MODELLO PER LA PRESENTAZIONE DEL PROGETTO DI RICERCA

a) Titolo del progetto

Enhancing software fairness in binary and multi-class classification tasks

b) Proponente (PI)

Antinisca Di Marco

c) Posizione accademica del proponente

Professoressa Associata in Informatica presso il Dipartimento DISIM

d) Curriculum vitae del proponente (max 5000 caratteri – circa 2 pagine) con elenco delle pubblicazioni più significative (max 10) nel periodo 2021-25, relative al tema del progetto. Indicatori ASN 2024/26 alla scadenza del bando (solo per i progetti di ricerca) relativamente alla fascia superiore del Settore concorsuale e del Settore scientifico disciplinare di appartenenza.

Antinisca Di Marco is an Associate Professor in Computer Science at the University of L'Aquila, L'Aquila, Italy, since October 2017. She holds and has held many institutional roles, the most relevant also regarding the pro:

- Vice-president of Multidisciplinary Internal Review Board, University of L'Aquila (since March 2025)
- Ideator and Coordinator of PINKAMP, a summer school to attract high school girls in STEM studies (2018-present). (www.pinkamp.disim.univaq.it)
- Representative of the University of L'Aquila in the Assembly of the European Open Science Cloud Association (since March 2024).
- Member of the Scientific Advisory Board of CeTemps Center of Excellence, University of L'Aquila, Italy. Decreto Rettorale Rep. n. 1090 prot. n. 105623 del 26.09.2023 – extended until to June 30th 2026 – Decreto Rettorale prot. n.0171543 del 29.10.2025
- Equal opportunities representative of the DISM department (from Oct.24 to Sept. 25)
- Director of the University of L'Aquila Node of the InfoLife CINI Laboratory. The laboratory deals with Bioinformatics and Systems Biology research topics (since 2015)
- Member of the Organizing Committee of the Professional Master "Post-disaster technical-administrative management in local authorities" of University of L'Aquila (from 2019 to 2022)
- Coordinator of the Committee designing the new interdepartmental Master Degree in Applied Data Science (LM Data Science), University of L'Aquila, Italy.

MODELLO PER LA PRESENTAZIONE DEL PROGETTO DI RICERCA

▪ Head of the Master Study Programme in Applied Data Science, University of L'Aquila (Oct2018-Oct2021).

She received his PhD from the University of L'Aquila (Italy) in June 2005. She spent 6 monhs at University College London as Visiting Researcher from March to August 2005. Her research interests fall into the broad area of data science and software engineering, applying them to multidisciplinary contexts (such as disaster recovery and resiliency, health, bioinformatics, digital twin, equal opportunity) having a specific focus on quality aspects (e.g., AI fairness, systems sustainability, software performance).
She has published in top-tier conferences and journals, such as ICSE, SANER, JSS, IST, and IEEE TSE. Her work is centred on real-world problems that require software engineering, software quality engineering, AI fairness, digital twins, applied AI, modeling and simulation.
Moreover, she has been involved in the organisation of top-tier and highly relevant international conferences (ICPE, FSE, PERCOM, and ASE)

RESEARCH PROJECTS COORDINATION. She is and has been:
▪Univaq Scientific Coordinator for DIRECT: Digital twIns foR EmergenCy support - Innovation Grant (18 months, € 57.500,00 for UNIVAQ)
▪Scientific coordinator of the Software Infrastructure, Bioinformatics and eHealth research unit of UNIVAQ for LIFEMAP: DALLA PATOLOGIA PEDIATRICA ALLE MALATTIE CARDIOVASCOLARI E NEOPLASTICHE NELL'ADULTO project (Feb. 2023- Jan. 2026, € 1,092,000 for UNIVAQ): Italian project Call T3-AN-14
▪Co-Principal Investigator (Co-PI) and Co-Coordinator: "SoBigData.it - Strengthening the Italian RI for Social Mining and Big Data Analytics" (01.11.2022 – 28.02.2026, €3.5M for UNIVAQ) – Decreto Direttoriale n. 107 del 20.06.2022 - Avviso n. 3264 del 28/12/2021 "Rafforzamento e creazione di IR nell'ambito del PNRR".
▪Leader of WP5: Analysis of citizen preparedness and ecosystems resilience to disasters of Spoke 5 on Environment and natural disaster of the "National Centre for HPC, Big Data and Quantum Computing" Research program CN00000013, PNRR MUR – M4C2 – Investimento 1.4 - Avviso "Centri Nazionali" - D.D. n. 3138 del 16 dicembre 2021
▪Co-PI and Unit Coordinator: SoBigData RI PPP- EU project - Call HORIZON-INFRA-2021-DEV-02 (grant n. 101079043, 01.10.2022- 30.09.2025, €52K for UNIVAQ)

▪Coordinator of L'Aquila participant (2020-2024) - SoBigData++ EU project - Call H2020-INFRAIA-2018-2020 (grant n. 871042).

▪Co-PI, Scientific Coordinator of research IT infrastructure, leader of WP1 and WP2, member of the Scientific Board and of the operational technical committee of the national research project Territori Aperti (Dec. 2018-Dec. 2024, grant of 1.8M - www.territoriaperti.univaq.it). Leader of the team that developed the disaster preparedness toolkit (Territori Aperti Toolkit – Territori Aperti Toolkit).

SELECTED PUBLICATIONS

- **How do generative models draw a software engineer? An empirical study on implicit bias of open-source image generation models.** G. d'Aloisio, T. Fadahunsi, A. Di Marco, F. Sarro, Information and Software Technology, Volume 190, 2026, https://doi.org/10.1016/j.infsof.2025.107956.
- **Towards early detection of algorithmic bias from dataset's bias symptoms: An empirical study,** G. d'Aloisio, C. Di Sipio, A. Di Marco, D. Di Ruscio, Information and Software Technology, Volume 188, 2025, https://doi.org/10.1016/j.infsof.2025.107905.
- **How fair are we? From conceptualization to automated assessment of fairness definitions.** d' Aloisio, G., Di Sipio, C., Di Marco, A. et al., Softw Syst Model (2025). https://doi.org/10.1007/s10270-025-01277-2
- **Uncovering gender gap in academia: A comprehensive analysis within the software engineering community** A D'Angelo, G d'Aloisio, F Marzi, A Di Marco, G Stilo, Journal of Systems and Software 217, 112162
- **Debiaser for Multiple Variables to enhance fairness in classification tasks**, G. d'Aloisio, A. D'Angelo, A. Di Marco, G. Stilo, Information Processing & Management, Volume 60, Issue 2, 2023, https://doi.org/10.1016/j.ipm.2022.103226.

Full list of publications: https://scholar.google.it/citations?user=QVzuSyIAAAAJ&hl=it

[Software] MANILA https://sobigdata.d4science.org/group/sobigdata.it/manila-univaq

ASN2024/2026 indicators – IRIS simulation on 12/01/2026:

COMMISSARIO	Valore	INDICATORE	Soglia	Stato
	21	Numero articoli ultimi 10 anni	11	✓
	2375	Numero citazioni ultimi 15 anni	391	✓
	15	H index ultimi 15 anni	11	✓
	La simulazione ASN per il ruolo di Commissario ha esito positivo?			SI

MODELLO PER LA PRESENTAZIONE DEL PROGETTO DI RICERCA

e) Eventuali componenti del gruppo di ricerca (solo per i progetti di ricerca di base)

- Giordano d'Aloisio (M) - postdoctoral researcher
- Payel Patra (F) - phd candidate in ICT

f) Settore di ricerca ERC di riferimento per la proposta (indicare anche due sotto-settori)

- PE6 - Computer Science and Informatics: Informatics and information systems, computer science, scientific computing, intelligent systems:
  - PE6_11 Machine learning, statistical data processing and applications using signal processing (e.g., speech, image, video);
  - SH3_2 Inequalities, discrimination, prejudice;

g) Abstract (max 1000 caratteri)

The widespread use of AI and ML models in sensitive areas raises significant concerns about fairness. While the research community has introduced various methods for bias mitigation in binary classification tasks, the issue remains under-explored in multi-class classification settings. To address this limitation, in this project, we aim to address the issue of bias-mitigation in multi-class classification settings.

We investigate how to formulate the problem of fair learning in multi-class classification as a multi-objective problem between effectiveness (i.e., prediction correctness) and multiple linear fairness constraints. Next, we will propose an in-process bias-mitigation algorithm to solve this task with binary and multi-class datasets under multiple fairness definitions. Finally, we will conduct an extensive empirical evaluation of the proposed approach to assess its effectiveness, considering, among others, gender as the primary binary and multi-class variable.

h) Descrizione del progetto (max 8.000 caratteri, compresi eventuali riferimenti bibliografici. È consentito inserire figure nella proposta. Le figure non concorrono alla determinazione del calcolo del numero dei caratteri.)

a. State of the art

With the increasing adoption of AI- and ML-based software systems in sensitive domains such as healthcare, finance, and education, it is critical to ensure that they act in an *unbiased* and *ethical* way. In other words, they must be *fair*. The relevance of software fairness has been highlighted in recent years not only in the research literature [1,2] but also in regulations, such as the European Union's recently introduced AI Act. In general, an ML system is said to be *biased* if it systematically

discriminates or privileges individuals or groups identified by a set of *sensitive features* (e.g., gender, ethnicity, or religion) [1].

To address this issue, the research community has proposed several methods for *bias mitigation* at different processing levels [1,3]. Specifically, there are *pre-processing* methods (i.e., methods that operate on the dataset before training an ML model); in-processing methods (i.e., methods that modify the training behaviour of ML models to make it fairer); and post-processing methods (i.e., methods that modify an already trained ML model).

However, the majority of them can be applied only to binary classification tasks. Instead, several examples of multi-class classification approaches have been applied in sensitive domains such as education [4], food [5], and health [6]. Ensuring that these systems behave in a fair and unbiased way is paramount, also to achieve some of the United Nations Sustainable Development Goals, (SDG), e.g., SDG 2 (zero hunger), SDG 3 (good health and well-being), SDG 4 (quality education), and SDG 5 (gender equality).

One of the first approaches proposed for multi-class bias mitigation is the *Blackbox* post-processing approach by Putzel et al. [7], which extends the *Equalized Odds* binary bias-mitigation algorithm to the multi-class setting. A similar approach is the *Demographic Parity* post-processing approach proposed by Denis et al. [8], where the predictions are instead optimised under the *Demographic Parity* fairness definition.

Quadros et al. [9] adapted a series of pre-processing and post-processing bias mitigation methods for binary classification to the multi-class classification task and tested them on a wage discrimination dataset.

Finally, one of the most recent approaches specifically designed for multi-class bias mitigation is the pre-processing *Debiaser for Multiple Variables (DEMV)* algorithm proposed by d'Aloisio et al. [10]. This algorithm extends the *Sampling* algorithm of Kamiran et al. to the multi-class setting and has been shown to overcome existing bias mitigation methods for multi-class classification.

However, from the review of the state of the art, we observe that no in-processing methods exist for multi-class classification tasks. Indeed, increasing the diversity of bias mitigation methods is needed to effectively address more diverse application contexts (i.e., ML models, datasets, metrics) [11].

b. Objectives

Starting from the review of the state of the art, this project aims to propose an in-processing bias mitigation method for binary and multi-class classification. Specifically, we aim to answer the following research questions:

- **RQ1:** *How can we formulate the problem of fairness in multi-class classification as a multi-objective problem between multiple fairness definitions and prediction effectiveness?*
- **RQ2:** *How can we design an algorithm to solve the problem formulated in RQ1?*
- **RQ3:** *To what extent will the algorithm be able to mitigate bias while keeping a high prediction effectiveness in a multi-class classification context?*
- **RQ4:** *To what extent will the algorithm be able to mitigate bias while keeping a high prediction effectiveness in a binary classification context?*
- **RQ5:** *How will the proposed algorithm compare against existing bias mitigation methods in the multi-class classification tasks?*
- **RQ6:** *How will the proposed algorithm perform in the intersectional-fairness scenario (i.e., with groups identified by multiple sensitive variables)?*

c. Methodology

To answer **RQ1** and **RQ2,** we aim to extend the *Exponentiated Gradient* in-processing algorithm proposed by Agarwal et al. [12]. In their work, the authors formulate a multi-objective optimisation problem to train a binary classifier under specific fairness constraints. Next, they present an Exponentiated Gradient (EG) method to solve this optimisation task. This project will extend the original EG algorithm to the multi-class classification setting and to the simultaneous optimisation of multiple fairness constraints, making it more general and practical for real-world use cases.

To answer **RQ3, RQ4,** and **RQ5**, we will perform an extensive empirical evaluation of the proposed approach, benchmarking it with several datasets, methods and metrics. Specifically, we plan to replicate the evaluation process followed in [10] to evaluate DEMV, which includes a comprehensive collection of datasets (binary and multi-class) and metrics. Eventually, we will benchmark the proposed approach against DEMV and, if available at the time of the evaluation, more recent multi-class bias mitigation methods. Finally, to answer **RQ6**, we will extend the algorithm to the intersectional fairness scenario, enabling optimisation under fairness definitions that simultaneously incorporate multiple sensitive variables. In the empirical evaluation, we will consider *gender* as the primary sensitive variable, intended in both biological terms (for binary classification) and gender identity (for multi-class classification). In the intersectional evaluation, we will complement *gender* with other sensitive variables, such as *ethnicity* or *religion*.

d. Working plan

The duration of the project is 8 months, from May to December 2026.

The working plan is divided as follows:

- T1: Multi-objective problem formulation (needed for **RQ1**)
- T2: Algorithm design and implementation (needed for **RQ2**)
- T3: Algorithm (and intersectional extension) evaluation(needed for **RQ3-4-5**)
- T4: Extension to intersectional fairness (needed for **RQ4-5-6**)

The timeline for each task is detailed in the table below:

	May	**June**	**July**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
T1								
T2								
T3								
T4								

e. References

[1] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, e A. Galstyan, «A Survey on Bias and Fairness in Machine Learning», *ACM Comput. Surv.*, vol. 54, fasc. 6, pp. 1–35, lug. 2021, doi: 10.1145/3457607.

[2] Chen, Z., Zhang, J. M., Hort, M., Harman, M., & Sarro, F. (2024). Fairness testing: A comprehensive survey and analysis of trends. ACM Transactions on Software Engineering and Methodology, 33(5), 1-59.

[3] Hort, M., Chen, Z., Zhang, J. M., Harman, M., & Sarro, F. (2023). Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey. *ACM Journal on Responsible Computing*, 3631326. https://doi.org/10.1145/3631326

[4] Yanes, N., Mostafa, A. M., Ezz, M., and Almuayqil, S. N. (2020). A machine learning-based recommender system for improving students learning experiences. IEEE Access, 8:201218–201235.898

[5] Meenachi, L., Ramakrishnan, S., Sivaprakash, M., Thangaraj, C., and Sethupathy, S. (2022). Multi Class Ensemble Classification for Crop Recommendation. In 2022 International Conference on Inventive Computation Technologies (ICICT), pages 1319–1324. ISSN: 2767-7788.837

[6] Zhang, J., Cao, P., Gross, D. P., and Zaiane, O. R. (2013). On the application of multi-class classification in physical therapy recommendation. Health Information Science and Systems, 1(1):15.904

[7] Putzel, P., & Lee, S. (2022). Blackbox post-processing for multiclass fairness. arXiv preprint arXiv:2201.04461.

[8] Denis, C., Elie, R., Hebiri, M., and Hu, F. (2024). Fairness guarantees in multi-class classification with demographic parity. Journal of Machine Learning Research, 25(130):1–46.
[9] Quadros, A., Magalhães, S., Mol, D., Lima, J., Vieira, A., and Brandão, W. (2025). Multi-class bias mitigation methods for classification without discrimination. SN Computer Science, 6(8):1–17.
[10] d'Aloisio, G., D'Angelo, A., Di Marco, A., & Stilo, G. (2023). Debiaser for Multiple Variables to enhance fairness in classification tasks. Information Processing & Management, 60(2), 103226. https://doi.org/10.1016/j.ipm.2022.103226
[11] Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2018). A comparative study of fairness-enhancing interventions in machine learning. https://doi.org/10.48550/arXiv.1802.04422
[12] Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., and Wallach, H. (2018b). A reduction approach to fair classification. In International conference on machine learning, pages 60–69. PMLR.

i) Elementi di originalità e innovazione della proposta e impatto in termini di rilevanza dell'avanzamento nella ricerca di base per la comunità scientifica di riferimento (max 3000 caratteri)

The proposed project holds significant relevance due to the following reasons:

- Addressing bias in ML- and AI-based software systems is gaining considerable relevance not only in literature, but also in policies and regulations.
- Despite the adoption of multi-class ML-based approaches in several sensitive domains, the topic of bias mitigation in the multi-class context is still underexplored. Indeed, to the best of our knowledge, no in-processing bias mitigation method for multi-class classification has been proposed so far.
- Different from previous approaches that are tailored to specific fairness definitions, our approach will be designed to address multiple fairness constraints simultaneously, making it more general and practical for real-world use cases.
- Our proposed approach will also be extended to the intersectional fairness scenario, which is another topic that is underexplored in the fairness literature.
- In the experiment, we focus on gender aspects by considering *gender* as the primary, both as a binary (i.e., biological sex) and as a multi-class (i.e., that considers sexual orientation) sensitive variable. While the first is well-known and studied in literature, the last is not yet adequately addressed, even if it is of paramount importance when human rights are considered.

MODELLO PER LA PRESENTAZIONE DEL PROGETTO DI RICERCA

j) Impatto del progetto in riferimento alle tematiche di genere (facoltativo, max 3000 caratteri)

Bias is generally exposed towards individuals or groups identified by some sensitive features, among which is gender. There are several examples of real ML-based systems which have shown to expose a bias toward women, e.g., the recruitment instrument employed by Amazon, which penalised women candidates for IT job positions (https://www.bbc.co.uk/news/technology-45809919), or the Facebook job advertising system not showing specific job advertisements to women (https://www.theguardian.com/world/2025/nov/05/facebook-job-ads-algorithm-is-sexist-french-equality-watchdog-rules). Therefore, proposing a novel bias mitigation method for the multi-class classification context can be an additional stepping stone towards AI- and ML-based system that treat women and people of LGBTQ+ groups fairer.

Piano di spesa

Voce di spesa	*Importo (Euro)*
Borse di ricerca (art.2 del Regolamento per il conferimento di borse di ricerca attualmente in vigore)	6000 €
Rinnovo assegni di ricerca	5000 €
Materiali di consumo	
Attrezzature, strumentazioni, software	
Missioni	4000 €
Acquisto prodotti ritenuti necessari per la realizzazione del progetto (es. materiale librario, licenze per l'accesso a banche dati, ecc.)	
Pubblicazioni, organizzazione di convegni e workshop	