Adoption of Deep Learning Models and its Applications in Dementia Research

Jonatan Reyes

A Thesis

in

The Department

of

Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy (Software Engineering) at

Concordia University

Montréal, Québec, Canada

August 2024

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Jonatan Reyes**

Entitled: **Adoption of Deep Learning Models and its Applications in Dementia Research**

and submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy (Software Engineering)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

_____ Chair
Dr. Andrea Schiffauerova

_____ External Examiner
Dr. Amber Simpson

_____ Examiner
Dr. Christophe Grova

_____ Examiner
Dr. Thomas Feves

_____ Supervisor
Dr. Marta Kersten-Oertel

Approved by _____
Dr. Leila Kosseim, Program Director
Department of Computer Science and Software Engineering

_____ 2024

_____
Dr. Mourad Debbabi, Dean
Gina Cody School of Engineering and Computer Science

Abstract

Adoption of Deep Learning Models and its Applications in Dementia Research

Jonatan Reyes, Ph.D.

Concordia University, 2024

Artificial Intelligence (AI) is at the forefront of the Fourth Industrial Revolution, fundamentally transforming industries and societies through unprecedented automation and data-driven applications. The Fourth Industrial Revolution is characterized by a fusion of software and hardware improvements, creating a seamless blend of the physical, digital, and biological spheres. These improvements make it possible for AI to leverage and process vast amounts of information to generate actionable insights, and perform complex tasks more quickly and more accurately than humans, leading to more informed decisions and efficient processes.

Despite its success and promising results in other domains, the adoption and integration of AI innovations in healthcare has been complex and slow. Few AI innovations have met with success and have been incorporated into daily practice. This thesis addresses technological, legal and ethical issues that must be mitigated before AI-based systems can be fully adopted and trusted into clinical trials and workflows. We identify an opportunity to further the state-of-the-art of AI solutions and their adoption in healthcare through privacy-preserving aggregation algorithms and human-centered evaluations of transparency in clinical decision support systems.

In particular, this dissertation explores advanced methodologies in Federated Learning (FL) for improving collaborative learning, data privacy, and decision-making across various domains. We improve the core FL aggregation algorithm for better handling the learning of distributed heterogeneous data sources, with a method named Precision-weighted Federated Learning. We perform extensive evaluations with benchmark datasets on resource-constrained environments to measure

its limits and perform additional tests on clinical data to enhance the quality of clinical assessment analysis, validating its utility.

Our research also aims to understand how to visualize AI model outputs to enhance transparency in clinical decision support systems. We conduct extensive evaluations to assess the impact of visualizing AI uncertainty and personal traits on decision-making, supporting the design of AI outputs that are interpretable by clinicians. Initially, we explore the effects in low-risk gaming scenarios, followed by an examination of AI uncertainty representation in high-stakes clinical decision-making, particularly in Alzheimer's disease prognosis.

In summary, this dissertation presents significant advancements in FL and clinical decision support systems. We address some of the current limitations and challenges of adopting AI systems, and demonstrate improvements in collaborative learning, data privacy, and human-AI decision-making. These findings offer valuable insights for designing robust, efficient, and trustworthy AI and FL systems. We believe that user-centered design practices will eventually play a more prominent role in the development of AI tools and technologies, becoming the driving force behind moving innovations from the laboratory to the clinic.

Acknowledgements

Thanks to God, all the glory to your name.

Thanks to Marta, my mentor.

Thanks to Cynthia, my crown, *mi conjunto*.

Thanks to Benjamin, Rebecca and Matthias, my arrows.

Thanks to my parents, siblings, and in-laws for their support and prayers.

Thanks to Xiao, Anil, Mina, Dominik, Alireza, Lisa, Cecile, all members of my committee, and everyone else who helped produce this thesis.

Contributions of Authors

I am the first author of all 4 manuscripts included in this dissertation, listed here in order of appearance. My contribution in this work was significant as I developed all of the artifacts, hereby described in this manuscript, such as algorithms, data collection and preparation, design of surveys, analysis of results and the creation of all manuscripts. The contribution of co-authors include supervision, software development, technical discussing, and reviewing of manuscripts.

(1) **Reyes, J.**, Di Jorio, L., Low-Kam, C., & Kersten-Oertel, M. (2021). Precision-weighted federated learning. arXiv preprint arXiv:2107.09627. ***Patent pending: **Reyes, J.**, Kersten-oertel, M., Cecile, L. K., Di Jorio, L., Lacaille, P., & Chandelier, F. (2024). U.S. Patent Application No. 18/273,255.

Contributions: Guarantors of integrity of the study: all authors; study and design concepts: all authors; software development: J.R.; data preparation and analysis: J.R.; supervision: L. Di J., M.K-0.; manuscript preparation: J.R.; manuscript revision: all authors; editing and final version: all authors.

(2) **Reyes, J.**., Noroozi, A., Xiao, Y., & Kersten-Oertel, M. Bridging the Gaps: Imputation of Parkinson's Disease Clinical Assessments with Federated Learning. *Submitted to IEEE Journal of Biomedical Imaging (2024).*

Contributions: Guarantors of integrity of the study: all authors; study and design concepts: all authors; software development: J.R. and A.N.; data collection: Y.X.; data preparation and analysis: J.R.; supervision: M.K.-O. and Y.X.; manuscript preparation: J.R.; manuscript revision: all authors; editing and final version: J.R. and M.K.-O.

(3) **Reyes, J.**., Ludera, D., Batmaz A. U., & Kersten-Oertel, M. Game On: How Human Perception of AI Uncertainty Shapes Decision-Making. *Submitted to Frontiers in Computer Science (2024)*.

Contributions: Guarantors of integrity of the study: all authors; study and design concepts: all authors; software development: J.R.; data collection: J.R., A.U.B. & M.K.-O.; data preparation and analysis: J.R.; supervision: A.U.B. & M.K.-O.; manuscript preparation: J.R.; manuscript revision: J.R., A.U.B. & M.K.-O.; editing and final version: J.R. and M.K.-O.

(4) **Reyes, J.**, Masoumi, M., Batmaz A. U., & Kersten-Oertel, M. Towards Trustworthy Predictions of Alzheimer's Disease under AI Uncertainty. Submitted to MICCAI UNSURE Workshop and extended for journal paper TBD.

Contributions: Guarantors of integrity of the study: all authors; study and design concepts: all authors; software development: J.R. and M.M.; data collection: J.R., M.M. & M.K.-O.; data preparation and analysis: J.R. and M.M.; supervision: A.U.B. & M.K.-O.; manuscript preparation: J.R.; manuscript revision: J.R., A.U.B. & M.K.-O.; editing and final version: J.R. and M.K.-O.

Furthermore, during the thesis I was first or second author of the following papers:

(1) Aktar, M., **Reyes, J.**, Tampieri, D., Rivaz, H., Xiao, Y., & Kersten-Oertel, M. (2023). Deep learning for collateral evaluation in ischemic stroke with imbalanced data. International Journal of Computer Assisted Radiology and Surgery, 18(4), 733-740.

(2) **Reyes, J.**, El-Mufti, N., Gorman, S., Xie, D., & Kersten-Oertel, M. (2022, September). User-Centered Design for Surgical Innovations: A Ventriculostomy Case Study. In Workshop on the Ethical and Philosophical Issues in Medical Imaging (pp. 51-62). Cham: Springer Nature Switzerland.

(3) **Reyes, J.**., Xiao, Y., & Kersten-Oertel, M. (2021). Data Imputation and Reconstruction of Distributed Parkinson's Disease Clinical Assessments: A Comparative Evaluation of Two

Aggregation Algorithms. In Clinical Image-Based Procedures, Distributed and Collaborative Learning, Artificial Intelligence for Combating COVID-19 and Secure and Privacy-Preserving Machine Learning: 10th Workshop, CLIP 2021, Second Workshop, DCL 2021, First Workshop, LL-COVID19 2021, and First Workshop and Tutorial, PPML 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27 and October 1, 2021, Proceedings 2 (pp. 163-173). Springer International Publishing.

(4) Léger, É., **Reyes, J.**, Drouin, S., Popa, T., Hall, J. A., Collins, D. L., & Kersten-Oertel, M. (2020). MARIN: an open-source mobile augmented reality interactive neuronavigation system. International journal of computer assisted radiology and surgery, 15(6), 1013-1021.

(5) Léger, É., **Reyes, J.**, Drouin, S., Collins, D. L., Popa, T., & Kersten-Oertel, M. (2018). Gesture-based registration correction using a mobile augmented reality image-guided neurosurgery system. Healthcare technology letters, 5(5), 137-142.

Abbreviations

Aβ	Amyloid-β protein
AD	Alzheimer's Disease
AEs	stacked denoising autoencoders
AI	Artificial intelligence
ANOVA	Analysis of Variance
AR	Augmented Eeality
BJLO	Benton Judgement of Line Orientation
BNT	Boston Naming Test Score
CCPA	California Consumer Privacy Act
CDSS	Clinical Decision-making Support Systems
CSF	Cerebrospinal Fluid
CWT	Cyclic Weight transfer
DL	Deep learning
DNNs	Deep Neural Networks
DTI	Disposition to Trust Inventory
EHR	Electronic Health Records
EMA	European Medicines Agency
ESS	Epworth Sleepiness Scale
EU AI	European Regulation on Artificial Intelligence
FCAE	Fully Connected Autoencoder
FDA	Food & Drug Administration

FedAvg	Federated Averaging
FedProx	FL Optimization
FL	Federated learning
GAAIS	General Attitudes towards Artificial Intelligence Scale
GBD	Global Burden of Disease
GDP	gross domestic product
GDPR	General Data Protection Regulation
GDS	Geriatric Depression Scale
HCI	Human-Computer Interaction
HIPAA	Health Insurance Portability and Accountability Act
HNY	Hoehn Yahr
HVLT	Hopkins Verbal Learning Test
IID	Independent and Identical Distributions
Industry 4.0 or 4IR	Fourth Industrial Revolution
IR	Internal Representation
LFS	Lexical Fluency Score
LNS	Letter-Number Sequencing
MCI	Mild Cognitive Impairment
MDS-UPDRS	Movement Disorder Society-Unified Parkinson's Disease Rating Scale
ML	Machine learning
MMSE	Mini-Mental State Examination
MoCA	Montreal Cognitive Assessment
MRI	Magnetic Resonance Imaging
MSE	Mean Squared Error

MSE	Mean Squared Error
MSE-ADL	Modified Schwab & England Activities of Daily Living
MTL	Multi-Task Learning
NN	Neural Networks
Non-IID	Non-Identical and Non-Independent
PD	Parkinson's Disease
PET	Positron Emission Tomography
PIGD	Postural Instability and Gait Difficulty
PPMI	Parkinson Progression Marker Initiative
PW	Precision-weighted Federated Learning
QUIP	Questionnaire for Impulsive-Compulsive Disorders
RBDQ	REM Behaviour Disorder Questionnaire
ReLU	Rectified Linear Unit
SCOPA-AUT	Scale for Outcomes in Parkinson's Disease for Autonomic symptoms
SDMT	Symbol Digit Modalities Text
SFT	Semantic Fluency Test
SGD	stochastic gradient descent
SHAP	SHapley Additive exPlanations
SPECT	Photon Emission Computed Tomography
STAI	State-Trait Anxiety Inventory
UPDRS	Unified Parkinson's Disease Rating scale
US	United States
XAI	explainable AI

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Artificial intelligence (AI), a field where computers mimic or replicate human problem-solving and decision-making abilities, stands at the forefront of the Fourth Industrial Revolution (Industry 4.0) [3]. It is profoundly shaping people's work and lives through automation, text-mining, personalized experiences, enhanced healthcare diagnostics, improved educational tools, advancements in autonomous transportation, financial innovations, entertainment applications, and beyond. The unpredictable growth in the availability of data as well as improvements in hardware technologies have enabled AI technologies to succeed in many applications [4].

In particular, *deep learning* (DL), a subset of AI, has gain widespread attention due to its data processing capabilities and accurate predictions. The success of DL algorithms stems from their ability to process data and identify key features in their raw form without human intervention. In contrast, *machine learning* (ML), its predecessor, requires human expertise to refine raw data into useful features for a model to differentiate between classes (i.e. feature engineering). These differences are illustrated in Figure 1.1. DL's capability to autonomously process complex data allows it to handle tasks that are difficult or even impossible for humans to interpret directly from the raw data, such as recognizing intricate patterns in images or understanding complex speech signals. This has made DL particularly effective in multiple fields including computer vision, speech recognition, natural language processing [5, 6, 7] and other domains where large amounts of complex data are involved. Thus, the adoption of DL-based applications, holds significant promise in shaping the future of various disciplines, economies, and industries — particularly in critical sectors like healthcare.

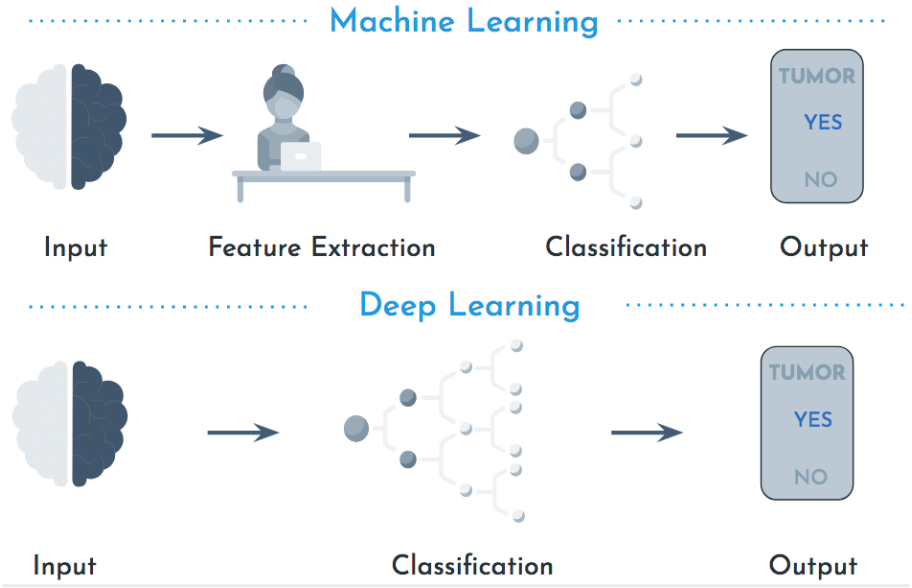

Figure 1.1: Comparison between machine learning and deep learning.

DL is driving a transformation in the healthcare industry. Biomedical research generates a variety of data sources in the form of clinical imaging, electronic health records, sensor data, and genetic information. This wealth of data has significantly enhanced data-driven processes and workflows, converting patient data into valuable insights for healthcare providers, spanning diagnostics to treatment. The combination of clinical data, hardware acceleration, substantial storage capacity, and DL methods has resulted in major advancements in numerous medical disciplines, including radiology, oncology, cardiology, ophthalmology, neurology, internal medicine and general practice decision-making.

However, the adoption and integration of AI innovations into clinical practice is complex. As evidence, the number of papers describing AI methods for medical applications increased from 596 in 2010 to 12,422 in 2019, but only 64 received regulatory approval in the US in 2019 [8]. A similar trend was observed in 2022, where only 139 devices were approved that year [9]. This underscores the need for AI solutions to adhere to rigorous standards and ethical guidelines established by regulatory bodies. Key agencies responsible for protecting the public health and, in this context, overseeing the development of AI medical solutions include the US Food & Drug Administration (FDA), the European Medicines Agency (EMA), General Data Protection Regulation (GDPR), and

the European regulation on artificial intelligence (EU AI) Act. These bodies ensure that AI innovations comply with various regulations and standards, emphasizing transparency, safety reporting, and socio-demographic representation. Technological, legal and ethical issues must be mitigated, before AI-based systems can be fully adopted and trusted into clinical trials and workflows.

Addressing technical and legal considerations

Federated Learning (FL) [10] emerges as a promising solution to address some of the technological and legal issues associated with the adoption of AI-based systems in domains where data privacy is particularly important, such as healthcare. FL is a framework that enables collaborative learning across a group of participants without compromising the privacy of its participants. In FL learning applications, only the model updates are shared, rather than sensitive raw data. The original FL method was aimed at maximizing the user experience of mobile applications by learning user behaviors across distributed learning models deployed to a group of mobile devices, resulting in more accurate next-word predictions and improved face/voice recognition [10].

In a clinical context, each medical center functions as a learning device, with neural networks trained on local patient data that remains within the center. This ensures compliance with GDPR data protection regulations [11] and exemplifies a privacy-preserving method gradually being adopted in medical applications. The literature shows that FL has been gradually applied to multiple tasks such as predictive healthcare [12], organ and pathology segmentation [13], medical imaging classification [14], and biomedical data curation [15]. Its gradual implementation not only suggests a promising solution to the analysis and aggregation of sensitive patient data distributed across multiple clinical institutions, but it also addresses the challenge of data volume commonly encountered in clinical studies using deep learning solutions.

Addressing ethical considerations

A substantial amount of resources are invested into healthcare to support, strengthen, and transform healthcare systems around the world. In 2011, the US reported an expenditure of 18% (3.2 trillion) of its gross domestic product (GDP) into healthcare [16], Their healthcare expenses are expected to grow 7.5 percent by 2023 [17]. Despite the large investments and enthusiasm about the

potential for AI to transform the healthcare industry, the literature reports that AI technologies in healthcare had the slowest rate of adoption in 2019 [18]. Notably, it is believed that poor health outcomes are attributed to sub-optimal decision-making [19, 20]. Decision-making is a central activity in healthcare, requiring clinicians to make high-risk decisions based on prior knowledge, the reliability of evidence and recommendations, and their own judgement [21]. AI has the potential to enhance decision-making under certain conditions. The GDPR prohibits solely automated decision-making and processing of health data, except with patient consent or for public interest reasons [22]. This regulation ensures that the AI system becomes a tool to support human decisions, rather than simply replace human decisions with algorithms [23].

As a response to both of these issues, the computer science and human-computer interaction (HCI) communities have produced significant research work, literature reviews, and special issues to improve clinical application based on FL and clinical decision-making based on AI [24]. Figure 1.2 shows the growing interest among the community on both of these topics over the years. With respect to clinical decision-making, the focus of these efforts is to slowly shift the development of technologies from technology-centered design [25], which focuses on algorithm, to human-centered design, or "human-AI" [26], which prioritizes the usability of systems. This is particularly important in the medical domain, where severe consequences can resulting in direct adverse effects on patients health and treatment. To that extent, transparent, informed, and rational clinical decision-making is of utmost importance for the successful adoption of AI in healthcare.

1.1 Motivation

Despite the promising results achieved with deep learning (DL), several issues have limited its broader impact in healthcare. These limitations are related to the inherent characteristics of medical data, such as its complexity, poor annotation, unstructured nature, high dimensionality, and heterogeneity. Other challenges hindering the adoption of DL in medical settings include architectural limitations such as interpretability, catastrophic forgetting, the vanishing gradient problem, model compression, and overfitting. Moreover, ethical concerns related to the lack of transparency, trust in AI solutions, and decision-making further contribute to the cautious adoption of DL in healthcare.

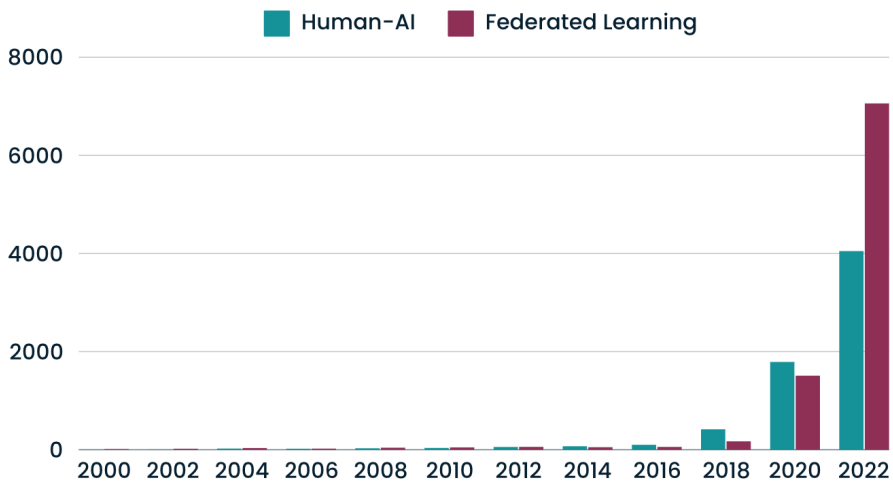

Figure 1.2: Relevance of Federated Learning and human-AI decision-making over the years among the research community. Data compiled from Google Scholar in June 2024.

Given the technical challenges of current DL methods and the compliance of FL with data protection regulations, we identify an opportunity to further the state-of-the-art of AI solutions and their adoption in clinical settings. We posit that this can be achieved through a combination of improved FL-based algorithms and extensive studies on human-AI interactions in decision-making. Although FL is a viable tool to support privacy-preserving distributed training, it has several limitations related to the network bandwidth, additional privacy guarantees and model protection, reproducibility and interpretability and communication bottlenecks [27]. In particular, the literature shows that the core averaging algorithm's performance on heterogeneous data is still an active area of research [28, 29, 30]. Therefore, in the development of the work in this thesis, we were motivated to (1) creating novel aggregation strategies to improve the learning of distributed machine and deep learning models, especially for heterogeneous data sources, and (2) finding the balance between confidence and caution in clinical decision-making through extensive evaluations of trust and decision-making via human-AI interaction research.

1.2 Objectives and Contributions

This dissertation aims to enhance the adoption of AI-based technologies. Our contributions are centered on the development of privacy-preservation algorithms and comprehensive analyses of

the effect of AI uncertainty and its impact on trust and decision-making, while using neurological disorders as case studies. Specifically, a novel FL-based aggregation algorithm was developed and validated through clinical assessments of Parkinson's disease. We also conducted extensive evaluations about the interpretability of AI uncertainty using various visual formats and their effects on trust and decision-making for both low-stakes and high-stakes scenarios, i.e. Alzheimer's disease prognosis. The main contributions of this dissertation are:

(1) A privacy-preserving aggregation algorithm.

- The introduction of Precision-weighted Federated Learning, a novel algorithm that leverages the inherent heterogeneity in the training data to perform the aggregation of distributed models. We perform exploratory evaluations based on performance and reliability with benchmark datasets, simulating different data distributions (Chapter 3).
- The demonstration of the clinical utility of the proposed algorithm. We extend performance evaluations between various FL-based aggregation algorithms for the imputation of distributed clinical assessments. With a downstream analysis, we measure the performance of these algorithm in the classification of PD patients based on symptom progression (Chapter 4).

(2) Human-centered evaluations of transparency in clinical decision support systems.

- The assessment of the impact of visual representations of AI uncertainty in low-stake decision-making. We measure changes in human perceptions (e.g. change of decision, trust in AI, confidence in decisions) in response to visual uncertainty in static gaming scenarios among individuals with different attitudes towards AI.(Chapter 5).
- The evaluation of different elements to assist in the understanding (e.g. description of AI's decision-making process and visual methods to show uncertainty) of AI uncertainty in high-stake decision-making. We investigate how individual's perceptions are influenced when additional information about the AI model is provided when making predictions of Alzheimer's disease progression (Chapter 6).

In summary, the work presented in this thesis address two of the challenging problems preventing a smooth transition of AI innovations from research to clinical practice, concerning data protection and transparency. Firstly, we leverage the data privacy-preserving mechanisms in the FL framework and develop a novel variance-based algorithm for aggregating models. To validate the clinical utility of the method, we present extensive evaluations about the generalization of distributed models and predictive power in a downstream analysis. Secondly, we encourage the development of AI tools and methods promoting transparent solutions. In particular, we assess the human perception of trustworthy AI systems in both low and high risk AI systems. Next-generation medical tools can profoundly impact healthcare , but safety and data privacy, the approach to clinical decision-making, and the transparency of clinical innovations need to be considered. These factors can collectively enable compliance with regulatory standards as well as the successful adoption of AI systems.

1.3 Thesis Outline

The remainder of this dissertation is organized as follows. Chapter 2 provides an overview of deep learning, Federated Learning, and AI decision-support systems in healthcare. Chapter 3 introduces a novel Federated Learning-based aggregation methods and presents evaluations using benchmark datasets. Chapter 4 provides in-depth evaluations of the proposed algorithm in a clinical setting, along with a downstream analyses to validate its clinical utility. Chapter 5 assesses the interpretability of AI solutions under uncertainty in low-stake decisions scenarios. In Chapter 6, we extend the evaluations of AI interpretability to its impact on trust and decision-making in clinical scenarios and specifically in Alzheimer's disease. Lastly, we conclude and present future directions of our work in Chapter 7.

Chapter 2

Background

This chapter begins with an introduction to Artificial Intelligence (AI) and continues with a discussion about Federated Learning (FL). This is followed by a discuss on clinical decision support systems and their associations with human-AI interactions. Lastly, we briefly present the area of application of this thesis, neurodegenerative diseases focusing on Parkinson's Disease (PD) and Alzheimer's Disease (AD).

2.1 Artificial Intelligence

Artificial Intelligence (AI) was officially introduced by John McCarthy at a 1956 conference at Dartmouth College, where it was defined as "the science and engineering of making intelligent machines" [31]. Initially, the goal was to explore how machines could perform cognitive tasks (e.g. multi-step reasoning, understand natural language, create designs, and reason about their own reasoning) [32]. Consequently, the field lies at the intersection of cognitive science and computer science. Due to its practical successes, particularly in machine learning, AI has generated significant research interest over the decades.

Machine learning (ML) is a sub-discipline of artificial intelligence (AI)(Figure 2.1). Its general-purpose algorithms are based on the "learning by example" principle and are intended to replicate human intelligence. The intuition behind the basic machine learning concepts and operations were originated after observing how neurons are interconnected and activated in the brain. Hence the

name neural networks. A typical neural network architecture contains units called *neurons* that transmit a signal in response to a stimulus based on the decisions for which these neurons are governed. The most common activation functions are Sigmoid, Tanh, Rectified Linear Unit (ReLU), and Leaky ReLU. These activation functions normalize the output signal and compute the gradient values of the *network parameters* (weights or bias) in single non-linear processing *hidden* layer. The fine-tuning process of the gradients is carried out by the *back-propagation* algorithm, which calculates the partial derivatives of a given cost function with respect to the network parameters. Thus, the back-propagation step allows the neural network to "learn".

One of the major limitations of traditional machine learning is the need for human intervention and expertise in the refinement of the raw data into useful features to help the model distinguish between classes. For example, engineers would craft a feature extractor that would clean and curate the raw data to facilitate the machine learning algorithm to find patterns, and thus, to perform a more accurate classification analysis [33]. This technique is called *feature engineering*. Conventional feature engineering techniques are often limited in their ability to learn new patterns of features from the raw data directly, but this particular limitation is addressed with deep learning.

Conceptually, *deep learning (DL)* lies within the broader field of machine learning (Figure 2.1). The success of DL algorithms is attributed to the ability to process natural data and to discover key features in their raw form without human intervention. As the input data is transformed through multiple hidden layers, the representation of features is learned with multiple levels of abstraction [33]. Miotto *et al.* [34] further differentiates deep learning from machine learning in the unrestricted number of hidden layers, connections, and their capability to learn meaningful abstractions of the inputs, which makes deep learning algorithms more powerful than its predecessor. Interestingly, deep learning paradigms, coupled with the availability of hardware acceleration and phenomenal storage capacity, have resulted in major advances to many prediction tasks, including computer vision, natural language processing, and speech recognition.

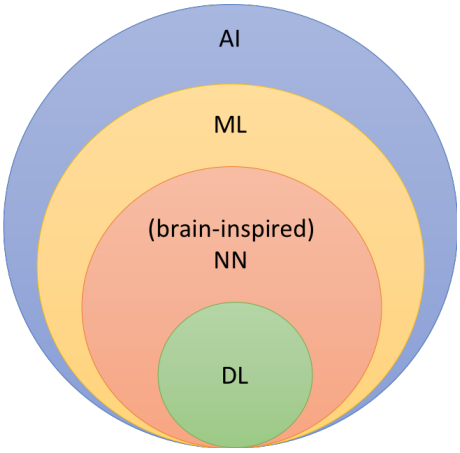

Figure 2.1: AI Taxonomy: Artificial Intelligence (AI), Machine Learning (ML), Neural Networks (NN), and Deep Learning (DL)

2.1.1 Challenges in Machine and Deep Learning

We describe the challenges and shortcoming of machine and deep learning for clinical applications. Clinical data is often not ready for deep learning analysis as it can be incomplete, imbalanced, or ambiguous, preventing its safe utilization for decision support and appropriate delivery of care [35]. Moreover, DL approaches require large amounts of data for fine-tuning model parameters, which is not always possible to compile due to data availability, privacy and security concerns. Clinical data is also highly complex and heterogeneous due to the variations in modalities, hardware configurations, acquisition protocols, genetic evolution, and demographics [36, 37]. As such, the assumption of enough high-quality training data that is balanced in both the number of samples and classes, *independent and identical distributions (IID)*, does not hold.

As well as data-related issues, there are other common limitations specific to applying DL models in health related fields. It is crucial for clinicians to be confident about the soundness of the AI recommendation, alongside the final prediction. Clinicians need to assess both the reliability and limitations of the model (uncertainty scaling) and decide whether or not to trust in the algorithmic advice. It is also important to understand the patterns and structures of features identified by models (interpretability) to mitigate the *black-box* effect in machine/deep learning networks. This can help clinicians recognize the phenotypic properties of diseases driving the predictions [34]. Other challenges are related to the tendency of AI models to forget old information when processing new

data (catastrophic forgetting), difficulty updating weights with very small gradients (vanishing gradient problem) or very large gradients (exploding gradient problem), and the complexity of the architecture due to the large number of parameters (model compression), which affects the model's portabibiliy. Additionally, AI models often struggle to generalize to new, unseen data (overfitting) and may perform poorly in target domain (underspecifications) [38, 39]. As such, the inherent characteristics in the training data as well as the architecture and capabilities of AI models contributes to biases during the training phase.

To address some of the previous issues, multiple techniques have been developed to enhance model learning. *Transfer learning* allows a model to apply knowledge from a previous learned task to a new, related task. However, this transfer of knowledge can sometimes negatively impact performance, when the new task is less related or more complex to the previously learned task. In such cases, additional human expertise is often needed to create a proper mapping that translates representations between tasks [40]. A special type of transfer learning is *domain adaptation*, which relaxes mismatches between data distributions in feature space, allowing a trained model to generalize to the domain of interest [41]. Despite this, domain adaptation faces challenges due to data volume, imbalanced samples and classes, and less explored tasks such as object detection, pose estimation, and time series analysis.

Early attempts to satisfy the demand of data volume, involved a remote central server that aggregated data from different locations. This method, known as *centralized learning*, is convenient as the size of the training data is increased; however, security concerns associated with transferring data to the central server have led to the development of new data aggregation techniques that emphasize data privacy and security. As a result, new methods combine data from geographically distributed sources while protecting data privacy, making it particularly valuable to clinical applications.

2.2 Federated Learning

Collaborative learning enables two or more individuals to learn something together. In the context of computer science, this concept has been extended to distributed machine learning. Early attempts to enable collaborative learning involved a centralized data center where raw data was

collected, analyzed and processed from each participant. However, when at least one participant, or even the centralized data center, cannot be trusted, a major concern about this technique is data privacy. To accommodate for this, McMahan *et al.* [42] introduced *Federated Learning (FL)*, a technique that allows the training of decentralized machine and deep learning in a federation of mobile devices. This technique not only allows collaborative learning but also encourages data privacy as the raw data never leaves the device.

FL works in rounds of communication through a distributed batch of mobile devices to learn a shared global model. At the beginning of each round, a server sends the initial shared global model to every client. Then, every client uses the shared model to compute stochastic gradient descent (SGD) optimizations with the local data and the resulting update (i.e. DNN weights) is sent to the server for further processing. After receiving all individual updates, the central server aggregates these updates via the *Federated Averaging (FedAvg)* algorithm and updates the shared global model and the round of communication repeats again. As more and more rounds of communications are performed across clients, the model learns a better representation of the data distribution and thus performance of the shared global model is optimized. When compared to the performance of a centralized data center, federated learning not only achieves better accuracy metrics but also enhances the overall user's experience according to their device usage. Furthermore, when a new client joins the round of communication, the global model contains enough information from other clients that there is no need to re-train the model as it can be used immediately on the new device. Thus, Federated Learning is a promising solution for the analysis of privacy-sensitive data distributed across a wide distribution of clients. Figure 2.2 shows how this algorithm employs an iterative model averaging scheme in a FL framework.

2.2.1 The FedAvg algorithm

FL traditionally achieves the training of distributed machine/deep learning models with the *Federated Averaging* algorithm. The complete pseudo-code describing the FedAvg method is given in Algorithm 1.

The amount of computation depends on these hyper-parameters: the learning rate η to train

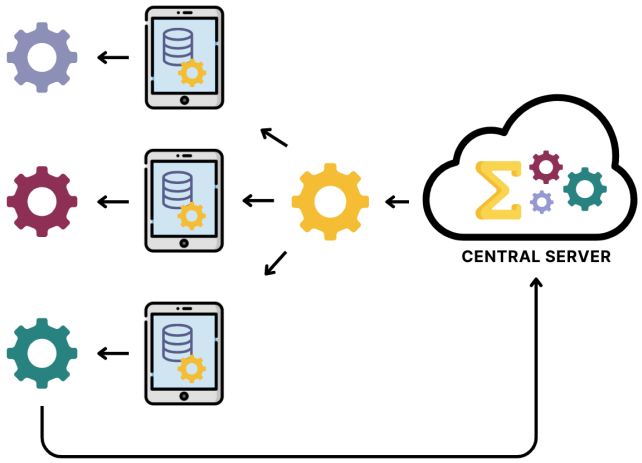

Figure 2.2: Federated Learning framework

client's models; the client fraction *C* indicating the number of clients participating in the computation on each round; the appropriate number of training passes (epochs) *E* over its local data; and the most effective local mini-batch size *B*. After training, each client shares local stochastic gradient descent (SGD) updates while minimizing the objective function described in Equation 1:

$$\min_{w \in R^d} f(w) \quad \text{with} \quad f(w) = \frac{1}{n} \sum_{i=1}^{n} f_i(w) \tag{1}$$

where n is the number of data samples and $f_i(w) = \ell(x_i, y_i; w)$ is the loss of the prediction on example (x_i, y_i) made with model parameter w. Since it is expected that the data will be partitioned over K clients, the objective function can be rewritten in terms of $\mathcal{P}_k$ (Equation 2); that is, the set of data samples on a given client k, and its size $n_k = |\mathcal{P}_k|$ as:

$$f(w) = \sum_{k=1}^{K} \frac{n_k}{n} F_k(w) \quad \text{with} \quad F_k(w) = \frac{1}{n_k} \sum_{i \in \mathcal{P}_k} f_i(w), \tag{2}$$

The aforementioned stochastic gradient descent optimization is expressed as $g_k = \nabla F_k(w_t)$, where each client k takes a step on the gradient descent with the current model parameters w_t using its local data. An equivalent update is shown in Equation 3, denoting all clients' parameters

Algorithm 1 FederatedAveraging: The K clients are indexed by k; B is the local minibatch size, E is the number of local epochs, and η is the learning rate.

Server executes:
  initialize w_0
  for each round $t = 1, 2, \dots$ **do**
    $m \leftarrow \max(C \cdot K, 1)$
    $S_t \leftarrow$ (random set of m clients)
    for each client $k \in S_t$ **in parallel do**
      $w_{t+1}^k \leftarrow \text{ClientUpdate}(k, w_t)$
    end for
    $w_{t+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} w_{t+1}^k$
  end for

function CLIENTUPDATE(k, w) ▷ Run on client k
  $\beta \leftarrow$ (split $\mathcal{P}_k$ into batches of size B)
  for each local epoch i from 1 to E **do**
    for each batch $b \in \beta$ **do**
      $w \leftarrow w - \eta \nabla \ell(w; b)$
    end for
  end for
  return w to server
end function

being updated in parallel with a fixed learning rate η. After local clients updates are computed, the parameters of each model are transferred to the server, where the server applies Equation 4. In this step, the server updates the global model with the weighted average of parameters across all clients.

$$w_{t+1}^k \leftarrow w_t - \eta g_k \text{ for all } k \tag{3}$$

$$w_{t+1} \leftarrow w_t - \eta \sum_{k=1}^{K} \frac{n_k}{n} g_k \tag{4}$$

2.3 Clinical Decision Support Systems

The literature shows a number of research work on computer-assisted tools as clinical decision support systems (CDSS). CDSSs are software tools designed to augment clinical decision-making

by allowing clinicians or patients to enter specific characteristics or symptoms of a person into a machine with enough clinical knowledge, to receive and decide on the results from personalized assessments or recommendations [43, 44]. Since their first use in the 1980s, CSSSs have aimed to assist physicians in their making complex decisions. Today, CDSSs continue to provide this support through digital mediums such as electronic health records (EHR) and other computer-based clinical workflows. Initially, conventional CDSS relied on decision rules explicitly coded based on medical experience (knowledge-based systems), but after the introduction of AI into healthcare, machine learning replaced these explicit rules (non-knowledge-based systems), expanding their level of performance [45].

Recent technological advancements have transformed the landscape for CDSSs. The widespread use of mobile computing, AI, and wireless devices enables the collection and analysis of vast amounts of data, enhancing the role of CDSS in the decision-making process. However, the internal mechanisms of machine learning and deep learning models often remain opaque (e.g., black-box). While this lack of transparency may not be a significant issue in other fields, it is critical in healthcare. Clinicians need to understand the reasoning behind AI-generated recommendations or predictions to judge whether or not to follow algorithmic advice, especially in cases of disagreement or high-stake decision making [34, 1]. As a result, many research papers, groups, and communities advocate for greater transparency in the design phase of AI models to clearly explain their underlying principles [46].

2.3.1 Decision-Making and Trust

Typically, the adoption of assistive AI systems is limited by a lack of trust of humans into an AI's prediction [47]. Trust is defined as the degree to which a person or group of people relies on or has confidence in the dependability of someone or something to fulfill their promise [48]. Thus, the development of AI must conform to the principles of trustworthy AI. This is critical for decision-makers in the healthcare domain, in situations when algorithmic advice can support their decisions. Transparency in AI decisions is essential before it can be used for patient care, ensuring that the AI is trusted when it is accurate and identified as untrustworthy when it is not [49]. There

are two interconnected areas within the field of AI that seek to enhance AI transparency, *human-AI* and *explainable AI (XAI)*. Human-AI research primarily aims to augment human capabilities and enhance decision-making processes by studying the usability, design, and overall experience of interacting with AI system. On the other hand, XAI focuses more on accurately estimating and communicating an AI model's reliability to decision-makers, providing a clearer interpretation of the AI's behavior [50]. In this works presented in this manuscript, we focus our contributions to human-AI research.

Human-AI, also known as "Human-Centered AI" is a design paradigm that emerged among the community of AI researchers as a result of the growing interest in augmenting human capabilities. This synergy would result in enhanced human decision-making process with AI assistance. As explained by Shneiderman [51], the collaboration between humans and computers involves people working alongside technology, in charge of technology, where better questions and more confident decisions can be made based on the abundant information provided by the system. The success of human-AI has lead the development of AI-driven decision support systems across various domains, such as law and civic, medicine and healthcare, finance, education, leisure, and others.

To enhance decision-making through design choices, significant efforts have been made to establish guidelines for developing AI technologies with a focus on human factors. For instance, after reviewing 20 years of research literature, Amershi [52] proposed 18 design guidelines categorized by different stages of system interaction to evaluate the development of emerging design ideas, whereas Horvitz [53] provided guidelines for the evaluation of the usability of such systems. Additionally, Lai*et al.* [1] identified common evaluation metrics for developing human-AI decision-making systems, classifying them based on their intended goals. The first category, related to the decision task, includes metrics such as performance (efficacy), speed of task completion (efficiency), task-level satisfaction, and mental demand. The second category involves evaluations of the AI itself, focusing on the user's perception of trust, fairness, usability, satisfaction, and understanding of AI predictions. These guidelines and design choices provide standardized measurements of model performance and reliability in normal, stressed, and adversarial situations, leading to improved decision-making and user experience.

2.3.2 Uncertainty

Research in organizational behavior and psychology indicates that the adoption of new technology is influenced by a users perceptions and beliefs about the technology [54]. This insight has motivated prompted researchers to explore people's perceptions of AI uncertainty as a metric for model adoption. If the AI technology is not believed to perform as well as or better than the intended user, its usability will be limited, potentially leading to its rejection. To address this issue, previous work has focused on bridging the gap between skepticism and model acceptance by designing CDSS that display AI uncertainty in their predictions and measuring its impact on decision-making through visual representations of AI uncertainty, user studies, and augmented reality environments [55, 56, 57, 58]. These techniques have been shown to provide users with a clearer understanding of AI confidence levels, thereby facilitating the adoption of algorithmic recommendations as tools to support human decisions. Figure 2.3 provides a few examples representing AI uncertainty.

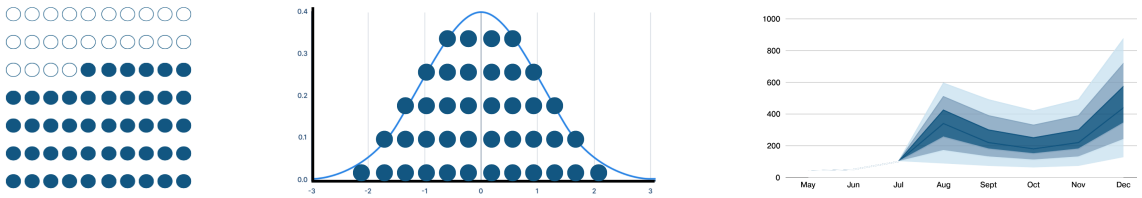

Figure 2.3: Different ways to represent AI uncertainty: icon array chart (left), quantile dot plot (middle), and predictive distribution (left)

2.4 Neurodegenerative Diseases

Neurological disorders mainly affect the nervous system, including the brain, spinal cord, and peripheral nerves. Examples of these disorders include multiple sclerosis, traumatic brain injury, stroke and diverse forms of dementia. *Neurodegenerative diseases* a subset of neurological disorders, specifically involve the progressive degeneration of neurons, leading to symptoms such as cognitive decline, motor dysfunction, and other neurological impairments. *Dementia* is a progressive neurodegenerative disease that primarily affects memory, reasoning, language, physical and

cognitive abilities. Most forms of dementia have a significant impact on the autonomy of individuals and their daily activities, leading to physical, psychological, and cognitive impairments. In this manuscript, we focus our works on Alzheimer's disease and Parkinson's disease.

2.4.1 Alzheimer's Disease

Alzheimer's disease (AD) is the most common form of dementia. Similar to the accumulation of fatty deposits in arteries, AD induces an abnormal accumulation of protein fragments β*-amyloid* outside neurons, as well as, twisted fibers of the protein *tau* inside the neurons, causing damage to the brain tissue and the neurons, mostly in the Hippocampus and temporal grey matter, stop functioning and die [59]. The effects of the damage are irreversible, and worsen over time. The global statistics of Alzheimer's estimate that 1 out of 85 individuals will be living with AD by 2025 [60], which is an imminent concern for present and future generations. In the United States, morbidity significantly increased between 2000 and 2017, with Alzheimer's-related deaths rising by 145%, making it the sixth leading cause of death in the country [59].

The prodromal stage of AD is *Mild Cognitive Impairment (MCI)*. MCI is a syndrome that expresses a greater cognitive decline than the expected for an individual's age and education level. Some patients can perform everyday tasks without noticing the syndrome as it does not interfere with activities of daily life, but in other instances MCI develops memory complaints and deficits (e.g., amnestic mild cognitive impairment). In fact, more than half of patients with MCI are at higher risk of developing AD within 5 years [61, 62]. As such, MCI currently constitutes the earliest possible traceable stage for AD and, therefore, it is critical to develop new tools that can provide better statistical estimations of these converting factors, or even better, to identify new ones for an early detection of AD.

2.4.2 Parkinson's Disease

Parkinson's disease (PD) is the second most prevalent neurodegenerative disesase. It is believed that dopamine-deprivation in the substantia nigra and basal ganglia is responsible for the development of this disease [63]. Both areas of the brain are responsible for the coordination motor movement and reward functions, which leads to classic motor symptoms such as tremor, rigidity,

and bradykinesia, which affects patients daily functions severely. In addition, psychiatric issues, including compulsive behaviors, depression, cognitive decline, and sleep disorders, can also affect PD patients [64, 65]. Approximately 40% of those with PD develop dementia [66]. According to the Global Burden of Disease (GBD) study [67] published in 2016, 6.1 million individuals experience PD around the world and, based on published prevalence studies, it is estimated that this number will significantly increase, potentially reaching 9.3 million people by 2030 [68].

2.4.3 Clinical Assessments and Medical Imaging

In practice, clinicians perform periodic evaluations to assess the degree and course of neurodegeneration. One of the first and most basic types of screening currently used in the clinic are non-computerized assessments where each screening test is administered by a certified therapist. The information obtained from screening testing help in determining the state and severity of motor and non-motor symptoms. Cognitive assessments most commonly used by clinicians are the Montreal Cognitive Assessment, Mini-Mental State Examination, Clock Drawing Test, among many others [69]. To assess motor symptoms, the most common standardized rating scale is the Unified Parkinson's Disease Rating scale (UPDRS), Hoehn and Yahr staging, and the Schwab and England rating of activities of daily living [70]

Depending on the clinical protocol, additional biomedical imaging evaluations may be performed on the patient at the time of the visit. A variety of image modalities have been used for the detection of abnormal neurodegeneration. It is common for magnetic resonance imaging (MRI) to be used to quantify the degree of neurodegeneration as it provides detailed images of brain tissues. Alternatively, positron emission tomography (PET) scans have been used to quantify the number of twisted fibers inside neurons, measure the levels of phosphorylated tau and cortical tau in the cerebrospinal fluid (CSF), identify neurodegeneration, and monitor the levels of glucose metabolism in the brain. Similarly, amyloid PET imaging and CSF testings of the amyloid protein can be utilized to measure the levels of deposits of protein fragments. More recently, additional image modalities, such as structural and functional MRI, and amyloid Florbetapir (^{18}F) tracers in PET imaging have facilitated the discrimination of damaged gray tissue linked to dementia [71]. Figure 2.4 contains three MRI scans, (left) normal cognitive decline, (middle) neurodegeneration on one patient with

Alzheimer's disease, and Parkinson's disease (right).

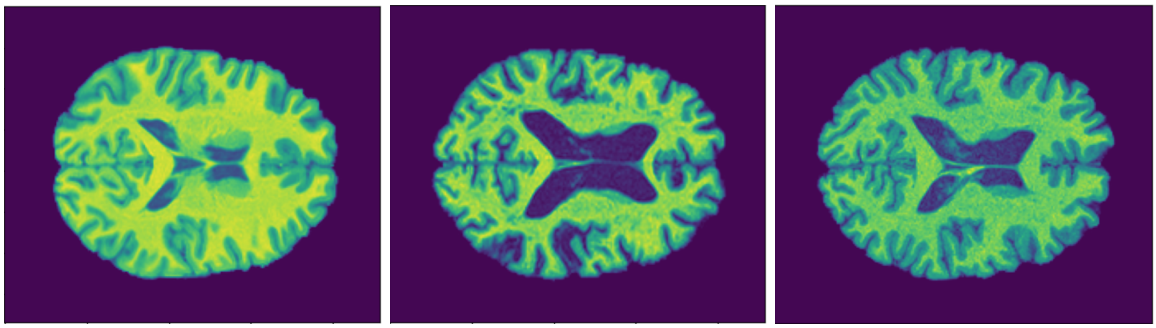

Figure 2.4: Comparison of normal cognitive decline (left) and neurodegeneation with Alzheimer's disease (middle) and Parkinson's disease (right). *Images courtesy of Simon Crête.*

2.5 Summary

In this chapter, we provided a brief introduction to AI. We described the differences between traditional ML and BL development, discussed their current challenges as highlighted in the literature, and included some of the methods proposed to address these limitations. Specifically, we presented an overview of the FL framework and its core aggregation algorithm, the Federated Averaging method. Additionally, we introduced CDSS and reviewed literature demonstrating their improvements in decision-making and trust in AI, particularly through evaluations of AI uncertainty. Finally, we provided an overview of Alzheimer's disease and Parkinson's disease, using them as case studies.

Chapter 3

Precision-weighted Federated Learning

Preface

This chapter is the first of two aimed at enhancing data privacy in AI model development, focusing on the technical and legal challenges of AI adoption in healthcare. In this first chapter, we introduce the Precision-weighted Federated Learning (PW) method, a novel aggregation algorithm for training distributed machine/deep learning models. Unlike traditional Federated Learning, PW uses the inverse of variance estimations from individual clients as a weighting factor, similar to the meta-analysis method in genetics. In the next chapter (Chapter 4), we explore how PW can be used to improve the quality of data across multiple clinical centers. This is to highlight the proposed method's utility in clinical practice.

In this chapter, we introduce the PW algorithm, effectively handles highly heterogeneous and sensitive data sources, crucial for applications like clinical data, which are often sensitive and varied. PW maintains data privacy by ensuring sensitive information remains secure and decentralized within each participant's data storage while still enabling distributed learning. This method addresses data heterogeneity and privacy challenges and complies with data protection regulations, thus supporting the broader adoption and trust of AI technologies in sensitive domains such as healthcare.

This work contains proprietary information. It resulted in the filing of a US patent [1] and a

[1] Co-inventor of US provisional patent application entitled "METHOD OF AND SYSTEM FOR PROVIDING AN

paper [72] to be submitted to IEEE Systems, Man and Cybernetics Society and currently on arXiv.

AGGREGATED MACHINE LEARNING MODEL IN A FEDERATED LEARNING ENVIRONMENT AND DETERMINING RELATIVE CONTRIBUTION OF LOCAL DATASETS THERETO" that was filed on January 19th of 2021

Abstract

Federated Learning using the Federated Averaging algorithm has shown great advantages for large-scale applications that rely on collaborative learning, especially when the training data is either unbalanced or inaccessible due to privacy constraints. We hypothesize that Federated Averaging underestimates the full extent of heterogeneity of data when the aggregation is performed. We propose *Precision-weighted Federated Learning*[2] a novel algorithm that takes into account the second raw moment (uncentered variance) of the stochastic gradient when computing the weighted average of the parameters of independent models trained in a Federated Learning setting. With Precision-weighted Federated Learning, we address the communication and statistical challenges for the training of distributed models with private data and provide an alternate averaging scheme that leverages the heterogeneity of the data when it has a large diversity of features in its composition. Our method was evaluated using three standard image classification datasets (MNIST, Fashion-MNIST, and CIFAR) with two different data partitioning strategies (independent and identically distributed (IID), and non-identical and non-independent (non-IID)) to measure the performance and speed of our method in resource-constrained environments, such as mobile and IoT devices. The experimental results demonstrate that we can obtain a good balance between computational efficiency and convergence rates with Precision-weighted Federated Learning. Our performance evaluations show 9% better predictions with MNIST, 18% with Fashion-MNIST, and 5% with CIFAR-10 in the non-IID setting. Further reliability evaluations ratify the stability in our method by reaching a 99% reliability index with IID partitions and 96% with non-IID partitions. In addition, we obtained a $20x$ speedup on Fashion-MNIST with only 10 clients and up to $37x$ with 100 clients participating in the aggregation concurrently per communication round. The results indicate that Precision-weighted Federated Learning is an effective and faster alternative approach for aggregating private data, especially in domains where data is highly heterogeneous.

[2] A US provisional patent application has been filed for protecting at least one part of the innovation disclosed in this article

3.1 Introduction

Machine learning based on distributed deep neural networks (DNNs) has gained significant traction in both research and industry [33, 73], with many applications in IoT, for mobile devices, and in the automobile sector. For example, IoT devices and sensors can be protected from web attacks during the exchanging of data between the device and web services (or data stores) in the cloud [74]. Mobile devices use distributed learning models to assist in vision tasks for automatic corner detection in photographs [75], prediction tasks for text entry [76], and recognition tasks for image matching and speech recognition [33]. Alternatively, modern automobiles utilize distributed machine learning models to improve drivers' experience, vehicle's self-diagnostics and reporting capabilities [77].

Despite the benefits provided by distributed machine learning, data privacy and data aggregation are raising concerns addressed in various resource-constrained domains. For example, the communication costs incurred when updating deep learning models in mobile devices is expensive for most users as their internet bandwidths are typically low. In addition, the data used during the training of models in mobile devices is privacy-sensitive, and operations of raw data outside the portable devices are susceptible to attacks. One solution is using secure protocols [78] or differential-privacy guarantees [79, 80] to ensure that data is transferred between clients and servers safely. Another solution is to use data aggregation for distributed DNNs mitigating the need for transferring data to a central data store. With this solution, the learning occurs at the client level where models are optimized locally across the distributed clients. This approach is termed *Federated Learning* [42].

McMahan *et al.* [42] introduced the notion of Federated Learning in a distributed setting of mobile devices. Their developed *Federated Averaging* algorithm uses numerous communication rounds where all participating devices send their local learning parameters, i.e. DNN weights, to be aggregated in a central server in order to create a global shared model. Once the global model is computed, it is distributed to every client replacing the current deep learning model. Since only the global model is communicated in these rounds, data aggregation is achieved, even though the client's raw data never leaves the device. Given such a setup, individual clients can collaboratively learn an averaged shared model without compromising confidentiality. This makes Federated Learning a

promising solution to the analysis of privacy-sensitive data distributed across multiple clients.

In McMahan *et al.'s* original paper the local learning parameters on each client are aggregated by the central server and the global model is maintained with the *weighted average* of these parameters [42]. There are potentially a few statistical shortcomings identified with this type of averaging method. If we consider that the aggregation of weights across multiple clients is similar to a meta-analysis which synthesizes the effects of diversity across multiple studies then variation across the population should be considered. Meta-analysis is a quantitative method that combines results from different studies on the same topic in order to draw a general conclusion and to evaluate the consistency among study findings [81, 82]. There is compelling evidence that demonstrates a misleading interpretation of results and a reduction of statistical power when combining data from different sources without accounting for variation across the sources [83, 84, 85].

3.1.1 Hypotheses

In this chapter, we build on the work of McMahan *et al.*[42], and propose the *Precision-weighted Federated Learning* algorithm, a novel *variance-based* averaging scheme to aggregate model weights across clients. The proposed method penalizes the model uncertainty at the client level to improve the robustness of the centralized model, regardless of the data distribution: independent and identically distributed (IID) or non-identical and non-independent (non-IID). Our approach makes use of the uncentered variance of the gradient estimator from the Adam optimizer [86] to compute the weighted average at each communication step (Figure 3.1).

We hypothesize that the Federated Averaging algorithm underestimates the full extent of heterogeneity on domains where data is complex with a large diversity of features in its composition. More specifically, we hypothesized that: (1) Precision-weighted Federated Learning can leverage individual intra-variability when averaging multiple sources to improve performances when the training data is highly-heterogeneous across sources, and (2) it can harness individual inter-variability when averaging multiple sources to accelerate the learning process, especially when data is highly-heterogeneous across sources.

To test our hypothesis we compared the performance of the original Federated Averaging algorithm against the Precision-weighted method in a number of image classification tasks using MNIST

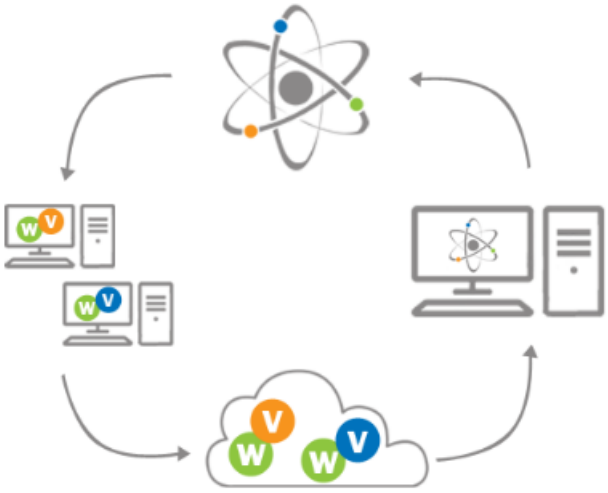

Figure 3.1: Aggregation of weights and variance via Precision-weighted Federated Learning: local models are trained across clients *(Left)*, weights and variances are aggregated by the central server *(Bottom)*, a centralized model is computed *(Right)* and the aggregated weights are redistributed across clients *(Top)*.

[87], Fashion-MNIST [88], and CIFAR-10 [89] datasets.

3.1.2 Contributions

The contributions of this chapter are threefold: (1) We propose a novel algorithm for the averaging of distributed models using the estimated variance of the stochastic gradient computed independently by each client in a Federated Learning environment; (2) We provide extensive evaluations of the method using benchmark image classifications datasets demonstrating its robustness to unbalanced and non-IID data distributions; and (3) We compare the method to Federated Averaging on empirical experiments, and with fewer communication rounds we obtain comparable accuracy on IID distributions, greater accuracy on non-IID distributions, and more stable accuracy over communication rounds over all distributions.

3.2 Related Work

There is an increasing concern for aggregation of private data in the data mining domain, particularly when models require access to a client's data in order to improve their accuracy [90, 91].

Data privacy and data aggregation are thus concerns that are actively being investigated for both centralized [92, 93] and decentralized (or distributed) [42] data environments.

The method proposed in this chapter is dedicated to the aggregation of weights for DNNs with decentralized data. It is, therefore, important to observe the communication challenges addressed in previous work, mainly the security and protection of data, and the reduction of the steps needed in communication cycles. Bonawitz *et al.* [78] proposed a complementary approach to Federated Learning: a communication-efficient secure aggregation protocol for high-dimensional data. In Bonawitz *et al.*'s work, Federated Learning was used in the training of DNN models for mobile devices using secure aggregation algorithms to protect the data residing on individual mobile devices. On the other hand, Konečnỳ *et al.* [94] presented two optimization algorithms (structural and sketched updates) to reduce the communication cost in the training of deep neural networks in a federation of participant mobile devices.

As well as communication challenges, the aggregation of data in decentralized environments is impacted by statistical challenges, especially when the training data is non-IID. Smith *et al.* [95] highlighted the fact that data across a DNN is often non-IID distributed; that is, each participant updating the shared model in a Federated Learning setting generates a distinct distribution of data. One way to handle data heterogeneity is by using multi-task learning (MTL) frameworks. Smith *et al.* created the MOCHA framework, which enables the analysis of data variability in a Federated MTL. However, as noted by Zhao *et al.* [96], the Federated MTL is not comparable with the original work on Federated Learning as the proposed framework does not apply to non-convex deep learning models. In the same paper, Zhao *et al.* proposed a data-sharing strategy to improve test-accuracy when data is non-IID. This method requires a small subset of data consisting of a uniform distribution to be shared across clients. Albeit promising results can be achieved with this method, the shared subset of data may not always be available, especially when data is highly sensitive in nature. Other methods explore the statistical challenges of Federated Learning by creating synthetic data using Dirichlet distributions with different concentration parameters. This technique allows the creation of more realistic non-IID data distributions at the client level, which are used to examine the effects on aggregations carried out with the Federated Learning algorithm [96, 30].

There is a diverse body of work that further explores collaborative learning, data sharing and

data preservation across multiple data centers. Note that all of these methods are substantially different than the original work on Federated Learning. Although some yield comparable or even better results than Federated Learning they lack empirical observations with non-IID data. Chang *et al.* [97] addressed the problem of distributed learning on medical data and compared five heuristics: separate training on subsets, training on pooled data, weight averaging, and weight transfer (single and cyclical transfer). Of all these heuristics, training on pooled data has the best prediction performance and training on cyclical weight transfer achieved comparable testing accuracy to that of centrally trained models. Xu *et al.* [98] introduced a collaborative deep learning (co-learning) method for the training of a shared global model using a cyclical learning rate schedule mixed with an incremental number of epochs. Their results demonstrate that the method is comparable with data centralized learning. Lalitha *et al.* [99] trained a model over a network of devices without a centralized controller. However, the users could communicate locally with their closest neighbors. The performance of the proposed algorithm on two users matches the performance of an algorithm trained by a central user with access to all data. They left a full empirical evaluation for future research. Chen *et al.* [100] proposed a Federated Meta-Learning framework for the training of recommended systems. The framework permits data sharing at the algorithm level, preserves data privacy, and reports an increase of 12.4% in accuracy compared with previous results. Kim *et al.* [101] addressed the problem of catastrophic forgetting (the ability of neural networks to learn new tasks while discarding knowledge about previous learned tasks) in a distributed learning environment on clinical data and introduced an approach for knowledge preservation. Similarly, Bui *et al.* [102] unify continual learning and Federated Learning in a partitioned variational inference framework. Vepakomma *et al.* [103], introduced *split learning*, which addresses challenges specific to health data, such as different modalities across clients, no label sharing and semi-supervised learning.

In the field of genetics, genome-wide association studies aim to identify genetic variants associated to phenotypes of interest. As the effect of these variants on phenotypes is usually moderate, individual hospital studies are under-powered to detect them with confidence and a growing number of consortia are created to combine data across studies. As patient genotypes are privacy sensitive, these consortia use *meta-analyses* to aggregate summary statistics from multiple studies. This

increases the statistical power of finding a mutation related to a phenotype, while protecting the privacy of individual genotypes. Lin and Zang [85] demonstrated that meta-analyses achieve comparable efficiency as analyses of pooled individual participants under mild assumptions. This proximity with the distributed learning setting motivated us to create the Precision-weighted Federated Learning, an averaging approach that considers a meta-analysis weighting scheme in the aggregation of the effects of the variances from the weights generated during training of the neural network.

3.3 Precision-weighted Federated Learning

The Precision-weighted Federated Learning approach combines the weights from each client into a globally shared model where the aggregation is achieved by averaging the weights by the inverse of their estimated variance. We will use the same notations than the Federated Averaging algorithm [42] to describe the implementation of the proposed method. We consider the general objective

$$\min_{w \in R^d} f(w) \quad \text{with} \quad f(w) = \frac{1}{n} f_i(w), \tag{5}$$

for $i = 1, ..., n$, where n is the number of data examples and $f_i(w) = \ell(x_i, y_i; w)$ is the loss of the prediction on example (x_i, y_i) made with model parameters w. If the data is partitioned over K clients, McMahan *et al.* rewrite the objective of Equation 5 as follows:

$$f(w) = \sum_{k=1}^{K} \frac{n_k}{n} F_k(w) \quad \text{with} \quad F_k(w) = \frac{1}{n_k} \sum_{i \in \mathcal{P}_k} f_i(w), \tag{6}$$

where $\mathcal{P}_k$ is the set of indexes of data examples on client k and $n_k = |\mathcal{P}_k|$. Under a uniform distribution of training examples over the clients, the *IID assumption*, the expectation of the client-specific loss $F_k(w)$ is $f(w)$. In a non-IID setting however, this result does not hold [42].

The corresponding stochastic gradient descent for optimization with a fixed learning rate η consists in computing the gradient

$$g_k \quad = \quad \nabla F_k(w_t) \tag{7}$$

for each client k at iteration t, and applying the two successive updates

$$w_{t+1}^k \leftarrow w_t - \eta g_k \text{ for all } k \tag{8}$$

$$w_{t+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} w_{t+1}^k. \tag{9}$$

With Precision-weighted Federated Learning the global update of Equation 9 is replaced by

$$w_{t+1} \leftarrow \sum_{k=1}^{K} \frac{\left(v_{t+1}^k\right)^{-1}}{\sum_{k=1}^{K} \left(v_{t+1}^k\right)^{-1}} w_{t+1}^k \tag{10}$$

where v_{t+1}^k denotes the variance of the maximum likelihood estimator of weight w at iteration $t+1$ for client k. This inverse variance weighting scheme used in Equation 10 corresponds to the fixed effect model used in meta-analyses. Intuitively, this method allows taking into consideration the uncertainty of each client into the aggregated result and uses the estimated variance to penalize the model uncertainty at the client level: models with high estimated variance across clients have a smaller impact on the aggregation result at the current communication round. Although v_{t+1}^k is inversely proportional to the sample size, it is a more nuanced summary as it captures additional uncertainty about the client's weights. The complete algorithm is provided in Algorithm 2.

3.3.1 The Precision-weighted Federated Learning algorithm

FL traditionally achieves the training of distributed machine/deep learning models with the *Federated Averaging* algorithm. The complete pseudo-code describing the FedAvg method is given in Algorithm 1.

To estimate the inverse of the variance of the maximum likelihood, we use the raw second moment estimate (uncentered variance) from the Adam optimizer [86], which approximates the diagonal of the Fisher information matrix [104]. Our experiments show that this approximation manages to capture the uncertainty of weights in practice.

Algorithm 2 Precision-weighted Federated Learning Algorithm. The K clients are indexed by k; B is the local minibatch size, E is the number of local epochs, and η is the learning rate.

Server executes:
  initialize w_0
  for each round $t = 1, 2, \dots$ **do**
    $m \leftarrow \max(C \cdot K, 1)$
    $S_t \leftarrow$ (random set of m clients)
    for each client $k \in S_t$ **in parallel do**
      $w_{t+1}^k, v_{t+1}^k \leftarrow \text{ClientUpdate}(k, w_t)$
    end for
    $w_{t+1} \leftarrow \sum_{k=1}^{K} \frac{\left(v_{t+1}^k\right)^{-1}}{\sum_{k=1}^{K}\left(v_{t+1}^k\right)^{-1}} w_{t+1}^k$
  end for

function CLIENTUPDATE(k, w) ▷ Run on client k
  $\beta \leftarrow$ (split $\mathcal{P}_k$ into batches of size B)
  for each local epoch i from 1 to E **do**
    for each batch $b \in \beta$ **do**
      $w \leftarrow w - \eta \nabla \ell(w; b)$
    end for
    if second half of last epoch **then** ▷ Weighted-Variance Callback
      $v \leftarrow v \frac{1}{V}$
    end if
  end for
  return w, v to server
end function

3.4 Methodology

We tested the Precision-weighted Federated Learning method under different data distributions for image classification tasks. The baseline we use is the Federated Averaging approach. Firstly, we explore the performance of our method in resource-constrained environments, applicable to areas where memory is limited, such as mobile and IoT devices. Next, we present a scenario in which we investigate the speedup of our method as a function of the number of clients participating in the aggregation of weights. Finally, we present the analysis for the generalization of the global model when the parameter variance is applied to the aggregation of parameters of all the models in the distributed learning process.

Since the statistics of the data are influenced by the way it is distributed across clients, we tested the proposed methodology with both IID and non-IID data distributions. To create these scenarios,

we distributed the training data across individual clients in two configurations (see Section 3.4.2). The complexity of the image recognition problems was increased in agreement with the methodology proposed by Scheidegge *et al.* [105] and therefore MNIST, Fashion-MNIST and CIFAR-10 were used as benchmarks. Furthermore, we utilized modest convolutional architectures to compare training speed and optimal convergence with our method and Federated Averaging and to explain the effects of variance in the generalization of the centralized model. All of the experiments were executed on an NVIDIA Tesla V100 Graphic Processing Unit.

3.4.1 Datasets

MNIST: The MNIST dataset consist of 70,000 gray-scale images (28 x 28 pixels in size) which are divided in 60,000 training and 10,000 test samples. The images are grouped in 10 classes corresponding to the handwritten numbers from zero to nine.

CIFAR-10: The CIFAR-10 dataset consists of 60,000 colored images (36 x 36 pixels in size) divided in a training set of 50,000 and a testing set of 10,000 images. Images in CIFAR-10 are grouped into 10 mutually exclusive classes of animals and vehicles: airplanes, automobiles, birds, cats, deer, dogs, frogs, horses, ships, and trucks.

Fashion-MNIST: The Fashion-MNIST dataset contains the same number of samples, image dimensions and number of classes (different labels) in its training and testing sets than MNIST, however, the images are of clothing (e.g. t-shirts, coats, dresses and sandals).

3.4.2 Data Distributions

IID: With IID data distribution the number of classes and the number of samples per class were assigned to clients with a uniform distribution. We shuffled the training data and created one partition per client with an equal number of samples per class. For example, 10 clients receive 600 samples per class. Figure 3.2 shows an example with 5 clients and 4 classes.

Non-IID: With this data partition, two classes are assigned per client at most. This is similar to the partition shown in [42] used to explore the limits of the Federated Averaging approach, which we now use to test and compare our algorithm under similar circumstances. In this extreme scenario, the number of samples per class per client is evenly distributed, creating a balanced scenario (Figure

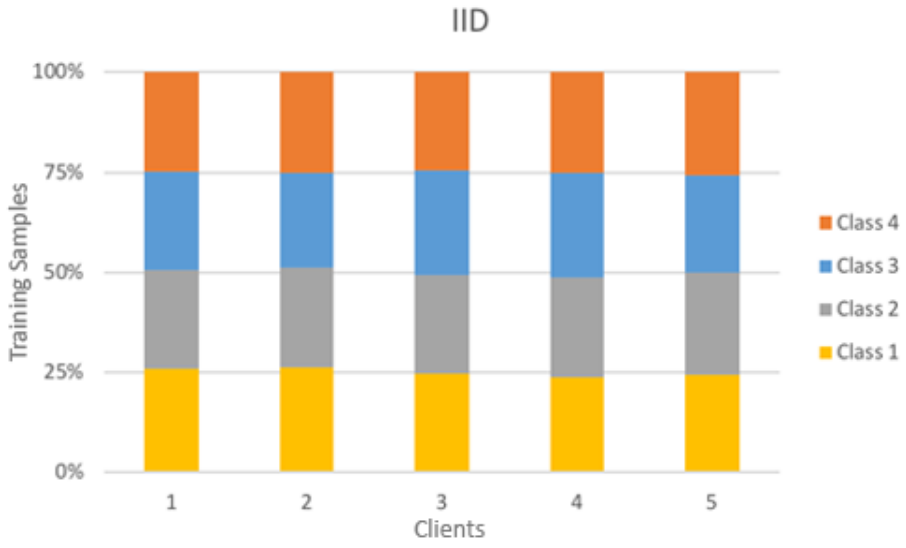

Figure 3.2: Example of a IID data distributions with 5 clients and 4 classes

3.3).

3.4.3 Convolutional Neural Networks

The architectures used in our experiments were CNNs trained from scratch. All artificial networks were based on the Keras sequential model, trained with the Adam optimizer and an objective function as defined by categorical cross-entropy.

For MNIST and Fashion-MNIST the architecture of the first artificial neural network consisted of two convolutional layers using 3x3 kernels (each with 32 convolution filters). A rectified linear unit (ReLU) activation is performed right after each convolution, followed with a 2x2 max pooling used to reduce the spatial dimension, a dropout layer used to prevent overfitting, a fully densely-connected layer (with 128 units using a ReLu activation), and leading to a final softmax output layer (600,810 total parameters). The network was trained from scratch using partitions of training data and the final model was evaluated using the testing set.

A second network was used to train our models using data from the CIFAR-10 dataset. The

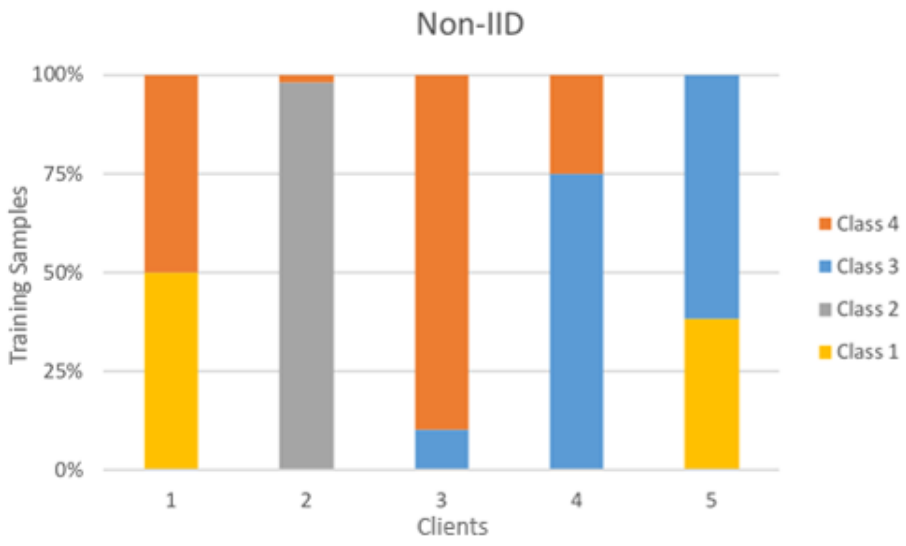

Figure 3.3: Example of a non-IID data distributions with 5 clients and 4 classes

architecture consisted of one 3x3 convolutional layer (with 32 convolution filter using a ReLu activation), followed with a 2x2 max pooling, a batch normalization layer; a second 3x3 convolutional layer (with 64 convolution filter using a ReLu activation), followed with, a batch normalization layer and a 2x2 max pooling; a dropout layer; one fully densely-connected layer (with 1024 and 512 units using a ReLu activation), another dropout layer; and a final softmax output layer (4,225,354 total parameters).

3.4.4 Adam and the Weighted-Variance Callback

A key component in the formulation of the weighted average algorithm is the estimation of the individual intra-variability expressed during the training of local data. As the training of the model proceeds, we capture the weights' variances via the second raw moment (uncentered variance) of the stochastic gradient descent from the Adam optimizer and use it in the construction of the Precision-weighted Federated Learning algorithm. In order to access the internal statistics of the model during training, we use a callback function that averages the variance estimators on the second half of the last epoch. The last epoch is chosen as it provides a more accurate prediction of the variance of the final weight.

3.5 Results

This section presents the results of our model predictions trained with the two aforementioned data partitioning strategies in Section 3.4 and demonstrates the limits and practical application of the proposed method. All of our experiments use a different random seeds to randomize the order of observations during the training of the local models. As noted in McMahan *et al.'s* paper, averaging federated models from different initial conditions leads to poor results. Thus, in order to avoid the drastic loss of accuracy observed on independent initialization of models for general non-convex objectives, each local model was trained using a shared random initialization for the first round of communication. After the first round of communication, all local models were initialized with the globally averaged model aggregated from the previous round.

3.5.1 Evaluating Computational Resources

Experiments with MNIST and Fashion-MNIST were conducted by using 500 rounds of communication, 1 epoch, and batch sizes (10, 25, 50, 100, and 200). Similarly, experiments with CIFAR-10 were executed for 500 rounds of communication, with 10 epochs, and batch sizes (10, 25, 50, 100, and 200). All of the training samples of each dataset were arranged among 10 clients.

The comparison results of test-accuracy between Federated Averaging and Precision-weighted Federated Learning aggregation methods using IID partitions is given in Table 3.1. Given this setup, test-accuracy scores are comparable with those obtained using Federated Averaging, however, our method is more stable. When we analyze the results of MNIST and Fashion-MNIST, we observe that test-accuracy values are consistent across batch sizes. The accuracy curves of Precision-weighted Federated Learning and the Federated Averaging for these datasets are show in Figures 3.4 and 3.5). Alternatively, CIFAR-10 models trained with $B = 10$ using Precision-weighted Federated Learning show an improvement of 12% (Figure 3.6) with more stable predictions. This improved accuracy on CIFAR-10 could indicate that there is greater heterogeneity in models trained on natural images than in models trained on grayscale images, even in an IID setting.

As discussed in the introduction section, we hypothesized improvements on the performance

Table 3.1: Comparison of test-accuracy results (IID data distributions)

	MNIST		Fashion-MNIST		CIFAR-10	
	FedAvg	PW	FedAvg	PW	FedAvg	PW
B = 10	0.99 ± 0.002	0.99 ± 0.002	0.93 ± 0.009	0.93 ± 0.008	0.69 ± 0.045	$\mathbf{0.77 \pm 0.019}$
B = 25	0.99 ± 0.002	0.99 ± 0.002	0.93 ± 0.010	0.93 ± 0.010	0.77 ± 0.004	0.77 ± 0.018
B = 50	0.99 ± 0.003	0.99 ± 0.003	0.93 ± 0.011	0.93 ± 0.011	0.76 ± 0.023	0.76 ± 0.013
B = 100	0.99 ± 0.004	0.99 ± 0.004	0.93 ± 0.013	0.93 ± 0.012	0.76 ± 0.014	0.76 ± 0.011
B = 200	0.99 ± 0.006	0.99 ± 0.006	0.93 ± 0.016	0.93 ± 0.015	0.76 ± 0.014	0.76 ± 0.011

Averaged results using 1 epoch (MNIST and Fashion-MNIST) and 10 epochs

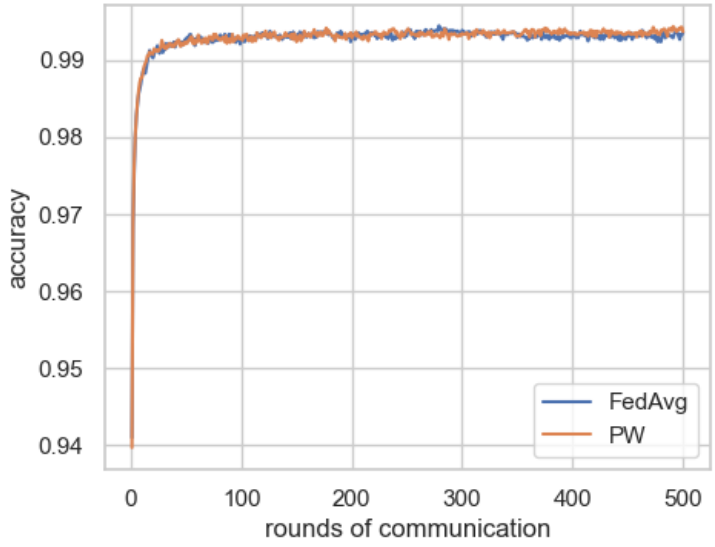

Figure 3.4: Test-accuracy for Federated Averaging (FedAvg) and Precision-weighted Federated Learning (PW) using IID data distributions with MNIST (B = 50)

of models whose training data is highly-heterogeneous in nature. The comparison results of performance using Non-IID data partitions are given in Table 3.2. As we observe, both methods perform poorly with a batch number of $B = 10$ and more notably in Precision-weighted Federated Learning, which is more sensitive to the noise present in the input images. This behavior of Federated Averaging is comparable with other related work in Federated Learning [96] and its effects are also visible in Precision-weighted Federated Learning. However, with larger batch sizes, higher test-accuracy and more stable predictions are obtained, starting from the first round irregardless of the dataset (Figure 3.7) This indicates that the estimations of variance are effectively used to computed a weighted average, resulting in more effective penalization of the model's uncertainty at the

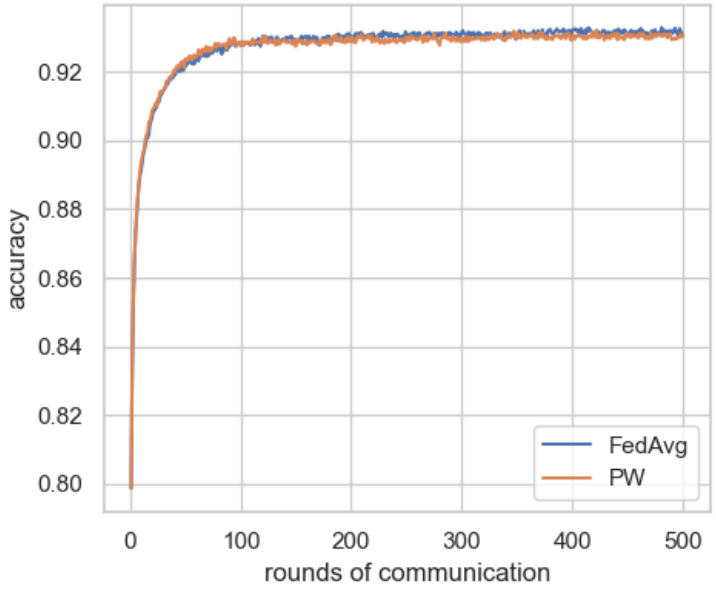

Figure 3.5: Test-accuracy for Federated Averaging (FedAvg) and Precision-weighted Federated Learning (PW) using IID data distributions with Fashion-MNIST (B = 50)

Table 3.2: Comparison of test-accuracy results (non-IID data distributions)

	MNIST		Fashion-MNIST		CIFAR-10	
	FedAvg	PW	FedAvg	PW	FedAvg	PW
B = 10	0.98 ± 0.026	0.98 ± 0.014	0.85 ± 0.028	0.82 ± 0.031	0.34 ± 0.052	0.16 ± 0.052
B = 25	0.97 ± 0.029	$\mathbf{0.98 \pm 0.028}$	0.85 ± 0.048	0.85 ± 0.024	0.51 ± 0.053	$\mathbf{0.53 \pm 0.054}$
B = 50	0.97 ± 0.053	$\mathbf{0.98 \pm 0.035}$	0.79 ± 0.048	$\mathbf{0.86 \pm 0.035}$	0.58 ± 0.048	$\mathbf{0.60 \pm 0.027}$
B = 100	0.95 ± 0.071	$\mathbf{0.98 \pm 0.055}$	0.77 ± 0.046	$\mathbf{0.86 \pm 0.040}$	0.56 ± 0.059	$\mathbf{0.59 \pm 0.041}$
B = 200	0.90 ± 0.083	$\mathbf{0.98 \pm 0.058}$	0.73 ± 0.052	$\mathbf{0.86 \pm 0.050}$	0.59 ± 0.07	0.59 ± 0.049

Averaged results using 1 epoch (MNIST and Fashion-MNIST) and 10 epochs (CIFAR-10)

client level. For MNIST, our method can obtain increases in test-accuracy of up to of 9% with $B = 200$. The results of Fashion-MNIST show the highest increment of 18% in the test-accuracy with $B = 200$. Similarly, the highest accuracy of CIFAR-10 improves by 5% with $B = 100$. These results demonstrate that our first hypothesis is confirmed only when models are trained with a batch size of $B = 25$ or higher.

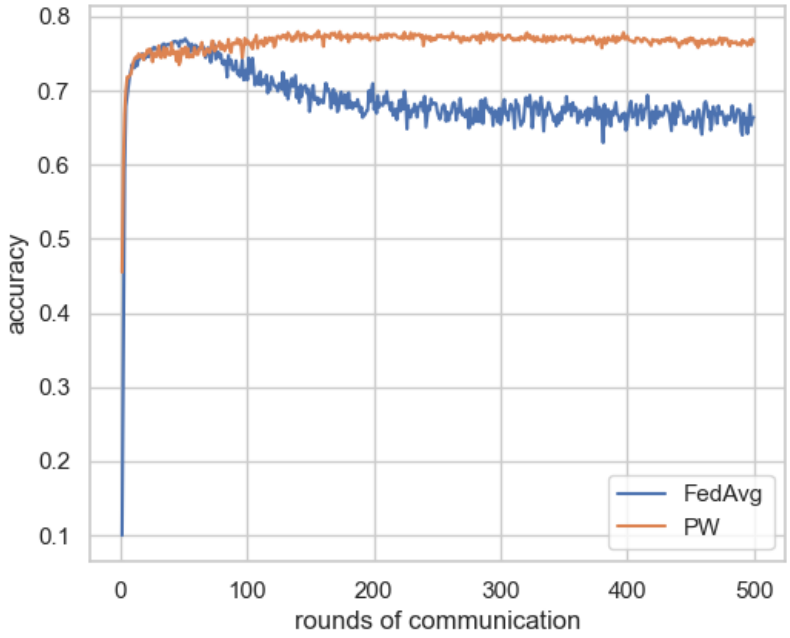

Figure 3.6: Test-accuracy for Federated Averaging (FedAvg) and Precision-weighted Federated Learning (PW) using IID data distributions with CIFAR-10 (B = 10)

3.5.2 Reliability

The reliability index is an important element to consider in the evaluations of the performance of machine learning systems. In this study, we compute the reliability index defined in [106] as the ratio of the standard deviation of the test-accuracy and mean value of the test-accuracy as shown in Equation 11.

$$\xi_k(\%) = \left(1 - \frac{\sigma_n}{\mu_n}\right) x100 \tag{11}$$

, where σ_n is the standard deviation and μ_n is the mean of test-accuracy scores per batch. Consequently, the overall system reliability index can be computed by averaging all of the reliability indexes as expressed in Equation 12. Table 3.4 quantifies the computed reliability index per batch size and shows the overall system stability, which confirms that Precision-weighted Federated Learning reaches optimal performance, except for CIFAR-10 in a non-IID. This is due to the sensitivity of

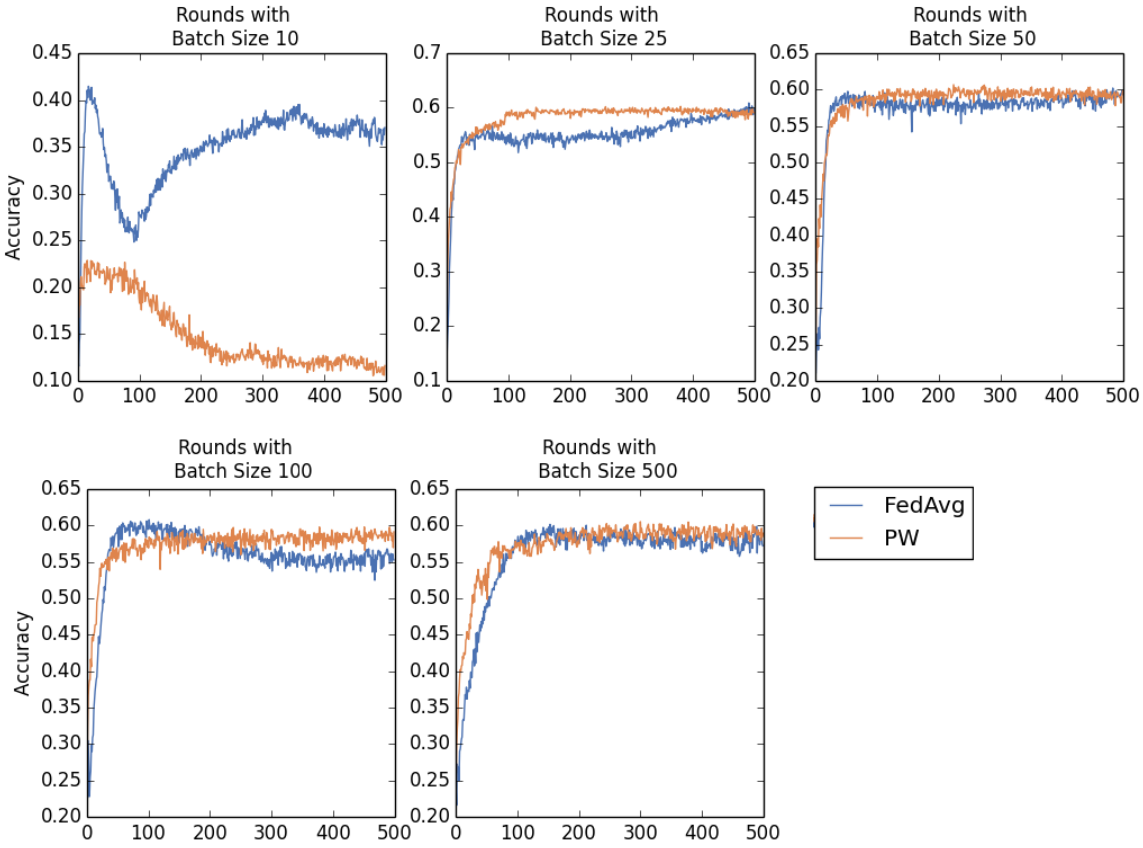

Figure 3.7: Test-accuracy increases as batch size is larger with non-IID partitions. Aggregation methods: Federated Averaging (FedAvg) and Precision-weighted Federated Learning (PW).

our method with small batch sizes, compromising performance.

$$\xi(\%) = \left(\frac{\sum_{k=1}^{K} \xi_k}{N} \right) \tag{12}$$

3.5.3 Increasing Participating Clients

Inspired by McMahan *et al.'s* original paper [42], we experiment with the client fraction C that controls the amount of multi-client parallelism. To this regard, we investigate the number of communication rounds necessary to achieve target test-accuracy of 75%, 80%, and 85% for models trained with Fashion-MNIST. For this purpose, the predictive models used a fixed batch size

Table 3.3: Reliability index across batches (IID data distributions)

	MNIST		Fashion-MNIST		CIFAR-10	
	FedAvg	PW	FedAvg	PW	FedAvg	PW
B = 10	99.80	99.80	99.03	99.14	93.47	97.52
B = 25	99.80	99.80	98.92	98.92	94.83	97.67
B = 50	99.70	99.70	98.81	98.81	96.99	98.29
B = 100	99.60	99.60	98.60	98.70	98.17	98.55
B = 200	99.40	99.40	98.27	98.38	98.29	98.55
	99.66	99.66	98.73	**98.79**	96.35	**98.11**

Table 3.4: Reliability index across batches (non-IID data distributions)

	MNIST		Fashion-MNIST		CIFAR-10	
	FedAvg	PW	FedAvg	PW	FedAvg	PW
B = 10	97.34	98.57	96.71	96.22	84.62	67.70
B = 25	97.02	97.13	94.34	97.17	89.57	89.87
B = 50	94.54	96.42	93.95	95.94	91.72	95.49
B = 100	92.56	94.37	94.00	95.37	89.52	93.09
B = 200	90.75	94.06	92.89	94.16	88.03	91.75
	94.44	**96.11**	94.38	**95.77**	88.69	87.58

$B = 100$ and epoch $E = 1$. The training data was split into 100 participants and evaluated speed for every 10, 20, 50, and 100 clients participating in the aggregation in parallel.

Table 3.5 provides the number of communication rounds needed to reach the aforementioned test-accuracy scores as well as their corresponding speedup. We observe a trend, which suggests that an increase in participants reduces the number of communication rounds regardless of the aggregation method evaluated. This behavior is in alignment with McMahan *et al.'s* work in [42]. Given this setup, Precision-weighted Federated Learning misses the first target with 10 and 50 clients, but it can reach subsequent target score up to $20x$ faster with 10 clients and $37x$ with 100 clients participating concurrently. Thus, we see that with a small client fraction ($C = 0.1$; that is 10 client per round), a good balance between computational efficiency and convergence rate can be obtained.

3.5.4 Variance Analysis

In this chapter we demonstrated that combining widely disparate sources can hide important features useful for discrimination, leading to limitations in the collaborative learning experience.

Table 3.5: Number of rounds and speedup relative to Federated Averaging to reach different test-accuracy values on Fashion-MNIST.)

	C = 0.1		C = 0.2		C = 0.5		C = 1.0	
ACC	FedAvg	PW	FedAvg	PW	FedAvg	PW	FedAvg	PW
75%	47	50	65	35 (19x)	21	23	17	12 (14x)
80%	149	125 (12x)	134	66 (20x)	153	57 (27x)	44	27 (16x)
85%	641	319 (20x)	671	225 (30x)	473	279 (17x)	286	78 (37x)

Owing to this, Precision-weighted Federated Learning considers the inverse of the estimated variance to compute a weighted average (Equation 10). As such, this algorithm operates under the assumption that weights with large variance estimations across sources reduces the quality of the analysis and therefore should have a smaller impact in the aggregation.

To explain the effects of variance in the generalization of the global model using Precision-weighted Federated Learning, we trained 4 clients with a fixed batch size $B = 10$ and epoch $E = 1$ for 100 communication rounds. The training data of CIFAR-10 was distributed uniformly among three clients with IID partitions and a single client with a non-IID partition (Client 1 in Figure 3.8). In this regard, three clients receive a large number of training samples per class, whereas one of the them receives a considerably small number of training samples (Clients 2, 3, and 4 in Figure 3.9). This is to maximize the expression of variation across clients.

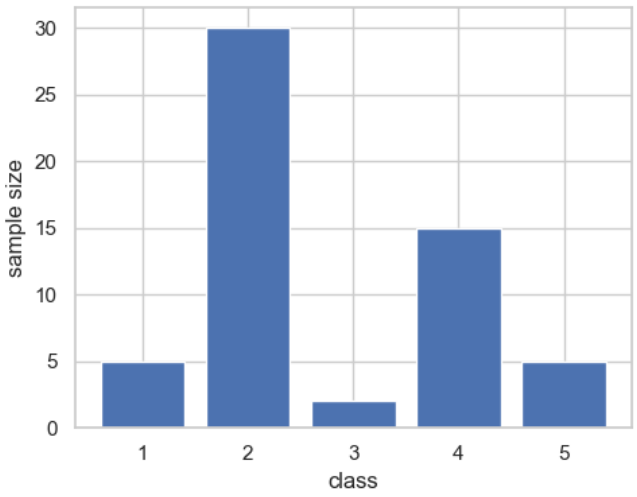

Figure 3.8: Class distribution per client. Client 1 using an non-IID unbalanced partition

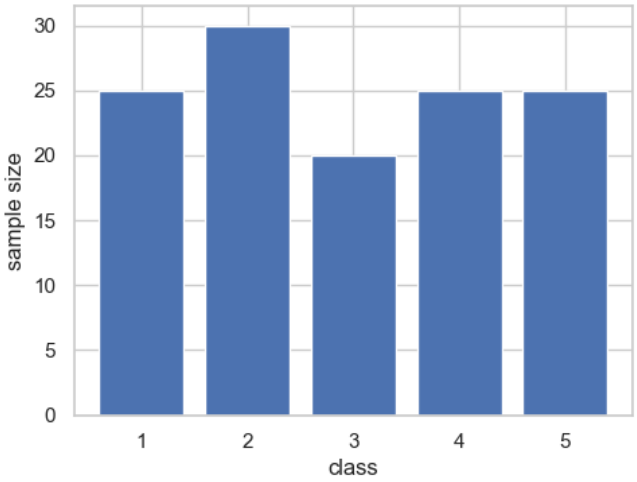

Figure 3.9: Class distribution per client. Clients 2, 3, 4 using an IID partition.

After model training and before the aggregation, we average the inverse of the estimated variance of the stochastic gradient per client before it is aggregated and plot it. Figure 3.10. Given a small number of training samples, the amount of intra-variability computed for Client 1 is significantly smaller than other models. Consequently, the inverse of these variances for this client is high and therefore the penalization of weights is greater. This behavior is evident since the beginning of the learning cycle and causes a reduction of the inverse of the variance as training continues. Alternatively, models with larger training samples provide weights with higher quality and their penalization is minimum. Figure 3.11 shows the inverse of the estimated variance per weight and client. With this view we can identify *conv2d/bias* and *conv2d/kernel* with the highest mean of the inverse variance. This suggests that the Adam optimizer could not capture the most prominent characteristics that make up the training data, for these layers, due to the limited number of training passes.

3.6 Discussion

Federated Learning is a promising solution to the analysis of privacy-sensitive data distributed globally across clients. At the core of Federated Learning is Federated Averaging, an aggregation algorithm that consolidates the weighted average of distributed machine learning models into a

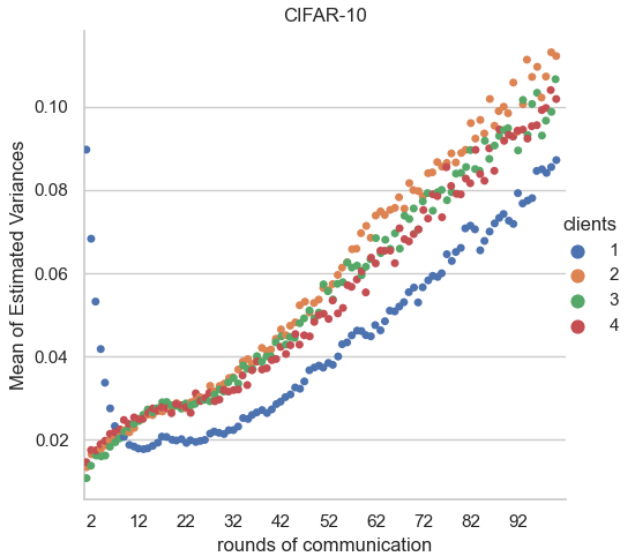

Figure 3.10: Effect of variance in the generalization of the global model. Each data point represents the mean of the inverse variances per client at a given communication round. Data points in the "Mean of Estimated Variance" graph were normalized between 0 and 1.

global model shared with every client participating in the learning cycle. In this chapter, we hypothesized that Federated Averaging underestimates the full extent of heterogeneity of data across participants, leading to a reduction in the statistical power and quality of predictions, and thus proposed Precision-weighted Federated Learning. Our method averages the weights of individual sources by the inverse of the estimated variance. When weighting machine learning models differently, it must be noted that different aggregation algorithms may yield different results under different circumstances. Our method shows the greatest advantages when the data is highly-heterogeneous across clients.

Our first hypothesis postulates that not accounting for variation across clients may lead to a reduction of statistical power when combining data form multiple sources. We confirmed this hypothesis by showing that models trained with batch size $B >= 25$ and Precision-weighted Federated Learning can obtain a 9% improvement with MNIST, 18% with Fashion-MNIST, and 5% with CIFAR-10 using non-IID partitions Nevertheless, the presented algorithm can still be improved. With a batch size $B = 10$, our method is sensitive to the noise introduce by individual sources,

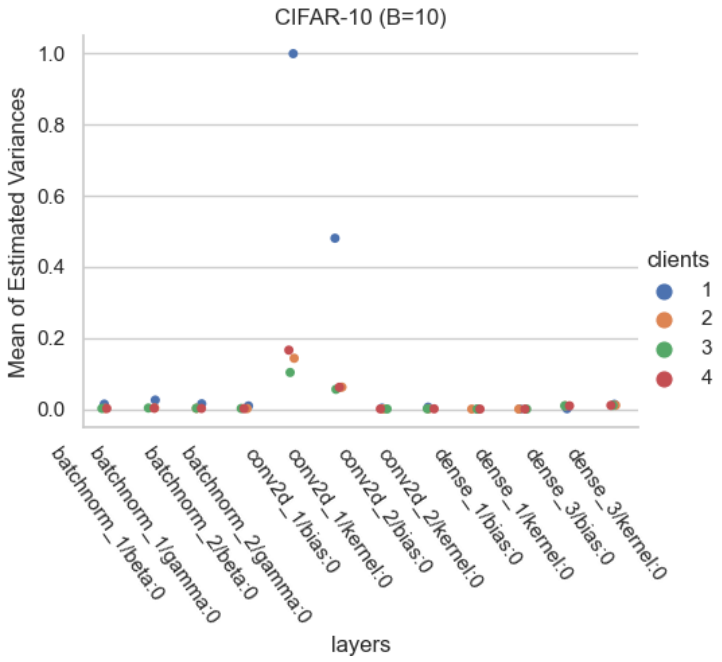

Figure 3.11: Category plots showing the dispersion of clients per layer at the first round. Data points in the "Mean of Estimated Inverse Variance" graph were normalized between 0 and 1.

degrading the performance of the method. These results prove the limits of our algorithm. Alternatively, when we compare our method with those models trained with IID partitions, our method shows comparable results to those of Federated Averaging. This suggest that the inter-variance estimations were small due to the large number of training samples and uniform distribution of classes among participants, leading to more confident predictions.

Our second hypothesis addresses convergence speed and supports the idea that the use of estimated variance can capture better representation of intricate features dispersed across sources, resulting in an acceleration of the learning process. We confirmed this hypothesis by demonstrating that our method can reach test-accuracy targets faster, and with less communication rounds between targets, than Federated Averaging. With Fashion-MNIST, we also obtained a $24x$ speedup (with only 10 clients trained in parallel) than Federated Learning. This suggest that our method reduces the communication costs required between rounds. Although it is possible to achieve higher test-accuracy by using more complex state-of-the-art architectures, our goal in this study was to explore the statistical challenges, especially when the training data is non-IID. Therefore, we measure the

performance of both aggregation method with simple network architectures.

Although the aggregation of model parameters, rather than raw individual client data, represents a significant step towards privacy preservation, the Precision-weighted Federated Averaging algorithm remains vulnerable to inference attacks, as the model parameters still contain information about data. This is a limitation of the general Federated Learning protocol and is not exclusive to our approach. Recently, Geyer *et al.* [107] and Truex *et al.* [108] introduced frameworks that preserve client-level differential privacy. However, Melis *et al.* demonstrated that privacy guarantees at the client-level are achieved at the expense of model performance and are only effective when the number of clients participating in the aggregation is significantly large, thousands or more [80]. Owing to this, we will examine the behavior and performance of the Precision-weighted Federated Learning scheme combined with Differential Private Federated Learning [79, 109] as a future work.

3.7 Conclusion

In this chapter, we presented an novel aggregation algorithm for computing the weighted average of distributed DNN models trained in a Federated Learning environment. It does not require sharing raw private data. Instead, this algorithm takes into consideration the second raw moment (uncentered variance) of the stochastic gradient estimated from the Adam optimizer to compute the weighted average of distributed machine learning models. Precision-weighted Federated Learning was benchmarked with MNIST, Fashion-MNIST and CIFAR using two data distribution strategies (IID and non-IID). When compared to Federated Averaging, this algorithm was shown to provide significant advantages when the data is highly-heterogeneous across clients, and showed comparable test-accuracy when the data is uniformly distributed across clients. Demonstrating that including the variability across models in the aggregation results in a more effective and faster option for averaging distributed machine learning models having complex data with a large diversity of features in its composition. With these advantages, Precision-weighted Federated Learning show promise in comprehensive exploratory analyses of sensitive biomedical data distributed across medical centers. Thus, in future work we will examine the feasibility of this method in medical image classification tasks.

Chapter 4

Bridging the Gaps: Imputation of Parkinson's Disease Clinical Assessments with Federated Learning

Preface

In Chapter 3, we introduced the Precision-weighted Federated Learning algorithm and explored its limits with benchmark datasets and simulations of various data distributions. In this chapter, we extend the evaluations of our previous contribution to the medical domain and demonstrate its clinical utility as we conduct a comparative analysis based on the performance of the proposed algorithm and other popular aggregation methods. We focus our evaluations on the task of data imputation of decentralized Parkinson's disease clinical assessments distributed while imposing data privacy. Specifically, we trained deep learning models in a collaborative learning environment to perform a data-driven imputation of Parkinson's clinical assessments and compare their results with traditional imputation strategies. To provide a comprehensive assessment of the task, we include a downstream analysis that validates the imputation results obtained with each of the different Federated Learning strategies in the classification of Parkinson's disease patients based on symptoms progression.

Our findings demonstrate that collaborative learning yields more secure and efficient outcomes

compared to traditional imputation methods or traditional learning strategies. The imputation of decentralized clinical assessments for Parkinson's disease has multiple implications. First, clinical assessments are siloed such that, when combined, they can enhance the statistical power of machine/deep learning models. Second, a larger volume of training data obtained from different populations and conditions can help mitigate biases during training that arise from incomplete data and smaller data sources. Third, sharing multi-center clinical data for collaborative model training allows for better generalizations, as models learn from clean, complete, and heterogeneous data sources. This is crucial for research aimed at identify disease sub-types and monitoring the progression of disease severity.

This chapter is based on the journal paper **Reyes, J.**, Noroozi, A., Xiao, Y., & Kersten-Oertel, M. Bridging the Gaps: Imputation of Parkinson's Disease Clinical Assessments with Federated Learning. Submitted to IEEE Journal of Biomedical Imaging (January 2024). A preliminary version of this work was presented at the Secure and Privacy-Preserving Machine Learning for Medical Imaging MICCAI 2021 workshop [15].

Abstract

Routine clinical assessments for Parkinson's disease are essential instruments in both clinical practice and research that are often used to identify disease sub-types and monitor the progression of disease severity. However, each clinic has limited access to information and the quality of these assessments is often degraded by the amount of missing information recorded at the time of each visit. The main objective of the work in this chapter is to evaluate how collaborative learning can improve the quality of decentralised Parkinson's disease clinical assessments while imposing data privacy. Specifically, we explore the impact of different aggregation strategies on the imputation of clinical data from 1,370 patients from the Parkinson Progression Marker Initiative (PPMI). To validate this study, we provide a downstream analysis where imputed data is used for the prediction of symptoms progression. We observed that a federated learning (FL) approach yields superior model performance based on imputation errors, when compared to a traditional learning strategies. These improvements can achieve 37.7%, 31.46%, and 13.86% lower imputation errors with low, moderate, and high degree of missing scores in the training data, respectively. In addition, we obtained better classification scores (2.98% AUC, 2.30% PR-AUC, 2.41% accuracy, and 6.09% F-1 score) than the centralized setting. However, significant improvements with FL imputations were not observed given the setup of the downstream analysis.

4.1 Introduction

Parkinson's disease (PD), a chronic progressive disorder affecting the central nervous system, is the second most common neurodegenerative disease after Alzheimer's. PD has increased significantly across the world and is currently the fastest growing disorder in the worldwide. In 1990, the Global Burden of Diseases (GDB) estimated a global, regional, and country level prevalence of Parkinson's of 2.5 million people. This number doubled to 6.2 million by 2016 [110], and based on these growing rates, recent studies estimate that more than 12.9 million people will live with the disease by 2040 [111].

Routine clinical assessments are an integral part of the care and management of patients with neurodegenerative disorders. The medical history, physical and psychiatric examinations, and neuroimaging data in the Parkinson's Progression Markers Initiative (PPMI) are examples of clinical

assessments used for developing data-driven solutions, such as the assignment of subjects to specific clinical PD subtypes, and the exploration of subtype-specific prognosis [112, 113, 114, 115]. One of the common challenges with clinical assessments concerns the availability of complete medical records, especially when pooling data from multiple centers for disease-related studies (e.g. PPMI). The issue of missing scores over time is a common problem that arises from missing paper records and the limitations of clinical protocols, which can sometimes prevent certain tests from being conducted during a patient's visit [116, 117]. Thus, it is not uncommon for there to be missing clinical scores, particularly in follow-up visits.

Data imputation is proposed as a solution to the problem of missing values. In this technique, missing values are replaced with estimations based on the interpretation of contextual information and population distribution [118]. Classic imputation methods include replacement with the mean/mode of non-missing values, and the use of an indicator variable, such as 0, in the presence of missing-values [119]. Often, these strategies are simple but expensive, as the quality of the statistical analyses can be degraded due to a reduction in the size of the data source, the introduction of sample bias [120, 121] and inevitable inaccurate standard errors on the estimation of the population distribution [122].

Single and multiple imputation methods can be used to impute missing values. Single imputation relies on a single interpretation from the contextual information of the existing data [122, 123, 124, 125], while multiple imputation estimates associations between missing and non-missing values, accounting for estimation uncertainty [126, 127]. A common shortcoming of single imputation is the underestimation of the precision (standard error) of the entire population [119, 128, 122]. Similarly, the need of enough domain knowledge to model the distribution of missing values is the main limitation of multiple imputation. Both methods can lead to inaccurate estimations.

Deep learning-based techniques have also been developed to capture dependencies and patterns in data, guiding the imputation method. For example, the k-nearest neighbors, random forest, and stacked denoising autoencoders (AEs) algorithms have demonstrated to be effective in the estimation of missing values in medical research [129, 130, 131, 116]. In this chapter, we explore the task of imputation of missing clinical values with deep learning when training of models occurs with data that is subject to privacy protections available at multiple clinical centers. Generally, when

learning from various data sources, more training data is beneficial, but patient privacy imposes additional challenges. For instance, sharing raw patient data among clinical centers may violate laws such as the General Data Protection Regulation (GDPR) of the European Union, the California Consumer Privacy Act (CCPA),and Health Insurance Portability and Accountability Act (HIPAA) [132], and the re-identification of patients through model-inversion and inference attacks [133, 134] limit the ability to perform collaborative learning activities, such as disease sub-typing, biomarker identification, and early disease diagnosis.

Federated Learning (FL) is a collaborative learning technique that has shown advantages when training data is decentralized and inaccessible due to privacy constraints. With FL the model is able to leverage all available data without sharing the information between clients (e.g. clinics). At each cycle, a global model is distributed to the clients, trained with the local data and individual synchronized stochastic gradient descent (SGD) updates are sent directly to a remote location (server) for aggregation or to intermediary worker nodes [135]. Then, the updated global model is sent to each client and the cycle repeats. McMahan *et al.* [10] originally introduced the notion of FL in the context of mobile devices, however, due to the privacy and ownership challenges of medical data, FL is being increasingly used in the clinical domain [136].

In the FL framework, the aggregation method is key in combining the parameters of neural networks from each client. Its objective is to ensure that the global model is maintained in a state that generalizes well across participants. The most widely used algorithm is Federated Averaging (FedAvg) [10], which uses a weighted average to penalize the weights of networks. Since its introduction, multiple alternative methods have been designed to improve convergence rates with complex data having a large diversity of features in its composition.

In this study, we explore data imputation of distributed PD clinical assessments from the angle of FL (Fig 4.1). Specifically, we compare the performance of 7 FL frameworks as well as centralized learning for the imputation task. Further, to complement this analysis, we validate the effectiveness of each imputation solution through estimations of changes in motor and non-motor symptoms based on the Unified Parkinson's Disease Rating Scale (MDS-UPDRS) scores. We demonstrate the feasibility and practicality of the imputation of distributed clinical assessments with FL and its ability to obtain lower reconstruction and imputation errors as well as higher performance, compared

to a centralized learning and traditional imputation methods.

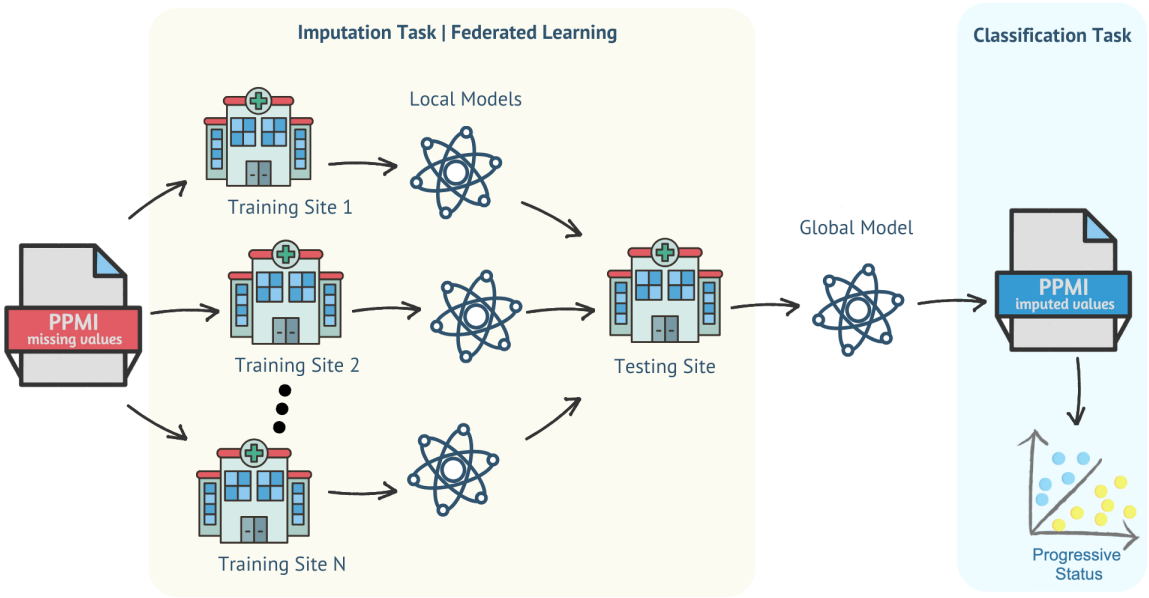

Figure 4.1: Framework utilized in this study. The PPMI database, with missing values, is split among multiple medical center. The imputation task is performed using a centralized and multiple FL strategies. To validate the results of the data imputation task, we predict symptoms progression, based on increase of MDS-UPDRS sub-scores at 12-moth after the first visit.

4.2 Methods

4.2.1 Data

We utilized the PPMI database, sponsored by Michael J. Fox Foundation for Parkinson Research, a comprehensive public multi-center database (www.ppmi-info.org/data), which includes longitudinal imaging, genetic, biosamples and clinical assessment data of large PD cohorts. We accessed longitudinal clinical evaluations from 2,347 participants in the PPMI curated dataset [137], in particular motor assessments, neuro-behavioral and neuro-psychiatric testing. Data was downloaded in October 2023.

Table 4.1 shows the 20 primary clinical assessments (and their sub-scores), imputed, including: Benton Judgement of Line Orientation (BJLO) Test, Boston Naming Test Score (BNT), Epworth Sleepiness Scale (ESS), Geriatric Depression Scale (GDS), Hoehn Yahr (HNY), Hopkins

Table 4.1: Clinical assessments used on the imputation task and their percentage of missing scores

Questionnaire	% missing value	Questionnaire	% missing value
BJLO	0.05	PIGD	0.19
BNT	0.73	QUIP	0.01
ESS	0.01	RBDSQ	0.01
GDS	0.01	SCOPA-AUT	0.01
HNY	0.18	SDMT	0.04
HVLT	0.04	SFT	0.04
LFS	0.70	STAI	0.01
LNS	0.04	MDS-UPDRS I	0.01
MoCA	0.04	MDS-UPDRS II	0.01
MSE-ADL	0.17	MDS-UPDRS III	0.19

Verbal Learning Test (HVLT), Lexical Fluency Score (LFS), Letter-Number Sequencing (LNS) , Montreal Cognitive Assessment (MoCA), Modified Schwab & England Activities of Daily Living (MSE-ADL) scale, postural instability and gait difficulty-predominant disease (PIGD), Questionnaire for Impulsive-Compulsive Disorders (QUIP), REM Behaviour Disorder Questionnaire (RBDQ), Scales for Outcomes in Parkinson's Disease - Autonomic Dysfunction (SCOPA-AUT), Symbol Digit Modalities Text (SDMT), Semantic Fluency Test (SFT), State-Trait Anxiety Inventory (STAI), , assessments from the Movement Disorder Society-Unified Parkinson's Disease Rating Scale (MDS-UPDRS) Part I: Non-Motor Aspects of Experiences of Daily Living, Part II: Motor Aspects of Experiences of Daily Living, and Part III: Motor Examinations , and included test results from 43 single measurements of photon emission computed tomography (SPECT) imaging, acquired at the baseline and subsequent visits. Generally, SPECT imaging tracks information about the metabolic rates of neurotransmitters and measures brain signals in different regions, which are commonly used in detecting metabolic abnormalities in the basal ganglia and is therefore included in our experiments.

4.2.2 Data pre-processing

For the imputation task, we scaled clinical scores with a min-max normalization method (min: 0, max: 1) based on individual score range. We also excluded patient information with a prodromal cohort status and removed categorical features from the analysis. For the prediction task, we impute

categorical values using the mode of the column. The final database contained clinical assessments of 1,370 patients (839 males/531 females, age at baseline (mean$\pm$SD) = 62.5$\pm$10.1).

4.2.3 Data splits

To emphasize the practical value of each FL experiment, we accounted for the non-identical and non-independent (non-IID) condition of each clinical center. We partitioned patients records in the PPMI database based on their respective *Site ID*, resulting in 52 centers. To create a hold-out test set, we pooled 20% of the total number of patients records into a single center, the remaining 27 centers were utilized for training and evaluating models.

4.2.4 Model selection

For the imputation task, we trained unoptimized Fully Connected Autoencoder (FCAE) models [116]. FCAEs are unsupervised learning models derived from denoising autoencoders. Their architecture consist of a custom masking layer, used to apply noise to entire modalities (clinical score and sub-scores), and a sequence of encoding and decoding layers made up of blocks of fully-connected and concatenation layers. These models were implemented with Keras 2.4.3 and Tensorflow 2.4.1, and their architecture is available in a public repository[1].

For the prediction task, we trained XGBClassifier models [138], which are based on decision trees. The popularity of these models in Parkinson's disease research has increased due to their predictive power and robustness when handling imbalanced data.

4.2.5 Model performance evaluations

To perform evaluations during the imputation task, we measure imputation errors of missing score and reconstruction errors of non-missing scores with the held-out test set by using two metrics (Equations 13 and 14). Since the missing scores in the ground-truth are unknown and unsuitable for measuring model performance, we introduced artificial missing clinical scores by hiding existing scores in the test set. These pseudo-missing values were then used as ground-truth. Formally, given a patient record N is composed of a sequence of $n \in \{0, ..., N-1\}$ clinical scores, where

[1] https://github.com/m-prl/PatiNAE

$k \in \{0, ..., N-1\}$ is a set of non-missing scores and $\mu \in \{0, ..., N-k-1\}$ is a set of missing scores in the original database, the additional pseudo-missing values $(\mu + \epsilon)$ is the total number of missing-scores in a patient record N, such that $k \in \{0, ..., N-(\mu+\epsilon)-1\}$.

Equation 13 measures the reconstruction error of non-missing scores:

$$A_1 = \frac{1}{K} \sum_{i=1}^{N} \sum_{j=1}^{M} (\hat{x}_j^i - x_j^i)^2 * P_j^i \tag{13}$$

, where K is the total number of known scores in the test set, N and M are the rows and columns in the database, x represents ground-truth values, and a mask P, which identifies pseudo-missing values in the test set. Similarly, Equation 14 quantifies the imputation error given the total number of missing clinical scores U.

$$A_2 = \frac{1}{U} \sum_{i=1}^{N} \sum_{j=1}^{M} (\hat{x}_j^i - x_j^i)^2 * (1 - P_j^i) \tag{14}$$

4.2.6 Centralized Learning

Early attempts to enable collaborative learning involved a centralized data center where raw data was collected, combined, analyzed, and processed from each participant on a single remote server. To simulate this setting, we trained instances of FCAEs by pooling patient records from the 27 training sets and evaluated their reconstruction and imputation performance on the test set. This served as a benchmark for evaluating the generalization of distributed models, which are described next.

4.2.7 Federated Learning and Aggregation Algorithms

In a FL framework, each model is trained with local data. Therefore, to simulate a FL environment, we instantiate 27 nodes and assigned a single partition of the PPMI data, each representing a single client in the learning process. This ensured that each clinical center had access to an independent local dataset. Another node was instantiated as the server, having exclusive access to the

held-out test set. With this setting, we explore and compare the performance a number of algorithms as described below.

Federated Simple Averaging (FedSimple)

With this setup, the aggregation algorithm computes the arithmetic mean of the network parameters. Equation 15 corresponds to simple average, where $w_t^k + 1$ represents the model weights of client k at iteration t, K signifies the total number of clients involved in the learning task.

$$w_{t+1} \leftarrow \frac{1}{K} \sum_{k=1}^{K} w_{t+1}^k \tag{15}$$

Federated Averaging (FedAvg)

McMahan *et al.* [10] introduced *Federated Learning*, a learning framework for distributed devices. At the core of this framework, the *Federated Averaging (FedAvg)* algorithm aggregates network parameters and maintains a global model that is shared across participants. The FedAvg algorithm computes the weighted average of all individual model updates, such that:

$$w_{t+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} w_{t+1}^k. \tag{16}$$

where $w_t^k + 1$ denotes the model weights of client k at iteration t, n_k represents the number of local training samples and n is the total number of samples.

Precision-weighted Federated Learning (PW)

Reyes *et al.* [72] proposed the Precision-weighted Federated Learning (PW) algorithm as a variance-based aggregation scheme for distributed machine/deep learning models. This algorithm differs from FedAvg in the way that individual local updates are aggregated. Instead of using the ratio of data samples as the multiplicative factor for weight update, PW takes into account local variance estimations, which are computed by the optimizer, and the update of the parameters of the

shared global model is made in proportion to the inverse of this variance, as shown in Equation 17:

$$w_{t+1} \leftarrow \sum_{k=1}^{K} \frac{\left(v_{t+1}^{k}\right)^{-1}}{\sum_{k=1}^{K}\left(v_{t+1}^{k}\right)^{-1}} w_{t+1}^{k} \tag{17}$$

where, w_{t+1}^{k} represents the model weights of client k at iteration t and v_{t+1}^{k} represents the variance of a given weight w at iteration t for client k. To estimate the inverse of the variance of the maximum likelihood, we propose the use of the raw second moment estimate (uncentered variance) from the Adam optimizer [86].

Imputation of Missing IoT Records (FedMiss)

Gkillas and Lalos [139] addressed the problem of data imputation in sensor recordings of distributed networks, such as IoT edge devices. With this aggregation method, missing rates of a given measurement are used to penalize local model updates. Equation 18 shows the aggregation operation performed at the server:

$$w_{t+1} \leftarrow \frac{1}{d} \sum_{k=1}^{d} q_k w_{t+1}^{k}. \tag{18}$$

where, w_{t+1}^{k} corresponds to model weights of client k at iteration t, q_k represents the rate of missing values in the dataset of each participant, and d corresponds to the total number of edge devices and the number of corresponding sensors within each device. In our study, we interpret edge devices as clinical centers, but we cannot conceptualize sensors. For that reason, we simply consider the number of participating clinical centers only.

Ditto

Li *et al.* [140] proposed Ditto as a general framework for personalized FL. Ditto's objectives consider fitting a single global model, w, across all local data in the network. The aim is to solve:

$$\min_{w} G(F_1(w), ..., F_K(w))$$

where $F_k(w)$ is the local objective for device k, and $G(.)$ is a function that aggregates the local objectives $F_k(w)_{k\in[K]}$ from each device. Ditto considers two 'tasks': the global objective (Global Obj) and the local objective $F_k(v_k)$, which aims to learn a model using only device k's data. To relate these tasks, they incorporate a regularization term that encourages the personalized models to be close to the optimal global model. The resulting bi-level optimization problem for each device $k \in [K]$ is given by:

$$\min_{v_k} h_k(v_k, w^*) := F_k(v_k) + \frac{\lambda}{2}||v_k - w^*||^2$$

$$s.t. w^* \in arg\min_{w} G(F_1(w), ..., F_K(w)).$$

where, the hyper-parameter λ controls the interpolation between local and global models.

Cyclic Weight Transfer (CWT)

Chang *et al.* [97] proposed the Cyclic Weight transfer (CWT) method. This learning method involves training a model on each client for a limited number of iterations and subsequently sharing the updated weights of such model with the next client. With CWT, a cycle completes when all clients are train with that model, then the cycle repeats.

Federated Learning Optimization (FedProx)

Li *et al.* [27] designed FL Optimization (FedProx), which is a generalization and re-parametrization of the FedAvg algorithm, that addresses the issues of training FL models exhibiting system and statistical (Non-IID) heterogeneity. This algorithm differs from the traditional FedAvg in that clients optimize a regularized loss with a proximal term. This term penalizes local updates to keep them closer to the global model, thus, accounting for heterogeneity associated with each local model.

More specifically, this method adds the proximal term to the original objective function defined in Equation 19.

$$\min_{w \in R^d} f(w) \quad \text{with} \quad f(w) = \frac{1}{n} \sum_{i=1}^{n} f_i(w) + \frac{\mu}{2} ||w - w^t||^2, \tag{19}$$

where w represents the model's weights being optimized, w^t is the global model's parameters at step t, μ is the proximal term's coefficient, which controls the regularization strength, and $f(w)$ is the local objective function. When $\mu = 0$, we obtain the original FedAvg algorithm. To handle the computation load over a wide variety of systems, local optimizations run for a device-determined number of epochs, instead of meeting a strict training deadline. This allows more clients to contribute to the aggregation algorithm depending on device resources.

4.2.8 Progressive and non-progressive status

For the predictions of PD symptoms progression, we define patients with a progressive or non-progressive status based on the increase in motor and non-motor symptoms. Since the data in the MDS-UPDRS scores is continuous, we convert the predictions of MDS-UPDRS into a classification task by using the method described by Sadaei *et al.* [115]. To that extent, we define patients with a progressive status as those who exhibit an increase in their MDS-UPDRS scores between time points, other cases are considered non-progressive. As such, the target variable chosen for the classification task was the PD progressive status computed on the 12-month visit for MDS-UPDRS subparts (MDS-UPDRS I, II, and III). It is interesting to perform predictions of disease progression with the MDS-UPDRS subparts scores as these indicate the severity of disease condition, based on motor and non-motor symptoms.

4.3 Experimental Results

4.3.1 Imputation of distributed clinical assessments

We provide the training setup as follows. Given a set of clinical examinations with missing values μ, we trained each FCAE with a pre-imputation strategy on the set of features by initializing missing scores with the mean value of their corresponding column. With this step, we bypassed the initialization step of FCAEs. Further, we split the training data into training (70%) and validation (30%) segments. The selection and optimization of hyper-parameters was carried out by using a Bayesian technique using a Gaussian process as a prior in the optimization. To provide fair comparisons, we utilize the same hyper-parameters across different implementations of collaborative learning algorithms in this study. The optimized training setup utilized for the imputation task includes the Adam optimizer with a learning rate of 1e-06, a batch size of 16, 120 epochs, and specifically for FCAEs, a drop out rate of 0.1 and an internal representation (IR) of 7. With a Mean Squared Error (MSE) loss function, we minimize the reconstruction and imputation errors. Further, we monitor training and reduce the learning rate when there is no improvement in learning after 10 epochs, and stop training when there is no absolute improvement after 10 training passes.

Effect of missing modalities during training

We explore the effects of having multiple degrees of noise in the training data in order to examine how its heterogeneity, expressed in the number of missing entries in clinical information, impacts the performance of aggregation methods. To that extent, we trained FCAEs models for every client with various artificial missing values introduced into the training data (e.g. corruption ratios) 10%, 30%, and 60%, and the optimized hype-parameters for 150 rounds of communication.

Figure 4.2 shows the reconstruction and imputation errors plots, based on the A1 and A2 metrics, respectively. These plots illustrate the performance of models up to communication round 75, allowing for a clearer appreciation of the models' convergence. As observed, that generalization improves more with FL models, compared to the central regime. We also noticed that FedMiss does not converge, primarily because the number of data sources available among participating clients is small compared to the large number of data sources from clients and sensors intended in the

original paper. For the rest of the FL algorithms, it takes less than 20 round of communication to outperform the central model. Table 4.3 provides more granular comparisons. Alternatively, CWT is the FL algorithms that presents lower imputation and reconstruction errors, except for A2 when the corruption ratio is high. Yet, CWT is the second best performing algorithm after centralized learning. These findings are in alignment with previous studies comparing FedAvg with CWT [11], as CWT shows certain advantages in performance over other FL algorithms. Here, we posit that the performance of CWT is attributed to the sharing of a single model's weights with the next client during training, allowing it to access the combined data from all clients multiple times, instead of using multiple models trained on individual data subsets. These results suggest that better generalization of data imputation can be obtained with distributed models, as demonstrated in our initial exploration of the effectiveness of FL algorithms in imputing distributed clinical assessments [15].

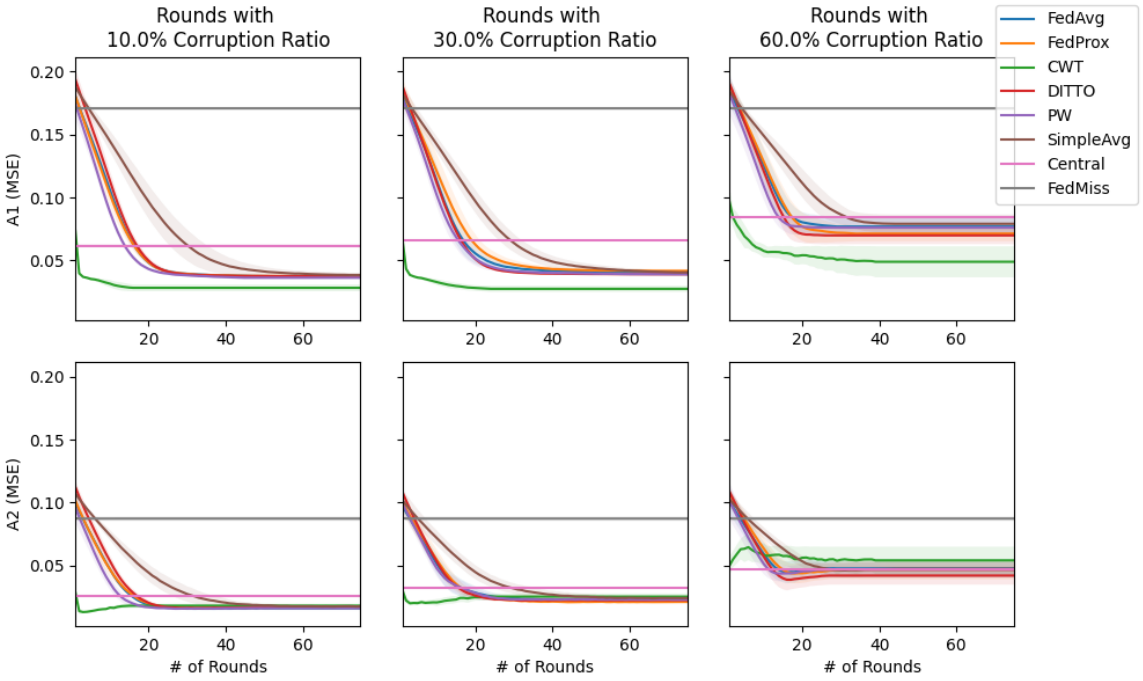

Figure 4.2: Performance of FL algorithms based on the reconstruction error (A1) and imputation errors (A2) with an increasing number of missing values (10%, 30%, and 60%) in the training set.

Impact of the number of clients participating concurrently in the learning process

In FL, the computational capabilities and availability of participating clients (e.g. clinical centers) can be challenging. The *client fraction* is a setting that allows a proportion of clients, relative to the total number of available clients, to actively operate at a given learning cycle. In the following experiments, we investigate the effects of having multiple clients participating on the learning process, concurrently. To do so, we configured the learning environment to select (20%, 50%, and 100%) random clients at each iteration. Then, we trained distributed FCAEs models with a fixed corruption ratio of 10% for 150 rounds of communication and used the same training hyper-parameters as described before.

Table 4.4 shows a summary of the performance of models based on the mean and standard deviation of the A1 and A2 (MSE) scores computed with the held-out test set. Interestingly, CWT can obtain lower reconstruction and imputation errors compared to models trained with a centralized learning approach, when the rate of participating clients is 20% and 50%. In general, we observe that better generalizations can be achieved when all participants remain active during training, except for DITTO and FedAvg. We presume that training with fewer participants improves the global regularization strategy in DITTO, encouraging local models to avoid overfitting at individual clients. Conversely, FedAvg demonstrates comparable performance with more stable imputations than models trained with all active clients in a centralized learning setting.

Table 4.3: Evaluation of model convergence utilizing centralized or FL, assessed by reconstruction error for known values (A1) and imputation error for missing values (A4), with varying levels of missing data (10%, 30%, and 60%) in the training dataset. We report the mean and standard errors of MSE errors obtained across multiple runs using different seeds. The FL strategy with the lowest MSE is highlighted in bold

	Corruption Ratio 10%		Corruption Ratio 30%		Corruption Ratio 60%	
	A1	A2	A1	A2	A1	A2
Central	0.0618 ± 0.0013	0.0260 ± 0.0012	0.0662 ± 0.0010	0.0321 ± 0.0011	0.0843 ± 0.0007	0.0469 ± 0.0004
CWT	$\mathbf{0.0363 \pm 0.0113}$	$\mathbf{0.0162 \pm 0.0041}$	$\mathbf{0.0353 \pm 0.0095}$	$\mathbf{0.0220 \pm 0.0035}$	$\mathbf{0.0576 \pm 0.0172}$	$\mathbf{0.0534 \pm 0.0097}$
DITTO	0.0742 ± 0.0487	0.0354 ± 0.0284	0.0727 ± 0.0461	0.0381 ± 0.0247	0.1089 ± 0.0416	0.0571 ± 0.0234
FedAvg	0.0683 ± 0.0434	0.0310 ± 0.0240	0.0665 ± 0.0413	0.0341 ± 0.0204	0.1147 ± 0.0383	0.0599 ± 0.0204
FedMiss	0.1708 ± 0.0013	0.0872 ± 0.0020	0.1708 ± 0.0014	0.0871 ± 0.0020	0.1709 ± 0.0013	0.0875 ± 0.0019
FedProx	0.0667 ± 0.0426	0.0311 ± 0.0239	0.0741 ± 0.0437	0.0353 ± 0.0229	0.1116 ± 0.0400	0.0608 ± 0.0212
PW	0.0648 ± 0.0411	0.0289 ± 0.0227	0.0646 ± 0.0404	0.0348 ± 0.0205	0.1079 ± 0.0380	0.0574 ± 0.0205
FedSimple	0.0748 ± 0.0462	0.0361 ± 0.0260	0.0743 ± 0.0432	0.0389 ± 0.0221	0.1205 ± 0.0348	0.0624 ± 0.0190

Table 4.4: Summary of model performance in terms of MSE reconstruction error and MSE imputation error, A1 and A2 respectively, with various clients participating concurrently in the learning process. Results are presented as the mean and standard deviation of A1/A2 errors obtained across multiple runs using different random seeds. The FL strategy with the lowest MSE is highlighted in bold.

	Participating Clients: 20%		Participating Clients: 50%		Participating Clients: 100%	
	A1	A2	A1	A2	A1	A2
Central	-	-	-	-	0.0618 ± 0.0013	0.0260 ± 0.0012
CWT	$\mathbf{0.0531 \pm 0.0362}$	$\mathbf{0.0224 \pm 0.0192}$	$\mathbf{0.0438 \pm 0.0235}$	$\mathbf{0.0186 \pm 0.0111}$	$\mathbf{0.0363 \pm 0.0113}$	$\mathbf{0.0162 \pm 0.0041}$
DITTO	0.0752 ± 0.0472	0.0351 ± 0.0265	0.0754 ± 0.0479	0.0347 ± 0.0268	0.0742 ± 0.0487	0.0354 ± 0.0284
FedAvg	0.0710 ± 0.0433	0.0322 ± 0.0236	0.0704 ± 0.0429	0.0310 ± 0.0232	0.0683 ± 0.0434	0.0310 ± 0.0240
FedMiss	0.1708 ± 0.0012	0.0873 ± 0.0018	0.1708 ± 0.0012	0.0874 ± 0.0018	0.1708 ± 0.0011	0.0872 ± 0.0020
FedProx	0.0752 ± 0.0461	0.0343 ± 0.0259	0.0744 ± 0.0458	0.0337 ± 0.0256	0.0667 ± 0.0426	0.0311 ± 0.0239
PW	0.0755 ± 0.0465	0.0341 ± 0.0253	0.0729 ± 0.0455	0.0319 ± 0.0250	0.0648 ± 0.0411	0.0289 ± 0.0227
FedSimple	0.0752 ± 0.0457	0.0355 ± 0.0259	0.0753 ± 0.0452	0.0348 ± 0.0254	0.0748 ± 0.0462	0.0347 ± 0.0260

Effects of Local Computation

In practice the batch size for machine and deep learning can vary based on the nature of the task, the amount of data, the architecture of the network, and the computational resources available. Often, batch sizes are often small for medical applications, due to the limited amount of private data stored within clinical centers, constrained computational resources, type of data modality (e.g. MRI), or high variability in the data. Owing to this, we design a set of experiments where we compare the impact of small batch sizes in model's performance. We trained models using the same configuration as in Section 4.3.1, and compared their MSE error with batch sizes 16 and 32.

Table 4.5 summarizes the imputation performance using small batch sizes. We observe that models trained in a FL setting are more robust when more heterogeneous samples are propagated through the network, an effect not seen in models trained in a centralized learning setting. Notably, CWT outperforms most of the FL aggregation algorithms, except when the number of missing clinical scores is high (corruption ration = 60%). These results suggests that, depending on the amount of missing values in the training data, more accurate imputations can be achieved with CWT and PW.

Table 4.5: Summary of model performance, based on imputation error (A2) only, when simulated clinical centers models are trained with small batch sizes 16 and 32. We report the mean and standard deviation of MSE error calculated during multiple runs using different seeds.

	Corruption Ratio 10%		Corruption Ratio 30%		Corruption Ratio 60%	
	A2 (B=16)	A2 (B=32)	A2 (B=16)	A2 (B=32)	A2 (B=16)	A2 (B=32)
Central	0.0260 ± 0.0012	0.0508 ± 0.0019	0.0321 ± 0.0011	0.0522 ± 0.0018	0.0469 ± 0.0004	0.0577 ± 0.0019
CWT	$\mathbf{0.0162 \pm 0.0041}$	$\mathbf{0.0131 \pm 0.0011}$	$\mathbf{0.0220 \pm 0.0035}$	$\mathbf{0.0207 \pm 0.0011}$	$\mathbf{0.0534 \pm 0.0097}$	0.0453 ± 0.0044
DITTO	0.0354 ± 0.0284	0.0157 ± 0.0016	0.0381 ± 0.0247	0.0214 ± 0.0019	0.0571 ± 0.0234	0.0436 ± 0.0029
FedAvg	0.0310 ± 0.0240	0.0156 ± 0.0021	0.0341 ± 0.0204	0.0225 ± 0.0011	0.0599 ± 0.0204	0.0489 ± 0.0121
FedMiss	0.0872 ± 0.0020	0.0871 ± 0.0019	0.0871 ± 0.0020	0.0873 ± 0.0021	0.0875 ± 0.0019	0.0870 ± 0.0020
FedProx	0.0311 ± 0.0239	0.0161 ± 0.0007	0.0353 ± 0.0229	0.0224 ± 0.0010	0.0608 ± 0.0212	0.0488 ± 0.0083
PW	0.0289 ± 0.0227	0.0149 ± 0.0014	0.0348 ± 0.0205	0.0221 ± 0.0009	0.0574 ± 0.0205	$\mathbf{0.0424 \pm 0.0088}$
FedSimple	0.0361 ± 0.0260	0.0172 ± 0.0008	0.0389 ± 0.0221	0.0248 ± 0.0013	0.0624 ± 0.0190	0.0472 ± 0.0085

4.3.2 Prediction of short-term disease trajectories

To evaluate the model's coherence, particularly regarding feature importance, we used the Select-K-Best method to identified the top 50 important features based on the F-1 score. The motor and non-motor symptoms identified in this study have been previously reported in the literature as being used in data-driven approaches for the identification of PD sub-types and disease trajectories [141, 114, 112]. The demographic and clinical examinations used to train XGBClassifier models include:

(1) Motor symptoms: PIGD, HNY.

(2) Non-motor symptoms: MDS-UPDRS I (np1anxs, np1apat, np1cog, np1dds, np1dprs, np1fatg, np1hall), GDS, MSE-ADL, QUIP (quip_any, quip_eat, quip_hobby, quip_pund, quip_sex, quip_walk), RBDSQ, SCOPA-AUT (scopa_gi, scopa_pm, scopa_therm, scopa_ur), STAI (stai, stai_trait).

(3) Neuropsychological features: MoCA, BJLO, HVLT (hvlt_discrimination, hvlt_immediaterecall, hvlt_retention, hvltrdly, hvltrec, hvltfprl), LNS, LFS, SDMT, SFT.

(4) Other features: SPECT (con_caudate, con_putamen, con_striatum, datscan_caudate_l, datscan_caudate_r, datscan_putamen_l, datscan_putamen_r, ips_caudate, ips_putamen, ips_striatum, lowput_ratio, mean_caudate, mean_putamen, and mean_striatum).

With the key features identified, we proceed to train XGBClassifier models to validate the impact of different imputation solutions classifying subjects based on their progressive status, as defined in Section 4.2.8. During training, we use the K-fold cross-validation method on imputed PPMI datasets, where K=5. Within each fold, 1/K of the samples were reserved as the test set, and the rest of samples were used for training models. To measure classification performance of aggregation algorithm with training data showing high levels of missing values, we configured a setting with corruption ratio of 60% , a client fraction of 100%, which allow all clinical centers to participate in the learning process, and a batch size of 32. No other optimization were considered. We measure classification performance based on accuracy, F1-score, Precision-Recall and ROC-AUC curves on each of the imputations produced with learning algorithms and included evaluations with imputations made with mean values from each column, which is considered the current practice.

The result of predictions of progressive status for MDS-UPDRS I, II, and III with clinical information from the 50 features, are shown in Table 4.6. While higher PR-AUC and ROC-AUC can be obtain with imputed values from models trained in a centralized learning, imputations with FedAvg and FedProx represents FL alternatives that achieve higher PR-AUC and ROC-AUC, respectively. In addition, we observed that the imputation of clincial assessments with PW may achieve both higher PR-AUC and ROC-AUC for MDS-UPDRS II, and III. Table 4.7 presents the performance metrics based on accuracy and F1-score metrics obtained when the training data was imputed with the different strategies. For MDS-UPDRS I, the averaging effect in FedMiss achieves higher accuracy, despite the low performance of models trained with FedMiss in our previous experiments; FedProx can exhibit higher F1-scores for MDS-UPDRS I, and better accuracy for MDS-UPDRS II; and PW achieves higher accuracy for MDS-UPDRS III as well as high F1-scores for MDS-UPDRS II, III. The choice of aggregation algorithm depends on the predicted variable. These results suggest that more accurate imputations may be obtained with FL algorithms learning from highly heterogeneous inputs.

To validate the statistical significance of the results, we conducted independent one-way ANOVAs for each metric. An ANOVA test revealed a significant effect on classification performance of MDS-UPDRS I based on the AUC metric ($F(9, 240) = 2.087, p = 0.031, \omega^2 = 0.038$). Post hoc testing

using Tukey's correction indicated that imputations produced with the a centralized learning strategy resulted in significantly higher AUC compared to those obtained with CWT (p = 0.036). No significant differences were observed for other metrics. Therefore, we conclude that using FL imputation strategies may lead to improvements in classification scores compared to a centralized learning setting, but no statistical significance was observed with the given configuration.

Table 4.6: Performance results based on (mean ± standard deviation) precision-recall (pr) and area under the roc curve (roc-auc) curve among the predictions of disease progressive status using federated and non-federated learning algorithms.

	MDS-UPDRS I		MDS-UPDRS II		MDS-UPDRS III	
	PR	ROC-AUC	PR	ROC-AUC	PR	ROC-AUC
Central	**0.5500 ± 0.0278**	**0.5572 ± 0.0303**	0.5629 ± 0.0678	0.6289 ± 0.0850	0.6307 ± 0.0357	0.5717 ± 0.0566
CWT	0.5291 ± 0.0347	0.5312 ± 0.0302	0.5617 ± 0.0656	0.6371 ± 0.0807	0.6245 ± 0.0341	0.5697 ± 0.0516
DITTO	0.5405 ± 0.0303	0.5420 ± 0.0261	0.5651 ± 0.0656	0.6357 ± 0.0788	0.6237 ± 0.0363	0.5666 ± 0.0553
FedAvg	0.5423 ± 0.0242	0.5480 ± 0.0274	0.5552 ± 0.0682	0.6283 ± 0.0905	0.6302 ± 0.0362	0.5734 ± 0.0566
FedMiss	0.5416 ± 0.0285	0.5487 ± 0.0289	0.5668 ± 0.0577	0.6340 ± 0.0748	0.6239 ± 0.0447	0.5685 ± 0.0628
FedProx	0.5378 ± 0.0336	0.5448 ± 0.0302	0.5709 ± 0.0665	0.6449 ± 0.0741	0.6304 ± 0.0333	0.5747 ± 0.0540
PW	0.5334 ± 0.0229	0.5326 ± 0.0276	**0.5758 ± 0.0613**	**0.6476 ± 0.0729**	**0.6322 ± 0.0267**	**0.5807 ± 0.0478**
FedSimple	0.5329 ± 0.0287	0.5346 ± 0.0244	0.5636 ± 0.0695	0.6358 ± 0.0825	0.6215 ± 0.0388	0.5710 ± 0.0544
Mean	0.5420 ± 0.0274	0.5418 ± 0.0260	0.5609 ± 0.0596	0.6292 ± 0.0784	0.6165 ± 0.0413	0.5686 ± 0.0656

4.4 Discussion

With this study, we address the issue of missing clinical data. In particular, we explored the clinical utility of FL for the tasks of data imputation of Parkinson's disease clinical assessments distributed across multiple centers. Our findings have implications for both researchers and clinicians. We demonstrate its utility by imputing real patient data from the PPMI database and validating the impact of different FL aggregation algorithms. The evidence that supports the aforementioned

Table 4.7: Performance results based on (mean ± standard deviation) accuracy and f-1 scores among the predictions of disease progressive status using federated and non-federated learning algorithms.

	MDS-UPDRS I		MDS-UPDRS II		MDS-UPDRS III	
	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Central	0.5254 ± 0.0253	0.4759 ± 0.1241	0.5853 ± 0.0602	0.5674 ± 0.0502	0.5641 ± 0.0430	0.6400 ± 0.0397
CWT	0.5200 ± 0.0253	0.4821 ± 0.0837	0.5918 ± 0.0542	0.5726 ± 0.0495	0.5653 ± 0.0356	0.6400 ± 0.0292
DITTO	0.5236 ± 0.0241	0.4985 ± 0.0586	0.5918 ± 0.0575	0.5749 ± 0.0455	0.5669 ± 0.0389	0.6387 ± 0.0262
FedAvg	0.5291 ± 0.0217	0.4925 ± 0.0898	0.5863 ± 0.0634	0.5673 ± 0.0452	0.5738 ± 0.0428	0.6458 ± 0.0308
FedMiss	**0.5334 ± 0.0202**	0.4729 ± 0.1486	0.5876 ± 0.0561	0.5577 ± 0.0709	0.5639 ± 0.0440	0.6362 ± 0.0301
FedProx	0.5305 ± 0.0207	**0.5049 ± 0.0632**	**0.5994 ± 0.0528**	0.5813 ± 0.0429	0.5714 ± 0.0366	0.6430 ± 0.0214
PW	0.5169 ± 0.0302	0.4824 ± 0.0673	0.5990 ± 0.0543	**0.5825 ± 0.0450**	**0.5773 ± 0.0344**	**0.6490 ± 0.0260**
FedSimple	0.5175 ± 0.0222	0.4911 ± 0.0603	0.5953 ± 0.0628	0.5795 ± 0.0538	0.5733 ± 0.0312	0.6447 ± 0.0250
Mean	0.5172 ± 0.0216	0.4773 ± 0.0897	0.5833 ± 0.0527	0.5747 ± 0.0315	0.5589 ± 0.0502	0.6243 ± 0.0279

clinical utility is based on extensive analyses on a distributed model's performance using seven FL aggregation algorithms. To the best of our knowledge, we are the first group evaluating imputation solutions of PPMI data in a FL setting.

In alignment with Tuladhar *et al.*'s findings in [142], we demonstrate that better generalization and lower MSE errors can be obtained with FL algorithms, compared to centralized learning and traditional imputation strategies. More recently, Danek *et al.* also demonstrated small performance gains (2% better AUC-PR than central models) with FL algorithms for the task of multi-omics Parkinson's disease prediction. In terms of the advantages obtained with learning from distributed data sources, we show that good performance across MDS-UPDRS subparts can be obtained while imposing data privacy. In this study, FL methods resulted in lower reconstruction and imputation errors compared to a centralized learning strategy.

An important outcome from the present study is that multi-center studies can be performed with FL. First, we can increase the models' generalization when independent models learn from multi-center clinical data, which is never transferred or pooled at a single medical, or research, center. Second, FL strives for unbiased learning from heterogeneous data sources. It is estimated that an ordinary medical center produces about 15TB to 20TB of new data every year [143], however it is a challenging task to consolidate raw biomedical data using traditional learning methods. With FL approaches the collaboration is enabled across different clinical centers. This is especially valuable for the analysis of rare diseases, where very few patients with rare conditions are seen at any single institution [144, 142]. Third, the iterative scheme in FL can benefit new medical centers joining the learning process at any point of the training or evaluation phase. This is possible since the intrinsic independence of distributed machine/deep models enables real-time continual learning as the aggregation of SGD updates and communication operations are synchronized and orchestrated by the server. Therefore, the learning cycle can be enriched with the new information provided by the joining center, leading to less biased decisions at any given iteration. These conclusions are consistent with the recent study performed with multi-omics data by Danek *et al.* [145]

We also highlight the improvements in the performance of distributed models trained from multi-institutional clinical assessments using FL. Our study utilized real clinical records, including longitudinal imaging, genetic, biosamples and clinical assessment data from the PPMI database

to explore the power and limitations of different FL frameworks. We demonstrated that lower reconstruction and imputation errors may be obtained with models trained in a FL setting. Furthermore, we observed higher accuracy, F1-scores, PR-AUC, and ROC-AUC when clinical assessments were imputed using FL algorithms. However, with 50 most important clinical assessments, a statistically significant improvement in imputations was only observed for non-motor symptoms in MDS-UPDRS I, using the centralized learning approach compared to the FL algorithm. Future research is needed to verify the effect of a different combination of features.

This study has potential limitations. The imputation scores and the prediction of the target variables were based on the same database. This limits the scalability of the predictive task to be applied to other datasets. Also, the data partitions contain a small number of patients, with a minimum of 12 patients. This limitation could bias and, subsequently, prevent local models from effectively learning a representation of the feature space. Further, our predictions may underestimated patient's disease trajectory since the experiments in this imputation task were carried out with numerical variables only. To learn meaningful categorical variables, we suggest the use of one-hot encoding or an entity embedding technique [146, 116]. For datasets with categorical features, embeddings or techniques like entity embeddings can be used to represent categorical variables in a continuous space, facilitating imputation. Similar to the study in [115], our study also exhibits a limitation in the use of MDS-UPDRS measurements having a mixture of ON and OFF medication effects. We do not adjust for medication status, instead we use unadjusted MDS-UPDRS scores in the prediction task.

Future work could extend these studies to complement the analyses of complex imaging modalities, such as X-ray, CT, PET, MRI or fMRI imaging. This exploration could be valuable, as the different images modalities might provide additional information needed to better understand the etiology and pathogenesis of the disease. This is particularly important for early identification of disease subtypes and trajectories, and planning of treatment strategies.

4.5 Conclusions

This chapter presents a systematic study assessing the performance of FL aggregation algorithms for handling missing clinical assessments, including evaluations with an independent subset of the PPMI database. We imputed data using seven aggregation algorithms and compared their performance against models trained using a centralized learning strategy. To validate the results of the imputation task, we performed a downstream analysis, aiming at predictions of PD symptoms progressions.

The results of the comparisons demonstrate better imputation performance with FL algorithms compared to the centralized counterpart. This is particularly important in sensitive domains where data privacy is a priority and evaluation of heterogeneity data sources is significant for an accurate prognosis, identification sub-typing, and personalized treatment plans.

4.6 Acknowledgements

The data that support the findings of the study are publicly available at www.ppmi-info.org. PPMI, a public-private partnership, is funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners, including AbbVie, Avid, Biogen, Bristol-Myers Squibb, Covance, GE Healthcare, Genentech, GlaxoSmithKline, Lilly, Lundbeck, Merck, Meso Scale Discovery, Pfizer, Piramal, Roche, Sano Genzyme, Servier, Teva, and UCB.

Chapter 5

Game On: How Human Perception of AI Uncertainty Shapes Decision-Making

Preface

After proposing a solution for technical and legal privacy issues in AI adoption in clinical trials and workflows (Chapter 3) and demonstrating its clinical utility in Chapter 4, the next two chapters focus on exploring how to visualize data to improve clinical decision support systems. We concentrate on visualizing uncertainty of AI models, emphasizing factors that can help understand and interpret the AI's model outcomes. Prior to studying decision-making in high-risk medical scenarios (in Chapter 5), we build a foundational understanding to identify effective visualization techniques and potential pitfalls in low-risk scenarios in this Chapter. Specifically, we utilized classic gaming scenarios as a proxy to investigate human-AI collaboration in decision-making. By using different visual methods with games, we explore individual's ability to better perceive AI uncertainty. We examined the impact of AI uncertainty on trust in AI, confidence in decisions, and decision changes among individuals with varying attitudes towards AI in situations with minimal potential harm.

This chapter contributes to human–computer interaction (HCI) research by revealing factors that influence trust in AI among people with different attitudes towards it. Based on our findings, we recommend designing AI outputs to cater to individuals' varying attitudes towards AI. First, AI

model outputs should be more transparent and provide informative feedback to ensure accountability in decision-making. Second, AI designers should take into account a person's specific attitudes towards AI to create personalized and engaging experiences. We demonstrate that more transparent AI solutions can increase trust in AI technologies in low-risk scenarios.

This chapter is based on the journal paper **Reyes, J.**., Ludera, D., Batmaz, A., & Kersten-Oertel, M. Game On: How Human Perception of AI Uncertainty Shapes Decision-Making. Submitted to PLOS ONE (June 2024).

Abstract

Decision-making based on AI can be challenging, particularly when factoring in the uncertainty associated with AI predictions. To investigate the impact of visualizing uncertainty in AI solutions, we considered human factors (e.g., visual perception and cognition) during the design of model outputs. We conducted a user study with 147 participants using static classic gaming scenarios as a proxy to show human-AI collaboration in decision-making. Our study measures changes in decisions, trust in AI, and decision-making confidence when uncertainty is visualized in a continuous format in comparison to a binary output of the AI model. We found that visualizing uncertainty significantly strengthens trust in AI for 58% of participants with negative attitudes towards AI, and 31% of these participants found the visualization of uncertainty useful. Additionally, size was identified as the visualization method that most impact in individuals' trust in AI and confidence in their decisions. We also found a strong association between gaming experience and decision changes when uncertainty was visualized, and a strong association between trust in AI and individuals' attitudes towards AI. Our study provides insights into understanding the psychology of participants, specifically how individuals perceive uncertainty in AI models. These findings provide significant implications for the design of human-AI based decision support systems.

5.1 Introduction

Artificial intelligence (AI), a field where computers are leveraged to mimic or reproduce the problem-solving and decision-making capabilities of the human mind, is having significant impacts on people's work and lives. The adoption of AI models for decision-making is significantly increasing in our daily lives from leisure, entertainment [147] and serious gaming [148, 149] to more sensitive domains, such as criminal justice, banking, or healthcare [150, 151, 152].

Until recently, the development of AI systems has mainly been driven by a "technology-centered approach", which focuses on algorithms rather than the development of useful AI systems that meet actual user needs [153, 25, 154]. However, neglecting the adoption of a "user-centered design" [155], "human-centered design" [26] or "human-AI" [25] approach, which prioritizes the usability and usefulness of these systems by focusing on users, their needs, and requirements, can

lead to limited use and uptake of these systems. One specific aspect of human-AI design is to consider how to display an AI model's information. Yet, few researchers have focused on how best to convey AI information to the user and how different visualizations can impact perception and cognition.

From the user perspective, users often rely on AI models without understanding the confidence of the AI's prediction almost to the point of delegating decisions to the automated systems completely [156, 157]. This can result in a false sense of confidence, ineffective decision-making and incorrect conclusions. Indeed, a clear interpretation of AI predictions' uncertainty (e.g., recommendation) is not trivial and can pose a challenge for many experts and non-experts, particularly in areas where there is high uncertainty. This is more evident when human factors (e.g., visual perception and cognition) are not considered in the design choices of presenting model results, which can lead to decision errors that can cause adverse effects on the users of those systems.

In this chapter, we sought to understand how visualizing an AI model's uncertainty affects decision-making, to identify common traits among people who accept machine judgment as support for their decisions, and to explore the ways humans manage decision-making under the exposure of algorithmic advice. Specifically, we focus on exploring the impact of visualization of uncertainty on people's decisions, trust in AI model's reliability, and confidence in decisions among people with different attitudes towards AI [158]. To answer these questions, we conducted a large exploratory study via online surveys using classical games. We designed a number of gaming scenarios, with and without visualization of AI uncertainty, where participants assessed situations to determine a move for a character in a game. We then measured the number of decision changes (when uncertainty visualization was used in a specific scenario versus when it was not used), and fluctuations in trust in AI and confidence in their decisions. In the last part of the study, we evaluated users' perceptions regarding the utility and preference of various visual representations of AI uncertainty, as a way to highlight the limitations of AI model outputs and promote transparency.

This chapter makes the following contributions: (1) We provide empirical evaluations on how visualizing AI uncertainty affects human factors, particularly trust in AI, confidence in decisions, and decision changes, considering individuals with varying attitudes towards AI. (2) We use static

classic game scenarios as proxies to study human-AI interaction in decision-making, through evaluations about the utility and preference of visual representations of AI uncertainties by using simple visual techniques like size, color saturation, and transparency.

5.2 Related Work

There are a number of research works that investigate new algorithms, improvements, applications, and the influence of AI. Indeed, recent studies have addressed questions about AI-based decision-support systems from the angle of human perceptions, including evaluations of factors such as risk, anxiety, fairness, usefulness, and trustworthiness [156, 159, 160, 161, 162]. The work in this chapter builds upon prior research at the intersection of data visualization, human-AI design, and decision support systems.

5.2.1 Data Visualization and Uncertainty

Data visualization is a representation technique that transforms datasets into visual components in order to obtain actionable insights. Mackinlay [163] addressed the importance of leveraging the human visual system and its perceptual capabilities and visual variables to create effective visual expressions of information. In our context, we use different visual representation of the AI model's output and measure people's perception of AI uncertainty as a way to improve decision-making confidence and alleviate the challenge of reasoning with uncertainty.

Previous research studies suggest three main categories for perceiving data uncertainty: color-oriented approaches (hue, saturation, or brightness), focus-based methods (mapping uncertainty to contour crispness, transparency, or resolution), and geometric mapping (e.g., sketchiness in rendering, distorting line marks) [164]. Blur has also been used to guide attention to in-focus regions in images, which can be considered to have more certainty. Also, heat maps are a commonly used color-oriented approach that is specifically useful in identifying regions of interest. Generally, the range between blue-green indicates low-interest regions, and the range between yellow-red indicates regions of high interest. Despite the benefits generated by heat maps, several researchers argued that such maps can be confusing due to the lack of perceptual ordering [165]. In addition, according to

Breslow *et al.*[166], an alternative to color heat maps is to use changes in contrast or luminance in a single hue, which allows one to compare relative values between high and low interest regions or regions with more or less certainty. Based on these previous works, we chose to use size, color saturation, and transparency as means to represent varying levels of uncertainty in AI's outputs.

5.2.2 Human-AI Decision Support Systems

As described by Jarrahi [23], human-AI research primarily aims to augment human capabilities and enhance decision-making processes rather than simply replacing humans in those decisions. To support this vision, a substantial body of research has explored the practical use of AI-driven decision support systems across various domains, including healthcare, productivity, performance evaluations, negotiation, law and civic affairs, finance, business, education, leisure and arts [167, 168, 169, 170, 171, 172, 173]. Although some research has focused on improving human-AI collaboration to support decision-making by building trust in AI, others have integrated visual representations of AI uncertainty.

Trust is defined as the degree to which a person or group of people relies on or has confidence in the dependability of someone or something to fulfill their promise [48]. Thus, establishing trust in AI is crucial for achieving the adoption of AI systems as decision support systems. While there is no consensus on how the broad conceptualization of trust should be measured, some works either utilize Mayer *et al.*'s dimensions of trust [174], or build their own self-assessment questions to measure trust.

Online surveys, specific-purpose applications, and simulations have been created as instruments for evaluating trust in AI. For example, Liu *et al.*'s work [159] investigated people's perception of trust, experience, and attitudes towards AI with emails written by AI language models. Trust was measured with Mayer *et al.*'s dimensions of trust [174], and the attitudes towards AI with the General Attitudes towards Artificial Intelligence Scale (GAAIS) [158]. Their findings suggested that trust in emails weakens when people are aware of AI's intervention, but it grows stronger when the content of the email involves relations between people. No significant correlations were found between AI attitudes and trust. Similarly, Zhang *et al.* [151] used an online survey to explore user's perceptions of trust, performance expectancy, and intentions regarding the quality of financial advice

provided by AI-driven advisors. Trust was measured with a Likert scale for the first part of the study and with a self-assessment based on three dimensions: cognitive trust in competency, cognitive trust in integrity, and emotional trust in the last part. The outcomes of the study suggested that human financial advisors were trusted more than AI-advisors, regardless of their expertise level and sex. Also, no significant differences between human and AI-advisors were found regarding performance expectancy and intention to hire. Cai *et al.* [161] developed an AI-system to assist clinicians in the search of anatomical images. Participants were assessed based on their trust in AI, perceived utility, workload, and preference between two interfaces were measured. Trust was measured with Mayer *et al.*'s dimensions of trust [174]. The study found a perceived increase in utility, trust in AI, and preference for the AI system over traditional interfaces. In the military domain, Gurney *et al.* [175] adapted an online simulation where an AI-agent provided recommendations to wear or not protective gear during reconnaissance missions. Trust was evaluated indirectly through compliance (participants followed the AI's recommendation at early stages of the mission) and directly with a subjective scale for attitudinal trust, namely the inventory (DTI) [176]. DTI measures perceptions of users in the AI-agent's abilities, safety promotion, and limitation. The study showed that early human behavior within the mission was a predictor of later compliance and mission success.

Another line of research focuses on improving decision-making by helping users understand the limits of AI through the visualizing of uncertainty in the predictions. This has been accomplished through interaction with the user interface and with the addition of visual cues into the AI output. Daradkeh and AbulHud [55] developed an interactive system that allows users to explore and compare the uncertainty and risk of AI predictions through adjustable bars. While this work highlights the importance of visualizing uncertainty through interactions, it does not delve into examining users' perceptions. Doula *et al.*'s work [56], compared the effect of displaying AI's uncertainty in an augmented reality (AR) environment. In this study, an AI-powered mobile application predicted the locations of sound sources behind walls and participants decided whether or not to follow the AI recommendations. Post-interviews revealed that the majority of participants would trust AI systems more when uncertainty is shown to the user. Marusich *et al.* [57]'s study assesses the utility of well-calibrated uncertainty in decision-making. With an online user study, the authors compared participant's accuracy and confidence in decisions, as well as the accuracy of AI

predictions using visual representations, such as needles and dotplots. Cassenti *et al.* [58]'s study aimed at identify the best ways to represent uncertainty. In an online survey, compliance with the AI recommendations for a convoy to pass or not a risky road were measured. Different representations of verbal and visual uncertainty were presented: text-based with probabilities, with frequencies, and graphical representation using a subjective logic triangle and beta distributions. This study also measured user's perceptions of trust in AI prediction with the Trust in Automation scale [177], a tool developed to measure the level of trust in automated systems.

5.2.3 Decision Making under Uncertainty in Gaming

In the context of gaming, uncertainty information has been studied from different perspectives. In Greis *et al.*'s work [178], the authors designed a web-based game to model risky situations in a farm and used four visual representations of the uncertainty in weather prediction probabilities. In their work, the authors conclude that more information presented on the screen leads people to take unnecessary risks. Alternatively, the gamification of real-life events has been used to explore the effects of uncertainty through simulations of natural disasters. Schueller *et al.* [148] designed three serious games with the objective of understanding how uncertainty in simulated crisis situations impacts the processing of early warnings and subsequent decision-making. The uncertainty information provided during the simulation was used to make predictions about the time and place of a hurricane touching down. Further, uncertainty in raw data has also been considered in the optimization of gaming applications. Jagtap *et al.* [149] designed an uncertainty-based decision support system, where the probability for the selection of the next move in a game increases as the uncertainty of the data is fed as input to the model.

5.3 Materials and Methods

We created an online survey to assess how visualizing AI uncertainty affects individuals with different attitudes towards AI. The study was approved by the Office of Research and Ethics and the Human Research Ethics Committee of our institution and complied with all requirements established by the corresponding governmental agencies overseeing research and ethics.

Prior to describing the study, we define a few key concepts that will remain consistent throughout this study:

- Decision change: a metric indicating whether a participant changes their response when presented with a different visual stimulus (e.g., visualization of uncertainty).
- Trust: the perceived amount of trust in AI solutions. We define solutions as the predictions, recommendations and decisions made by AI-based systems.
- Confidence: the perceived degree of confidence in decisions. Individuals with a higher degree of confidence will find decisions correct or appropriate given the available information.

5.3.1 Research Questions

We were interested in understanding the role of uncertainty visualization in decision-making from the angle of people's attitudes towards AI. As such, this study aimed to answer the following research questions:

- **RQ1:** Does visualization of uncertainty impact decision-making, trust, and confidence among people with different attitudes towards AI?
- **RQ2:** Do attitudes towards AI influence decision-making, trust in AI, and confidence in the decisions made differently?
- **RQ3:** How is the visualization of uncertainty perceived by people when making decisions?

To answer these questions, we developed an online survey where respondents chose the next gameplay move in one of three games (Pac-man, Minesweeper, and Soccer). Participants evaluated a total of 9 sets of gaming scenes with different levels of risk, each with and without uncertainty visualization. The order of games was randomly for each participant. Then, participants rated how visual cues support the perception of uncertainty and their overall experience with the visualization of uncertainty while making decisions. We believed that AI uncertainty information would impact people differently depending on their baseline opinions towards AI, leading to differences in decision-making and trust levels. To assess individual's opinions towards AI, we used the General

Attitudes towards Artificial Intelligence Scale (GAAIS) [158]. The GAAIS scale is used to measure extreme perspectives on AI systems. The original scale identifies two sub-scales: a positive scale including 20 items and a negative scale with 8 items.

Survey Study Design

The survey was designed in Qualtrics [1] and distributed it through Amazon Mechanical Turk [2]. Qualtrics is a system that facilitates data collection through online surveys, while Amazon Mechanical Turk is a crowdsourcing marketplace that allows a distributed workforce to perform virtual tasks. The survey had four parts: (1) Study purpose and consent, (2) a pre-test survey, including demographics and questions pertaining to AI, (3) a testing session with game scenarios visualized with and without uncertainty information, and (4) a post-test questionnaire. Specifically, participants were first informed about the purpose of the study and engaged in a user study where they completed self-reported assessments. To minimize the collection of missing data, we used built-in features on Qualtrics to verify each question was answered before continuing to the next page and set exactly to one answer per question. No time limit was imposed.

Pre-test Questionnaire

We collected information about the demographics of participants (age, gender, current occupation, and level of education), perception of risk as well as their experience with the usage of AI systems and specific arcade games. We also asked questions as to their attitudes towards AI using the General Attitudes towards Artificial Intelligence Scale (GAAIS) [158]. Similar to the work in [159], we used a short version of the GAAIS scale with two questions that highly correlate to each (positive and negative) attitude in the GAAIS scale and one more question, for each attitude, that is representative of the everyday use of AI. We determine the average for each category, classifying participants as having **positive attitudes** if their average positive score exceeds their average reversed negative score, and as having **negative attitudes** if it is the opposite.

[1] https://www.qualtrics.com
[2] Amazon.com, Inc. Bellevue, Washington, United States

Given this categorization, we can clearly distinguishing the impact of AI uncertainty on individuals who are either favorable or unfavorable towards AI. In addition, responses were recorded using a 5-point Likert scale with options in this configuration: left-to-right "strongly disagree; somewhat disagree; neutral; somewhat agree; strongly agree". Table 5.1 shows the questions in this questionnaire.

Table 5.1: Questions in the pre-test questionnaire

Category	**Questions**
AI	I am experienced with the use of AI systems
	AI systems used for decision-making are always accurate
	I am confident about using AI systems in my daily life
GAAIS Positive	There are many beneficial applications of AI
	I would trust an AI investment system with life savings
	AI can have positive impacts on people's well-being
GAAIS Negative	I find AI frightening
	AI might take control of people's lives
	People like me will suffer if AI is used more and more

Gaming Scenarios

We utilized classic games in the testing session as a proxy to show human-AI interaction in decision making. All gaming scenarios were constructed by humans, but we told participants that were generated by an AI-system. This design choice was motivated by previous works that aimed to control the quality of the output as a potential factor in human evaluations [159]. As such, we designed gaming scenarios with situations that motivate players to survive or win the game, while considering "the game's AI" probabilities and uncertainties associated with the opponent's actions, rather than suggestions for player's next gameplay moves. As a pre-condition for establishing trust, we informed participants that predictions are as reliable as the systems typically encountered in everyday life, such as movie recommendations, traffic updates, or weather forecasts.

Sets of gaming scenarios were configured as follows. First, participants assessed a situation within a designated gaming scenario with the AI's prediction shown in the binary format (without uncertainty information). Subsequently, participants decide their next gameplay move. After this, we presented the same gaming scene with uncertainty information visible on the screen and, based on the additional information, we recorded their next gameplay move. To illustrate uncertainty in

the AI's prediction, each scenario used a specific visualization method to signal probabilities. As mentioned above, there are three main categories of visualizing data and their uncertainty: color-oriented approaches, focus-based methods, and geometric mapping (e.g., sketchiness in rendering, distorting line marks) [164]. We used a mix of size, color saturation, and transparency as visual representations for uncertainty, a decision motivated by the work of Guo *et. al* [160](see Figure 5.1), and conducted an in-the-wild study. We selected 3 different classic games to showcase diverse gameplay styles: Pac-Man demands quick reflexes within a dynamic setting, Minesweeper requires logical deduction, and the Soccer game requires intuitive decision-making rather than prior gaming expertise. Figure 5.2, shows examples of three sets of gaming scenarios (Pac-Man, Minesweeper, and Soccer).

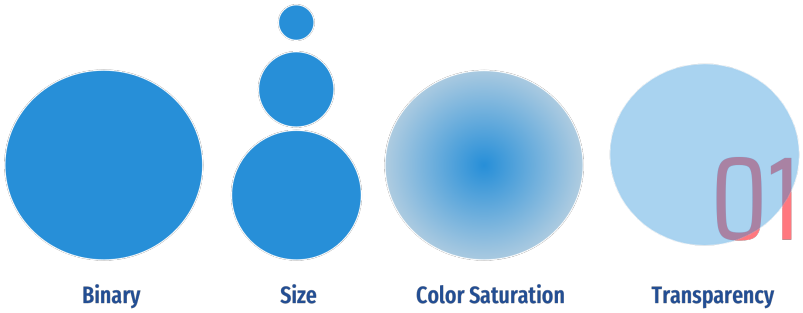

Figure 5.1: A binary visualization gives a model's output with only one label, number, or output. Alternatively, confidence/probability can be depicted with visuals cues (e.g., size, saturation, or transparency) in a non-binary format.

Pac-Man: In this game, the player controls the main character, Pac-Man. The ghosts, which try to kill Pac-Man, are controlled by the "computer". In a binary format, the AI shows the path with the highest probability for a ghost to take. Alternatively, when uncertainty is visualized, the AI presents up to four predictions, showing the range of probabilities using size as a visual cue. The thicker the arrows, the higher the probability of ghosts taking that path. We designed 4 sets of scenarios based on the Pac-Man game. Depending on the grid position, the player is then asked what direction Pac-Man should move in (up, down, left, right).

Minesweeper: In the Minesweeper game, the player needs to identify and avoid the location of mines, which were placed randomly on a static grid-based board. In this version, the "AI agent"

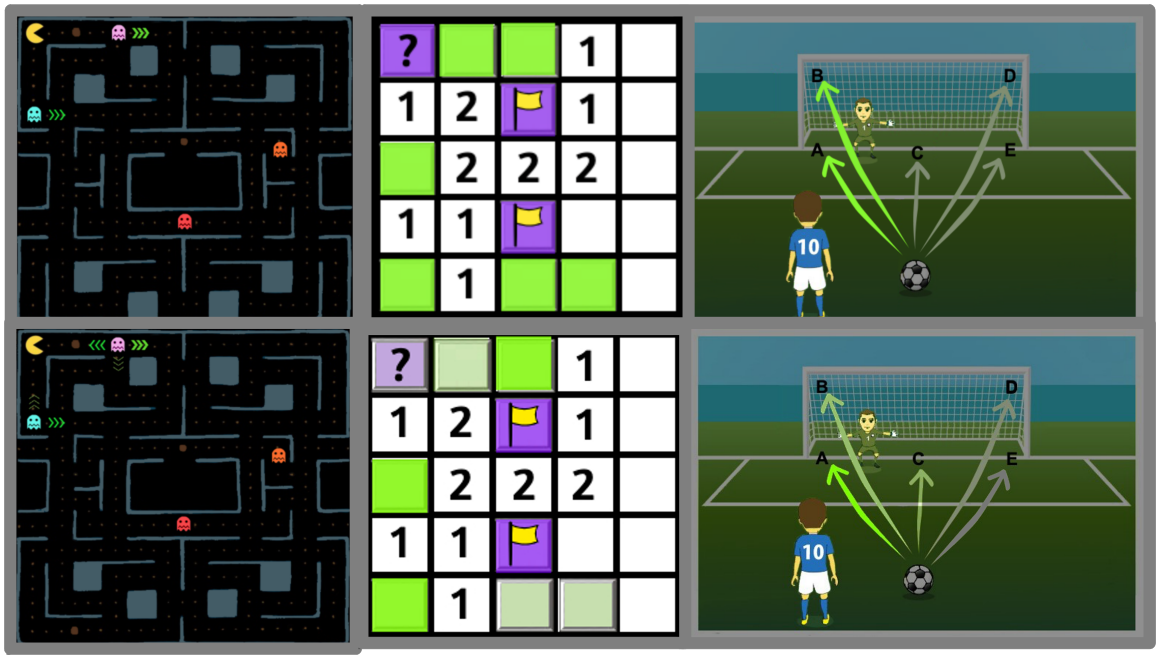

Figure 5.2: Examples of game scenes shown in the testing session. *Top row:* We show the prediction in a binary format; *Bottom row:* We convey uncertainty using different visual representations: size, color saturation, and transparency for Pac-Man, soccer game, and Minesweeper, respectively.

suggests safe and dangerous locations. Specifically, the AI will paint the squares using colors representing the likelihood of the presence or absence of a mine. In a binary format, squares are painted with only two colors representing the probable presence (green) or absence (purple) of a mine. In the uncertainty format, transparency is used as a visual cue to identify the confidence level of said prediction. The more vivid the color is, the more confidence the agent has in its prediction. In contrast, the duller the color is, the less confident the agent is about its prediction. We constructed 2 sets of scenarios.

Soccer Game: In the soccer game, the player is the striker and the computer controls the goalkeeper. The player must select one of the six targets (A-F) to attempt a penalty shot. Here, the "AI agent" monitors the goalkeeper's placement and movement in the net and makes predictions on where to kick the ball to have the highest chance to score. In a baseline case, images are shown with two shades of green, bright paths are more likely to score, while darker paths are less likely to score. Uncertainty is visualized with multiple shades of green. The brighter the path of the color

is, the higher the chances the player will score. Here, color saturation is used as the visual cue. We designed 3 soccer scenarios.

After making their selection for the next gameplay move, participants are required to answer three questions related to their decision. Table 5.2 shows the questions and metrics used in the study: "decision change" is the action of altering a previously made decision based on a different visualization; "trust" relates to the degree of reliability one can possess towards AI systems; "confidence" is the degree of self-assurance in the decision; and "usability", which quantifies whether participants found visual cue informative while performing the task. The questions in this assessment used a 10-point scale.

Table 5.2: Questions measuring decision change, trust in AI, confidence in decisions, and usability of visualization of uncertainty in the testing session.

Category	**Questions**
Decision	*Pac-Man:* in which direction would you move next?
	Minesweeper: consider the square with the question mark, would you mark it as safe?
	Soccer: which target are you shooting at?
Trust	How much trust do you have in this AI prediction?
Confidence	How confident are you with your decision?
Usability	Did you find the visualization of uncertainty informative in this task?

5.3.2 Post-test Questionnaire

Lastly, participants were asked about their opinions on the utility, intuition, and impression of the different visual cues. We also asked them to rank their preferred visualization method. The questions in this part of the assessment used a 10-point scale, which were averaged to compute the final score for each of these questions. Table 5.3 shows the questions in this questionnaire.

5.3.3 Recruitment

Data was compiled between February and June 2023 using Qualtrics [3], a cloud-based application that allows data collection through online surveys. For the dissemination of invitations of

[3]https://www.qualtrics.com

Table 5.3: Questions in the post-test questionnaire

Category	Questions
Visualization of Uncertainty	The visualization of uncertainty was useful when making decisions.
	The visualization of uncertainty was confusing when making decisions
	The visualization of uncertainty made me feel more confident in my decision.
	The visualization of uncertainty helped me take objective decisions.
Visual Representations	How intuitive was size as a visual cue?
	How intuitive was color saturation as a visual cue?
	How intuitive was transparency as a visual cue?
	Rate your preference towards size as a visual cue?
	Rate your preference towards color size as a visual cue?
	Rate your preference towards transparency as a visual cue?

participation, we used our institution's mailing lists and online communication mediums, such as LinkedIn and Twitter. In addition, to get a diverse and random population of users, we recruited workers using Amazon Mechanical Turk. We paid $1.00 USD for participation after the survey completion. JASP 0.17.2.1 software [4] was used to build contingency tables, measure relative frequencies, and report statistical analyses in this study.

5.4 Results

We collected data from 277 participants across the United States, Canada, Mexico, Australia, Turkey, Thailand, France, Poland, Norway, Germany and the United Kingdom. To ensure data quality, we removed responses from 86 Amazon Turk participants because they did not pass our validation checkpoints, suggesting that their responses could not be trusted. These checkpoints consisted of age verification according to the provided birth year and visual attention checks located at different points in the survey. An example of a visual attention checkpoint is the scenario where the Pac-Man is surrounded by ghosts with only one way to escape. We also removed 36 records as these did not meet the minimum completion time of 9 minutes, a threshold we set to exclude possible responses lacking careful consideration. Lastly, we removed 8 responses with missing values. In total, we analyzed responses from 147 participants. We analyzed participants' general characteristics using frequency analysis and descriptive statistics. Table 5.4 summarizes the demographics of

[4] https://jasp-stats.org

Table 5.4: Demographics of participants included in the study.

Characteristic	Quantity	Characteristic	Quantity
Participants	147	Education	
Age		Limited/No schooling completed	3
Min	19	Trade/technical/vocational training	8
Max	69	High school graduate/some college credit	15
Mean	32.3	Bachelor's degree	72
SD	9.9	Master's degree	43
Gender		Doctorate degree	6
Male	77		
Female	69		
Non-binary	1		

the participants.

5.4.1 Decision change

To look at decision change, we computed a binary adjustment score that indicates when a person made a change in their decisions in more than half of the scenarios. With this information, we create a contingency table that allows us to quantify and compare participant's who changed or did not change their decisions among attitudes towards AI.

Table 5.5 shows that 71% of participants hold a positive attitude towards AI (n=104) and 29% a negative attitude towards AI (n=43). Also, we noticed that 33% of all participants adjusted their responses after seeing the uncertainty of the AI prediction, among these 23% with a negative GAAIS attitude and 37% a positive GAAIS attitude. This tendency was expected as it was believed that the majority of people with a positive attitude towards AI would adhere to the AI model's recommendation, even in the absence of supplementary information. To measure the statistical significance of our findings, we analyze the observed frequencies of our binary data and perform a chi-square (χ^2) test. No significant associations were found between participant's attitude towards AI and decision change ($\chi^2(1) = 2.441, p = 0.118$).

Figure 5.3 summarizes the impact of the different types of visualization in decision change. Generally, we observe that size and transparency are the visual methods representing AI uncertainty with small number of changes in decisions, 24% and 11% decision change rate respectively,

Table 5.5: Shows the number of people, for each of the examined attitudes, who changed or not their decisions as a response of the visual uncertainty of the AI predictions.

	GAAIS Attitude		
	Negative	Positive	Total
Change	10	38	48
No Change	33	66	99
Total	43	104	147

compared to 56% found when the method used was color saturation. Figure 5.3 *(left)* shows that 27% individuals with a positive GAAIS attitude and 16% with a negative GAAIS attitude changed their responses with size as the visual cue. Figure 5.3 *(middle)* presents higher rates of decision change with color saturation among people with positive attitude towards AI (58%) and individuals with negative attitude towards AI (51%). With Figure 5.3 *(right)*, we can observe that 13% of participants with positive GAAIS attitude updated their responses and only 5% of the individuals with negative GAAIS attitude also changed their responses after seen the uncertainty with the transparency method. A (χ^2) test showed no significant association between size and decision change, (χ^2) (1) = 1.359, p = 0.243), between color saturation and decision change (χ^2) (1) = 0.294, p = 0.587) and between transparency and decision change (χ^2) (1) = 1.611, p = 0.204). Despite the trend in our results, we did not observe a statistical relationship between the types of representations of uncertainty and changes in decisions among individuals with different opinions towards AI.

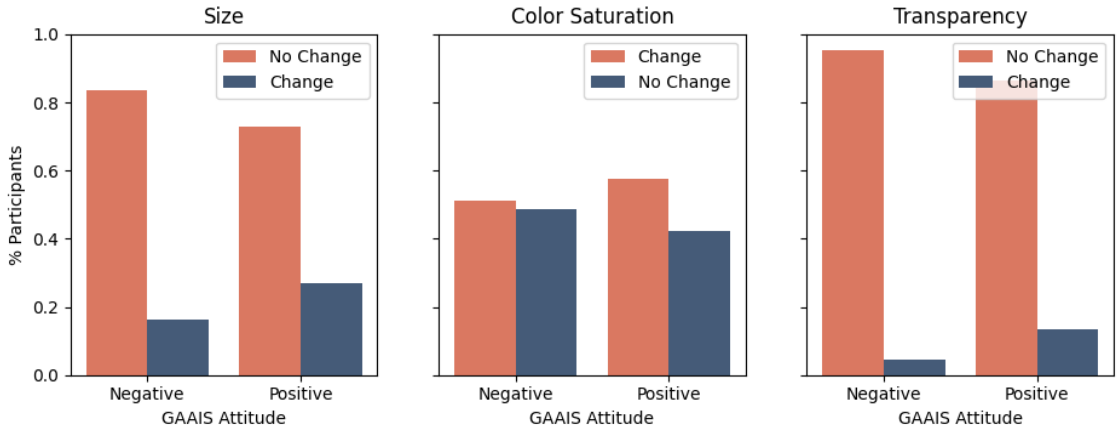

Figure 5.3: Illustrates the impact of the different types of visualization in decision change among GAAIS attitudes.

5.4.2 Trust

To evaluate trust in AI, we compute a trust score, which quantifies the impact strength of trust in AI for each participant based on the average differences observed between the pre- and post-uncertainty conditions.

A contingency table (not shown here) revealed that trust in AI increased in 58% among those with a negative attitude towards AI and in 43% of individuals with a positive attitude towards AI. We also compared trust strengths within groups of GAAIS attitudes. Figure 5.4 *(top)* demonstrates the impact of visualizing the uncertainty of predictions on trust in AI for each GAAIS attitude. Interestingly, we observe a more prominent impact in trust among those negatively inclined towards AI (M = 0.39, Mdn = 0.33, SD = 0.76) compared to participants with a positive GAAIS attitude (M = 0.05, Mdn = 0.0, SD = 0.56). A Welch two-samples t-test showed that trust in AI was significantly reinforced among participants with a negative GAAIS attitude than those with a positive GAAIS attitude, $t(61.291) = 2.651, p < .01$. $Cohen's\ d(0.51)$.

To complement these findings, we further explore the impact of the different visual representations of uncertainty (e.g. size, color saturation, and transparency) on participant's trust in AI. Each point in Figure 5.4 *(bottom)* represents a participant color-coded according to their GAAIS attitude. The point's position indicates whether participant's trust in AI increased (above zero) or decrease (below zero), and the magnitude of this change. We observed a greater impact in trust in AI among individuals with a negative GAAIS attitude when size (M = 1.00, Mdn = 0.00, SD = 3.472) and color saturation (M = 0.374, Mdn = 0.00, SD = 3.046) are used as visual representations, compared to transparency (M = -0.020, Mdn = 0.00, SD = 2.696). An independent one-way ANOVA found a statistically significant main effect, $F(2, 438) = 4.375, p < .001, \omega^2 = 0.214$. Post-hoc testing using Sheffe's correction revealed that size resulted in a greater impact compared to transparency ($p < .05$). However, there were no significant main effects between color saturation and either size ($p = .630$) or transparency ($p = .193$).

We conclude that using size as an indicator of the uncertainty in AI predictions, significantly impacts trust in AI, particularly among participants with a negative attitude towards GAAIS.

5.4.3 Confidence

We quantified the number of individuals whose confidence in their decisions increased when the uncertainty of the AI agent was displayed and found that 42% participants (n = 44) had a positive attitude towards AI, while 49% had a negative attitude towards AI (n = 21). Further, Figure 5.5 (top) illustrates the overall strength of confidence levels in participant's decisions when uncertainty is visualized. We notice a subtle difference between the means of confidence and GAAIS attitudes, where confidence in decisions has a larger impact on participants with negative attitudes towards AI (M = 0.12, Mdn = 0.0, SD = 0.59) than in the other group (M = 0.03, Mdn = 0.0, SD = 0.65). A Mann-Whitney U test was used to evaluate the significance of our findings. The results of the statistical analysis suggest that there is no significant differences between the means of the compared groups, $U = 2392.5, p = 0.505$.

Alternatively, we investigate the impact of the different types of visualization in confidence in people's decisions. Figure 5.5 (bottom) distinguishes differences across visual representations. We observe that confidence in decisions grows weaker (below zero) among participants, regardless of their attitudes towards AI. Specifically, we observe stronger confidence in decisions with size (M = 0.639, Mdn = 0.0, SD = 3.705), and color saturation (M = 0.279, Mdn = 0.0, SD = 3.201), compared to transparency (M = -0.401, Mdn = 0.0, SD = 2.640). An independent one-way ANOVA found a statistically significant main effect ($F(2, 438) = 4.375, p < .001, \omega^2 = 0.214$). Post hoc testing using Sheffe's correction revealed that representations with size resulted in significantly greater impact compared to transparency ($p < .05$). However, no significant differences in the strength of confidence in decision were observed between color saturation and either size ($p = .630$) or transparency ($p = .193$).

We conclude that fluctuations in confidence in decisions may be due to different factors rather than the individual's attitudes towards AI. We suspect that gaming experience may have caused the changes in confidence perceived, however this needs to be further explored. Interestingly, we found that certain visual representations of AI uncertainty can lead to more confidence in people's decisions, as we observe that significant stronger decisions were perceived when the uncertainty was represented with size.

5.4.4 Correlations

Inspired by the experimental design in Liu et al. [159], we constructed two separate regression models and evaluate each of the following dependent variables: changes in decisions, trust in AI and confidence in participant's decisions. The baseline model includes only demographic aspects (e.g. age, education, and gender). Model 2 introduces two supplementary factor, the GAAIS score and gaming experience. We compute the GAAIS score as the average positive and negative (reverse-coded) GAAIS attitudes and gaming experience was obtained from the pre-test questionnaire.

Table 5.6 shows the results of the two logistic regression models assessing the effects of GAAIS and gaming experience scores on the likelihood that participants have a change in decisions. For Model 1, the logistic regression was not statistically significant, $\chi^2(143) = 4.444$, $p = 0.931$. However, Model 2 shows to be statistically significant $\chi^2(141) = 14.313$, $p < .05$. It was also found that holding all other predictor variables constant, the odds of change in decision is higher (odds ratio = 1.859, $p < .01$) for those with more gaming experience when uncertainty is available. These results confirm our previous findings about the lack of association between GAAIS attitudes and changes in decisions. More importantly, they highlight a strong association between gaming experience and decision change.

Table 5.7 provides the results of the regression models predicting trust in AI. Model 1, was not statistically significant $\chi^2(143) = 1.172$, $p = 0.760$. On the contrary, Model 2 was statistically significant $\chi^2(141) = 12.289$, $p < .05$. It was also found that holding all other predictor variables constant, the odds of trust in AI predictions when uncertainty is available was 37% higher (odds ratio = 2.023, $p < 0.01$) for those with strong overall positive opinions towards AI (agree or strongly agree to the positive items and disagree or strongly disagree to the negative items in the GAAIS scale questions). We conclude that there exists a significant relationship between GAAIS scores, which is positively correlated to the increases of perceived trust in AI.

The results of the logistic regression models predicting confidence in participant's decisions (not shown here) found no significant associations between the variables under investigation in either model. Model 1, $\chi^2(143) = 3.659$, $p = 0.301$. and Model 2 $\chi^2(141) = 7.362$, $p = 0.195$. This absence of correlation suggests that the neither gaming experience nor GAAIS attitudes exhibit a

Table 5.6: Coefficients table from two logistic regression models predicting changes in decision after the uncertainty is visualized.

Model	R^2	Variables	Estimate	Standard Error	Odds Ratio	*p*
1	0.008	(Intercept)	−1.230	1.022	0.292	0.229
		Age	0.001	0.018	1.001	0.937
		Education	0.019	0.082	1.019	0.820
		Gender	0.202	0.345	1.224	0.558
2	0.077	(Intercept)	−2.256	1.890	0.105	0.233
		Age	0.006	0.019	1.006	0.748
		Education	−0.014	0.089	0.986	0.879
		Gender	0.385	0.365	1.470	0.292
		GAAIS Score	−0.404	0.380	0.668	0.288
		Gaming Experience	**0.620**	**0.222**	**1.859**	**0.005**

Table 5.7: Coefficients table showing the results of the logistic regression models predicting trust in AI as a result of the uncertainty visualized.

Model	R^2	Variable	Estimate	Standard Error	Odds Ratio	*p*
1	0.006	(Intercept)	−0.760	0.955	0.468	0.426
		Age	−0.001	0.017	0.999	0.964
		Education	0.026	0.076	1.026	0.736
		Gender	0.316	0.326	1.372	0.332
2	0.060	(Intercept)	−5.471	1.870	0.004	0.003
		Age	−0.005	0.018	0.995	0.774
		Education	−0.011	0.081	0.989	0.893
		Gender	0.402	0.348	1.495	0.247
		GAAIS Score	**1.162**	**0.372**	**3.196**	**0.002**
		Gaming Experience	0.307	0.187	1.359	0.102

relationship with the apparent trend in people's confidence in decisions.

Visual Perceptions of AI Uncertainty

To assess how AI uncertainty is perceived among individuals with different attitudes towards AI, we created a utility score based on the post-test questionnaire. This score includes the perception of uncertainty as useful, confusing (reverse-coded), and supportive of both objective and confident decisions.

Figure 5.6 presents the impact of the perceived utility of AI uncertainty among participants. We notice a lower perceived utility for the visualization of uncertainty in participants with a positive attitude towards AI (M = 6.91, Median = 6.875, SD = 1.10) compared to those with a negative attitude towards AI(M = 7.25, Median = 7.25, SD = 1.36). A Mann-Whitney U test showed that

participants with negative GAAIS attitudes perceive the visualization of uncertainty statistically with greater utility ($M = 7.372, Median = 8.0, SD = 2.65$) than people with a positive GAAIS attitude ($M = 5.09, Median = 5.00, SD = 2.56$), $U = 3302.5, p < .001$. Therefore, we conclude that the visualization of uncertainty in AI's predictions can be of greater utility to people with negative GAAIS attitude.

Further, we assessed the perceived value of the different visualization techniques. Table 5.8 presents our findings based on intuition, preference, and the amount of information perceived given the different GAAIS attitudes. We found that color saturation yielded higher intuition and preference; this is followed by transparency and size. We also observe that participants can perceive more information with size, followed by transparency.

We ran three two-way independent ANOVA tests, one for each factor measured. For intuition, we found a significant main effect for the specific visual representation of uncertainty ($F(2, 435) = 8.34, p < .001, \omega^2 = 0.032$). No significant difference was found for GAAIS attitudes, or significant interaction between GAAIS attitudes and visual representations. Scheffe's post hoc correction revealed that intuition was significantly higher in the representation of uncertainty with color saturation compared to size ($t = 4.059, p < .001$). Post hoc testing shows no significant difference between transparency and either color saturation or size.

For preference of representations of uncertainty, there were significant main effects for both GAAIS attitudes ($F(1, 435) = 5.279, p < .05, \omega^2 = 0.009$) and the types of visual representations of uncertainty ($F(2, 435) = 8.191, p < .001, \omega^2 = 0.031$). No significant significant interaction were found between GAAIS attitudes and visual representations. Post hoc testing with Scheffe's correction revealed a statistically significant difference between individuals with positive and negative GAAIS attitudes ($t = -2.298, p < .05$) as well as significant differences between using color saturation and size to represent uncertainty ($t = 4.044, p < .001$). Post hoc testing shows no significant difference between transparency and either color saturation or size.

Lastly, we assessed the level of information perceived with different representations of uncertainty and found a significant main effect for GAAIS Attitude ($F(1, 435) = 3.955, p < .05, \omega^2 = 0.007$). No significant difference was found for visual representations, or significant interactions

between GAAIS attitudes and visual representations. Scheffe's post hoc correction showed the perceived level of information to be significantly higher for individuals with a positive attitude towards AI compared to those with a negative attitude ($t = -1.989.044, p < .05$).

Table 5.8: Shows how intuitive the different representations of uncertainty were perceived.

	GAAIS	Size	Color Saturation	Transaprency
Intuition	Positive	6.35 ± 2.42	$\mathbf{7.00 \pm 2.10}$	6.70 ± 2.24
	Negative	5.65 ± 2.81	$\mathbf{7.42 \pm 2.11}$	6.28 ± 2.40
Preference	Positive	6.41 ± 2.61	$\mathbf{7.20 \pm 1.92}$	6.80 ± 2.32
	Negative	5.37 ± 3.26	$\mathbf{7.17 \pm 2.31}$	6.16 ± 2.61
Informative	Positive	$\mathbf{7.58 \pm 1.82}$	7.36 ± 1.78	7.37 ± 1.71
	Negative	$\mathbf{7.37 \pm 2.12}$	6.83 ± 2.25	6.92 ± 2.15

5.5 Discussion

To highlight the importance of visualizing uncertainty in human-AI collaboration, a few studies have investigated the effects of uncertainty in games and simulations [148, 149] and the humanistic factors that enable the utility and adoption of AI-based technologies [156, 160, 161, 162]. Our paper expands this line of research by examining participants' attitudes toward uncertainty in AI and its impact on the decision-making process. We utilize classic games as a tool to assess the impact of the visual perception of the uncertainty of AI outputs into decision change, trust and confidence in decisions, using simple simple techniques such as size, color saturation and transparency. Table 7.1 summarizes the findings of our study in relation to our research questions.

5.5.1 Decision Change, Trust in AI and Confidence in Decisions

According to our findings, the visualization of uncertainty has a noticeable impact on decision-making, leading at least one-forth of people in each group, regardless of their attitudes towards AI, to make a change. In general, up to 33% of all participants re-evaluated their choices based on the different visual representations. However, with the observed data, we could not verify a significant relationship between these decision changes and individuals' opinions towards AI. The visualization of uncertainty had a significant impact on people's trust in AI. We observed a significant higher impact in trust in AI among people with negative attitude towards AI relative to those with positive

attitude towards AI, when size was used to show AI uncertainty. We also observed an apparent impact on confidence in decisions among people with different attitudes towards AI, but could not confirmed a significant effect. Instead, further analysis revealed that significantly more confident decisions can be made when AI uncertainty is represented by size.

5.5.2 Correlations

Another important finding in our study is the identification of the factors that affect decision-making and trust in AI. We utilized logistic regression algorithms to explore the correlation between demographic information, GAAIS attitudes and gaming experience and each of the variables of interest. Our results indicate that gaming experience is a significant predictor variable for decision change. Increasing gaming experience was positively associated with an increase in decision change when uncertainty is visualized. This suggests that people with higher gaming skills are able to combine the visual information produced by the AI agent and their gaming experience to better recognize behavior patters in the explored domain and make informed decisions. Alternatively, we observed how personal traits (GAAIS attitudes) influence the way trust can be perceived when the uncertainty in AI predictions is evaluated. We found that high levels of positive opinions towards AI were associated with an increase in trust in AI predictions. These findings are reinforced by the idea that trust in human information interaction is influenced by individual characteristics such as memories, assumptions, perceptions, and heuristics of the trusting individuals (first pillar of trust) [179].

5.5.3 Usability

Based on feedback regarding the usability of AI's uncertainty representations in decision-making, we found that visual elements were considered helpful by many, particularly those with a negative GAAIS attitude. We discovered that 31% of participants with a negative GAAIS attitude found uncertainty visualization more useful, relative to the groups of people with positive GAAIS attitude. In addition, we examined the perceived value of different visual method to represent uncertainty based on intuition, preference and the amount of information provided. We found that color saturation was the most significantly intuitive and preferred approach, while size was perceived as providing

the most information about the degree of uncertainty. We believe that the high percentages of utility observed in the study reflects both the potential of representing uncertainty of AI solutions and people's appreciation of the additional information in their decision-making. This motivates us to perform more in-depth explorations of this research within a specialized domain, where both the scenarios and decisions involve higher risk and complexity.

5.5.4 Limitations

A limitation of our work is the use of static gaming scenarios as *toy experiments*. Since, the visual recommendations presented across these gaming scenarios were not generated by AI, a proper calibration of the uncertainty could not be estimated. Despite this, this design choice still allowed us to identify subtle differences between user's perceptions of AI's uncertainty and its effects in decision making. As such, the aforementioned implications might be somewhat limited due to the lack of AI and future research using well-calibrated uncertainty estimates is deemed.

5.5.5 Implications

These findings have multiple implications for AI designers. First, considering human factors, such as visual perception, to communicate the uncertainty of AI predictions can offer individuals more transparent and informative feedback. This encourages informed decision-making throughout their tasks. For example, this may be of particular relevance to serious and health games. AI designers should pay attention to determine what visual representations maximize the impact in the decision-making process and whether or not a combination of visual representations is necessary to achieve a similar effect. Further, designers can leverage the effectiveness of visualization of uncertainty among the different attitudes towards AI to create unique experiences that encourage engagement and satisfaction. For those individuals with positive attitudes towards AI, the use of visual representations to show the uncertainty associate with the AI prediction can increment trust levels. This is important in health-oriented applications. A person can be persuaded to change their exercise habits to reach their wellness goals given a range of possible outcomes and a confidence level accompanying AI's suggestions. On the other hand, to accommodate people who hold negative

attitudes toward AI, designers may need to incorporate additional features to alleviate any skepticism. Lastly, the strong association observed between gaming experience and decision changes highlights the importance of presenting the optimal level of uncertainty to people with different skills levels. By doing this, designers can leverage the adaptive capabilities of AI-based systems to enhance individual's experiences for both expert and non-expert.

It is also important to note that we do not endorse a specific type of visualization technique for communicating uncertainty nor do we intend to convince participants or users to place their trust in AI. Instead, this chapter encourages the evaluation of visual factors into the design of AI systems to alleviate the challenge of reasoning with uncertainty. Moreover, the findings here provide evidence demonstrating the influence of uncertainty visualization on decision-making in everyday situations. With this in mind, we understand that the problem and level of risk described in this study does not compare to those experienced in sensitive domains, as decisions are more complex, and a direct application might not be smooth. However, we hypothesize that the detection of large fluctuations in decision-making, trust, and confidence in low-risk situations suggests that they will be even more prominent in complex and risky decisions.

5.6 Conclusion

Recent improvements in AI have enabled machines with human-level perception capabilities, allowing their acceptance into many domains. However, many of these applications offer solutions that are not always transparent, accurate, or trustworthy. To address these shortcomings, we encourage the adoption of visual methods for a better interpretation of AI uncertainty, which can lead to more informed and grounded decision-making .

This chapter investigates human responses to algorithmic advice throughout the decision-making process in classic games, offering insights into how the visual representation of uncertainty impacts decision-making, trust in AI, and confidence in decisions among individuals with different attitudes towards AI. We demonstrated that the consideration of human factors into the representation of AI outputs impacts trust of people differently, but also leads to different outcomes depending on their experience. These findings motivate designers of decision-making systems to communicate

AI decision information to users (via visualization of uncertainty) and explore effective visual representations that may bring higher impact to perception and cognition. These efforts will result in AI systems and agents that are not only trustworthy but useful.

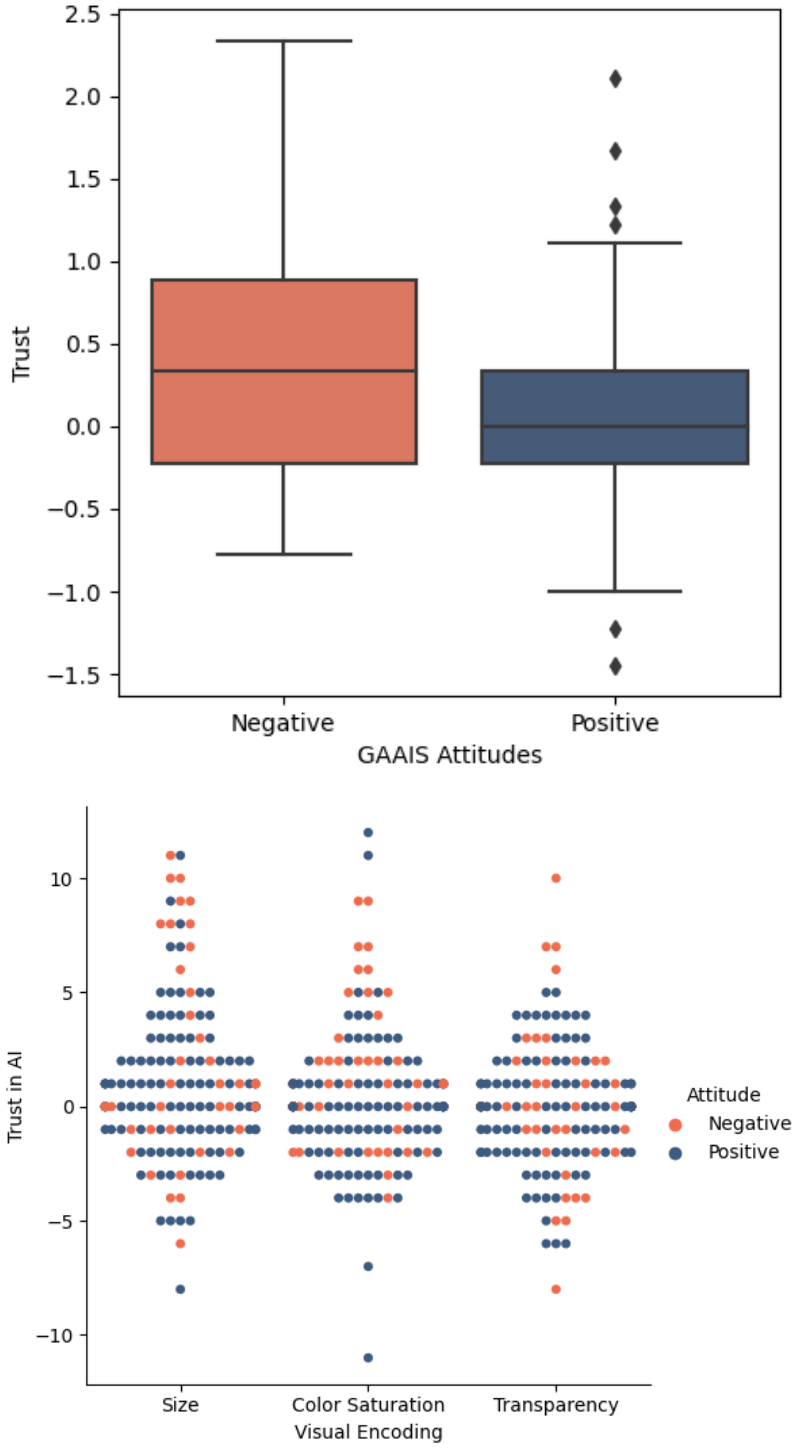

Figure 5.4: Illustrates the impact of visualization of uncertainty on trust in AI according to GAAIS attitudes (top). We show the impact of the different visual cues of uncertainty on participant's trust in AI (bottom).

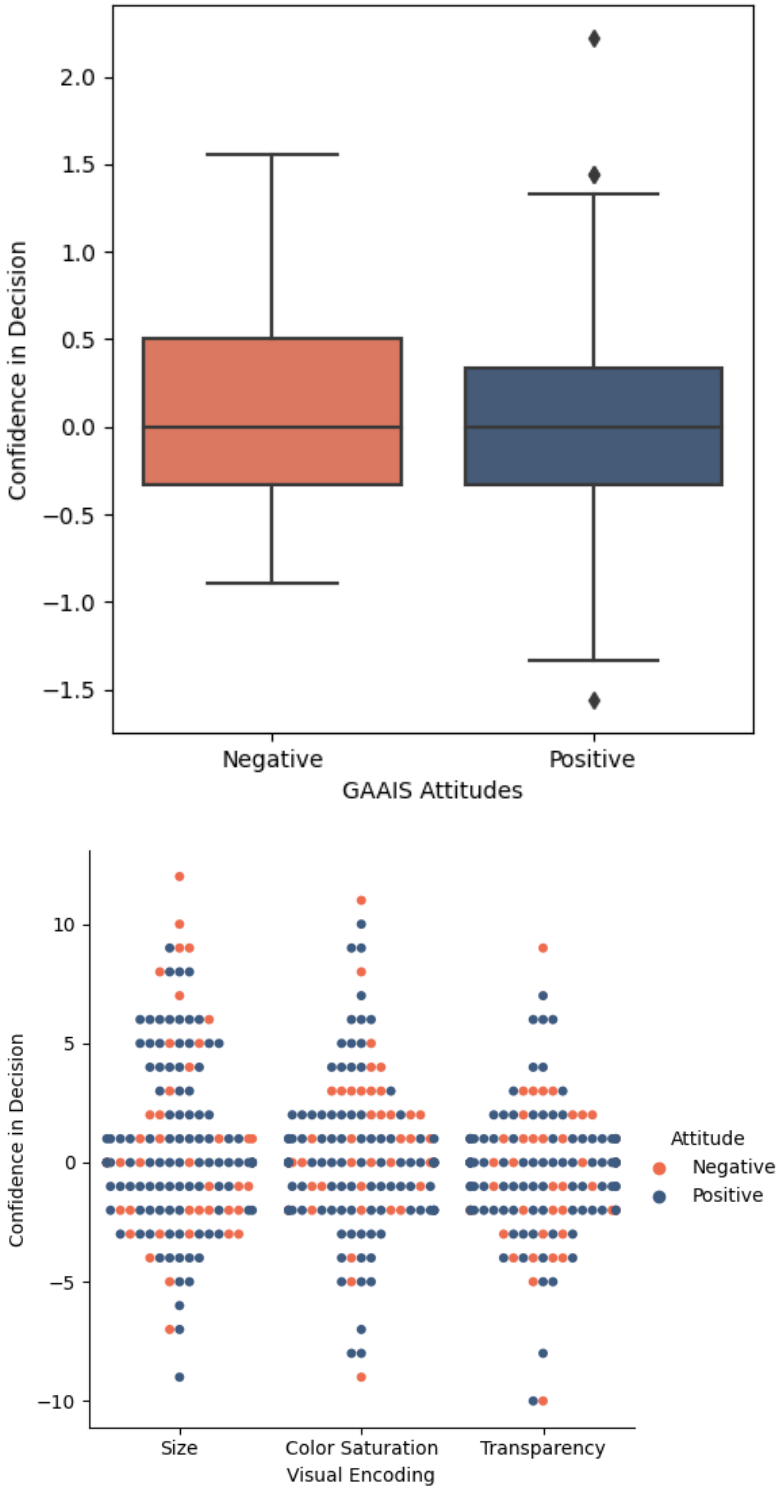

Figure 5.5: Illustrates the impact of visualization of uncertainty on trust in AI according to GAAIS attitudes (top). We show the impact of the different visual cues of uncertainty on participant's trust in AI (bottom).

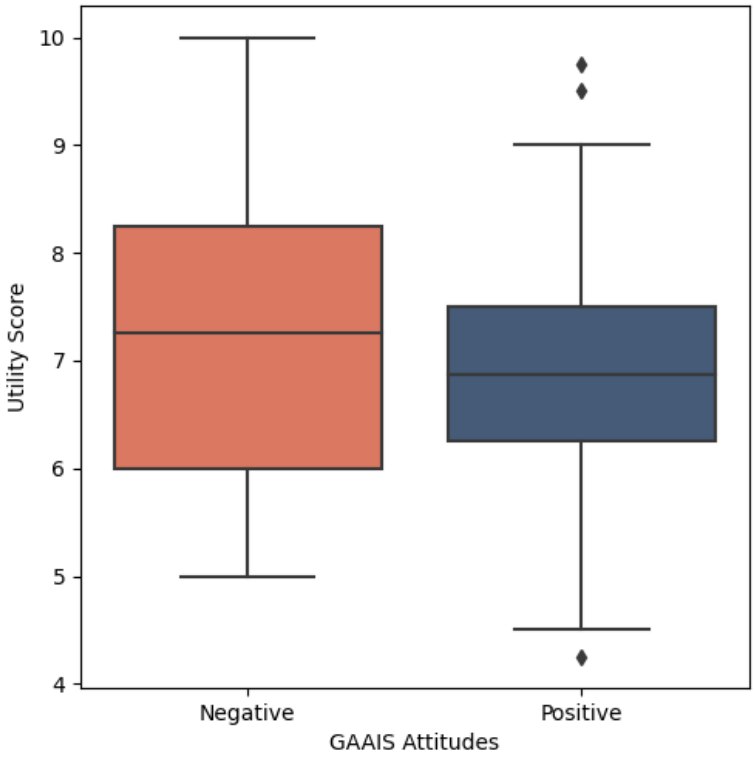

Figure 5.6: Box plot illustrating the average utility score of participants after seen the uncertainty of the model. This score considers how the uncertainty is perceived as useful, confusing (reverse-coded), and supportive of both objective and confident decisions

Chapter 6

Towards Trustworthy Predictions of Alzheimer's Disease under AI Uncertainty

Building on the analysis described in Chapter 5, we extend our investigation to human-AI collaboration in high-risk clinical decision-making. Specifically, we conduct a mixed-methods study with two tasks: (1) identifying factors needed for building trustworthy AI applications and (2) minimizing over-reliance on AI technologies by highlighting their limitations through uncertainty visualization. Both tasks focused on AI-generated predictions of Alzheimer's disease prognosis. In the first task, we used these predictions to elicit user opinions about the utility of AI uncertainty and their trust in the AI's predictions. We also examined whether the amount of information provided to the user about the AI's model development impacted trust. In the second task, we explored which visual method best conveys AI uncertainty to individuals with varying skill levels to determine the optimal representation of uncertainty.

This chapter makes contributions to the field of human–computer interaction (HCI), AI, and Alzheimer's disease research. Our findings revealed that providing information about the AI model development significantly enhances individual's perception of reliability. In addition, we confirmed

that AI uncertainty improves trust in high-stake decision-making and serves as a form of transparency in disease prognosis the evaluations. We also observed that individuals perceive uncertainty better with simple traditional visual methods, such as bar charts. Lastly, an important finding is that individuals tend to be overconfident when unaware of the AI model's uncertainty but start to question the AI's reliability when informed about it. This insight is critical in clinical decision-making and warrants further research.

This chapter was based on a paper that was submitted to the MICCAI UNSURE (June 2024) workshop and is in preparation for a journal paper **Reyes, J.**., Masoumi, M., Batmaz, A., & Kersten-Oertel, M. Towards Trustworthy Predictions of Alzheimer's Disease under AI Uncertainty.

Abstract

Dementia research, coupled with artificial intelligence (AI), has significantly advanced the discovery of disease etiology and biomarker in its early stages, offering the possibility of diagnosis and treatment before symptom onset. However, AI models are prone to biases and inconsistencies during the learning process, resulting in varying degrees of uncertainty in predictions, and ultimately, leading clinicians to automation bias. With a mixed-methods study, we train an AI model with clinical assessments and neuroimaging data from 1,123 patients to assess (1) how the level of detail about the AI model impacts the human-AI decision-making process and (2) how uncertainty impacts decision-making in high-stake decisions, such as Alzheimer's disease prognosis. Our findings indicate that human-AI decision-making process is perceived to be 42% more reliable, based on four facets of trust, when AI uncertainty is expressed in a continuous format rather than a binary format. When comparing representation methods, people showed 13% more trust in the binary (color/no color) than continuous (color saturation) format. These results confirm that information about AI uncertainty improves high-stake decision-making. Our findings suggest that people tend to be overconfident when they are unaware of the model's uncertainty, but they start to question the AI's reliability when they are informed about it.

6.1 Introduction

Artificial intelligence (AI) is increasingly being integrated into Alzheimer's disease (AD) research due to its potential to enhance patient care, prognosis, personalize treatments, and streamline clinical decision-making process [180, 181]. Particularly at earlier stages of the disease, it is vital to assess the degree of neurodegeneration with precision and confidence for timely intervention and treatment planning, potentially leading to the slowing of disease progression. AI-based systems, have gained attention in the dementia research community for their ability to automatically identify abnormal neurodegeneration, such as volume loss in the hippocampus or abnormal amyloid plaques deposits, that are believed to result in AD progression [182]. An automatic detection of pathologic signatures in early stages of the disease could enhance the reliability of the prognosis, facilitating early access to intervention [71]. Several clinical data-driven algorithms have been developed to automatically detect biomarkers for accurate AD diagnosis and prognosis. Typically, these AI-based

methods are trained and validated with longitudinal and cross-sectional data from the Alzheimer's Disease Neuroimaging Initiative (ADNI), the most comprehensive public dataset [183]. ADNI provides information from various modalities, magnetic resonance imaging (MRI), positron emission tomography (PET), cognitive assessments, and cerebrospinal fluid (CSF) testsacross its four versions (ADNI1, ADNI2, ADNI-GO, ADNI3). Given that the nature of clinical data which is highly heterogeneity and complex, AI models face limitations and uncertainties from the training data. As a result, the adoption of these algorithms in clinical practice has been slower compared to other domains. For a review of the contribution of AI-based systems in the classification of MCI and AD using the ADNI dataset, we direct the reader to Tanveer *et al.*'s work [180] and Zhao *et al.*'s work [181]. Although AI models outperform traditional algorithms using single or multiple biomarker modalities, they still face limitations and uncertainties in the input data [71]. Uncertainty, the lack of knowledge about an outcome, is classified into aleatoric (irreducible, noise in training data) and epistemic (due to inadequate or incomplete data) [184]. Popular techniques for quantifying uncertainty in AI include Monte Carlo sampling and Markov chain Monte Carlo. Alternatively, HCI and human-AI communities assess uncertainty by considering human factors such as visual cognition [56, 58]. In AD studies, inherent uncertainty in diagnoses, heterogeneity in pathological signatures, and the imbalance between progressive and non-progressive MCI individuals introduce biases during training, reducing prediction sensitivity and compromising accuracy and reliability [185, 71].

Decision-making based on AI can be challenging due to prediction uncertainty, the lack of transparency in model decisions, or an over-reliance on algorithmic advice. Uncertainty can serve as a proxy for scientific transparency, potentially increasing trust in high-stake decisions such as AD prognosis[186], but also a reduction on the over-reliance and over trust in the AI technology. Visualizing AI uncertainty alongside predictions provides users with an accurate representation of the AI model facilitating the adoption of AI-based tools by clinicians for high-stake decisions, making the statistical power of each prediction more transparent. This may lead to fewer false positives, mitigation of over-reliance and under-utilization of AI systems, and improvements in risk factor assessments.

In this study, we seek to facilitate the adoption of AI technologies in the clinic. To do so, we train Eslami *et al.*'s AI model to generate color-coded images depicting uncertainty using binary

(color/no color) and continuous (color saturation) formats. With these images, we measure individual's visual perceptions of AI uncertainty and its impact in high-stake decision-making, in particular with predictions of AD prognosis. We also measure the impact of different levels of details about the AI model on trust and assess various visual methods to find an optimal representation of AI uncertainty in predictions of AD trajectory. Given a number of AI-generated predictions of AD progression in an online survey, participant determine patient's AD predicted disease stage and describe their level of trust in the AI prediction and own decisions. The main contributions of the paper are as follows:

(1) AD/AI: We extend Eslami's *et al.* [187] qualitative exploration of visual AI uncertainty estimation to enhance algorithmic transparency in the predicting of AD progression, and improve the utility, safety, and reliance of AI solutions.

(2) AD/AI: We conduct an empirical analysis among clinicians and AI experts with the goal of elucidating how decision makers in the AD/AI domains integrate uncertainty as supplementary information into their decision-making processes.

(3) HCI/AI: We explore the relationship between the visualization of AI uncertainty and the processes of trust formation and decision reliance.

6.2 Previous Work

A number of studies have focused on the development of clinical decision support systems aiming at facilitating decision-making under uncertainty. However assessments to identify the most appropriate visual representations for AI uncertainty in high-stake decision-making, especially in Alzheimer's disease research, has not been fully explored. In this section, we present previous research aimed at integrating visual components of AI uncertainty within the areas of AI, human-AI and AD research.

6.2.1 Clinical Decision Support Systems

Clinical decision support systems (CDSS) represent any type of software that is directly involved in clinical decision-making, in which characteristics of individual patients are used to generate patient-specific assessments or recommendations that are then presented to clinicians for consideration [43]. According to Sutton *et al.*'s review article[45], using CDSS can bring multiple benefits at different levels of care (e.g. patient's safety, clinical management, administrative functions, diagnostic and patient decision support), but can also induce an automation bias, which refers to an over-reliance and over-trust in the AI solution.

As more AI applications are developed, the issue of over-reliance is gaining increasing attention. This is particularly significant because automation bias can prevent users from acquiring the skills and expertise that is typically developed with experience [188]. To empirically study this effect, Wysocki *et al.* assessed decision-making for patient admissions based of their likelihood of needing oxygen and the severity of COVID symptoms. The study revealed that respondents took less time to make decisions when some degree of AI explanations were provided, indicating over-reliance on the tool. This over-reliance led to a false impression of correctness, reducing a need to auditing the output [189]. This finding underscores the risk of inappropriate trust in AI systems, highlighting the potential dangers of over-reliance on AI in critical decision-making processes.

6.2.2 Human-AI and Decision-Making

Previous studies have addressed the problem of data visualization and representation of uncertainty in multiple domains. Some research studies offer literature reviews describing the success of integrating interactive visualizations to support decision-making, increase trust in AI models, and describe multiple approaches and issues of uncertainty visualization [190, 191, 192]. A line of research is concentrated on accurately estimating and communicating AI model's reliability to decision-makers, *explainable AI (XAI)*. AI researchers have extended the focus from model explainability to interpretability, in an area of AI called *human-AI*.

Human-AI research aims at augmenting human capabilities by incorporating human factors (e.g. visual perception and cognition) into the design of the AI solution. Generally, the literature shows a

growing interest in transforming uncertainty into trustworthy systems. For example, Chatzimparmpas *et al.* [191] presented a review of 200 papers dating back to 2008, describing how interactive visualization can be used to improve trust in AI models, including a few papers implementing interactive approaches that enabled a visual analysis of uncertainty. Bhatt *et al.* [186] described how uncertainty can serve as a useful form of transparency for decision makers. Zhao *et al.* [193] assessed the impact of uncertainty visualizations on reliance, trust and dependency on AI models. Prabhudesai *et al.* [157] explored perception, reasoning, and judgment of decision-makers with AI predictions displaying their uncertainty associated. Doula *et al.* [56] built an augmented reality (AR) mobile application to measure the impact of AI uncertainty on decision-making. These studies suggest that designing AI solutions with human-AI considerations can provide more comprehensive analysis of AI predictions as they become more usable solutions, where users can understand how predictions are made, interact with the system when new input is added, and decide to trust or not in the recommendation or prediction.

Recent studies have focused on human-AI collaboration to improve CDSS. Kniss [194] suggested the incorporation of understanding to not only highlight regions of interest in the brain, but to also provide a semantic meaning of the area of uncertainty highlighted. Lundström *et al.* [195] integrates probabilistic animations methods to expand the decision support of clinicians through the visualization of uncertainty into medical volumes. Yang *et al.* [196] revealed that model adoption of AI models in clinical settings is often hindered by a lack of perceived need, trust in AI, and insufficient user-centered human-computer interaction (HCI) considerations. More recent works by [197] and [198] provide comprehensive overviews of uncertainty-aware visualization in medical imaging and scientific analysis. They highlight the importance of visualizing uncertainty to improve understanding, communication, and decision-making processes, while also identifying challenges and proposing further research directions to advance the field.

In dementia research, only a limited number of studies have investigated the potential of AI-based systems for the early prediction of the risk of patients converting to AD. Eslami *et al.*'s study [187] stands out as a pioneering work in the visual assessment of the uncertainty of AI outputs for AD research. The authors present an intuitive color-coded visualization system, which integrates multiple biomarker modality information for the prediction of disease trajectory. The proposed AI

model was trained with data from ADNI QT-PAD [183], which is a data freeze subset of the ADNI 1/Go/2 cohorts, and was evaluated by three experts. Interestingly, this system not only presents the predictive outcomes of deep learning models, but also communicates the associated uncertainty levels visually. This approach contributes to the AD research by leveraging the visual depiction of uncertainty to improve the interpretability and reliability of AD predictions.

Lai *et al.* [1] suggested that AI predictions be accompanied by explanations detailing the rationale behind the solutions provided. More specifically, the authors recommended (1) displaying AI uncertainty to convey information about the prediction and (2) expanding details about the training data to provide additional insights into the AI models. We adopted these recommendations in designing our experiments, aiming to further explore visual AI uncertainty estimation to enhance algorithmic transparency in predicting Alzheimer's disease progression, building on Eslami *et al.*'s work [187].

6.3 Materials and Methods

We conducted an online user study to assess level of trust in AI predictions when making high-risk decisions under conditions of uncertainty, given different levels of model information. This study was approved by the Office of Research and Ethics and the Human Research Ethics Committee of our institution.

We aimed to answer the following research questions:

- How does the level of detail about the AI model and AI's uncertainty impact expert's trust in high-stake predictions?
- What visual format of uncertainty enhances interpretability of AI predictions in AD trajectory?

To answer these questions, we developed a web application with two main experimental tasks that presents different levels of detail about the AI model and predictions of AI disease progression. In the first task, participants inspect various AI-generated images of AD trajectories to determine the AD disease state at different clinical examinations. This is followed by questions about their

trust in the AI's decision making the process and confidence in their decisions. In the second task, participants assess different visual representations of uncertainty, provide feedback and rank their preference. We hypothesized that varying levels of information about how the model was trained and its uncertainty would impact individuals differently, resulting in differences in trust. Thus, we formed the following hypotheses:

- (H1) Individuals presented with more detailed information about the AI model will exhibit higher levels of trust in the AI's prediction compared to those given less information,
- (H2) Individuals exposed to more transparent AI systems, based on uncertainty, will trust the predictions more than those who receive less information about the model.
- (H3) Visual methods that continuously express AI uncertainty throughout the output are perceived as more interpretable and trustworthy than methods that summarize the uncertainty..

6.3.1 Data and model setup

We utilized patient data from ADNI QT-PAD. This database includes information from MRI/PET imaging, CSF biomarkers, genetic risk factors, and clinical assessments. Data can be obtained from the "Test Data/Data for Challenges" section of the LONI website [1]. We processed longitudinal evaluations from 1,123 participants in the ADNI QT-PAD dataset. To make the data ready for AI, we pre-processed, filtered, normalized, imputed, trained and evaluated clinical data as in Eslami *et al.* [187]'s paper. In addition, we implemented an instance of the *Machine Learning for Visualizing AD (ML4VisAD)* model to generate color-coded predictions, representing disease progression. The architecture of the ML4VisAD model is available in a public repository [2].

6.3.2 AI Model Output / Stimuli

Images (23x23 pixels) representing AD progression are color-coded every 5 pixels, indicating the diagnosis at the time of a each of the visits (e.g., green for cognitively normal patients, blue for mild cognitive impairment, and red for Alzheimer's disease). Images with a single color across

[1] ADNI QT-PAD data was downloaded on March 2024 from https://ida.loni.usc.edu. Additional information about the QT-PAD Challenge can be found at www.pi4cs.org/qt-pad-challenge

[2] https://github.com/mohaEs/ML4VisAD

all visits, represent patients with stable cognitive status (e.g., stable CN, stable MCI, and stable AD as shown in the first three images left-to-right in Figure 6.4), while images with two or three colors, indicate progression or conversion stages of the disease (4th image in Figure 6.4). In both cases, AI uncertainty is depicted in a binary format, either with color or no color (on the last 3 pixels of the image). When these images are processed by the ML4VisAD model, it generated a similar representation of the input image, showing the degree of uncertainty in the prediction. In this case, the uncertainty is expressed continuously across each visit (e.g., color saturation as shown in Figure 6.4 far right), where darker colors indicate greater uncertainty and more vivid colors indicate higher certainty in the prediction.

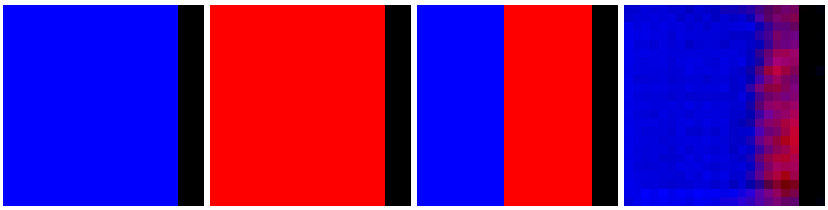

Figure 6.1: Examples of stable cognitive status across examinations (three images to the left) and converting stage (fourth image). These four images present uncertainty in a binary format (color/no color). At the far right, an example of converting stage with uncertainty expressed throughout the image using a continuous format (color saturation).

6.4 User-study design

We set up a user study where we used an explanatory sequential design, a variation of a mixed-methods design [199]. The first part of the user study consists of a consent form, pre-test questionnaire, two tasks, and a post-test questionnaire. This is followed by a survey designed to understand the rationale behind the responses provided during both tasks. Figure 6.2 shows an overview of our user study.

Pre-test questionnaire: We collected demographic data, including age, gender, level of education, current occupation and area of work/research. In addition, with a 10-point scale, we record the user's knowledge of AI development/use, data visualization, and Alzheimer's Disease.

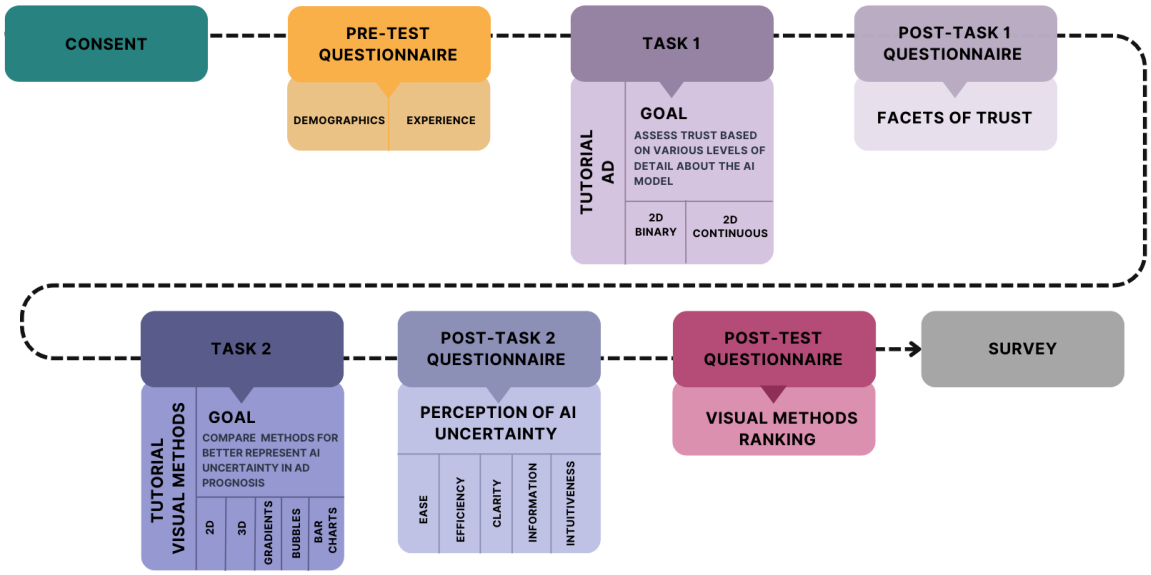

Figure 6.2: Overview of our mixed-methods study.

6.4.1 Task 1

We aimed to assess changes in trust given different levels of detail about the AI model and varying amounts of uncertainty in the model. To do so, we designed situations that required participants to determine state progression of AD based on the predictions provided by an AI system. With a fixed threshold, we categorized predictions based on the level of uncertainty computed per image. Uncertainty was quantified based on the accumulated L component of each image in a L-a-b format. In CIELAB space, this format expresses color space after a color-opponent theory, L-channel refers to lightness normalized from zero to one, and a* and b* are chromaticity coordinates, where a* represents the red/green, and b* the yellow/blue [200]. Figure 6.4 shows an example of predictions under each category: low uncertainty (a narrow degree of uncertainty between 0% and 10%), medium (a moderate level of uncertainty from 11% to 20%), and high (a wide amount of uncertainty in the prediction above 21%). This categorization is motivated by the clinical practice of conducting pre-screenings, where the computed degree of uncertainty determines whether expert intervention is required for further inspection. We selected 3 predictions from each category. Figure 6.3 illustrates a diagram of AI assistance elements considered in our study.

While all participants assessed the same visual stimulus, they were randomly presented with varying levels of detail about the AI model. A summary of the level of information provided is

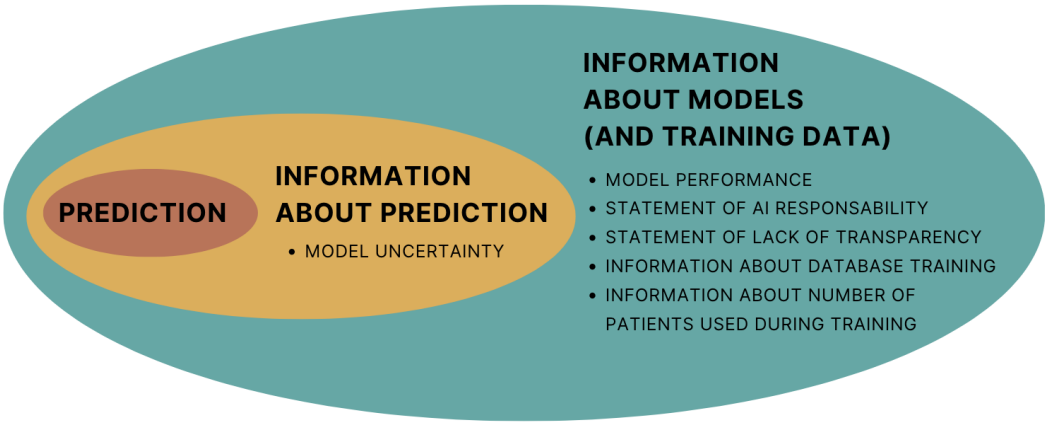

Figure 6.3: Diagram of AI assistance elements inspired by Lai *et al.* [1].

below.

- *Minimum Detail*: No details about the AI model are given.
- *Moderate Detail*: AI model details given include the dataset name, the number of subjects used to train the model, patient inclusion and exclusion criteria based on diagnosis at the first visit, and details about the model's hyper-parameters are available upon request.

The task followed a blocked within-subjects design to present two conditions of AI uncertainty. The first condition included 6 predictions with uncertainty in a binary format and 6 predictions with the uncertainty in a continuous format. Each prediction was followed by three questions: (1) At which evaluation would you consider this patient to have transitioned to another state in Alzheimer's disease (6, 12, 24-month visit or stable status)? (2) How confident are you in your decision? and (3) How much do you trust the human-AI decision-making process? Questions 2 and 3 were based on a 10-point scale. We randomized the predictions at the block-level to avoid bias towards a specific condition.

After each block, we measure overall trustworthiness at the condition-level by using an adapted scale from Ashoori and Weisz's paper [2]. This allows us to explore the different facets of trust about the evaluated human-AI decision-making process either with a binary or continuous format. We ignored the facet "understandability" as it showed poor reliability in the original study. The 4 facets of trust and questions we asked participants to rate with a 10-point scale are:

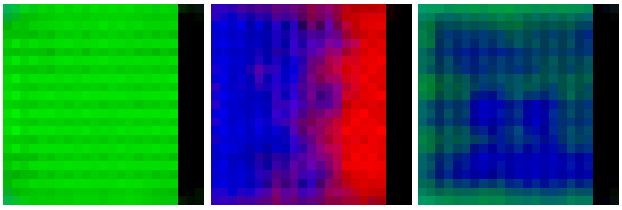

Figure 6.4: From left to right, examples of images with low, medium, and high uncertainty.

- *Trustworthiness*: 1) I believe the decision-making process is trustworthy (i.e. the model's outputs or predictions are reliable, consistent, and can be depended upon) and 2) I need more information about how the AI model was trained and tested in order to trust the design making process.
- *Reliability:* The visualization of the predictions of the system effectively acknowledges the AI model's limitations.
- *Technical Competence:* I can clearly identify the limitations of the AI model.
- *Personal Attachment:* This decision-making process can be integrated into my own research/clinical practice.

6.4.2 Task 2

The aim of this task is to explore different visual formats to represent AI uncertainty for predictions of AD disease trajectory. We use a blocked within-subjects design for this task. Each block consisted of 3 images that were selected based on their degree of uncertainty (low, medium, and high). The uncertainty was encoded using 5 different visual formats, as described below. We considered it important to preserve the original color-coding mapping to each clinical diagnosis. An example of each is shown in Figure 6.5. Following each visual rendering, we asked participants to answer three questions, similar as in Task 1: 1) At which evaluation would you consider this patient to have transitioned to another state in Alzheimer's disease (6, 12, 24-month visit or stable status)? 2) How confident are you in your decision? and 3) How much do you trust the human-AI decision-making process? At the end of each block, we asked participants to provide feedback about their ability to perceive uncertainty each of the visual format based on how easy, efficient, clear, intuitive

and informative they think the predictions are in that block. We randomized the blocks so that each participant evaluates a different visual formats in different order to avoid presentation bias.

- *2D (stimuli in Task 1):* We considered the 23x23 prediction from the DLML4VisAD model, transformed in the L-ab-format. The amount of information per pixel from the L-component is used to calculate the certainty and uncertainty of each image. In this case, if *L* is the normalized sum of the L-component across all pixels representing certainty in the image, the remaining (*1-L*) is used to represent its uncertainty.

- *3D model:* The 3D representation from the 2D image above, based on the L-component in the L-a-b format. The web app enabled users to interact with the 3D model through zoom, pan, and translations.

- *Bubbles*: We built 2d bubbles in JavaScript with the package highcharts [3]. The color of the bubble indicates the diagnosis at the time of the visit and the size of the bubble determines the amount of uncertainty estimated per visit. The bigger the bubble, the more uncertainty accumulated across those pixels representing a specific visit.

- *Bar chart*: A 2D bar generated in JQuery using the package skill.bars [4]. With the total certainty estimated from the L-a-b format, we create each bar per visit. In addition, to uncertainty of the prediction is colored with gray.

- *Gradients:* To generate gradient images for each visit, we begin by analyzing the 3D model and assign specific colors to individual pixels. Then, we group the pixels based on their assigned colors and calculate the average certainty and uncertainty within each group. At this point, we identify the number of color groups in each visit and create a corresponding colored gradient image. With a single dot, we represent the maximum likelihood for the prediction based on the average certainty.

[3] https://www.highcharts.com/demo/highcharts/bubble
[4] https://github.com/umarwebdeveloper/jquery-css-skills-bar

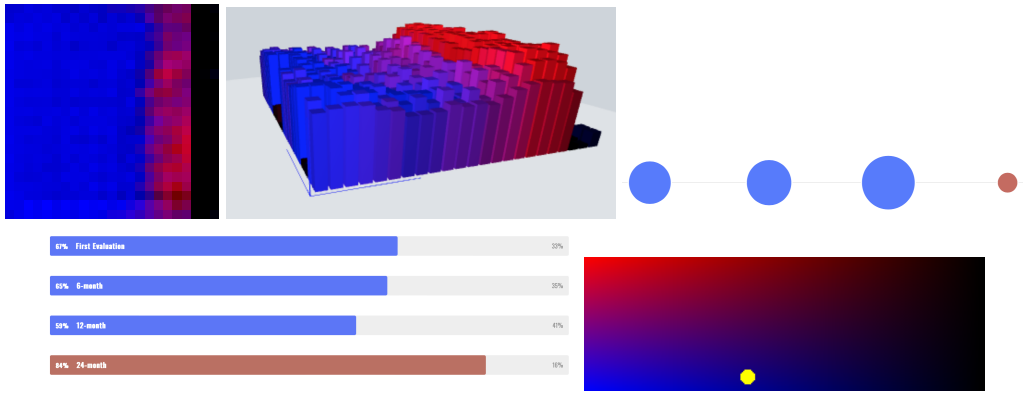

Figure 6.5: Visual formats used in Task 2. (top) 2D, 3D and bubbles, and (bottom) bars and gradients.

Post-test questionnaire and Survey:

At the end of both tasks, we ask participants to rank their preferred visual format and describe the reasons behind the selection of the most and least favorite visualization methods. This was followed by a survey, where participants completed a few open-ended questions: (1) How did the visualization of uncertainty influence your decision-making, if at all? and (2) What factors influenced your perception of uncertainty and trust in the AI predictions?

Recruitment, data collection and analysis:

We recruited participants with varying levels familiarity with Alzheimer's disease, and experience with data visualization and AI. Data was collected in June 2024 using a web app built in our laboratory [5]. For the dissemination of invitations of participation, we used our institution's mailing lists and online communication mediums such as LinkedIn and X (formerly Twitter). Although the survey was designed to be completed online, we invited participants to the lab to conduct it in a Think-Aloud manner, allowing us to answer any questions they had about the test. JASP [6] 0.17.2.1 was used to report statistical analyses.

[5] [link removed to preserve anonymity]
[6] https://jasp-stats.org

6.5 Results

We recruited 37 participants and analyzed their general characteristics using frequency analysis and descriptive statistics. The average participant age was 27.8 years (range: 21-38), with 59% male and 41% female. We recruited clinical practitioners and graduate students specializing in fields such as VR/AR, HCI, ML/DL, neuroimaging, psychology, computer vision, photonics, and formal methods. We also assessed their experience levels on a scale of 1 (no experience) to 10 (very experienced) in AI development (M=6.1, SD=2.63), AI usage (M=7.3, SD=1.77), data visualization (M=6.1, SD=2.09), and Alzheimer's Disease (M=3.08, SD=2.33).

6.5.1 Task 1

We report our findings about whether providing different levels of detail about the AI model will impact trust in the AI prediction. Figure 6.6 shows how the two levels of information about the AI model affect trust. We observed higher trust scores with a moderate level of information (M = 6.99, Mdn = 7.08, SD = 1.64) than with less information(M = 6.44, Mdn = 6.58, SD = 1.66). An independent t-test shows no significant effect for the level on information provided. This suggests that the level of information provided triggers certain confidence in the prediction, however the amount of information seems not to be appropriate or enough to cause a significant effect. Therefore, we cannot accept H1. To further understand this observation, we evaluated participant's opinions collected at the end of the user study. One participate expressed their need for more information about the design of the output: "I'd trust the predictions more if I know more about how the predictions are produced and visualized". Other participants, who were presented with the minimum amount of details, described the type of information that would have been helpful to reinforce trust: "I would have liked to know more about the model and subject data... it is hard to trust a model I know nothing about. " and "I just trust that the colours are informative for the AI prediction not knowing anything about the data, training and testing of the AI was not helpful."

We also assess how the uncertainty of the AI model affects the four facets of trust. Figure 6.7 shows that trust is generally perceived higher with uncertainty represented in a continuous visual format compared to binary, except in Personal Attachment. The highest increase in Reliability

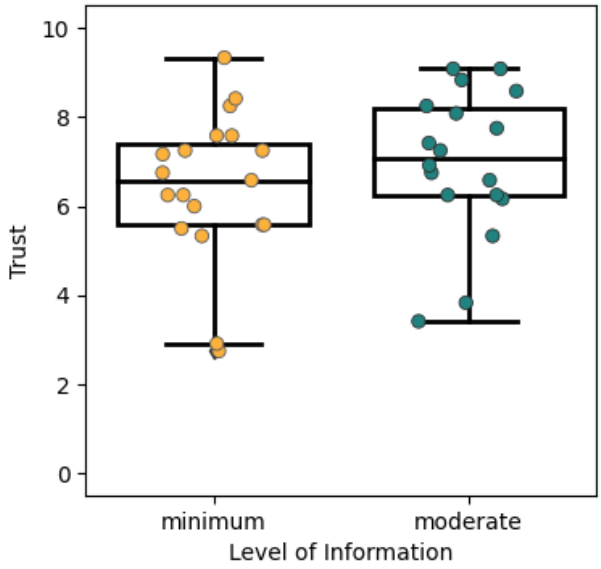

Figure 6.6: Comparing the impact of providing various levels of information on the human-AI's decision-making process.

(42%) was detected with a continuous format (M=6.919, Mdn=7.0, SD=1.991), compared to binary (M=4.865, Mdn=5.0, SD=2.830). Paired sample t-tests confirmed that there was a significant increase in the reliability of predictions with a continuous format over the binary format, $t(36) = -3.768, p < .001$. No other significant effects were observed. A qualitative analysis on participant's opinions provide further insight about the fluctuations of perceived trust in the human-AI decision-making process. One participant commented on model reliance: "If there is no uncertainty visualized, it is easy to be overly confident in the model. The more it is visualized, the more I can be trustworthy of the decisions". Another participant shared their opinions about the use of binary representations: "It caused me to think more thoroughly. Binary representation is very uninformative and is hard to trust."

Further, we analyzed individual's perception of overall trust after seeing the output images generated by the AI model. In Figure 6.8, we noticed that participants trust (13% more) AI predictions when the uncertainty is represented in a binary format (M = 7.19, Mdn = 7.33, SD = 2.52) rather than in a continuous format (M = 6.22, Mdn = 6.17, SD = 1.36). A paired sample t-test showed that the continuous format of AI uncertainty significantly decreases trust scores ($t(36) = 2.550, p < .05$)

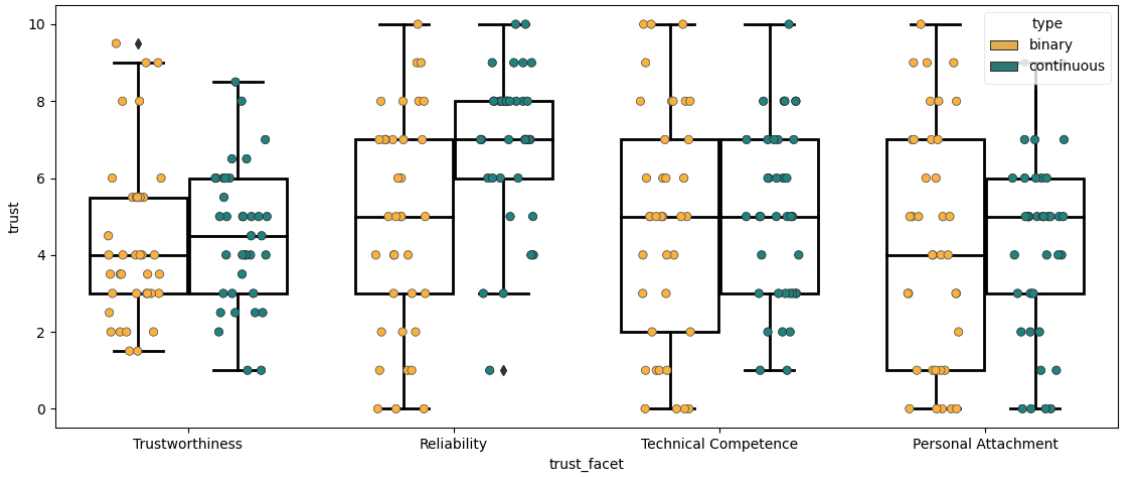

Figure 6.7: Multi-dimensional concept of trust facets according to Ashoori and Weisz [2].

compared to the binary representation. When looking at the qualitative results, we found that participants negatively perceived the use of color saturation, "Factors like blurred transition due to uncertainty being plotted, despite providing more information, made the decision boundaries vague and reduced my confidence.", "How noisy the [visual] data was affected my trust in the AI model making a meaningful prediction."

These results suggest that the representation of AI uncertainty in a 23x23 image using color saturation as a medium to represent uncertainty in a continuous format raises more concerns about the reliability of the prediction. Given that we observe more significant reliable predictions with AI's predictions shown in a continuous format, we partially accept H2.

6.5.2 Task 2

We present our findings on using various visual formats to represent AI uncertainty. We were motivated to separate this analysis based on participants' experience in data visualization due to the multiple studies that describe difficulties for individuals to interpret uncertainty, including domain experts [201]. Owing to this, we set a fixed threshold to classify participants based on their level of experience in data visualization provided in the pre-test questionnaire, experienced individuals (experience greater than 5) and inexperienced otherwise.

Figure 6.9 (top) shows how inexperienced participants in data visualization perceive trust across

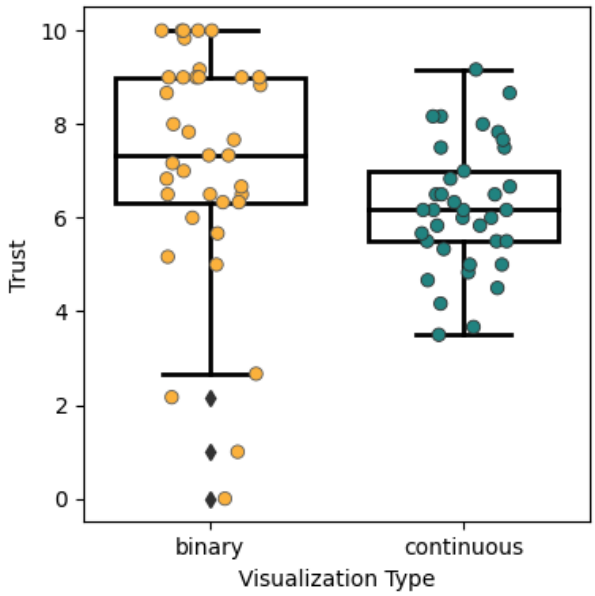

Figure 6.8: Comparing the effect of two visual methods for AI uncertainty on trust, while using a 23x23 image as stimulus.

different visual formats. We noticed that these participants trust more in the AI prediction with 3D models (M = 7.1, Mdn = 7.0, SD = 2.25), followed by gradients (M = 6.95, Mdn = 7.0, SD = 1.84), bar charts (M = 6.95, Mdn = 6.0, SD = 2.03), 2D (M = 5.87, Mdn = 6.33, SD = 2.17) and bubbles (M = 4.9, Mdn = 5.0, SD = 2.03). An independent one-way ANOVA, which showed a significant effect on the type of visual format of AI uncertainty on trust ($F(4, 60) = 2.671$, $p = .041$), however, no statistical significance between the groups was found with a post hoc test. On the contrary, Figure 6.9 (bottom) shows that experienced participants trust more in the AI prediction with bar charts (M = 7.25, Mdn = 7.33, SD = 1.59), followed by 2D (M = 6.8, Mdn = 6.67, SD = 1.4), gradients (M = 6.78, Mdn = 7.0, SD = 2.02), 3D model (M = 6.62, Mdn = 7.33, SD = 2.09) and bubbles (M = 6.19, Mdn = 6.67, SD = 2.12). An independent one-way ANOVA did not report a statistical significance. These results suggest that trust in AI can increase among inexperienced people in data visualization, significantly with 3D models rather than with bubbles; however these differences are not strong or consistent enough to be detected by more stringent post-hoc tests. On the other hand, we found that the type of visualization method does not benefit nor harm an individual's trust in AI for those with more experience in data visualization.

We also investigate the preferred visual method to represent AI uncertainty. In Figure 6.10 (top),

we observe that the majority of individuals with less experience in data visualization preferred the 2D method (N = 9, 47%), followed by 3D (N = 5, 26%), bubbles (N = 3, 16%), and bar charts (N = 2, 11%). No one selected gradients as their first option. The least preferred methods were bubbles (N = 8, 42%), 3D (N = 5, 26%), gradients (N = 4, 21%), and 2D (N = 2, 11%). We ran a Friedman's test to validate the results, but no significant effect was found. Alternatively, Figure 6.10 (bottom) presents how people experienced in data visualization ranked visual methods according to their preference. We noticed that the majority of people in this group preferred 2D (N = 8, 33%) and bar charts (N = 8, 33%) alike to represent AI uncertainty, followed by gradients (N = 4, 17%), and 3D (N = 4, 8%) and bubbles (N = 4, 8%). The least preferred methods were bubbles (N = 10, 42%), 3D (N = 6, 25%), gradients (N = 5, 21%), 2D (N = 2, 8%), and bar charts (N = 1, 4%). A second Friedman's test did not find a significant effect. Our findings show that participants tend to prefer visual methods that express AI uncertainty continuously throughout its output, such as 2D, regardless of an individuals experience with data visualization. However, we could not confirm a statistical effect in our analysis.

To investigate the reason behind participant's preferences in visual methods, we explore the ability to perceive AI uncertainty based on the following factors: ease, efficiency, clarity, intuitiveness, and informativeness. Figure 6.11 (top) shows how inexperienced participants in data visualization perceive each of these factors. We observe that individuals in this group can perceive uncertainty more easily with bar charts (M = 5.77, Mdn = 6.0, SD = 1.36), more efficiently with bar charts (M = 5.38, Mdn = 6.0, SD = 1.61), clearer with gradients (M = 5.38, Mdn = 5.0, SD = 1.26), more intuitively and informatively with 3D models (M = 5.92, Mdn = 6.0, SD = 1.61) and (M = 5.38, Mdn = 6.0, SD = 1.04), respectively. We can also observe and compare those visualizations that rated poorly on each of these factors. We noticed that inexperienced participants perceive less information about the uncertainty of AI with bubbles, across all factors. A two-way independent ANOVA was used to examine the effect of the type of visual format and the ability of inexperienced individuals in data visualization to perceive uncertainty on trust. There was a significant main effect for on the type of visual format used ($F(4, 305) = 4.560$, $p=.001$). A post hoc test did not detect a statistical significant effect among the groups.

Figure 6.11 (bottom) shows that experienced individuals in data visualization can understand

uncertainty more easily with bar charts (M = 5.04, Mdn = 6.0, SD = 1.99), more efficiently with bar charts (M = 4.78, Mdn = 6.0, SD = 2.15), clearer with bar charts (M = 4.74, Mdn = 5.0, SD = 1.98), more intuitively with bar charts (M = 4.70, Mdn = 5.0, SD = 1.82) and more informatively with 2D (M = 5.13, Mdn = 5.0, SD = 1.49). Notably, the visual method using 3D and bubbles was rated low as it was perceived to provide less AI uncertainty information among those experienced participants in data visualization, along with the visual format using gradients on the rest of the factors for this group. Despite these apparent effects, a two-way independent ANOVA did not report a significant difference. It should also be noted that none of the visualization methods tested performed particularly well, receiving fairly neutral ratings (4-6 out of 10) across all aspects.

We conclude that the visualization of AI uncertainty is better perceived with simple representations by the group of individuals with higher experience in data visualization, but more elaborated visual outputs seem to have the same effect on the group of people with less experience in data visualization, which lead us to partially accept H3.

6.6 Discussion

Building on Eslami *et al.* [187]'s work on AI uncertainty estimations in Alzheimer's disease prognosis, we conducted a mixed-methods study to measure trust in AI predictions using binary (color/no color) and continuous (color saturation) formats and assessed multiple visual formats to represent AI uncertainty. We also surveyed participants to understand the reasons that drove their choices. While some studies explore AI uncertainty in relation to user trust using Mayer *et al.*'s dimensions of trust [174], our research goes further by assessing the multi-dimensional concept of trust, according to Ashoori and Weisz [2] from the angle of human-AI decision-making processes. In addition, our focus on AD predictions under uncertainty is a unique contribution to the field.

According to our findings the level of information about the AI model and AI's uncertainty makes a difference in trust among our participants, though it was not significant. However, we posit that including more detailed references about model training and evaluation could make a significant difference and plan to explore this in future work. Moreover, the results from the facets of trust revealed that individuals reflected on both the benefits and limitations of each condition with

an inclination to more transparent representations of AI.

Finding the balance between confidence and caution is crucial in clinical decision-making, where over-reliance on AI without recognizing its limitations can have serious consequences. Our findings indicate that representing AI uncertainty in a binary format enhances perceived reliability of AI predictions, while continuous representations may introduce ambiguity and reduce user confidence. In addition to our results, participant's comments suggest that visualizing uncertainty may help users avoid overconfidence in AI predictions.

When analyzing participants' preferred visual methods, we found that experienced individuals in data visualization consistently preferred 2D and bar charts for trusting, representing, and interpret AI uncertainty. However, this effect was not consistent among inexperienced individuals in data visualization. While this group did prefer the 2D method for visualizing AI uncertainty, they did not find it to be the most understandable nor trustworthy. Instead, they reported that bar charts, gradients, and 3D visuals enhanced their trust and ability to perceive AI uncertainty across the five factors we evaluated. This inconsistency suggests that familiarity with the 2D method from Task 1, or the interactions allowed onto the 3D model may have biased their preferences. Bubbles ranked the lowest across all measurements on trust, preference, and their ability to perceive AI uncertainty.

This study has several limitations. First, our responses are biased as most respondents are experienced in AI with limited knowledge of AD. Second, allowing participants to adjust their responses could bias results. For example, when uncertainty in a single sample is high, participants may adjust other responses as a result of that single experience. Third, the amount of information about patients and models as well as the visual representations of AI uncertainty were limited. Fourth, all visual methods in Task 2 are static expect the 3D model, which allows interactions (e.g. zoom, pan, translation). This interactivity may have introduced an unintended bias towards the 3D visualization, potentially causing inexperienced participants to vary in their perceptions of AI uncertainty, trust, and preferences.

Future research will involve collecting more responses from participants, in particular clinical experts, and incorporating additional patient and AI model information to address previous limitations. During our analysis of perceived trust among participants inexperienced in data visualization, we found a significant effect using ANOVA. However, the lack of significance in post hoc tests

suggests the need to increase sample sizes in each group to detect these differences. Therefore, recruiting more participants will be a focus of our future work.

6.7 Conclusion

Our study examines perceptions of AI predictions under uncertainty in AD prognosis. Through a user study, we assessed the impact of varying levels of AI model detail about and uncertainty on trust. While different levels of information did not significantly affect trust, participants perceived predictions as more reliable when uncertainty was represented in the output. In summary, AI uncertainty enhances trust in predictions, potentially leading to more trustworthy and interpretable AI solutions and greater adoption of AI technologies as high-stake decision support systems.

These findings highlight the importance of carefully balancing the amount and type of information provided about AI models to build user trust, as merely increasing information is not always effective. Representing AI uncertainty in a continuous format can enhance perceived reliability but also risks introducing ambiguity, underscoring the need for clear and intuitive visualizations, such as bar charts. Ultimately, human-AI design and context-specific customization are important for improving trust and decision-making in human-AI collaborations.

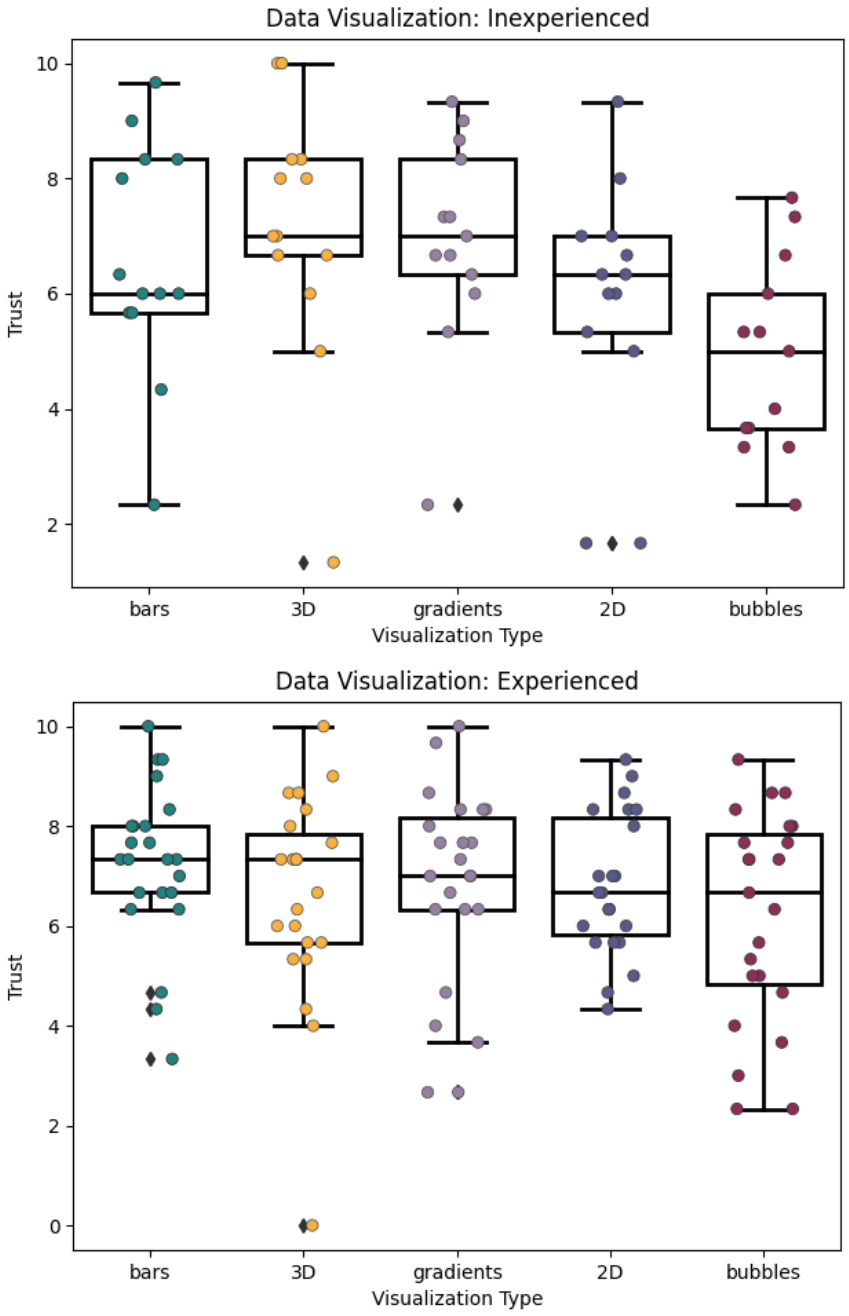

Figure 6.9: Users perceptions of trust given various visual formats among inexperienced participants (top) and experienced participants (bottom) in data visualization.

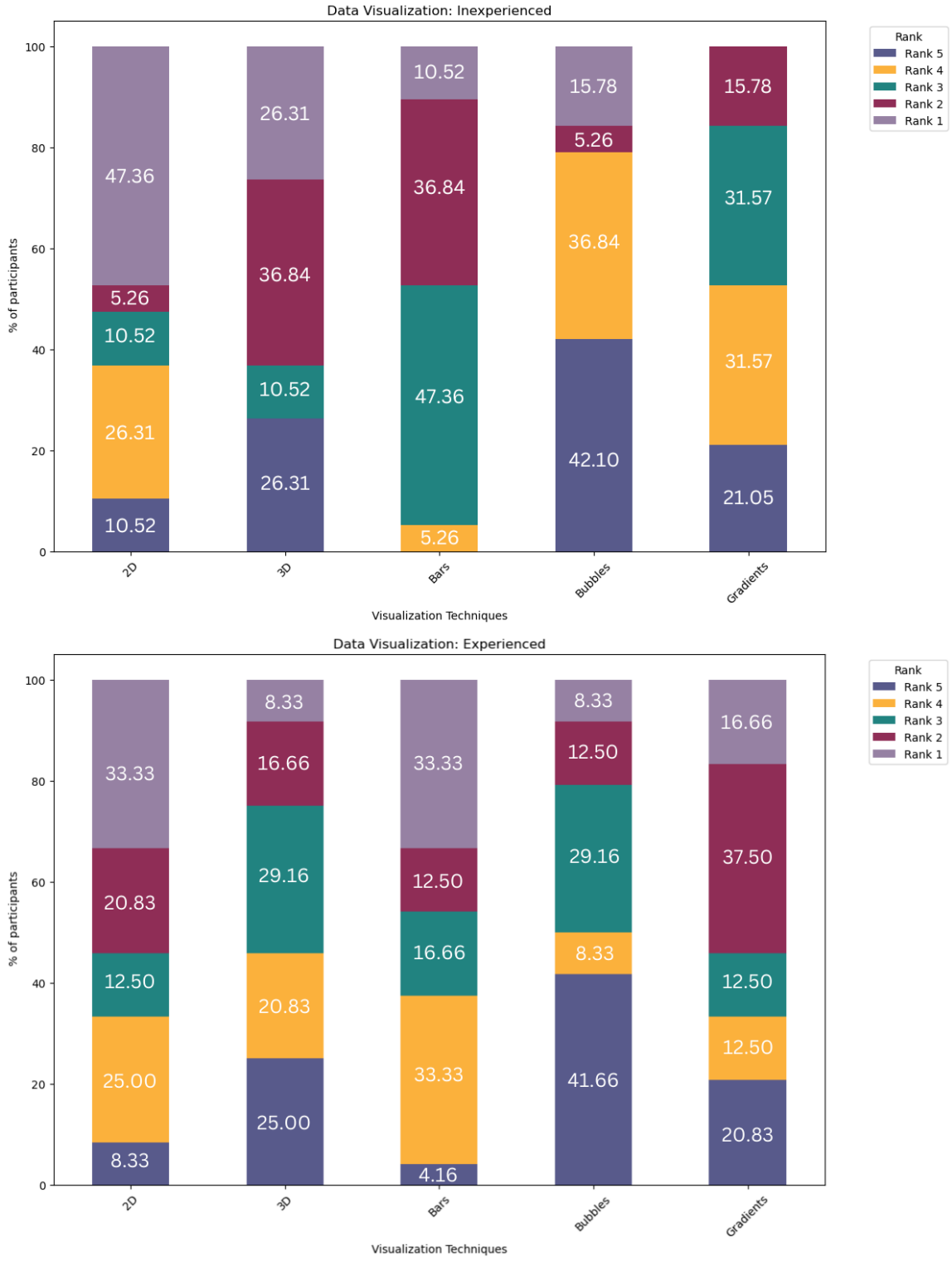

Figure 6.10: Comparing the preferred visual method to represent AI uncertainty between participants with low experience (top) and high experience (bottom) in data visualization.

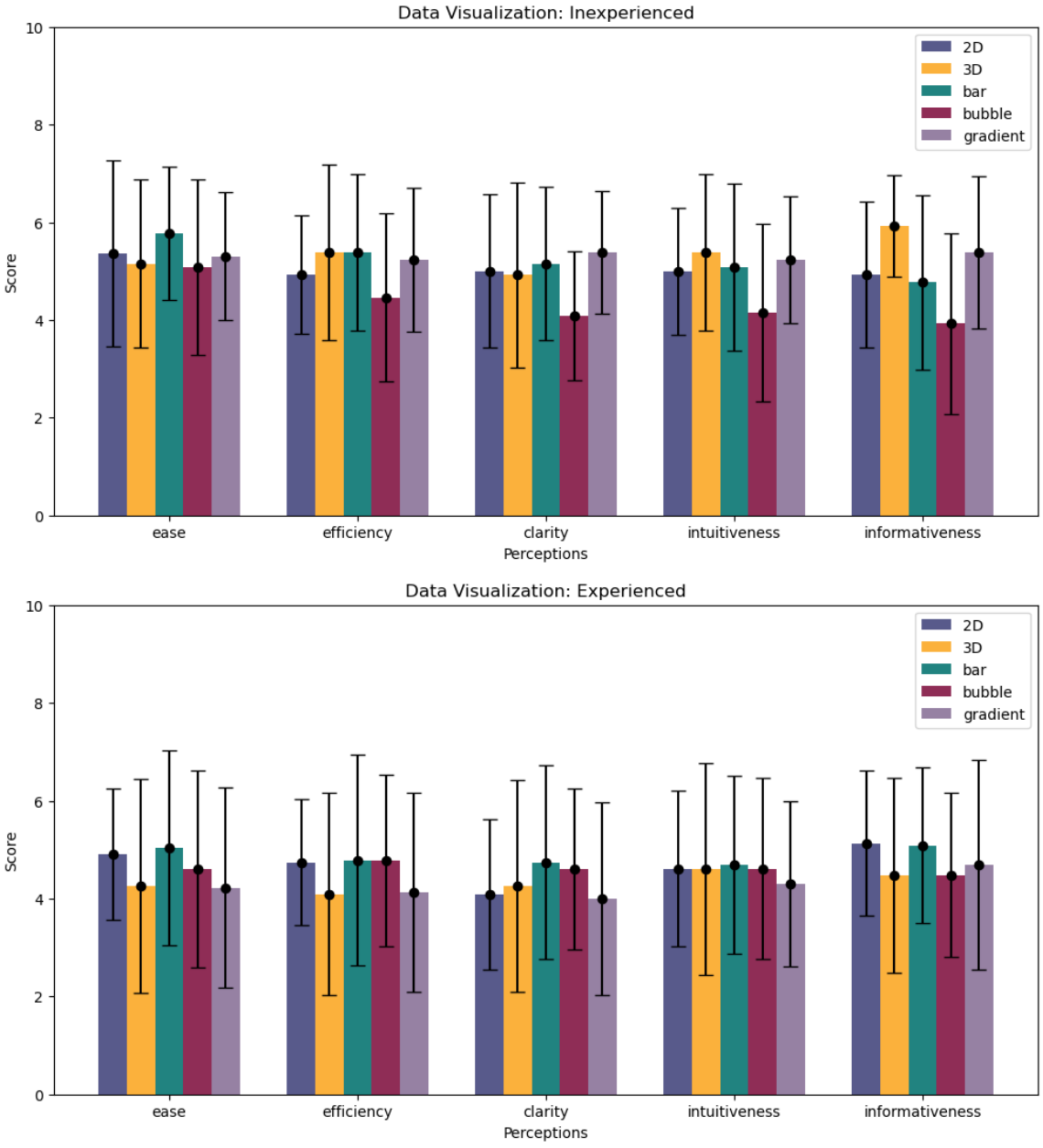

Figure 6.11: Comparing the ability to perceive AI uncertainty across various visual methods based on a 10-point scale. Participants with low experience in data visualization are shown on the top, and those with experienced participants are shown on the bottom.

Chapter 7

Conclusions and Future Work

While many domains have significantly benefited from Industry 4.0, the impact of the transformation in healthcare has been slower due to various technological, legal, and ethical issues. Although DL-based technologies offer unprecedented opportunities for clinical solutions and decision-making, addressing the aforementioned challenges is crucial for a successful digital transformation in this domain. Without these measures, the transition of technologies from research laboratories to daily practice will remain slow, limited, or even disregarded. Thus, further research is needed to provide end-to-end solutions that are secure, trusted, and compliant with stringent data protection regulations and guidelines. This will enable a smoother adoption of AI technologies in clinical practice.

Given the challenges of current deep learning methods, there is an opportunity to enhance clinical acceptance of AI technologies through privacy-preserving mechanisms and trustworthy solutions. This thesis addresses these aspects, first to leverage the potential of federated learning as a privacy-preserving framework, with improvements on the core aggregation algorithm for better handling the learning of distributed heterogeneous data sources, and second in its consideration of human factors in the design of novel technologies to optimize the output in a transparent and interpretable way. At the same time, the considerations of human factors in the design and development of AI solutions can lead to usable, trustworthy, and interpretable AI systems, which is valuable for enhancing confidence in AI support-decision systems particularly in the clinical domain.

7.1 Summary of Findings

This thesis explored how to enhance the adoption of AI-based technologies through privacy-preserving mechanisms and trustworthy solutions. To do so, our contributions addressed technical, legal, and ethical challenges of AI innovations through the formulation of a Federated Learning (FL)-based aggregation algorithm. In addition, we conducted thorough qualitative and quantitative analyses to demonstrate that the transparency of AI outputs is needed to enhance the interpretability and adoption of AI solutions. In addition to these contributions, a number of research questions were answered. Table 7.1 summarizes all research questions explored in this thesis and their conclusions.

7.1.1 A privacy-preserving aggregation algorithm

The motivation behind the research in Chapter 3 is to address the technical and legal privacy issues of deep learning (DL) methods, such as data volume, quality of data, and data privacy-preservation. We turn to the FL framework to address these issues, since it enables fast and secure collaboration. However, the core Federated Averaging (FedAvg) algorithm in FL, has been shown to inadequately account for data heterogeneity across different clients, reducing the statistical power and prediction quality of models trained in a Federated Learning setting. Our research focused on developing an improved method, Precision-weighted Federated Learning (PW), which aggregates models by the inverse of the estimated variance. The substantial advantage of the PW method lies in its ability to handle heterogeneous data more efficiently, speeding up the training process and leading to better generalizations, especially with diverse datasets. This is highly relevant to clinical applications.

In Chapter 4, we applied the algorithm to a practical clinical use case: imputing missing clinical data across multiple centers. Our evaluations compared seven FL aggregation algorithms against centralized learning. To validate our findings, we performed a downstream analysis to classify Parkinson's disease patients based on symptom progression. The results demonstrated that FL algorithms can achieve better generalization without pooling data from multiple clinical centers and enable collaborative learning as data from additional centers is incorporated. Specifically, we confirmed that the PW algorithm leads to better generalizations, highlighting its clinical utility.

Research Question	Finding
Can a meta-analysis weighting scheme be integrated to aggregate summary statistics in a federated learning framework?	Yes. A novel method was designed, developed, tested with benchmark datasets and found to obtain comparable results in an IID distribution and, depending on the training set up, better performance in non-IID distributions than the original FL aggregation algorithm.
Can we improve the quality of curated clinical data for more consistent prediction of disease progression status with a meta-analysis weighting scheme?	Yes. We extended the evaluations of the proposed algorithm to include training distributed models with clinical data, demonstrating that better generalization can be obtained with our proposed aggregation algorithm.
Does visualization of uncertainty impact decision-making, trust, and confidence among people with different attitudes towards AI?	Yes, partially. A particular view of AI technologies affects trust differently. People with a negative attitude towards AI have a tendency to trust more in AI with a particular visual method. However, this effect did not hold for confidence in decisions and decision change.
Do attitudes towards AI influence decision-making, trust in AI, and confidence in the decisions made differently?	Yes, partially. Personal traits is a driving factor for trust in the AI technology. More positive attitude towards AI was correlated to higher level of trust in the AI technology. We did not observe this effect in our analysis of confidence in decisions and decision change
How is the visualization of uncertainty perceived by people when making decisions?	People find the assistance of visual elements useful when evaluating AI uncertainty, particularly those with a negative attitude towards AI.
How does the level of detail about the AI model and AI's uncertainty impact expert's trust in high-stake predictions?	People tend to be overconfident when unaware of AI's uncertainty, but awareness of AI's uncertainty leads to questioning its reliability. Providing information about AI's limits can improve output reliability with caution, as poorly communicating visual methods can negatively affect trust.
What visual format of uncertainty enhances interpretability of AI predictions in AD trajectory?	We observed that people favored visual methods expressing AI throughout its output. However, we could not confirm its significance.

Table 7.1: Summary of research questions and findings in this dissertation.

7.1.2 Human-centered evaluations of transparency in clinical decision support systems

After addressing the technical and legal privacy issues in AI clinical acceptance, we tackle the ethical challenges aiming at developing solutions to make AI decisions understandable and transparent to clinicians. Thus, we focus this research to understand how to visualize data to improve human-AI decision making. Since previous research indicated that visualizing uncertainty can influence the utility and adoption of AI-based technologies, Chapter 5 expands on this by examining how participants' attitudes towards AI and the visualization of its uncertainty affect their decision-making, trust in AI, and confidence in their decisions. According to our findings, personal traits can influence trust in AI. This becomes a strong case for conducting further evaluations in high-stake decision-making.

The work presented in Chapter 6 builds on Eslami *et al.* [187]'s work on AI uncertainty in Alzheimer's disease prognosis, aiming to measure trust in AI predictions through various visual formats. The goal is to understand how different representations of AI uncertainty affect human-AI decision-making processes among experts, with a focus on improving trust and interpretability in clinical practice. Our findings suggested that visualizing AI uncertainty can help users avoid overconfidence and lead to more trustworthy AI solutions. More importantly, it highlights the need for clear and intuitive visualizations to balance confidence and caution in clinical decision-making.

7.2 Future Work

Many open questions remain or build on the research provided in this manuscript. In this section, we present several areas that warrant further exploration.

One avenue for future research is to enhance the performance and defense mechanisms of our algorithm. We identified that our developed method is sensitive to the noise in the training data. Specifically, PW's performance diminishes with small batch sizes. This allows further investigation on detection mechanisms that can be integrated to identify and correct noise before aggregation. Other strategies such as constraints or penalties could serve as a form of regularization in these scenarios. On the other hand, the FL framework faces challenges related to inference attacks. To

address these security issues, various protocols have been proposed to provide an extra layer of protection at the client-level [107, 108, 80], or at the framework-level [202] ensuring collaborative learning with privacy guarantees. Strengthening these privacy measures is essential to comply with data protection laws and conduct regular security audits.

Moreover, providing external validation to demonstrate the stability of our clinical analyses could further strengthen our findings. Our current design choice, which involved using the same database to split data into train, validation, and test sets, limited the scalability of the predictive task. Future studies could replicate our research using the Parkinson's Disease Biomarkers Program (PDBP) databasee [203], which offers clinical data suitable for Federated Learning evaluations.

In addition, we have limited the evaluations of the proposed method to clinical data. We could extend the evaluations of the proposed algorithm with distributed medical imaging data (e.g. X-ray, CT, PET, MRI or fMRI). This would enable us to perform exploratory analysis in the identification of symptomatic progression of Alzheimer's disease using the available imaging data in the ADNI dataset. Another interesting application could be to explore the introduction of unintended biases into the learning process. It is important to develop discrimination-aware learning frameworks to avoid potential biases towards a certain attribute (e.g. gender or racial) in the population. Failure in accounting for these differences might result in discriminatory outcomes

Furthermore, incorporating feedback from clinicians, medical trainees, and neurodegeneration researchers would strengthen our assessments of transparency in clinical decision support systems. Their insights would ensure the provided visualizations address real-world clinical needs, making our findings more applicable and relevant to clinical practice. Another interesting improvement to this research work is to conduct longitudinal studies. This would allow a better understanding of the long-term usage of the intended AI solution and identify areas that are useful in practice and those needed improvement.

Finally, considering factors that prevent the adoption of AI technologies is crucial for a smooth transition from research to daily practice. While this manuscript addresses common issues hindering AI adoption in clinics, AI developers and designers should also focus on integrating technology without disrupting established clinical workflows. Additionally, improving AI uncertainty evaluations could involve methods for well-calibrated uncertainty or those offering better interpretation

of AI models, such as SHAP (SHapley Additive exPlanations), to provide clear explanations for predictions.

Bibliography

[1] V. Lai, C. Chen, A. Smith-Renner, Q. V. Liao, and C. Tan, "Towards a science of human-ai decision making: An overview of design space in empirical human-subject studies," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1369–1385, 2023.

[2] M. Ashoori and J. D. Weisz, "In ai we trust? factors that influence trustworthiness of ai-infused decision-making processes," *arXiv preprint arXiv:1912.02675*, 2019.

[3] A. D. Maynard, "Navigating the fourth industrial revolution," *Nature nanotechnology*, vol. 10, no. 12, pp. 1005–1006, 2015.

[4] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," in *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27*, pp. 270–279, Springer, 2018.

[5] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[6] K. Chowdhary and K. Chowdhary, "Natural language processing," *Fundamentals of artificial intelligence*, pp. 603–649, 2020.

[7] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Computational intelligence and neuroscience*, vol. 2018, no. 1, p. 7068349, 2018.

[8] S. Benjamens, P. Dhunnoo, and B. Meskó, "The state of artificial intelligence-based fda-approved medical devices and algorithms: an online database," *NPJ digital medicine*, vol. 3, no. 1, p. 118, 2020.

[9] V. Murali, B. A. Adewale, C. J. Huang, M. T. Nta, P. O. Ademiju, P. Pathmarajah, M. K. Hang, O. Adesanya, R. O. Abdullateef, A. O. Babatunde, *et al.*, "Health disparities and reporting gaps in artificial intelligence (ai) enabled medical devices: A scoping review of 692 us food and drug administration (fda) 510k approvals," *medRxiv*, pp. 2024–05, 2024.

[10] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*, pp. 1273–1282, PMLR, 2017.

[11] E. Darzidehkalani, M. Ghasemi-Rad, and P. Van Ooijen, "Federated learning in medical imaging: part ii: methods, challenges, and considerations," *Journal of the American College of Radiology*, vol. 19, no. 8, pp. 975–982, 2022.

[12] S. Boughorbel, F. Jarray, N. Venugopal, S. Moosa, H. Elhadi, and M. Makhlouf, "Federated uncertainty-aware learning for distributed hospital ehr data," *arXiv preprint arXiv:1910.12191*, 2019.

[13] C. Shen, P. Wang, H. R. Roth, D. Yang, D. Xu, M. Oda, W. Wang, C.-S. Fuh, P.-T. Chen, K.-L. Liu, *et al.*, "Multi-task federated learning for heterogeneous pancreas segmentation," in *Clinical Image-Based Procedures, Distributed and Collaborative Learning, Artificial Intelligence for Combating COVID-19 and Secure and Privacy-Preserving Machine Learning: 10th Workshop, CLIP 2021, Second Workshop, DCL 2021, First Workshop, LL-COVID19 2021, and First Workshop and Tutorial, PPML 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27 and October 1, 2021, Proceedings 2*, pp. 101–110, Springer, 2021.

[14] B. L. Y. Agbley, J. P. Li, A. U. Haq, E. K. Bankas, C. B. Mawuli, S. Ahmad, S. Khan, and A. R. Khan, "Federated fusion of magnified histopathological images for breast tumor

classification in the internet of medical things," *IEEE Journal of Biomedical and Health Informatics*, 2023.

[15] J. Reyes, Y. Xiao, and M. Kersten-Oertel, "Data imputation and reconstruction of distributed parkinson's disease clinical assessments: A comparative evaluation of two aggregation algorithms," in *Clinical Image-Based Procedures, Distributed and Collaborative Learning, Artificial Intelligence for Combating COVID-19 and Secure and Privacy-Preserving Machine Learning: 10th Workshop, CLIP 2021, Second Workshop, DCL 2021, First Workshop, LL-COVID19 2021, and First Workshop and Tutorial, PPML 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27 and October 1, 2021, Proceedings 2*, pp. 163–173, Springer, 2021.

[16] D. M. Berwick and A. D. Hackbarth, "Eliminating waste in us health care," *Jama*, vol. 307, no. 14, pp. 1513–1516, 2012.

[17] J. A. Fiore, A. J. Madison, J. A. Poisal, G. A. Cuckler, S. D. Smith, A. M. Sisko, S. P. Keehan, K. E. Rennie, and A. C. Gross, "National health expenditure projections, 2023–32: Payer trends diverge as pandemic-related policies fade: Study examines national health expenditure projections, 2023–32," *Health Affairs*, pp. 10–1377, 2024.

[18] S. Makridakis, "The forthcoming artificial intelligence (ai) revolution: Its impact on society and firms," *Futures*, vol. 90, pp. 46–60, 2017.

[19] A. A. Abujaber, A. J. Nashwan, and A. Fadlalla, "Harnessing machine learning to support evidence-based medicine: a pragmatic reconciliation framework," *Intelligence-Based Medicine*, vol. 6, p. 100048, 2022.

[20] B. Djulbegovic, S. Elqayam, and W. Dale, "Rational decision making in medicine: implications for overuse and underuse," *Journal of evaluation in clinical practice*, vol. 24, no. 3, pp. 655–665, 2018.

[21] A. A. Abujaber, A. J. Nashwan, and A. Fadlalla, "Enabling the adoption of machine learning in clinical decision support: a total interpretive structural modeling approach," *Informatics in Medicine Unlocked*, vol. 33, p. 101090, 2022.

[22] J. Meszaros, J. Minari, and I. Huys, "The future regulation of artificial intelligence systems in healthcare services and medical research in the european union," *Frontiers in Genetics*, vol. 13, p. 927721, 2022.

[23] M. H. Jarrahi, "Artificial intelligence and the future of work: Human-ai symbiosis in organizational decision making," *Business horizons*, vol. 61, no. 4, pp. 577–586, 2018.

[24] T. O. Andersen, F. Nunes, L. Wilcox, E. Coiera, and Y. Rogers, "Introduction to the special issue on human-centred ai in healthcare: Challenges appearing in the wild," 2023.

[25] W. Xu, M. J. Dainoff, L. Ge, and Z. Gao, "Transitioning to human interaction with ai systems: New challenges and opportunities for hci professionals to enable human-centered ai," *International Journal of Human–Computer Interaction*, vol. 39, no. 3, pp. 494–518, 2023.

[26] S. Oviatt, "Human-centered design meets cognitive load theory: designing interfaces that help people think," in *Proceedings of the 14th ACM international conference on Multimedia*, pp. 871–880, 2006.

[27] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE signal processing magazine*, vol. 37, no. 3, pp. 50–60, 2020.

[28] Y. Kang, B. Li, and T. Zeyl, "Fedrl: Improving the performance of federated learning with non-iid data," in *GLOBECOM 2022-2022 IEEE Global Communications Conference*, pp. 3023–3028, IEEE, 2022.

[29] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," in *International conference on machine learning*, pp. 5132–5143, PMLR, 2020.

[30] T.-M. H. Hsu, H. Qi, and M. Brown, "Measuring the effects of non-identical data distribution for federated visual classification," *arXiv preprint arXiv:1909.06335*, 2019.

[31] C. Collins, D. Dennehy, K. Conboy, and P. Mikalef, "Artificial intelligence in information systems research: A systematic literature review and research agenda," *International Journal of Information Management*, vol. 60, p. 102383, 2021.

[32] P. Langley, "Artificial intelligence and cognitive systems," 2012.

[33] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[34] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Briefings in bioinformatics*, vol. 19, no. 6, pp. 1236–1246, 2018.

[35] M. Beauchemin, C. Weng, L. Sung, A. Pichon, M. Abbott, D. L. Hershman, and R. Schnall, "Data quality of chemotherapy-induced nausea and vomiting documentation," *Applied Clinical Informatics*, vol. 12, no. 02, pp. 320–328, 2021.

[36] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[37] D. Ravì, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G.-Z. Yang, "Deep learning for health informatics," *IEEE journal of biomedical and health informatics*, vol. 21, no. 1, pp. 4–21, 2016.

[38] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: concepts, cnn architectures, challenges, applications, future directions," *Journal of big Data*, vol. 8, pp. 1–74, 2021.

[39] A. D'Amour, K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, M. D. Hoffman, *et al.*, "Underspecification presents challenges for credibility in modern machine learning," *Journal of Machine Learning Research*, vol. 23, no. 226, pp. 1–61, 2022.

[40] L. Torrey and J. Shavlik, "Transfer learning," in *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pp. 242–264, IGI global, 2010.

[41] A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia, "A brief review of domain adaptation," *Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020*, pp. 877–894, 2021.

[42] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, *et al.*, "Communication-efficient learning of deep networks from decentralized data," *arXiv preprint arXiv:1602.05629*, 2016.

[43] K. Kawamoto, C. A. Houlihan, E. A. Balas, and D. F. Lobach, "Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success," *Bmj*, vol. 330, no. 7494, p. 765, 2005.

[44] D. J. Power, "Understanding data-driven decision support systems," *Information Systems Management*, vol. 25, no. 2, pp. 149–154, 2008.

[45] R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker, "An overview of clinical decision support systems: benefits, risks, and strategies for success," *NPJ digital medicine*, vol. 3, no. 1, p. 17, 2020.

[46] F. Magrabi, E. Ammenwerth, J. B. McNair, N. F. De Keizer, H. Hyppönen, P. Nykänen, M. Rigby, P. J. Scott, T. Vehko, Z. S.-Y. Wong, *et al.*, "Artificial intelligence in clinical decision support: challenges for evaluating ai and practical implications," *Yearbook of medical informatics*, vol. 28, no. 01, pp. 128–134, 2019.

[47] P. Schmidt, F. Biessmann, and T. Teubner, "Transparency and trust in artificial intelligence systems," *Journal of Decision Systems*, vol. 29, no. 4, pp. 260–278, 2020.

[48] "American Psychological Association dictionary of psychology." https://dictionary.apa.org/trust. Accessed: 2024-06-11.

[49] Z. Salahuddin, H. C. Woodruff, A. Chatterjee, and P. Lambin, "Transparency of deep neural networks for medical image analysis: A review of interpretability methods," *Computers in biology and medicine*, vol. 140, p. 105111, 2022.

[50] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable ai: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, 2020.

[51] B. Shneiderman, *Human-centered AI*. Oxford University Press, 2022.

[52] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, *et al.*, "Guidelines for human-ai interaction," in *Proceedings of the 2019 chi conference on human factors in computing systems*, pp. 1–13, 2019.

[53] E. Horvitz, "Principles of mixed-initiative user interfaces," in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pp. 159–166, 1999.

[54] I. Ajzen, "The theory of planned behaviour is alive and well, and not ready to retire: a commentary on sniehotta, presseau, and araújo-soares," *Health psychology review*, vol. 9, no. 2, pp. 131–137, 2015.

[55] M. Daradkeh and B. Abul-Huda, "Incorporating uncertainty into decision-making: An information visualisation approach," in *Decision Support Systems VII. Data, Information and Knowledge Visualization in Decision Support Systems: Third International Conference, ICDSST 2017, Namur, Belgium, May 29-31, 2017, Proceedings 3*, pp. 74–87, Springer, 2017.

[56] A. Doula, L. Schmidt, M. Mühlhäuser, and A. S. Guinea, "Visualization of machine learning uncertainty in ar-based see-through applications," in *2022 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pp. 109–113, IEEE, 2022.

[57] L. R. Marusich, J. Z. Bakdash, Y. Zhou, and M. Kantarcioglu, "Using ai uncertainty quantification to improve human decision-making," *arXiv preprint arXiv:2309.10852*, 2023.

[58] D. N. Cassenti, L. M. Kaplan, and A. Roy, "Representing uncertainty information from ai for human understanding," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 67, pp. 177–182, SAGE Publications Sage CA: Los Angeles, CA, 2023.

[59] A. Association *et al.*, "2018 alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 14, no. 3, pp. 367–429, 2018.

[60] R. Brookmeyer, E. Johnson, K. Ziegler-Graham, and H. M. Arrighi, "Forecasting the global burden of alzheimer's disease," *Alzheimer's & dementia*, vol. 3, no. 3, pp. 186–191, 2007.

[61] S. Gauthier, B. Reisberg, M. Zaudig, R. C. Petersen, K. Ritchie, K. Broich, S. Belleville, H. Brodaty, D. Bennett, H. Chertkow, *et al.*, "Mild cognitive impairment," *The lancet*, vol. 367, no. 9518, pp. 1262–1270, 2006.

[62] R. C. Petersen, G. E. Smith, S. C. Waring, R. J. Ivnik, E. G. Tangalos, and E. Kokmen, "Mild cognitive impairment: clinical characterization and outcome," *Archives of neurology*, vol. 56, no. 3, pp. 303–308, 1999.

[63] B. R. Bloem, M. S. Okun, and C. Klein, "Parkinson's disease," *The Lancet*, vol. 397, no. 10291, pp. 2284–2303, 2021.

[64] J. Jankovic, "Parkinson's disease: clinical features and diagnosis," *Journal of neurology, neurosurgery & psychiatry*, vol. 79, no. 4, pp. 368–376, 2008.

[65] V. Voon and S. H. Fox, "Medication-related impulse control and repetitive behaviors in parkinson disease," *Archives of neurology*, vol. 64, no. 8, pp. 1089–1096, 2007.

[66] M. Emre, "Dementia associated with parkinson's disease," *The Lancet Neurology*, vol. 2, no. 4, pp. 229–237, 2003.

[67] V. L. Feigin, E. Nichols, T. Alam, M. S. Bannick, E. Beghi, N. Blake, W. J. Culpepper, E. R. Dorsey, A. Elbaz, R. G. Ellenbogen, *et al.*, "Global, regional, and national burden of neurological disorders, 1990–2016: a systematic analysis for the global burden of disease study 2016," *The Lancet Neurology*, vol. 18, no. 5, pp. 459–480, 2019.

[68] E. a. Dorsey, R. Constantinescu, J. Thompson, K. Biglan, R. Holloway, K. Kieburtz, F. Marshall, B. Ravina, G. Schifitto, A. Siderowf, *et al.*, "Projected number of people with parkinson disease in the most populous nations, 2005 through 2030," *Neurology*, vol. 68, no. 5, pp. 384–386, 2007.

[69] K. K. Tsoi, J. Y. Chan, H. W. Hirai, S. Y. Wong, and T. C. Kwok, "Cognitive tests to detect dementia: a systematic review and meta-analysis," *JAMA internal medicine*, vol. 175, no. 9, pp. 1450–1458, 2015.

[70] J. S. Perlmutter, "Assessment of parkinson disease manifestations," *Current protocols in neuroscience*, vol. 49, no. 1, pp. 10–1, 2009.

[71] S. Mathotaarachchi, T. A. Pascoal, M. Shin, A. L. Benedet, M. S. Kang, T. Beaudry, V. S. Fonov, S. Gauthier, P. Rosa-Neto, A. D. N. Initiative, *et al.*, "Identifying incipient dementia individuals using machine learning and amyloid imaging," *Neurobiology of aging*, vol. 59, pp. 80–90, 2017.

[72] J. Reyes, L. Di Jorio, C. Low-Kam, and M. Kersten-Oertel, "Precision-weighted federated learning," *arXiv preprint arXiv:2107.09627*, 2021.

[73] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep learning applications and challenges in big data analytics," *Journal of Big Data*, vol. 2, no. 1, p. 1, 2015.

[74] Z. Tian, C. Luo, J. Qiu, X. Du, and M. Guizani, "A distributed deep learning system for web attack detection on edge devices," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 3, pp. 1963–1971, 2019.

[75] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *European conference on computer vision*, pp. 430–443, Springer, 2006.

[76] T. Yang, G. Andrew, H. Eichner, H. Sun, W. Li, N. Kong, D. Ramage, and F. Beaufays, "Applied federated learning: Improving google keyboard query suggestions," *arXiv preprint arXiv:1812.02903*, 2018.

[77] M. Johanson, S. Belenki, J. Jalminger, M. Fant, and M. Gjertz, "Big automotive data: Leveraging large volumes of data for knowledge-driven product development," in *Big Data (Big Data), 2014 IEEE International Conference on*, pp. 736–741, IEEE, 2014.

[78] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical secure aggregation for federated learning on user-held data," *arXiv preprint arXiv:1611.04482*, 2016.

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

[80] L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov, "Exploiting unintended feature leakage in collaborative learning," in *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 691–706, IEEE, 2019.

[81] L. V. Hedges and I. Olkin, *Statistical methods for meta-analysis*. Academic Press, 1985.

[82] S. Nakagawa and E. S. Santos, "Methodological issues and advances in biological meta-analysis," *Evolutionary Ecology*, vol. 26, no. 5, pp. 1253–1274, 2012.

[83] J. P. Ioannidis, N. A. Patsopoulos, and E. Evangelou, "Heterogeneity in meta-analyses of genome-wide association investigations," *PloS one*, vol. 2, no. 9, p. e841, 2007.

[84] S. I. Bangdiwala, A. Bhargava, D. P. O'Connor, T. N. Robinson, S. Michie, D. M. Murray, J. Stevens, S. H. Belle, T. N. Templin, and C. A. Pratt, "Statistical methodologies to pool across multiple intervention studies," *Translational behavioral medicine*, vol. 6, no. 2, pp. 228–235, 2016.

[85] D. Lin and D. Zeng, "Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data," *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, vol. 34, no. 1, pp. 60–66, 2010.

[86] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations*, 2014.

[87] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[88] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[89] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," tech. rep., Citeseer, 2009.

[90] R. Agrawal and R. Srikant, *Privacy-preserving data mining*, vol. 29. ACM, 2000.

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

[92] L. Xu, C. Jiang, J. Wang, J. Yuan, and Y. Ren, "Information security in big data: privacy and data mining," *IEEE Access*, vol. 2, pp. 1149–1176, 2014.

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

[94] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," *arXiv preprint arXiv:1610.05492*, 2016.

[95] V. Smith, C.-K. Chiang, M. Sanjabi, and A. S. Talwalkar, "Federated multi-task learning," in *Advances in Neural Information Processing Systems*, pp. 4424–4434, 2017.

[96] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-iid data," *arXiv preprint arXiv:1806.00582*, 2018.

[97] K. Chang, N. Balachandar, C. Lam, D. Yi, J. Brown, A. Beers, B. Rosen, D. L. Rubin, and J. Kalpathy-Cramer, "Distributed deep learning networks among institutions for medical imaging," *Journal of the American Medical Informatics Association*, 2018.

[98] K. Xu, H. Mi, D. Feng, H. Wang, C. Chen, Z. Zheng, and X. Lan, "Collaborative deep learning across multiple data centers," *arXiv preprint arXiv:1810.06877*, 2018.

[99] A. Lalita, S. Shekhar, T. Javidi, and F. Koushanfar, "Fully decentralized federated learning," in *NeurIPS Workshop on Bayesian Deep Learning*, 2018.

[100] F. Chen, Z. Dong, Z. Li, and X. He, "Federated meta-learning for recommendation," *arXiv preprint arXiv:1802.07876*, 2018.

[101] H.-E. Kim, S. Kim, and J. Lee, "Keep and learn: Continual learning by constraining the latent space for knowledge preservation in neural networks," *arXiv preprint arXiv:1805.10784*, 2018.

[102] T. D. Bui, C. V. Nguyen, S. Swaroop, and R. E. Turner, "Partitioned variational inference: A unified framework encompassing federated and continual learning," in *NeurIPS Workshop on Bayesian Deep Learning*, 2018.

[103] P. Vepakomma, O. Gupta, T. Swedish, and R. Raskar, "Split learning for health: Distributed deep learning without sharing raw patient data," *arXiv preprint arXiv:1812.00564*, 2018.

[104] R. Pascanu and Y. Bengio, "Natural gradient revisited," in *Proceedings of the 3rd International Conference on Learning Representations*, 2014.

[105] F. Scheidegger, R. Istrate, G. Mariani, L. Benini, C. Bekas, and C. Malossi, "Efficient image dataset classification difficulty estimation for predicting deep-learning accuracy," *arXiv preprint arXiv:1803.09588*, 2018.

[106] M. Maniruzzaman, M. J. Rahman, M. Al-MehediHasan, H. S. Suri, M. M. Abedin, A. El-Baz, and J. S. Suri, "Accurate diabetes risk stratification using machine learning: role of missing value and outliers," *Journal of medical systems*, vol. 42, no. 5, pp. 1–17, 2018.

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

[108] S. Truex, N. Baracaldo, A. Anwar, T. Steinke, H. Ludwig, and R. Zhang, "A hybrid approach to privacy-preserving federated learning," *arXiv preprint arXiv:1812.03224*, 2018.

[109] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. Quek, and H. V. Poor, "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3454–3469, 2020.

[110] E. R. Dorsey, A. Elbaz, E. Nichols, N. Abbasi, F. Abd-Allah, A. Abdelalim, J. C. Adsuar, M. G. Ansha, C. Brayne, J.-Y. J. Choi, *et al.*, "Global, regional, and national burden of

parkinson's disease, 1990–2016: a systematic analysis for the global burden of disease study 2016," *The Lancet Neurology*, vol. 17, no. 11, pp. 939–953, 2018.

[111] E. R. Dorsey and B. R. Bloem, "The parkinson pandemic—a call to action," *JAMA neurology*, vol. 75, no. 1, pp. 9–10, 2018.

[112] K. Dhima, *A Data-Driven Approach for Deriving Parkinson's Disease Subtypes and Related Trajectories of Cognitive and Motor Function*. PhD thesis, 2018.

[113] A. Dadu, V. Satone, R. Kaur, S. H. Hashemi, H. Leonard, H. Iwaki, M. B. Makarious, K. J. Billingsley, S. Bandres-Ciga, L. J. Sargent, *et al.*, "Identification and prediction of parkinson's disease subtypes and progression using machine learning in two cohorts," *npj Parkinson's Disease*, vol. 8, no. 1, p. 172, 2022.

[114] S.-M. Fereshtehnejad, Y. Zeighami, A. Dagher, and R. B. Postuma, "Clinical criteria for subtyping parkinson's disease: biomarkers and longitudinal progression," *Brain*, vol. 140, no. 7, pp. 1959–1976, 2017.

[115] H. J. Sadaei, A. Cordova-Palomera, J. Lee, J. Padmanabhan, S.-F. Chen, N. E. Wineinger, R. Dias, D. Prilutsky, S. Szalma, and A. Torkamani, "Genetically-informed prediction of short-term parkinson's disease progression," *npj Parkinson's Disease*, vol. 8, no. 1, p. 143, 2022.

[116] M. Peralta, P. Jannin, C. Haegelen, and J. S. Baxter, "Data imputation and compression for parkinson's disease clinical questionnaires," *Artificial Intelligence in Medicine*, vol. 114, p. 102051, 2021.

[117] M. C. Brumm, A. Siderowf, T. Simuni, E. Burghardt, S. H. Choi, C. Caspell-Garcia, L. M. Chahine, B. Mollenhauer, T. Foroud, D. Galasko, *et al.*, "Parkinson's progression markers initiative: A milestone-based strategy to monitor parkinson's disease progression," *Journal of Parkinson's disease*, no. Preprint, pp. 1–18, 2023.

[118] B. Efron, "Missing data, imputation, and the bootstrap," *Journal of the American Statistical Association*, vol. 89, no. 426, pp. 463–475, 1994.

[119] G. J. Van der Heijden, A. R. T. Donders, T. Stijnen, and K. G. Moons, "Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example," *Journal of clinical epidemiology*, vol. 59, no. 10, pp. 1102–1109, 2006.

[120] P. D. Allison, *Missing data*. Sage publications, 2001.

[121] Z. Zhang, "Missing data imputation: focusing on single imputation," *Annals of translational medicine*, vol. 4, no. 1, 2016.

[122] D. A. Newman, "Longitudinal modeling with randomly and systematically missing data: A simulation of ad hoc, maximum likelihood, and multiple imputation techniques," *Organizational research methods*, vol. 6, no. 3, pp. 328–362, 2003.

[123] C. K. Enders, "A primer on the use of modern missing-data methods in psychosomatic medicine research," *Psychosomatic medicine*, vol. 68, no. 3, pp. 427–436, 2006.

[124] J. M. Engels and P. Diehr, "Imputation of missing longitudinal data: a comparison of methods," *Journal of clinical epidemiology*, vol. 56, no. 10, pp. 968–976, 2003.

[125] C. M. Musil, C. B. Warner, P. K. Yobas, and S. L. Jones, "A comparison of imputation techniques for handling missing data," *Western Journal of Nursing Research*, vol. 24, no. 7, pp. 815–829, 2002.

[126] D. B. Rubin, "Multiple imputations in sample surveys-a phenomenological bayesian approach to nonresponse," in *Proceedings of the survey research methods section of the American Statistical Association*, vol. 1, pp. 20–34, American Statistical Association Alexandria, VA, USA, 1978.

[127] D. B. Rubin, *Multiple imputation for nonresponse in surveys*, vol. 81. John Wiley & Sons, 2004.

[128] A. R. T. Donders, G. J. Van Der Heijden, T. Stijnen, and K. G. Moons, "A gentle introduction to imputation of missing values," *Journal of clinical epidemiology*, vol. 59, no. 10, pp. 1087–1091, 2006.

[129] J. M. Jerez, I. Molina, P. J. García-Laencina, E. Alba, N. Ribelles, M. Martín, and L. Franco, "Missing data imputation using statistical and machine learning methods in a real breast cancer problem," *Artificial intelligence in medicine*, vol. 50, no. 2, pp. 105–115, 2010.

[130] J. S. Vetter, K. Schultebraucks, I. Galatzer-Levy, H. Boeker, A. Brühl, E. Seifritz, and B. Kleim, "Predicting non-response to multimodal day clinic treatment in severely impaired depressed patients: a machine learning approach," *Scientific Reports*, vol. 12, no. 1, pp. 1–12, 2022.

[131] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*, pp. 1096–1103, 2008.

[132] E. Vayena, A. Blasimme, and I. G. Cohen, "Machine learning in medicine: addressing ethical challenges," *PLoS medicine*, vol. 15, no. 11, p. e1002689, 2018.

[133] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pp. 1322–1333, 2015.

[134] 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)*, pp. 3–18, IEEE, 2017.

[135] M. Li, D. G. Andersen, J. W. Park, A. J. Smola, A. Ahmed, V. Josifovski, J. Long, E. J. Shekita, and B.-Y. Su, "Scaling distributed machine learning with the parameter server," in *11th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 14)*, pp. 583–598, 2014.

[136] M. J. Sheller, B. Edwards, G. A. Reina, J. Martin, S. Pati, A. Kotrotsou, M. Milchenko, W. Xu, D. Marcus, R. R. Colen, *et al.*, "Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data," *Scientific reports*, vol. 10, no. 1, pp. 1–12, 2020.

[137] K. Marek, D. Jennings, S. Lasch, A. Siderowf, C. Tanner, T. Simuni, C. Coffey, K. Kieburtz, E. Flagg, S. Chowdhury, *et al.*, "The parkinson progression marker initiative (ppmi)," *Progress in neurobiology*, vol. 95, no. 4, pp. 629–635, 2011.

[138] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

[139] A. Gkillas and A. S. Lalos, "Missing data imputation for multivariate time series in industrial iot: A federated learning approach," in *2022 IEEE 20th International Conference on Industrial Informatics (INDIN)*, pp. 87–94, IEEE, 2022.

[140] T. Li, S. Hu, A. Beirami, and V. Smith, "Ditto: Fair and robust federated learning through personalization," in *International Conference on Machine Learning*, pp. 6357–6368, PMLR, 2021.

[141] F. Faghri, S. H. Hashemi, H. Leonard, S. W. Scholz, R. H. Campbell, M. A. Nalls, and A. B. Singleton, "Predicting onset, progression, and clinical subtypes of parkinson disease using machine learning," *bioRxiv*, p. 338913, 2018.

[142] A. Tuladhar, S. Gill, Z. Ismail, N. D. Forkert, A. D. N. Initiative, *et al.*, "Building machine learning models without sharing patient data: A simulation-based analysis of distributed learning by ensembling," *Journal of biomedical informatics*, vol. 106, p. 103424, 2020.

[143] L. Yue, D. Tian, W. Chen, X. Han, and M. Yin, "Deep learning for heterogeneous medical data analysis," *World Wide Web*, vol. 23, no. 5, pp. 2715–2737, 2020.

[144] A. Denis, L. Mergaert, C. Fostier, I. Cleemput, and S. Simoens, "A comparative study of european rare disease and orphan drug markets," *health Policy*, vol. 97, no. 2-3, pp. 173–179, 2010.

[145] B. P. Danek, M. B. Makarious, A. Dadu, D. Vitale, P. S. Lee, A. B. Singleton, M. A. Nalls, J. Sun, and F. Faghri, "Federated learning for multi-omics: A performance evaluation in parkinson's disease," *Patterns*, vol. 5, no. 3, 2024.

[146] Y. Ma and Z. Zhang, "Travel mode choice prediction using deep neural networks with entity embeddings," *IEEE Access*, vol. 8, pp. 64959–64970, 2020.

[147] T. Kulesza, S. Stumpf, M. Burnett, and I. Kwan, "Tell me more? the effects of mental model soundness on personalizing an intelligent agent," in *Proceedings of the sigchi conference on human factors in computing systems*, pp. 1–10, 2012.

[148] L. Schueller, L. Booth, K. Fleming, and J. Abad, "Using serious gaming to explore how uncertainty affects stakeholder decision-making across the science-policy divide during disasters," *International Journal of Disaster Risk Reduction*, vol. 51, p. 101802, 2020.

[149] V. Jagtap, P. Kulkarni, and P. Joshi, "Uncertainty-based decision support system for gaming applications," *Journal of Intelligent & Fuzzy Systems*, no. Preprint, pp. 1–17, 2023.

[150] C. McKay, "Predicting risk in criminal procedure: actuarial tools, algorithms, ai and judicial decision-making," *Current Issues in Criminal Justice*, vol. 32, no. 1, pp. 22–39, 2020.

[151] L. Zhang, I. Pentina, and Y. Fan, "Who do you choose? comparing perceptions of human vs robo-advisor in the context of financial services," *Journal of Services Marketing*, vol. 35, no. 5, pp. 634–646, 2021.

[152] A. Madani, B. Namazi, M. S. Altieri, D. A. Hashimoto, A. M. Rivera, P. H. Pucher, A. Navarrete-Welton, G. Sankaranarayanan, L. M. Brunt, A. Okrainec, *et al.*, "Artificial intelligence for intraoperative guidance: using semantic segmentation to identify surgical anatomy during laparoscopic cholecystectomy," *Annals of surgery*, 2020.

[153] B. Shneiderman, "Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered ai systems," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 10, no. 4, pp. 1–31, 2020.

[154] N.-n. Zheng, Z.-y. Liu, P.-j. Ren, Y.-q. Ma, S.-t. Chen, S.-y. Yu, J.-r. Xue, B.-d. Chen, and F.-y. Wang, "Hybrid-augmented intelligence: collaboration and cognition," *Frontiers of Information Technology & Electronic Engineering*, vol. 18, no. 2, pp. 153–179, 2017.

[155] C. Abras, D. Maloney-Krichmar, J. Preece, *et al.*, "User-centered design," *Bainbridge, W. Encyclopedia of Human-Computer Interaction. Thousand Oaks: Sage Publications*, vol. 37, no. 4, pp. 445–456, 2004.

[156] T. Araujo, N. Helberger, S. Kruikemeier, and C. H. De Vreese, "In ai we trust? perceptions about automated decision-making by artificial intelligence," *AI & society*, vol. 35, pp. 611–623, 2020.

[157] S. Prabhudesai, L. Yang, S. Asthana, X. Huan, Q. V. Liao, and N. Banovic, "Understanding uncertainty: How lay decision-makers perceive and interpret uncertainty in human-ai decision making," in *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pp. 379–396, 2023.

[158] A. Schepman and P. Rodway, "Initial validation of the general attitudes towards artificial intelligence scale," *Computers in human behavior reports*, vol. 1, p. 100014, 2020.

[159] Y. Liu, A. Mittal, D. Yang, and A. Bruckman, "Will ai console me when i lose my pet? understanding perceptions of ai-mediated email writing," in *Proceedings of the 2022 CHI conference on human factors in computing systems*, pp. 1–13, 2022.

[160] S. Guo, F. Du, S. Malik, E. Koh, S. Kim, Z. Liu, D. Kim, H. Zha, and N. Cao, "Visualizing uncertainty and alternatives in event sequence predictions," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2019.

[161] C. J. Cai, E. Reif, N. Hegde, J. Hipp, B. Kim, D. Smilkov, M. Wattenberg, F. Viegas, G. S. Corrado, M. C. Stumpe, *et al.*, "Human-centered tools for coping with imperfect algorithms during medical decision-making," in *Proceedings of the 2019 chi conference on human factors in computing systems*, pp. 1–14, 2019.

[162] Y. Kwak, J.-W. Ahn, and Y. H. Seo, "Influence of ai ethics awareness, attitude, anxiety, and self-efficacy on nursing students' behavioral intentions," *BMC nursing*, vol. 21, no. 1, pp. 1–8, 2022.

[163] J. Mackinlay, "Automating the design of graphical presentations of relational information," *Acm Transactions On Graphics (Tog)*, vol. 5, no. 2, pp. 110–141, 1986.

[164] N. Boukhelifa, A. Bezerianos, T. Isenberg, and J.-D. Fekete, "Evaluating sketchiness as a visual variable for the depiction of qualitative uncertainty," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2769–2778, 2012.

[165] D. Borland and R. M. T. Ii, "Rainbow color map (still) considered harmful," *IEEE computer graphics and applications*, vol. 27, no. 2, pp. 14–17, 2007.

[166] L. A. Breslow, R. M. Ratwani, and J. G. Trafton, "Cognitive models of the influence of color scale on data visualization tasks," *Human factors*, vol. 51, no. 3, pp. 321–338, 2009.

[167] G. Phillips-Wren, "Ai tools in decision making support systems: a review," *International Journal on Artificial Intelligence Tools*, vol. 21, no. 02, p. 1240005, 2012.

[168] J. M. Malof, M. A. Mazurowski, and G. D. Tourassi, "The effect of class imbalance on case selection for case-based classifiers: An empirical study in the context of medical decision support," *Neural Networks*, vol. 25, pp. 141–145, 2012.

[169] E. I. Papageorgiou, A. T. Markinos, and T. A. Gemtos, "Fuzzy cognitive map based approach for predicting yield in cotton crop production as a basis for decision support system in precision agriculture application," *Applied Soft Computing*, vol. 11, no. 4, pp. 3643–3657, 2011.

[170] A. Monteserin and A. Amandi, "Argumentation–based negotiation planning for autonomous agents," *Decision Support Systems*, vol. 51, no. 3, pp. 532–548, 2011.

[171] N. Taghezout and P. Zaraté, "An agent-based simulation approach in an idss for evaluating performance in flow-shop manufacturing system," *Intelligent Decision Technologies*, vol. 5, no. 3, pp. 273–293, 2011.

[172] V. Lai, C. Chen, Q. V. Liao, A. Smith-Renner, and C. Tan, "Towards a science of human-ai decision making: a survey of empirical studies," *arXiv preprint arXiv:2112.11471*, 2021.

[173] L. Bellaiche, R. Shahi, M. H. Turpin, A. Ragnhildstveit, S. Sprockett, N. Barr, A. Christensen, and P. Seli, "Humans versus ai: whether and why we prefer human-created compared to ai-created artwork," *Cognitive Research: Principles and Implications*, vol. 8, no. 1, pp. 1–22, 2023.

[174] R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An integrative model of organizational trust," *Academy of management review*, vol. 20, no. 3, pp. 709–734, 1995.

[175] N. Gurney, D. V. Pynadath, and N. Wang, "Measuring and predicting human trust in recommendations from an ai teammate," in *International Conference on Human-Computer Interaction*, pp. 22–34, Springer, 2022.

[176] D. H. McKnight, V. Choudhury, and C. Kacmar, "Developing and validating trust measures for e-commerce: An integrative typology," *Information systems research*, vol. 13, no. 3, pp. 334–359, 2002.

[177] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an empirically determined scale of trust in automated systems," *International journal of cognitive ergonomics*, vol. 4, no. 1, pp. 53–71, 2000.

[178] M. Greis, P. E. Agroudy, H. Schuff, T. Machulla, and A. Schmidt, "Decision-making under uncertainty: How the amount of presented uncertainty influences user behavior," in *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*, pp. 1–4, 2016.

[179] L. Fell, A. Gibson, P. Bruza, and P. Hoyte, "Human information interaction and the cognitive predicting theory of trust," in *Proceedings of the 2020 conference on human information interaction and retrieval*, pp. 145–152, 2020.

[180] M. Tanveer, B. Richhariya, R. U. Khan, A. H. Rashid, P. Khanna, M. Prasad, and C.-T. Lin, "Machine learning techniques for the diagnosis of alzheimer's disease: A review," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 16, no. 1s, pp. 1–35, 2020.

[181] Q. Zhao, H. Xu, J. Li, F. A. Rajput, and L. Qiao, "The application of artificial intelligence in alzheimer's research," *Tsinghua Science and Technology*, vol. 29, no. 1, pp. 13–33, 2023.

[182] H. Rusinek, M. J. De Leon, A. E. George, L. A. Stylopoulos, R. Chandra, G. Smith, T. Rand, M. Mourino, and H. Kowalski, "Alzheimer disease: measuring loss of cerebral gray matter with mr imaging.," *Radiology*, vol. 178, no. 1, pp. 109–114, 1991.

[183] "Alzheimer's disease neuroimaging initiative." http://adni.loni.usc.edu./. Accessed: 2024-03-21.

[184] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?," *Advances in neural information processing systems*, vol. 30, 2017.

[185] S. F. Eskildsen, P. Coupé, D. García-Lorenzo, V. Fonov, J. C. Pruessner, D. L. Collins, A. D. N. Initiative, *et al.*, "Prediction of alzheimer's disease in subjects with mild cognitive impairment from the adni cohort using patterns of cortical thinning," *Neuroimage*, vol. 65, pp. 511–521, 2013.

[186] U. Bhatt, J. Antorán, Y. Zhang, Q. V. Liao, P. Sattigeri, R. Fogliato, G. Melançon, R. Krishnan, J. Stanley, O. Tickoo, *et al.*, "Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 401–413, 2021.

[187] M. Eslami, S. Tabarestani, and M. Adjouadi, "A unique color-coded visualization system with multimodal information fusion and deep learning in a longitudinal study of alzheimer's disease," *Artificial Intelligence in Medicine*, vol. 140, p. 102543, 2023.

[188] S. G. Sutton, V. Arnold, and M. Holt, "How much automation is too much? keeping the human relevant in knowledge work," *Journal of emerging technologies in accounting*, vol. 15, no. 2, pp. 15–25, 2018.

[189] O. Wysocki, J. K. Davies, M. Vigo, A. C. Armstrong, D. Landers, R. Lee, and A. Freitas,

"Assessing the communication gap between ai models and healthcare professionals: Explainability, utility and trust in ai-driven clinical decision-making," *Artificial Intelligence*, vol. 316, p. 103839, 2023.

[190] W. Berger, H. Piringer, P. Filzmoser, and E. Gröller, "Uncertainty-aware exploration of continuous parameter spaces using multivariate prediction," in *Computer Graphics Forum*, vol. 30, pp. 911–920, Wiley Online Library, 2011.

[191] A. Chatzimparmpas, R. M. Martins, I. Jusufi, K. Kucher, F. Rossi, and A. Kerren, "The state of the art in enhancing trust in machine learning models with the use of visualizations," in *Computer Graphics Forum*, vol. 39, pp. 713–756, Wiley Online Library, 2020.

[192] A. Kamal, P. Dhakal, A. Y. Javaid, V. K. Devabhaktuni, D. Kaur, J. Zaientz, and R. Marinier, "Recent advances and challenges in uncertainty visualization: a survey," *Journal of Visualization*, vol. 24, no. 5, pp. 861–890, 2021.

[193] J. Zhao, Y. Wang, M. V. Mancenido, E. K. Chiou, and R. Maciejewski, "Evaluating the impact of uncertainty visualization on model reliance," *IEEE Transactions on Visualization and Computer Graphics*, 2023.

[194] J. M. Kniss, "Managing uncertainty in visualization and analysis of medical data," in *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 832–835, IEEE, 2008.

[195] C. Lundström, P. Ljung, A. Persson, and A. Ynnerman, "Uncertainty visualization in medical volume rendering using probabilistic animation," *IEEE transactions on visualization and computer graphics*, vol. 13, no. 6, pp. 1648–1655, 2007.

[196] Q. Yang, J. Zimmerman, A. Steinfeld, L. Carey, and J. F. Antaki, "Investigating the heart pump implant decision process: opportunities for decision support tools to help," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 4477–4488, 2016.

[197] C. Gillmann, D. Saur, T. Wischgoll, and G. Scheuermann, "Uncertainty-aware visualization in medical imaging-a survey," in *Computer Graphics Forum*, vol. 40, pp. 665–689, Wiley Online Library, 2021.

[198] D. Weiskopf, "Uncertainty visualization: Concepts, methods, and applications in biological data visualization," *Frontiers in Bioinformatics*, vol. 2, p. 793819, 2022.

[199] J. Schoonenboom and R. B. Johnson, "How to construct a mixed methods research design," *Kolner Zeitschrift fur Soziologie und Sozialpsychologie*, vol. 69, no. Suppl 2, p. 107, 2017.

[200] J. Seymour, "Why does the cielab a* axis point toward magenta instead of red?," *Color Research & Application*, vol. 45, no. 6, pp. 1040–1054, 2020.

[201] J. Hullman, "Why authors don't visualize uncertainty," *IEEE transactions on visualization and computer graphics*, vol. 26, no. 1, pp. 130–139, 2019.

[202] P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes," in *International conference on the theory and applications of cryptographic techniques*, pp. 223–238, Springer, 1999.

[203] "Parkinson's disease biomarkers program." https://pdbp.ninds.nih.gov/. Accessed: 2024-06-09.