Mohammad Mahdi **Maheri**

☎ (+44) 753 324468 | ✉ m.maheri23@imperial.ac.uk | mammadmaheri7

SUMMARY

I am a third-year PhD student in computer science at Imperial College London under Hamed Haddadi, specializing in privacy and security for user-centered systems. My research focuses on designing frameworks for **data privacy on edge devices**, exploring areas like **differential privacy, zero-knowledge proofs, personalized machine learning, and generative models**. I am particularly interested in assessing risks of deploying large language models on edge devices to mitigate **AI risks** and promote **trustworthy machine learning**.

Education

Imperial College London
London, United kingdom

PhD in privacy and security of user-centered systems
2023 - now

- Awarded a full scholarship based on academic excellence and research potential
- Researching about Machine learning Privacy and Security
- Developed expertise in Zero-knowledge Proof and Differential Privacy
- Supervised by **Prof. Hamed Haddadi**.

Sharif University of Technology
Tehran, Iran

M.S. in Software Engineering
2020 - 2023

- Graduated with First Rank, M.S. GPA: 20/20, Sharif University of Technology

Shahid Beheshti University
Tehran, Iran

B.S. in Computer Engineering
2016 - 2020

- Graduated with First Rank, B.Sc. GPA: 19.07/20, Shahid Beheshti University
- Appreciated as the best student among computer engineering students
- I got a Direct master's offer from Sharif University of Technology (The best Engineering University in Iran)

Solaha High School
Tehran, Iran

DIPLOMA IN MATHEMATICS AND PHYSICS DISCIPLINE
2012 - 2016

- GPA (19.83 / 20)

Publications

WARP: Weight Teleportation for Attack-Resilient Unlearning Protocols.
Mohammad M Maheri*, Xavier Cadet, Peter Chin, Hamed Haddadi*. The Fourteenth International Conference on Learning Representations (ICLR), 2026. [arXiv]

ZK-APEX: Zero-Knowledge Approximate Personalized Unlearning with Executable Proofs.
Mohammad M Maheri*, Sunil Cotterill, Alex Davidson, Hamed Haddadi*. 9th Annual Conference on Machine Learning and Systems (MLSys) 2026. [arXiv]

TeleSparse: Practical Privacy-Preserving Verification of Deep Neural Networks.
Mohammad M Maheri*, Hamed Haddadi, Alex Davidson*. Proceedings on Privacy Enhancing Technologies (PoPETs), 2025. [DOI] [arXiv] [Code]

Client Clustering Meets Knowledge Sharing: Enhancing Privacy and Robustness in Personalized Peer-to-Peer Learning.
Mohammad M Maheri*, Denys Herasymuk, Hamed Haddadi*. IEEE Annual Congress on Artificial Intelligence of Things (IEEE AIoT), 2025. [arXiv] [Code]

Verifiable Unlearning on Edge.
Mohammad M Maheri*, Alex Davidson, Hamed Haddadi*. Poster, IEEE European Symposium on Security and Privacy (EuroS&P), 2025. [DOI] [arXiv]

An Early Experience with Confidential Computing Architecture for On-Device Model Protection.
Sina Abdollahi, ***Mohammad M Maheri****, Sandra Siby, Marios Kogias, Hamed Haddadi*. Accepted to the 8th Workshop on System Software for Trusted Execution (SysTEX), 2025. [arXiv] [Code]

GuardNet: Graph-Attention Filtering for Jailbreak Defense in Large Language Models.
Javad Forough, ***Mohammad M Maheri****, Hamed Haddadi*. Preprint, 2025. [arXiv]

GuaranTEE: Towards Attestable and Private ML with CCA.
*Sandra Siby, Sina Abdollahi, **Mohammad M Maheri**, Marios Kogias, Hamed Haddadi*. EuroMLSys (Workshop on Machine Learning and Systems, co-located with EuroSys), 2024. [DOI] [arXiv] [Code]

P4: Towards Private, Personalized, and Peer-to-Peer Learning.
***Mohammad M Maheri**, Sandra Siby, Sina Abdollahi, Anastasia Borovykh, Hamed Haddadi*. Preprint, 2024. [arXiv]

Privacy Challenges in Meta-Learning: An Investigation on Model-Agnostic Meta-Learning.
*Mina Rafiei, **Mohammad M Maheri**, Hamid R. Rabiee*. Preprint, 2024. [arXiv]

ClusterSeq: Enhancing Sequential Recommender Systems with Clustering based Meta-Learning.
***Mohammad M Maheri**, Reza Abdollahzadeh, Bardia Mohammadi, Mina Rafiei, Jafar Habibi, Hamid R. Rabiee*. Preprint, 2023. [arXiv]

RESEARCH EXPERIENCE

WARP: Weight Teleportation for Attack-Resilient Unlearning Protocols [arXiv]

2025

- Developed unlearning-specific privacy audits, including MIA and DRA, to quantify residual leakage after unlearning.
- Proposed *WARP*, a plug-in teleportation **defense** using prediction-preserving neural network symmetries to improve privacy
- Reduced attacker advantage substantially, with up to **64%** lower black-box AUC and **92%** lower white-box AUC across **six** unlearning algorithms.
- Provided an **information-theoretic** analysis supporting the privacy gains from symmetry-based perturbations.
- Supervised by **Prof. Hamed Haddadi** and **Prof. Peter Chin**.

ZK-APEX: Zero-Knowledge Approximate Personalized Unlearning with Executable Proofs [arXiv]

2025

- **Zero-shot** personalized unlearning via sparse masking with curvature-aware, closed-form (Group-OBS) compensation.
- Designed and implemented a Halo2 **ZK-SNARK** that certifies unlearning *operator compliance* via linear KKT-style certificates while preserving privacy of client data and personalized parameters.
- Reduced proving cost: up to $\sim$**10**$^7\times$ faster and $\sim$**350**$\times$ lower peak memory vs. optimization-based ZK baselines.
- Provided theoretical support, including a lower-bound analysis for forget-set loss increase and the optimality/uniqueness argument
- Supervised by **Prof. Hamed Haddadi** and **Prof. Alex Davidson**.

TeleSparse: Practical Privacy-Preserving Verification of Deep Neural Networks [Paper]

2024

- Designed a ZK-SNARK-friendly verification framework that leverages sparsity and neural network symmetries to reduce proving overhead while preserving model utility.
- Implemented the full pipeline in **Halo2** and evaluated across multiple architectures and datasets to validate practicality at scale.
- Achieved **67% memory reduction** and **46% faster** proof generation with negligible accuracy degradation.
- Supervised by **Prof. Hamed Haddadi** and **Prof. Alex Davidson**.

GuaranTEE: Towards Attestable and Private ML with CCA [Paper]

2024

- Co-developed *GuaranTEE*, a framework for **attestable** and **privacy-preserving** ML inference on edge devices.
- Integrated Arm **Confidential Computing Architecture (CCA)** to provide hardware-backed isolation and secure execution.
- Built and released a working prototype demonstrating feasibility on resource-constrained platforms.

Client Clustering Meets Knowledge Sharing: Enhancing Privacy and Robustness in Personalized Peer-to-Peer Learning [Paper]

2023–2024

- Developed a privacy-preserving personalized P2P learning framework with **differential privacy** guarantees.
- Proposed a lightweight decentralized clustering mechanism to handle **client heterogeneity** and improve personalization.
- Implemented differentially private knowledge distillation for co-training with minimal utility loss.
- Demonstrated up to **40%** accuracy improvement over strong baselines and validated deployment on **Raspberry Pi**.

GuardNet: Graph-Attention Filtering for Jailbreak Defense in Large Language Models [arXiv]

2025 (preprint)

- Developed a pre-inference **LLM jailbreak** defense based on **graph-attention** filtering.
- Implemented **hierarchical** filtering with prompt-level detection and token-level localization for selective masking.

Privacy Challenges in Meta-Learning: An Investigation on Model-Agnostic Meta-Learning [arXiv]

2022–2023

- Developed a **membership inference attack** framework for meta-learning, characterizing privacy leakage in MAML-style pipelines.
- Studied mitigation strategies (e.g., differential privacy) and analyzed their privacy–utility trade-offs in the meta-learning setting.

ClusterSeq: Enhancing Sequential Recommender Systems with Clustering-Based Meta-Learning [arXiv]

2021–2023

- Developed a clustering-based meta-learning approach for sequential recommendation to address the **cold-start** problem.
- Combined **meta-learning** with adversarial learning to improve **personalization** under sparse, non-i.i.d. user histories.
- Improved recommendation performance in data-efficient settings through few-shot adaptation.

Human Activity Recognition

2020

- Built smartphone-sensor activity classification pipelines using classical ML (scikit-learn) with careful preprocessing and evaluation.
- Designed hardware-friendly feature sets via linear statistical features and feature selection.
- Benchmarked computational costs across model families and compared accuracy–efficiency trade-offs.

Professional Experience

Fanavaran Mosbat Company

Tehran, Iran

SOFTWARE ENGINEER & SOFTWARE ARCHITECTURE — *2019 - 2020*

- Member of web DevOps team
- Research about software testing
- Research about modern software architecture such as micro-services
- Backend developer of some web app projects, using Laravel and Vue.js
- Handling some tasks of customer management by NoSQL databases

Achar-Farance Startup

Tehran, Iran

CO-FOUNDER & SOFTWARE ENGINEER — *2018 - 2019*

- Implemented RESTful API server for freelance services application
- Design and implement SQL database
- developed as web application by Laravel and Bootstrap

Teaching Experience

TEACHING ASSISTANT-SHIP

Deep Learning, Prof.Kainz	*Spring 2025*
Internet of Things and Applications, Prof.Haddadi	*Fall 2024*
Stochastic Process, Dr.Rohban	*Spring 2022*
Linear Algebra, Prof.Rabiee	*Fall 2021, Spring 2022*
Artificial Intelligence, Dr.Rohban	*Fall 2021*
Computational Intelligence, Dr.Malek	*Fall 2020*
Soft/Hard Codesign, Dr.Mahdiani	*Spring 2020, Fall 2020*
Design and Analysis of Algorithms, Dr.Ghavamizadeh	*Spring 2020*
Theory of Languages and Machines, Dr.Ghavamizadeh	*Spring 2018,Fall 2018,Fall 2019*
Advanced Programming, Dr.Vahidi	*Fall 2017*

OTHER

Shahid Beheshti University, Organizing ”Motion Graphic” workshop	*Spring 2017*

Skills

• Technical (proficient):	Python, PyTorch, Scikit-learn, PHP, Laravel, , C/C++, Java, Git, Design patterns, Verilog, VHDL
• Technical (familiar):	Databases, Docker, microservices, software testing, Apache server, Rust, Node.js, Python, JavaScript
• Theoretical (familiar):	Machine Learning, Deep Learning, NLP, LLMs, Algorithm Design, Distributed Databases, Statistical Analysis
• Languages:	Persian (mother tongue), English (upper-intermediate)

Notable Projects

Expert Finding in Stackoverflow

RECOMMENDER SYSTEM FOR SUGGESTING MOST SUITABLE PERSON TO ANSWER THE QUESTION — *2021*

- Developed by Tensorflow
- Using deep learning approaches for solving the problem

Offensive Detection

CLASSIFICATION OF OFFENSIVE TWEETS *2020*

- Detecting offensive tweets crawled from twitter
- Modeled by recurrent Neural Networks and Naive Bayes models (using skit-learn, Keras)
- Using NLP techniques such as word embedding and transfer learning
- Achieved 94% accuracy in identifying offensive content.

Proposal Manager

SYSTEM FOR MANAGE UPLOAD AND ARBITRATION OF STUDENT'S PROPOSALS *2020*

- Developed by Agile methodology
- Android application with Laravel backend deployed on Heroku

Crisis Manager

WEB APPLICATION FOR MANAGE AND RECORD REQUIRED EMERGENCIES STUFF *2019*

- Developed with React and Node.js
- CI/CD pipeline implemented
- Deployed on Heroku

Tic-tac-toe player

INTELLIGENT PLAYER OF TIC-TAC-TOE GAME (USING GENETIC ALGORITHM) *2018*

- Able to prevent losing in more than 99 percentages of different games

P++ Compiler

COMPILE SEMI C++ PROGRAM TO LLVM (USING JAVA, JFLEX, CUP PARSER) *2017*

- Support all C++ features
- Generate Optimized machine code from LLVM code

Instagram Simulator

WINDOWS APPLICATION LIKE INSTAGRAM (USING JAVA, SOCKET PROGRAMMING) *2016*

- Client-Server architecture
- Implement by Java
- Socket programming used to communicate with other nodes