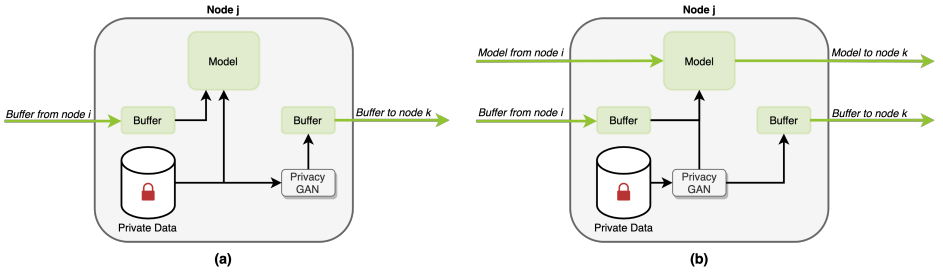


Figure 8.5: Privacy-enhanced alternative architectures. *(a)* FedER-A *configuration ("Buffer-only sharing"): a local node model is trained on real data, but only a buffer of synthetic samples is shared with other nodes. (b)* FedER-B *("Synthetic-only training"): Even within the dataset owner node, models are trained on synthetic data only.*

tests by proposing two FedER variants aiming to increase the level of privacy preservation: a) *FedER-A*: models are not shared among nodes — only synthetic buffers are sent and received; b) *FedER-B*: models are trained only using synthetic data, even on local nodes. Fig. 8.5 shows the internal architecture of each node in the two variants. Results obtained with these alternative privacy-enhanced configurations are provided in Table 8.9. It can be noted that FedER-A (i.e., "buffer-only sharing") configuration achieves comparable performance to our standard FedER (82.76 vs 83.41), but, remarkably, it outperforms all existing federated learning methods on the same datasets (compare Table 8.5 with the node performance block in Table 8.9). The FedER-B (i.e., "synthetic-only training") configuration, instead, performs slightly worse than the other two configurations, but is still on par with existing federated methods.

Config	Node Performance			Node Convergence	
	Shenzhen	Montgomery	Mean	Shenzhen	Montgomery
FedER	80.15	86.67	83.41	80.54 ±1.20	85.67 ±2.36
FedER-A	83.54	82.00	82.76	78.84 ±6.64	81.00 ±3.30
FedER-B	74.15	81.33	77.74	73.61 ±4.68	80.40 ±3.60

Table 8.9: Classification accuracy of the proposed privacy-enhanced strategies in the* non-i.i.d. *setting*. FedER-A: only buffers are shared (Fig. 8.5-a). FedER-B: models are trained on synthetic data only (Fig. 8.5-b). Node performance measures how each node model performs on its own private dataset, while node convergence assesses how a node model performs on other federation nodes.*

8.4.5 Communication and computational performance

We conclude the experimental analysis by measuring *communication* and *computational* costs.

As for *communication costs*, compared to state-of-the-art approaches, FedER requires additional transmission of synthetic images between nodes at each round. Tab. 8.10 reports per-node communication costs for state-of-the-art models (the table reports FedAvg, but the same values apply for FedProx and FedBN) and for FedER, in its full formulation and in the FedER-A variant, where only buffers of synthetic data are shared. The main cost for state-of-the-art models lies in the transfer of the model, and depends on the specific architecture (we included ResNet-18 and ResNet-152 as representative examples of different model scales). Values for our approach are reported for buffers

of size 512 containing 256×256 images, and depend on the color space. For our full FedER model, the increment in communication costs is significant but not excessive. However, if we take into consideration the variant where only synthetic data are exchanged (i.e., FedER-A), which still performs better than state-of-the-art methods (Tab. 8.5 and Tab. 8.9), communication overhead becomes significantly less than model-sharing approaches.

As for *computational costs* of federated training, FedER incurs the same overhead for parameter optimization and aggregation as state-of-the-art methods. Additionally, before federated training starts, FedER requires that each node trains a local privacy-preserving GAN off-line; this, however, does not affect online federated learning costs, as it is carried out only once at the very beginning of the whole procedure.

Furthermore, we argue that, in the medical domain, the number of institutions in a federation is relatively low and it is reasonable to assume that nodes can benefit from a powerful communication network and computing infrastructure: thus, the overhead introduced by FedER is tolerable, in light of the methodological advantages and the obtained performance and generalization capabilities showed by the resulting models.

8.5 Discussion

We propose FedER, a decentralized federated learning framework that replaces traditional parameters averaging with a more principled feature integration approach based on the combination of experience replay and privacy-preserving generative models. In FedER, nodes

	Tuberculosis		Melanoma	
	ResNet-18	**ResNet-152**	**ResNet-18**	**ResNet-152**
FedAvg FedProx FedBN	45 MB	230 MB	45 MB	230 MB
FedER	65 MB	250 MB	105 MB	290 MB
FedER-A	20 MB	20 MB	60 MB	60 MB

Table 8.10: Communication results comparison

communicate with each other by sharing local models and buffers of synthetic samples; local model updates are carried out in a way that encourages the reuse and adaptation of features learned by other nodes, thus avoiding potentially disruptive effects due to blind feature averaging. Experimental results show that our method outperforms significantly state-of-the-art server-based approaches in a non-*i.i.d.* scenario, which is a typical setting in the medical domain. Additionally, quantitative and qualitative analysis shows that our privacy-preserving generation approach is able to synthesize samples that are significantly different from real data, while correctly supporting the learning of discriminative features. In the future, we aim at investigating some unexplored properties of our method: for instance, unlike all other existing methods based on parameter averaging is required, our approach does not strictly require that all nodes share the same model architecture. Model heterogeneity could therefore be employed to create a shared ensemble and combine different feature learning capabilities.

8.6 Publications

M. Pennisi, F. Proietto Salanitri, G. Bellitto, B. Casella, M. Aldinucci, S. Palazzo, and C. Spampinato, "Feder: Federated learning through experience replay and privacy-preserving data synthesis," Computer Vision and Image Understanding, p. 103882, 2023. [133].

Part IV

CONCLUSION

As we wrap up this thesis, it's important to think back on the journey we've taken, the insights we have gathered, and the implications of our findings. The research we've presented has covered different aspects of Medical Imaging Analysis, offering both new perspectives and reaffirming established paradigms. We now distill the essence of the conducted research and consider its broader implications for the field and potential avenues for future exploration.

CHAPTER

NINE

CONCLUSIONS

The field of medical imaging analysis has been evolving over the years, driven by the search for more accurate, efficient, and privacy-friendly methodologies. This thesis has traversed a road map in this field, beginning with standard centralized approaches for medical image analysis and culminating in the exploration of federated learning techniques.

The initial phases of our research were rooted in centralized methodologies. These approaches, while powerful and effective, primarily rely on the consolidation of data at a single location or server. We have attempted to address fundamental tasks in the field of medical image analysis by proposing innovative and effective methodologies. An emphasis on the interpretability of deep learning models has been at the core of all the devised approaches. In a field as critical as medical imaging, where patient health is crucial, it is essential that our models not only perform accurately, but also provide understandable and usable information for medical professionals. Our

journey started with an AI-driven system, designed for the pressing challenge of COVID-19 assessment by CT scans. Achieving a sensitivity of 90.3% and specificity of 93.5%, the system went beyond just accurate lesion categorization. It harnessed the power of explainable AI techniques, offering radiologists insights into its decision-making, thereby bolstering trust and interpretability. While we faced hurdles such as a limited sample size, the robustness of our approach was underscored by a thorough confidence level analysis. We then moved to the development of models specifically designed for the pancreas, an organ that presents significant challenges in medical imaging analysis. The complexity of accurately identifying and analyzing the pancreas is highlighted by the fact that pancreatic cancer is one of the leading causes of mortality worldwide. We designed a Transformer-based neural network tailored for MRI imaging to tackle the classification challenges associated with Intraductal Papillary Mucosal Neoplasms (IPMN), a precursor to pancreatic cancer. Our innovative approach not only showcased the superior capabilities of Transformer architectures in identifying complex patterns within MRI data, thereby enhancing classification accuracy, but also highlighted their innate advantage in terms of interpretability. Transformers have demonstrated to be more generalizable for IPMN risk stratification than CNNs. Furthermore, our transformer-based classifier is able to provide meaningful insights, focusing predictions on the most relevant regions of the image related to cysts, which in turn bolstered the robustness of the automated diagnosis. Finally, we introduced PanKNet, a novel Hierarchical 3D fully-convolutional network for pancreas segmentation in both MRI and CT scans. Given the pancreas's complex structure and its close proximity to other organs, achieving accurate segmentation

in medical imaging is notoriously challenging. Our approach emphasizes a hierarchical decoding strategy, which facilitates a coarse-to-fine segmentation process. This method combines multi-scale features: deeper features provide intermediate segmentation masks capturing fine details, while initial features offer coarse segmentation outlines. We evaluated PanKNet performance using CT scans from the NIH CT-Pancreas benchmark, achieving a leading Dice score of 88.01%. For MRI scans, we attained a Dice score of 77.46%, setting state-of-the-art results on the segmentation of the MRI pancreas. Furthermore, the PanKNet architecture is adaptable for 3D segmentation in several medical areas. However, as with all technological advancements, centralized approaches are not without their limitations. First and foremost, the availability of large data within a single node to enhance the generalization capabilities of trained models. However, in the field of medical imaging, this is often unfeasible. The process of annotating medical images is time consuming, requiring expert knowledge and precision. Furthermore, smaller medical centers or institutions might only have access to a limited number of samples, making it challenging to curate a comprehensive dataset. This lack of data, combined with the inherent complexities of medical imaging, underscores the need for collaborative and decentralized solutions, leading us to the domain of federated learning. Federated learning, with its decentralized nature, allows for data analysis at the source, sidestepping the need for data consolidation. This not only addresses privacy concerns but also offers a scalable solution for medical imaging analysis across diverse and distributed datasets.

In the research reported in this thesis, we initially focused on data-centric methods, leveraging the capabilities of GANs to approach fed-

erated learning from a data perspective. Our primary endeavor was to delve into the latent space of GANs, aiming to aggregate features derived from real samples encoded within this space. This exploration led to the development of techniques that produce synthetic images while preserving data privacy. Instead of the traditional federated learning approach of gradient aggregation, which poses privacy risks due to potential input reconstruction, we introduced a synthetic data aggregation method. In this approach, each node generates a synthetic dataset, and the central model is trained using the combined synthetic datasets. The obtained results are promising, with performance metrics comparable to training on the combined original datasets. Such method not only facilitates the aggregation of synthesized information in a federated setting, but also highlights the potential benefits of utilizing GAN latent space to bolster federated learning. Building on these results, we introduced PLAN, an innovative approach to navigate this space. This method "*walks*" from one aggregated point to another within the latent space, ensuring equidistance, privacy, and class consistency. Through this, we devised a strategy that generates synthetic data with enhanced privacy considerations, refining traditional latent space aggregation techniques. Experimental results on two medical image analysis tasks shows PLAN's resilience against membership inference attacks, while supporting model training with performance metrics comparable to those achieved with real data. Moreover, when PLAN is combined with state-of-the-art *k-anonymity* methods, we observe a reduced performance drop, while upholding privacy-preservation properties.

In our final exploration, we delved into a more reliable setting of federated learning: the decentralized approach. We introduced

FedER, a comprehensive framework that seamlessly integrates the principles of continual Learning and federated Learning to address the challenges posed by distribution shifts (both in space and in time). Distinct from traditional federated methods that rely on parameter averaging, FedER adopts a principled feature integration approach. This is achieved by combining experience replay with privacy-preserving generative models. In the FedER paradigm, nodes communicate by sharing both local models and buffers of synthetic samples. The local model updates are orchestrated in a way that promotes the reuse and adaptation of features learned by other nodes, thereby sidestepping the potential pitfalls of blind feature averaging. Our experimental results show the superiority of FedER, as it significantly outperforms state-of-the-art centralized approaches, especially in non-*i.i.d.* scenarios, a common setting in the medical domain. Moreover, both quantitative and qualitative analyses reveal that our privacy-preserving generation approach can synthesize samples that are markedly distinct from real data, yet effectively support the learning of discriminative features. A unique aspect of FedER, setting it apart from other methods reliant on parameter averaging, is its flexibility in accommodating nodes with varied model architectures. Distinct from our previous two endeavors, this framework ensures rigorous data privacy through the deployment of a specially trained GAN, emphasizing privacy preservation.

The potential for future developments is huge. As we continue to scale our techniques, accommodating the ever-growing datasets in medical imaging becomes crucial. Ensuring that methods like PLAN remain robust and effective, especially when compared with a multitude of identities, will be a focal point of our research. By delving deeper into the latent space, we envision a comprehensive explo-

ration of trajectories of varying lengths, aiming to strike a balance between optimizing privacy and enhancing data augmentation capabilities. Furthermore, the inherent flexibility of FedER offers an unique and promising solution to address challenges in federated learning. Its ability to seamlessly integrate nodes with different architectures, not only underscores its adaptability, but also opens up avenues for innovative research. Indeed, by harnessing this model heterogeneity, we may create shared ensemble models without moving the data from its origin. The ensemble models would allow for learning diverse features, offering a holistic solution that captures the nuances of varied datasets. As we move forward, the fusion of these approaches and the exploration of their synergies will undoubtedly shape the next frontier in medical imaging research.

BIBLIOGRAPHY

[1] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 2016, pp. 308–318.

[2] M. A. Abdou, "Literature review: Efficient deep neural networks techniques for medical image analysis," *Neural Computing and Applications*, vol. 34, no. 8, pp. 5791–5812, 2022.

[3] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," *arXiv preprint arXiv:2005.00928*, 2020.

[4] D. Abramian and A. Eklund, "Refacing: Reconstructing anonymized facial features using GANS," in *16th IEEE International Symposium on Biomedical Imaging, ISBI 2019, Venice, Italy, April 8-11, 2019*. IEEE, 2019, pp. 1104–1108.

[5] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Advances in Neural Information Processing Systems 31*, S. Bengio,

H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Curran Associates, Inc., 2018, pp. 9505–9515. [Online]. Available: http://papers.nips.cc/paper/8160-sanity-checks-for-saliency-maps.pdf

[6] Y. Alaluf, O. Patashnik, and D. Cohen-Or, "Restyle: A residual-based stylegan encoder via iterative refinement," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6711–6720.

[7] M. Aldinucci, S. Rabellino, M. Pironti, F. Spiga, P. Viviani, M. Drocco, M. Guerzoni, G. Boella, M. Mellia, P. Margara, I. Drago, R. Marturano, G. Marchetto, E. Piccolo, S. Bagnasco, S. Lusso, S. Vallero, G. Attardi, A. Barchiesi, A. Colla, and F. Galeazzi, "HPC4AI, an AI-on-demand federated platform endeavour," in *ACM Computing Frontiers*, Ischia, Italy, May 2018. [Online]. Available: https://iris.unito.it/retrieve/handle/2318/1765596/689772/2018_hpc4ai_ACM_CF.pdf

[8] Z. Allam and D. S. Jones, "On the coronavirus (covid-19) outbreak and the smart city network: universal data sharing standards coupled with artificial intelligence (ai) to benefit urban health monitoring and management," in *Healthcare*, vol. 8, no. 1. Multidisciplinary Digital Publishing Institute, 2020, p. 46.

[9] R. S. Antunes, C. André da Costa, A. Küderle, I. A. Yari, and B. Eskofier, "Federated learning for healthcare: Systematic review and architecture proposal," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 13, no. 4, pp. 1–23, 2022.

[10] S. Armato, G. McLennan, L. Bidaut, M. McNitt-Gray, C. Meyer, A. Reeves, H. MacMahon, R. Engelmann, R. Roberts, A. Starkey, P. Caligiuri, D. Aberle, M. Brown, R. Pais, D. Qing, P. Batra, C. Jude, I. Petkovska, A. Biancardi, B. Zhao, C. Henschke, D. Yankelevitz, D. Max, A. Farooqi, E. Hoffman, E. van Beek, A. Smith, E. Kazerooni, P. Bland, G. Laderach, G. Gladish, R. Munden, L. Quint, L. Schwartz, B. Sundaram, L. Dodd, C. Fenimore, D. Gur, N. Petrick, J. Freymann, J. Kirby, B. Hughes, A. Casteele, S. Gupte, M. Sallam, M. Heath, M. Kuhn, E. Dharaiya, R. Burns, D. Fryd, M. Salganicoff, V. Anand, U. Shreter, S. Vastagh, B. Croft, and L. Clarke, "The lung image database consortium, (lidc) and image database resource initiative (idri):: a completed reference database of lung nodules on ct scans," *Medical Physics*, vol. 38, no. 2, pp. 915–931, Feb. 2011.

[11] H. Asaturyan, A. Gligorievski, and B. Villarini, "Morphological and multi-level geometrical descriptor analysis in ct and mri volumes for automatic pancreas segmentation," *Computerized Medical Imaging and Graphics*, vol. 75, pp. 1–13, 2019.

[12] H. X. Bai, R. Wang, Z. Xiong, B. Hsieh, K. Chang, K. Halsey, T. M. L. Tran, J. W. Choi, D.-C. Wang, L.-B. Shi *et al.*, "Ai augmentation of radiologist performance in distinguishing covid-19 from pneumonia of other etiology on chest ct," *Radiology*, p. 201491, 2020.

[13] X. Bai, C. Fang, Y. Zhou, S. Bai, Z. Liu, Q. Chen, Y. Xu,

T. Xia, S. Gong, X. Xie *et al.*, "Predicting covid-19 malignant progression with ai techniques," *MedRxiv*, pp. 2020–03, 2020.

[14] D. Bermejo-Peláez, S. Y. Ash, G. R. Washko, R. S. J. Estépar, and M. J. Ledesma-Carbayo, "Classification of interstitial lung abnormality patterns with an ensemble of deep convolutional neural networks," *Scientific reports*, vol. 10, no. 1, pp. 1–15, 2020.

[15] A. Bischoff-Grethe, I. B. Ozyurt, E. Busa, B. T. Quinn, C. Fennema-Notestine, C. P. Clark, S. Morris, M. W. Bondi, T. L. Jernigan, A. M. Dale, G. G. Brown, and B. Fischl, "A technique for the deidentification of structural brain MR images," *Hum Brain Mapp*, vol. 28, no. 9, pp. 892–903, Sep 2007.

[16] T. S. Brisimi, R. Chen, T. Mela, A. Olshevsky, I. C. Paschalidis, and W. Shi, "Federated learning of predictive models from federated electronic health records," *International journal of medical informatics*, vol. 112, pp. 59–67, 2018.

[17] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096*, 2018.

[18] L. Brunese, F. Mercaldo, A. Reginelli, and A. Santone, "Explainable deep learning for pulmonary disease and coronavirus covid-19 detection from x-rays," *Computer Methods and Programs in Biomedicine*, vol. 196, p. 105608, 2020.

[19] P. Buzzega *et al.*, "Dark experience for general continual learning: a strong, simple baseline," *NeurIPS*, 2020.

[20] J. Cai, L. Lu, Y. Xie, F. Xing, and L. Yang, "Improving deep pancreas segmentation in ct and mri images via recurrent neural contextual learning and direct loss function," *arXiv preprint arXiv:1707.04912*, 2017.

[21] J. Cai, L. Lu, Z. Zhang, F. Xing, L. Yang, and Q. Yin, "Pancreas segmentation in mri using graph-based decision fusion on convolutional neural networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 442–450.

[22] S. Candemir *et al.*, "Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration," *IEEE TMI*, vol. 33, no. 2, pp. 577–590, 2013.

[23] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*. Springer, 2020, pp. 213–229.

[24] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *CVPR*, 2017, pp. 6299–6308.

[25] A. M. Carrington, P. W. Fieguth, H. Qazi, A. Holzinger, H. H. Chen, F. Mayr, and D. G. Manuel, "A new concordant partial auc and partial c statistic for imbalanced data in the evaluation of machine learning algorithms," *BMC medical informatics and decision making*, vol. 20, no. 1, pp. 1–12, 2020.

[26] K. H. Cha, L. Hadjiiski, H.-P. Chan, A. Z. Weizer, A. Alva, R. H. Cohan, E. M. Caoili, C. Paramagul, and R. K. Samala, "Bladder cancer treatment response assessment in ct using radiomics with deep-learning," *Scientific reports*, vol. 7, no. 1, pp. 1–12, 2017.

[27] D. Chen, T. Orekondy, and M. Fritz, "Gs-wgan: A gradient-sanitized approach for learning differentially private generators," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 673–12 684, 2020.

[28] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[29] J. Chen, Y. He, E. C. Frey, and Y. Li, "Yong du. vit-v-net: Vision transformer for unsupervised volumetric medical image registration," *arXiv preprint arXiv:2104.06468*, 2021.

[30] M. Chung, A. Bernheim, X. Mei, N. Zhang, M. Huang, X. Zeng, J. Cui, W. Xu, Y. Yang, Z. A. Fayad *et al.*, "Ct imaging features of 2019 novel coronavirus (2019-ncov)," *Radiology*, vol. 295, no. 1, pp. 202–207, 2020.

[31] N. C. Codella *et al.*, "Skin lesion analysis toward melanoma detection," in *IEEE ISBI*, 2018.

[32] J. P. Cohen, P. Morrison, L. Dao, K. Roth, T. Q. Duong, and M. Ghassemi, "Covid-19 image data collection: Prospective predictions are the future," *arXiv 2006.11988*, 2020.

[33] M. Combalia *et al.*, "Bcn20000: Dermoscopic lesions in the wild," *arXiv:1908.02288*, 2019.

[34] J. E. Corral, S. Hussein, P. Kandel, C. W. Bolan, U. Bagci, and M. B. Wallace, "Deep learning to classify intraductal papillary mucinous neoplasms using magnetic resonance imaging," *Pancreas*, vol. 48, no. 6, pp. 805–810, 2019.

[35] K. Ćosić, S. Popović, M. Šarlija, I. Kesedžić, and T. Jovanovic, "Artificial intelligence in prediction of mental health disorders induced by the covid-19 pandemic among health care workers," *Croatian Medical Journal*, vol. 61, no. 3, p. 279, 2020.

[36] Y. Dai, Y. Gao, and F. Liu, "Transmed: Transformers advance multi-modal medical image classification," *Diagnostics*, vol. 11, no. 8, p. 1384, 2021.

[37] I. Dayan *et al.*, "Federated learning for predicting clinical outcomes in patients with COVID-19," *Nature medicine*, vol. 27, 2021.

[38] M. Delange *et al.*, "A continual learning survey: Defying forgetting in classification tasks," *IEEE PAMI*, 2021.

[39] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[40] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-

training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[41] S. S. Diwangkara and A. I. Kistijantoro, "Study of data imbalance and asynchronous aggregation algorithm on federated learning system," in *2020 International Conference on Information Technology Systems and Innovation (ICITSI)*. IEEE, 2020, pp. 276–281.

[42] L. Dong, S. Xu, and B. Xu, "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5884–5888.

[43] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[44] L. T. Duong, N. H. Le, T. B. Tran, V. M. Ngo, and P. T. Nguyen, "Detection of tuberculosis from chest x-ray images: Boosting the performance with vision transformer and transfer learning," *Expert Systems with Applications*, vol. 184, p. 115519, 2021.

[45] D. Evans *et al.*, "A pragmatic introduction to secure multi-party computation," *Foundations and Trends in Privacy and Security*, vol. 2, no. 2-3, pp. 70–246, 2018.

[46] I. Feki *et al.*, "Federated learning for COVID-19 screening

from chest x-ray images," *Applied Soft Computing*, vol. 106, p. 107330, 2021.

[47] H. Fereidooni, S. Marchal, M. Miettinen, A. Mirhoseini, H. Möllering, T. D. Nguyen, P. Rieger, A.-R. Sadeghi, T. Schneider, H. Yalame *et al.*, "Safelearn: Secure aggregation for private federated learning," in *2021 IEEE Security and Privacy Workshops (SPW)*. IEEE, 2021, pp. 56–62.

[48] M. Gao, U. Bagci, L. Lu, A. Wu, M. Buty, H.-C. Shin, H. Roth, G. Z. Papadakis, A. Depeursinge, R. M. Summers *et al.*, "Holistic classification of ct attenuation patterns for interstitial lung diseases via deep convolutional neural networks," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 6, no. 1, pp. 1–6, 2018.

[49] Z. Gao, F. Wu, W. Gao, and X. Zhuang, "A new framework of swarm learning consolidating knowledge from multi-center non-iid data for medical image segmentation," *IEEE TMI*, pp. 1–1, 2022.

[50] L. Gazit, J. Chakraborty, M. Attiyeh, L. Langdon-Embry, P. J. Allen, R. K. Do, and A. L. Simpson, "Quantification of ct images for the classification of high-and low-risk pancreatic cysts," in *Medical Imaging 2017: Computer-Aided Diagnosis*, vol. 10134. International Society for Optics and Photonics, 2017, p. 101340X.

[51] J. Geiping *et al.*, "Inverting gradients-how easy is it to break privacy in federated learning?" *NeurIPS*, 2020.

[52] I. Goodfellow *et al.*, "Generative adversarial nets," *NeurIPS*, 2014.

[53] I. J. Goodfellow *et al.*, "An empirical investigation of catastrophic forgetting in gradient-based neural networks." *arXiv:1312.6211*, 2013.

[54] M. Gorris, S. A. Hoogenboom, M. B. Wallace, and J. E. van Hooft, "Artificial intelligence for the management of pancreatic diseases," *Digestive Endoscopy*, vol. 33, no. 2, pp. 231–241, 2021.

[55] W. Gu, S. Bai, and L. Kong, "A review on 2d instance segmentation based on deep neural networks," *Image and Vision Computing*, vol. 120, p. 104401, 2022.

[56] S. Guo, T. Zhang, G. Xu, H. Yu, T. Xiang, and Y. Liu, "Topology-aware differential privacy for decentralized image classification," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[57] A. N. Hanania, L. E. Bantis, Z. Feng, H. Wang, E. P. Tamm, M. H. Katz, A. Maitra, and E. J. Koay, "Quantitative imaging to evaluate malignant potential of ipmns," *Oncotarget*, vol. 7, no. 52, p. 85776, 2016.

[58] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "Unetr: Transformers for 3d medical image segmentation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 574–584.

[59] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[60] M. Heusel *et al.*, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *NeurIPS*, 2017.

[61] J. Hofmanninger, F. Prayer, J. Pan, S. Rohrich, H. Prosch, and G. Langs, "Automatic lung segmentation in routine imaging is a data diversity problem, not a methodology problem," *arXiv preprint arXiv:2001.11767*, 2020.

[62] A. Holzinger, "From machine learning to explainable ai," in *2018 world symposium on digital intelligence for systems and machines (DISA)*. IEEE, 2018, pp. 55–66.

[63] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 4, p. e1312, 2019.

[64] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *arXiv preprint arXiv:1709.01507*, vol. 7, 2017.

[65] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[66] L. Huang, R. Han, T. Ai, P. Yu, H. Kang, Q. Tao, and L. Xia, "Serial quantitative chest ct assessment of covid-19: Deep-

learning approach," *Radiology: Cardiothoracic Imaging*, vol. 2, no. 2, p. e200075, 2020.

[67] P. Huang, T. Liu, L. Huang, H. Liu, M. Lei, W. Xu, X. Hu, J. Chen, and B. Liu, "Use of chest ct in combination with negative rt-pcr assay for the 2019 novel coronavirus but high clinical suspicion," *Radiology*, vol. 295, no. 1, pp. 22–23, 2020.

[68] S. Hussein, P. Kandel, J. E. Corral, C. W. Bolan, M. B. Wallace, and U. Bagci, "Deep multi-modal classification of intraductal papillary mucinous neoplasms (ipmn) with canonical correlation analysis," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018, pp. 800–804.

[69] J. Irvin *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *AAAI*, 2019.

[70] S. Jaeger *et al.*, "Automatic tuberculosis screening using chest radiographs," *IEEE TMI*, vol. 33, no. 2, pp. 233–245, 2013.

[71] ——, "Two public chest x-ray datasets for computer-aided screening of pulmonary diseases," *Quantitative imaging in medicine and surgery*, vol. 4, no. 6, p. 475, 2014.

[72] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," in *CVPRW 2017*. IEEE, 2017, pp. 1175–1183.

[73] M. Jeon, H. Park, H. J. Kim, M. Morley, and H. Cho, "k-salsa: k-anonymous synthetic averaging of retinal images via local style

alignment," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXI*. Springer, 2022, pp. 661–678.

[74] J. Jia, A. Salem, M. Backes, Y. Zhang, and N. Z. Gong, "Memguard: Defending against black-box membership inference attacks via adversarial examples," in *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*, 2019, pp. 259–274.

[75] S. Kadhe, N. Rajaraman, O. O. Koyluoglu, and K. Ramchandran, "Fastsecagg: Scalable secure aggregation for privacy-preserving federated learning," *arXiv preprint arXiv:2009.11248*, 2020.

[76] P. Kairouz *et al.*, "Advances and open problems in federated learning," *Foundations and Trends in Machine Learning*, vol. 14, 2021.

[77] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, "Training generative adversarial networks with limited data," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 104–12 114, 2020.

[78] T. Karras *et al.*, "Analyzing and improving the image quality of stylegan," in *CVPR*, 2020.

[79] ——, "Training generative adversarial networks with limited data," in *NeurIPS*, 2020.

[80] S. D. Karthik, Maggie, "Aptos 2019 blindness detection," 2019. [Online]. Available: https://kaggle.com/competitions/aptos2019-blindness-detection

[81] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[82] Y.-Y. Ke, T.-T. Peng, T.-K. Yeh, W.-Z. Huang, S.-E. Chang, S.-H. Wu, H.-C. Hung, T.-A. Hsu, S.-J. Lee, J.-S. Song *et al.*, "Artificial intelligence approach fighting covid-19 with repurposing drugs," *Biomedical Journal*, 2020.

[83] E. Kerfoot, J. Clough, I. Oksuz, J. Lee, A. P. King, and J. A. Schnabel, "Left-ventricle quantification using residual u-net," in *International Workshop on Statistical Atlases and Computational Models of the Heart*. Springer, 2018, pp. 371–380.

[84] L. U. Khan, W. Saad, Z. Han, E. Hossain, and C. S. Hong, "Federated learning for internet of things: Recent advances, taxonomy, and open challenges," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 3, pp. 1759–1799, 2021.

[85] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM computing surveys (CSUR)*, vol. 54, no. 10s, pp. 1–41, 2022.

[86] N. Khosravan, A. Mortazi, M. Wallace, and U. Bagci, "Pan: Projective adversarial network for medical image segmentation,"

in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 68–76.

[87] J. Kotia, A. Kotwal, R. Bharti, and R. Mangrulkar, "Few shot learning for medical imaging," *Machine learning algorithms for industrial applications*, pp. 107–132, 2021.

[88] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[89] T. Kuwahara, K. Hara, N. Mizuno, N. Okuno, S. Matsumoto, M. Obata, Y. Kurita, H. Koda, K. Toriyama, S. Onishi *et al.*, "Usefulness of deep learning analysis for the diagnosis of malignancy in intraductal papillary mucinous neoplasms of the pancreas," *Clinical and translational gastroenterology*, vol. 10, no. 5, 2019.

[90] A. Lalitha *et al.*, "Peer-to-peer federated learning on graphs," *arXiv:1901.11173*, 2019.

[91] R. LaLonde, I. Tanner, K. Nikiforaki, G. Z. Papadakis, P. Kandel, C. W. Bolan, M. B. Wallace, and U. Bagci, "Inn: inflated neural networks for ipmn diagnosis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 101–109.

[92] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[93] M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana, "Certified robustness to adversarial examples with differential privacy," in *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2019, pp. 656–672.

[94] J. Lee and R. M. Nishikawa, "Detecting mammographically occult cancer in women with dense breasts using deep convolutional neural network and radon cumulative distribution transform," *Journal of Medical Imaging*, vol. 6, no. 4, pp. 044 502–044 502, 2019.

[95] H. Li, Q. Lü, G. Chen, T. Huang, and Z. Dong, "Convergence of distributed accelerated algorithm over unbalanced directed networks," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, pp. 1–12, 2019.

[96] L. Li, L. Qin, Z. Xu, Y. Yin, X. Wang, B. Kong, J. Bai, Y. Lu, Z. Fang, Q. Song *et al.*, "Artificial intelligence distinguishes covid-19 from community acquired pneumonia on chest ct," *Radiology*, 2020.

[97] Q. Li, Z. Wen, Z. Wu, S. Hu, N. Wang, Y. Li, X. Liu, and B. He, "A survey on federated learning systems: Vision, hype and reality for data privacy and protection," *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[98] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *Proceedings of Machine Learning and Systems*, vol. 2, pp. 429–450, 2020.

[99] W. Li *et al.*, "Privacy-preserving federated brain tumour segmentation," in *International workshop on machine learning in medical imaging*. Springer, 2019, pp. 133–141.

[100] X. Li, M. Jiang, X. Zhang, M. Kamp, and Q. Dou, "Fedbn: Federated learning on non-iid features via local batch normalization," *arXiv:2102.07623*, 2021.

[101] X. Lian *et al.*, "Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent," *NeurIPS*, 2017.

[102] W. Liang, J. Yao, A. Chen, Q. Lv, M. Zanin, J. Liu, S. Wong, Y. Li, J. Lu, H. Liang *et al.*, "Early triage of critically ill covid-19 patients using deep learning," *Nature communications*, vol. 11, no. 1, pp. 1–7, 2020.

[103] L. Lin and Z. Hou, "Combat covid-19 with artificial intelligence and big data," *Journal of travel medicine*, vol. 27, no. 5, p. taaa080, 2020.

[104] W. Lin, Y. Xu, B. Liu, D. Li, T. Huang, and F. Shi, "Contribution-based federated learning client selection," *International Journal of Intelligent Systems*, vol. 37, no. 10, pp. 7235–7260, 2022.

[105] H. Liu, F. Liu, J. Li, T. Zhang, D. Wang, and W. Lan, "Clinical and ct imaging features of the covid-19 pneumonia: Focus on pregnant women and children," *Journal of infection*, 2020.

[106] S. Liu, X. Yuan, R. Hu, S. Liang, S. Feng, Y. Ai, and Y. Zhang, "Automatic pancreas segmentation via coarse location and ensemble learning," *IEEE Access*, vol. 8, pp. 2906–2914, 2020.

[107] Y. Lu, X. Huang, Y. Dai, S. Maharjan, and Y. Zhang, "Blockchain and federated learning for privacy-preserved data sharing in industrial iot," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 6, pp. 4177–4186, 2019.

[108] B. Luo, X. Li, S. Wang, J. Huang, and L. Tassiulas, "Cost-effective federated learning design," in *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*. IEEE, 2021, pp. 1–10.

[109] A. Madani, M. Moradi, A. Karargyris, and T. Syeda-Mahmood, "Semi-supervised learning with generative adversarial networks for chest x-ray classification with ability of data domain adaptation," in *2018 IEEE 15th International symposium on biomedical imaging (ISBI 2018)*. IEEE, 2018, pp. 1038–1042.

[110] F. Mahmood, R. Chen, and N. J. Durr, "Unsupervised reverse domain adaptation for synthetic medical images via adversarial training," *IEEE transactions on medical imaging*, vol. 37, no. 12, pp. 2572–2581, 2018.

[111] P. Malhotra, S. Gupta, D. Koundal, A. Zaguia, W. Enbeyle *et al.*, "Deep neural networks for medical image segmentation," *Journal of Healthcare Engineering*, vol. 2022, 2022.

[112] Y. Man, Y. Huang, J. Feng, X. Li, and F. Wu, "Deep q learning driven ct pancreas segmentation with geometry-aware u-net,"

IEEE Transactions on Medical Imaging, vol. 38, no. 8, pp. 1971–1980, 2019.

[113] B. McMahan *et al.*, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[114] B. Meden, Ž. Emeršič, V. Štruc, and P. Peer, "k-same-net: k-anonymity with generative deep neural networks for face deidentification," *Entropy*, vol. 20, no. 1, p. 60, 2018.

[115] X. Mei, H.-C. Lee, K.-y. Diao, M. Huang, B. Lin, C. Liu, Z. Xie, Y. Ma, P. M. Robson, M. Chung *et al.*, "Artificial intelligence–enabled rapid diagnosis of patients with covid-19," *Nature Medicine*, pp. 1–5, 2020.

[116] F. Milletari, N. Navab, and S. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*, 2016, pp. 565–571.

[117] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv:1411.1784*, 2014.

[118] Y. Mo, Y. Wu, X. Yang, F. Liu, and Y. Liao, "Review the state-of-the-art technologies of semantic segmentation based on deep learning," *Neurocomputing*, vol. 493, pp. 626–646, 2022.

[119] S. Mohanty, M. H. A. Rashid, M. Mridul, C. Mohanty, and S. Swayamsiddha, "Application of artificial intelligence in covid-

19 drug repurposing," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 2020.

[120] J. H. Moltz, L. Bornemann, J.-M. Kuhnigk, V. Dicken, E. Peitgen, S. Meier, H. Bolte, M. Fabel, H.-C. Bauknecht, M. Hittinger *et al.*, "Advanced segmentation techniques for lung nodules, liver metastases, and enlarged lymph nodes in ct scans," *IEEE Journal of selected topics in signal processing*, vol. 3, no. 1, pp. 122–134, 2009.

[121] V. Mothukuri, R. M. Parizi, S. Pouriyeh, Y. Huang, A. Dehghantanha, and G. Srivastava, "A survey on security and privacy of federated learning," *Future Generation Computer Systems*, vol. 115, pp. 619–640, 2021.

[122] P. Nardelli, D. Jimenez-Carretero, D. Bermejo-Pelaez, G. R. Washko, F. N. Rahaghi, M. J. Ledesma-Carbayo, and R. S. J. Estépar, "Pulmonary artery–vein classification in ct images using deep learning," *IEEE transactions on medical imaging*, vol. 37, no. 11, pp. 2428–2440, 2018.

[123] M. Nasr, R. Shokri, and A. Houmansadr, "Machine learning with membership privacy using adversarial regularization," in *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*, 2018, pp. 634–646.

[124] N. Navab, J. Hornegger, W. M. Wells, and A. Frangi, *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Ger-*

many, October 5-9, 2015, Proceedings, Part III. Springer, 2015, vol. 9351.

[125] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," *Advances in neural information processing systems*, vol. 14, 2001.

[126] M.-Y. Ng, E. Y. Lee, J. Yang, F. Yang, X. Li, H. Wang, M. M.-s. Lui, C. S.-Y. Lo, B. Leung, P.-L. Khong *et al.*, "Imaging profile of the covid-19 infection: radiologic findings and literature review," *Radiology: Cardiothoracic Imaging*, vol. 2, no. 1, p. e200034, 2020.

[127] P. E. Oberstein and K. P. Olive, "Pancreatic cancer: why is it so hard to treat?" *Therapeutic advances in gastroenterology*, vol. 6, no. 4, pp. 321–337, 2013.

[128] E. S. of Radiology (ESR) communications@ myesr. org Emanuele Neri Nandita de Souza Adrian Brady Angel Alberich Bayarri Christoph D. Becker Francesca Coppola Jacob Visser, "What the radiologist should know about artificial intelligence–an esr white paper," *Insights into imaging*, vol. 10, pp. 1–8, 2019.

[129] W. H. Organization, "Novel coronavirus (2019-ncov): situation report, 8," World Health Organization, Tech. Rep., 2020.

[130] J. Pang *et al.*, "Collaborative city digital twin for the covid-19 pandemic: A federated learning solution," *Tsinghua science and technology*, vol. 26, no. 5, pp. 759–771, 2021.

[131] M. Pennisi, S. Palazzo, and C. Spampinato, "Self-improving classification performance through gan distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021, pp. 1640–1648.

[132] M. Pennisi, F. Proietto Salanitri, S. Palazzo, C. Pino, F. Rundo, D. Giordano, and C. Spampinato, "Gan latent space manipulation and aggregation for federated learning in medical imaging," in *International Workshop on Distributed, Collaborative, and Federated Learning*. Springer, 2022, pp. 68–78.

[133] M. Pennisi, F. P. Salanitri, G. Bellitto, B. Casella, M. Aldinucci, S. Palazzo, and C. Spampinato, "Feder: Federated learning through experience replay and privacy-preserving data synthesis," *Computer Vision and Image Understanding*, p. 103882, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S107731422300262X

[134] C. S. Perone, P. Ballester, R. C. Barros, and J. Cohen-Adad, "Unsupervised domain adaptation for medical imaging segmentation with self-ensembling," *NeuroImage*, vol. 194, pp. 1–11, 2019.

[135] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. (2018) Language models are unsupervised multitask learners. [Online]. Available: https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf

[136] J.-F. Rajotte, S. Mukherjee, C. Robinson, A. Ortiz, C. West, J. M. L. Ferres, and R. T. Ng, "Reducing bias and increasing

utility by federated generative modeling of medical images using a centralized adversary," in *Proceedings of the Conference on Information Technology for Social Good*, 2021, pp. 79–84.

[137] R. Ratcliff, "Connectionist models of recognition memory: constraints imposed by learning and forgetting functions." *Psychological review*, vol. 97, no. 2, p. 285, 1990.

[138] S. Ribaric and N. Pavesic, "An overview of face de-identification in still images and videos," in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 04, 2015, pp. 1–6.

[139] P. Richardson, I. Griffin, C. Tucker, D. Smith, O. Oechsle, A. Phelan, and J. Stebbing, "Baricitinib as potential treatment for 2019-ncov acute respiratory disease," *Lancet (London, England)*, vol. 395, no. 10223, p. e30, 2020.

[140] A. Robins, "Catastrophic forgetting, rehearsal and pseudorehearsal," *Connection Science*, vol. 7, no. 2, pp. 123–146, 1995.

[141] D. Rolnick *et al.*, "Experience replay for continual learning," *NeurIPS*, 2019.

[142] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[143] H. R. Roth, L. Lu, A. Farag, H.-C. Shin, J. Liu, E. B. Turkbey, and R. M. Summers, "Deeporgan: Multi-level deep convolu-

tional networks for automated pancreas segmentation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2015, pp. 556–564.

[144] H. R. Roth, L. Lu, A. Farag, A. Sohn, and R. M. Summers, "Spatial aggregation of holistically-nested networks for automated pancreas segmentation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2016, pp. 451–459.

[145] H. R. Roth, L. Lu, N. Lay, A. P. Harrison, A. Farag, A. Sohn, and R. M. Summers, "Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation," *Medical image analysis*, vol. 45, pp. 94–107, 2018.

[146] A. G. Roy *et al.*, "Braintorrent: A peer-to-peer environment for decentralized federated learning," *arXiv:1905.06731*, 2019.

[147] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature machine intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[148] F. Rundo, C. Spampinato, G. L. Banna, and S. Conoci, "Advanced deep learning embedded motion radiomics pipeline for predicting anti-pd-1/pd-l1 immunotherapy response in the treatment of bladder cancer: Preliminary results," *Electronics*, vol. 8, no. 10, p. 1134, 2019.

[149] A. Sadilek, L. Liu, D. Nguyen, M. Kamruzzaman, S. Serghiou, B. Rader, A. Ingerman, S. Mellem, P. Kairouz, E. O. Nsoesie

et al., "Privacy-first health research with federated learning," *NPJ digital medicine*, vol. 4, no. 1, p. 132, 2021.

[150] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[151] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[152] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[153] A. A. A. Setio, F. Ciompi, G. Litjens, P. Gerke, C. Jacobs, S. J. Van Riel, M. M. W. Wille, M. Naqibullah, C. I. Sánchez, and B. van Ginneken, "Pulmonary nodule detection in ct images: false positive reduction using multi-view convolutional networks," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1160–1169, 2016.

[154] M. J. Sheller, G. A. Reina, B. Edwards, J. Martin, and S. Bakas, "Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation," in

International MICCAI Brainlesion Workshop. Springer, 2018, pp. 92–104.

[155] F. Shi, J. Wang, J. Shi, Z. Wu, Q. Wang, Z. Tang, K. He, Y. Shi, and D. Shen, "Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for covid-19," *IEEE reviews in biomedical engineering*, 2020.

[156] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[157] N. Shoham *et al.*, "Overcoming forgetting in federated learning on non-iid data," *arXiv:1910.07796*, 2019.

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

[159] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[160] A. C. Society, "Cancer facts & figures," *American Cancer Society*, 2021.

[161] J. Song and J. C. Ye, "Federated cyclegan for privacy-preserving image-to-image translation," *arXiv:2106.09246*, 2021.

[162] S. Suganyadevi, V. Seethalakshmi, and K. Balasamy, "A review on deep learning in medical image analysis," *International Journal of Multimedia Information Retrieval*, vol. 11, no. 1, pp. 19–38, 2022.

[163] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 843–852.

[164] T. Sun, D. Li, and B. Wang, "Decentralized federated averaging," *arXiv:2104.11375*, 2021.

[165] 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.

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

[167] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," *arXiv preprint arXiv:2012.12877*, 2020.

[168] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific data*, vol. 5, no. 1, pp. 1–9, 2018.

[169] L. Tunstall, L. Von Werra, and T. Wolf, *Natural language processing with transformers.* " O'Reilly Media, Inc.", 2022.

[170] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[171] H. Wang *et al.*, "Federated learning with matched averaging," *arXiv:2002.06440*, 2020.

[172] W. Wang, Q. Song, R. Feng, T. Chen, J. Chen, D. Z. Chen, and J. Wu, "A fully 3d cascaded framework for pancreas segmentation," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 2020, pp. 207–211.

[173] X. Wang, Y. Han, C. Wang, Q. Zhao, X. Chen, and M. Chen, "In-edge ai: Intelligentizing mobile edge computing, caching and communication by federated learning," *Ieee Network*, vol. 33, no. 5, pp. 156–165, 2019.

[174] X. Wang *et al.*, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *CVPR*, 2017.

[175] Y. Wang, G. Gong, D. Kong, Q. Li, J. Dai, H. Zhang, J. Qu, X. Liu, and J. Xue, "Pancreas segmentation using a dual-input v-mesh network," *Medical Image Analysis*, vol. 69, p. 101958, 2021.

[176] Z. Wang, M. Song, Z. Zhang, Y. Song, Q. Wang, and H. Qi, "Beyond inferring class representatives: User-level privacy leak-

age from federated learning," in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 2019, pp. 2512–2520.

[177] T. Wink and Z. Nochta, "An approach for peer-to-peer federated learning," in *2021 51st Annual IEEE/IFIP DSN-W*, 2021.

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

[179] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *ECCV*, 2018, pp. 305–321.

[180] Y. Xie, J. Zhang, C. Shen, and Y. Xia, "Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*. Springer, 2021, pp. 171–180.

[181] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *NIPS*, 2015.

[182] C. Xu, J. Ren, D. Zhang, Y. Zhang, Z. Qin, and K. Ren, "Ganobfuscator: Mitigating information leakage under gan via differential privacy." *IEEE Trans. Inf. Forensics Secur.*, vol. 14, no. 9, pp. 2358–2371, 2019.

[183] J. Xu, B. S. Glicksberg, C. Su, P. Walker, J. Bian, and F. Wang, "Federated learning for healthcare informatics," *Journal of Healthcare Informatics Research*, vol. 5, pp. 1–19, 2021.

[184] Q. Yang *et al.*, "Federated machine learning: Concept and applications," *ACM TIST*, vol. 10, no. 2, pp. 1–19, 2019.

[185] J. Yoon, J. Jordon, and M. van der Schaar, "PATE-GAN: Generating synthetic data with differential privacy guarantees," in *International Conference on Learning Representations*, 2019.

[186] Q. Yu, L. Xie, Y. Wang, Y. Zhou, E. K. Fishman, and A. L. Yuille, "Recurrent saliency transformation network: Incorporating multi-stage visual cues for small organ segmentation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8280–8289.

[187] J. R. Zech *et al.*, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study," *PLoS medicine*, vol. 15, no. 11, 2018.

[188] L. Zhang, J. Xu, P. Vijayakumar, P. K. Sharma, and U. Ghosh, "Homomorphic encryption-based privacy-preserving federated learning in iot-enabled healthcare system," *IEEE Transactions on Network Science and Engineering*, 2022.

[189] L. Zhang, B. Shen, A. Barnawi, S. Xi, N. Kumar, and Y. Wu, "Feddpgan: federated differentially private generative adversarial networks framework for the detection of covid-19 pneumonia," *Information Systems Frontiers*, vol. 23, no. 6, pp. 1403–1415, 2021.

[190] R. Zhang *et al.*, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018.

[191] Y. Zhang, H. Liu, and Q. Hu, "Transfuse: Fusing transformers and cnns for medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*. Springer, 2021, pp. 14–24.

[192] B. Zhao *et al.*, "IDLG: Improved deep leakage from gradients," *arXiv:2001.02610*, 2020.

[193] N. Zheng, S. Du, J. Wang, H. Zhang, W. Cui, Z. Kang, T. Yang, B. Lou, Y. Chi, H. Long *et al.*, "Predicting covid-19 in china using hybrid ai model," *IEEE Transactions on Cybernetics*, 2020.

[194] Y. Zhou, L. Xie, E. K. Fishman, and A. L. Yuille, "Deep supervision for pancreatic cyst segmentation in abdominal ct scans," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2017, pp. 222–230.

[195] Y. Zhou, L. Xie, W. Shen, Y. Wang, E. K. Fishman, and A. L. Yuille, "A fixed-point model for pancreas segmentation in abdominal ct scans," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2017, pp. 693–701.

[196] H. Zhu, R. S. M. Goh, and W.-K. Ng, "Privacy-preserving weighted federated learning within the secret sharing framework," *IEEE Access*, vol. 8, pp. 198 275–198 284, 2020.

[197] L. Zhu *et al.*, "Deep leakage from gradients," *NeurIPS*, 2019.

[198] N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu *et al.*, "A novel coronavirus from patients with pneumonia in china, 2019," *New England Journal of Medicine*, 2020.