Jon Vadillo

ML Safety Researcher & Assistant Professor | University of the Basque Country (UPV/EHU)

jon.vadillo@ehu.eus Personal Website

Education

Bachelor's Degree in Computer Engineering — Graduation date: 07/16/2018
University of the Basque Country (UPV/EHU) — *San Sebastian, Gipuzkoa, Spain*

- Extraordinary Award to the Best Academic Record.
- **Average grade: 9.52/10**.
- 24 courses passed with Honors.
- Final project passed with Honors.

Master's Degree in Computational Engineering and Intelligent Systems — Graduation date: 10/17/2019
University of the Basque Country (UPV/EHU) — *San Sebastian, Gipuzkoa, Spain*

- **Average grade: 9.81/10**.
- Master's Thesis 'Universal Adversarial Examples in Speech Command Classification' graded 10/10 with Honors.

Ph.D. in Computer Science — Graduation date: 01/13/2023
University of the Basque Country (UPV/EHU) — *San Sebastian, Gipuzkoa, Spain*

- Thesis title: **"Broadening the Horizon of Adversarial Attacks in Deep Learning"**.
- Grade: Cum Laude.
- International PhD Mention.
- Four-month research stay at the **University of Oxford, UK**.

Publications

Vadillo, J., Santana, R., and Lozano, J. A. (2025). **Adversarial Attacks in Explainable Machine Learning: A Survey of Threats Against Models and Humans**. WIREs Data Mining Knowledge Discovery, 15(1):107719. DOI: 10.1002/widm.1567.

Vadillo, J., Santana, R., Lozano, J. A., and Kwiatkowska, M. (2024). **Uncertainty-Aware Explanations Through Probabilistic Self-Explainable Neural Networks**. arXiv preprint arXiv:2403.13740. Online: https://arxiv.org/abs/2403.13740.

Vadillo, J., Santana, R., and Lozano, J. A. (2023). **Extending Adversarial Attacks to Produce Adversarial Class Probability Distributions**. Journal of Machine Learning Research (JMLR), 24(15):1–42. **Featured at ICML 2024**.
Online: https://jmlr.org/papers/v24/21-0326.html.

Vadillo, J., Santana, R., and Lozano, J. A. (2022). **Analysis of Dominant Classes in Universal Adversarial Perturbations**. Knowledge-Based Systems, 236:107719. DOI: 10.1016/j.knosys.2021.107719.

Vadillo, J., and Santana, R. (2022). **On the Human Evaluation of Universal Audio Adversarial Perturbations**. Computers & Security, 112:102495. DOI: 10.1016/j.cose.2021.102495.

Vadillo, J., and Santana, R. (2021). **Universal Adversarial Examples in Speech Command Classification**. Proceedings of the XIX Conference of the Spanish Association for Artificial Intelligence (CAEPIA), pages 642-647. ISBN: 978-84-09-30514-8. Online: https://caepia20-21.uma.es/proceedings.html.

Garciarena, U., Vadillo, J., Mendiburu, A., Santana, R. (2021). **Adversarial Perturbations for Evolutionary Optimization**. Machine Learning, Optimization, and Data Science (LOD), vol. 13164, pages 408-422. DOI: 10.1007/978-3-030-95470-3_31.

Vadillo, J., Santana, R., Lozano, J. A. (2020). **Exploring Gaps in DeepFool in Search of More Effective Adversarial Perturbations**. Machine Learning, Optimization, and Data Science (LOD), vol. 12566, pages 215-227. DOI: 10.1007/978-3-030-64580-9_18.

Villalobos, K., Vadillo, J., Diez, B., Calvo, B., and Illarramendi, A. (2018). **I4TSPS: a Visual-Interactive Web System for Industrial Time-Series Pre-processing**. Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), pages 2012-2018. DOI: 10.1109/BigData.2018.8621887.

Invited Talks

Fifth Bilbao Data Science Workshop (BIDAS) 06/09/2023
Basque Center for Applied Mathematics (BCAM) *Bilbao, Spain*
- Title: *Adversial Examples in Deep Learning - Introduction and Recent Trends.*

Data Science and Artificial Intelligence Institute (DATAI) 04/26/2023
University of Navarra (UN) *Pamplona, Navarra, Spain*
- Title: *Broadening the Horizon of Adversarial Attacks in Deep Learning.*

Outreach and Dissemination

Vadillo, J., Santana, R., and Lozano, J. A. **Do You Trust Your Speech Recognition System? Fooling Deep Learning with (Inaudible) Audio Perturbations**. SIAM News.
- Dissemination article about the research carried out during my PhD, published in the official newsjournal of the Society for Industrial and Applied Mathematics (SIAM).

Research Projects

Project: A Computational Intelligence Approach to Insurance and Accident Data Processing *University of the Basque Country (UPV/EHU)* 12/30/2016 - 12/31/2020
- Funding granted: 227,117.00€.

Project: Comprehensive Advanced Data Analytics Solution for Optimal Support and Management *LKS S.Coop.* 09/03/2018 - 12/31/2018
- Funding granted: 229,800.00€.

Project: Surgical Block Leveling Module 09/03/2018 - 12/31/2019
LKS S.Coop.
- Funding granted: 238,000.00€.

Consolidated Group Support: Intelligent Systems Group 01/01/2022 - 12/31/2025
University of the Basque Country (UPV/EHU)
- Funding granted: 274,000.00€.

Consolidated Group Support: Intelligent Systems Group 01/01/2019 - 12/31/2021
University of the Basque Country (UPV/EHU)
- Funding granted: 170,000.00€.

Project: Cognitive Mechatronics for the Design of Industrial Machines 03/07/2024 - 12/31/2025
University of the Basque Country (UPV/EHU)
- Funding granted: 72,472€.
- Principal Investigator.

Work Experience

University of the Basque Country (UPV/EHU) 09/01/2023 – Currently
Assistant Professor *San Sebastian, Gipuzkoa, Spain*
- Faculty of Computer Science.
- Department of Computer Science and Artificial Intelligence.

LKS S.Coop. 09/03/2018 – 01/10/2020
Research Consultant *San Sebastian, Gipuzkoa, Spain*
- Participation in research projects related to advanced data analytics, machine learning and optimization in healthcare domains.

Savvy Data Systems 06/06/2017 – 09/18/2017
Internship *San Sebastian, Gipuzkoa, Spain*
- Internship (300 hours) in which I took part in projects related to the application of machine learning and business intelligence techniques for industrial data.
- Internship carried out during the third year of my Bachelor's Degree in Informatic Engineering, University of the Basque Country (UPV/EHU).

Grants and Scholarships

Postdoctoral Research Grant 01/14/2023 – 08/31/2023

Ministry of Science, Innovation and Universities - Spanish Government

- Postdoctoral Orientation Period. Reference: FPU19/03231 (extension).

Predoctoral Research Grant 11/01/2020 – 01/13/2023

Ministry of Science, Innovation and Universities - Spanish Government

- Reference: FPU19/03231.

Predoctoral Research Grant 01/20/2020 – 10/31/2020

Basque Government

- Reference: PRE_2019_1_0128.
- Voluntarily replaced on 11/01/2020 in favour of a FPU Predoctoral Research Grant awarded by the Ministry of Science, Innovation and Universities, Spanish Government (FPU19/03231).

Ikasiker University Scholarship Plan 12/05/2017 – 07/16/2018

Basque Government

- Scholarship awarded by the Basque Government for collaboration in Research Groups of the Basque University System. Collaboration carried out in the Interoperable Databases Group (BDI), Faculty of Computer Science, University of the Basque Country (UPV/EHU). Reference: IkasC_2017_1_0117.

Other Awards and Mentions

- **First Award to the Best PhD Project**
  Doctoral Consortium of the XIX Conference of the Spanish Association for Artificial Intelligence, 2021.
- **Third Best Paper Award**
  X Symposium of Theory and Applications of Data Mining - XIX Conference of the Spanish Association for Artificial Intelligence, 2021. Paper: "Universal Adversarial Examples in Speech Command Classification".
- **Extraordinary Award to the Best Academic Record of the Degree in Computer Engineering (Academic Year: 2017/18)**
  University of the Basque Country (UPV/EHU).
- **Honorable Mention in the International Competitive Programming Contest SWERC 2017 (Paris, France)**
  Participant representing the University of the Basque Country (UPV/EHU).

Teaching

Teaching conducted at the Faculty of Informatics of the University of the Basque Country (UPV/EHU) since the 2020-21 academic year. All courses taught are associated to the Department of Computer Science and Artificial Intelligence.

2020-21 Academic year

- **Search Heuristics** (Rating received from the students: 4.7 / 5).

2021-22 Academic year

- **Advanced Statistical Methods** (Rating received from the students: 4.3 / 5).
- **Operations Research** (Rating received from the students: 4.6 / 5).
- **Search Heuristics** (Rating received from the students: 4.5 / 5).

2022-23 Academic year

- **Advanced Machine Learning** (Rating received from the students: 4.8 / 5).

2023-24 Academic year

- **Advanced Statistical Methods** - Spanish (Rating received from the students: 4.9 / 5).
- **Advanced Statistical Methods** - Basque (Rating received from the students: 4.7 / 5).
- **Statistical Methods in Engineering** (Rating received from the students: 4.6 / 5).
- **Visualization and Virtual Environments** (Rating received from the students: 4.8 / 5).