UNIVERSITY OF MACERATA

PhD course in Law and Innovation

Cycle XXXVIII

Feasibility and Legal Compliance of AI Systems Across Domains: Five Case Studies in Industrial Decision Support, Municipal Tax Revenue, and Healthcare under EU Regulations 2016/679 and 2024/1689, and a Cohesive Impact Assessment Framework for DPIA and FRIA

SUPERVISORS:

Prof. Emanuele Frontoni

Prof. Simone Calzolaio

COORDINATOR:

Prof. Massimo Meccarelli

CANDIDATE:

Giovanna Migliorelli

Year 2026

Finanziato dall'Unione europea NextGenerationEU | Ministero dell'Università e della Ricerca | Italiadomani | unimc Università di Macerata

Abstract

This doctoral thesis investigates the design and development of Artificial Intelligence (AI) systems characterized by variable risk levels according to Art. 6 and Annex III in [1], with the aim of ensuring their compliance with legal and ethical standards, either imposed or recommended by the European Union, with particular reference to the regulatory frameworks [2] and [1], as well as the ethical guidelines [3].

The research alternates between theoretical exploration, both technological and regulatory, and operational phases in which acquired knowledge is implemented and translated into concrete and applicable results. The adopted approach is highly transversal, allowing exploration from Machine Learning to Deep Learning, encompassing a wide range of tasks (classification, regression, object detection), various data types (tabular data, time-series, XCA and US images), and different domains (industrial decision support, municipal tax revenue, and healthcare) characterized by varying degrees of risk and multiple compliance requirements. Compliance is implemented, often voluntarily, always respecting the principle of proportionality, without ever being redundant or obstructive, and it specifically addresses the following profiles: Privacy by Design (Considerando 78, Art. 25(1)) including Data Anonymization, Data Minimization, Computational and Output Privacy; Privacy by Default (Considerando 78, Art. 25(2)) in [2]; Technical Robustness (Art. 16), Data Governance (Art. 10), Transparency (Art. 13), Human Oversight (Art. 14) in [1].

From a methodological perspective, the scientific-technological and regulatory dimensions develop in parallel and are both considered from the earliest stages of the AI system life-cycle, in an extended "by design" perspective. The five case studies, which belong to the industrial decision support, tax revenue, and healthcare sectors, constitute the operational core of the research, each representing a specific combination of task, application domain, data type, and risk category. The results are presented in five publications, either accepted or under review.

The sixth contribution focuses on the operationalization of the regulatory frameworks [2][1], proposing Cohesive Impact Assessment (COHESIA), an integrated methodological framework that unifies the Data Protection Impact Assessment (DPIA) and the Fundamental Rights Impact Assessment (FRIA) into a semi-quantitative compliance model. COHESIA operationalizes the concept of trustworthy AI in concrete scenarios, supporting systematic evaluation of AI applications across different risk levels and regulatory dimensions, facilitating DPIA and FRIA comparison, both diachronically (across successive versions of the same document) and synchronically (across multiple DPIAs or FRIAs, or between the DPIA and the FRIA produced for the same system).

To demonstrate its practical utility, COHESIA is applied to the FAITH-RSDD AI system, designed to support municipal fiscal efficiency, as well as to a series of AI systems inspired by the other four case studies investigated in this thesis.

The need for a tool facilitating the coordinated drafting of DPIA and FRIA arises from the principle that these two documents should not be regarded as mere formal obligations to be completed only at deployment or produced in case of a data breach, but as operational instruments accompanying the entire system life-cycle [4][2][1], continuously guiding design choices to enhance the quality of the final product and foster a culture of responsible *know-how*.

To support this approach, key compliance requirements have also been evaluated for AI systems not classified as high-risk, or in contexts where they are not explicitly mandatory, in line with voluntary alignment (Considerando 7 and 27, Art. 95 in [1], Art. 40 in [2]) and [3]. In this context, it is necessary to determine, on a case-by-case basis, the optimal trade-off between regulatory compliance and operational utility, in order to maximize both without conflict, according to the principle of proportionality.

In conclusion, this research allows for an integrated experience and leads to the concrete development of AI systems oriented toward trustworthiness, a crucial ethical principle that enables stakeholders to adopt new AI-based solutions with confidence, awareness, and safety.

Keywords: trustworthy AI, legal compliance, GDPR, AI Act, DPIA, FRIA, fashion, fiscal administration, healthcare

Abstract

Il presente lavoro di tesi indaga la progettazione e lo sviluppo di sistemi di Intelligenza Artificiale (IA) caratterizzati da livelli di rischio variabili in base all'Art. 6 e Annesso III in [1], al fine di garantirne la conformità agli standard legali ed etici rispettivamente imposti e raccomandati dall'Unione Europea, con particolare riferimento ai quadri normativi [2] e [1], nonché alle linee guida etiche [3].

La ricerca alterna fasi di approfondimento teorico, sia tecnologico che normativo, a fasi operative in cui le conoscenze acquisite vengono implementate e tradotte in risultati concreti e applicabili. L'approccio adottato è improntato alla massima trasversalità e consente di spaziare dal Machine Learning al Deep Learning, includendo un'ampia gamma di task (classificazione, regressione, *forecast* di serie storiche, *object detection*), vari tipi di dato (dati tabulari, serie storiche, immagini radiografiche ed ecografiche) e domini differenti (settore moda, settore fiscale, settore sanitario) caratterizzati da diversi gradi di rischiosità e molteplici esigenze di *compliance*. La *compliance* è impostata, spesso su base volontaria, ma sempre nel rispetto del principio di proporzionalità, senza mai assumere carattere ridondante od ostativo, e riguarda in particolare i seguenti profili: *Privacy by Design* (Considerando 78, Art. 25(1)) inclusiva di *Data Anonymization*, *Data Minimization*, *Computational e Output Privacy*; *Privacy by Default* (Considerando 78, Art. 25(2)) in [2]; *Technical Robustness* (Art. 16), *Data Governance* (Art. 10), *Transparency* (Art. 13), *Human Oversight* (Art. 14) in [1].

Dal punto di vista metodologico, le dimensioni scientifico-tecnologica e normativa si sviluppano in parallelo e sono curate fin dalle prime fasi del ciclo di vita del sistema AI, in un'ottica "*by design*" estesa. I cinque casi di studio, appartenenti ai settori moda, fiscale, e sanitario, costituiscono il nucleo operativo della ricerca, ciascuno rappresentando una combinazione specifica di task, dominio applicativo, tipo di dato e categoria di rischio- I risultati sono raccolti in cinque pubblicazioni, accettate o in fase di valutazione.

Il sesto contributo si concentra sull'operazionalizzazione delle normative [2][1], proponendo il COHEsive Impact Assessment (COHESIA), un quadro metodologico integrato che unifica Data Protection Impact Assessment (DPIA) e Fundamental Rights Impact Assessment (FRIA) in un modello di conformità semi-quantitativo. COHESIA rende operativo il concetto di IA affidabile in scenari concreti, supportando la valutazione sistematica delle applicazioni di IA attraverso diversi livelli di rischio e dimensioni regolatorie, promuovendo un confronto tra DPIA e FRIA sia diacronico (tra versioni successive dei documenti) che sincronico (tra DPIA e FRIA redatte per sistemi di IA simili ma diversi o tra la DPIA e la FRIA redatte per lo stesso sistema, eventualmente da diversi operatori).

Per dimostrarne l'utilità pratica, COHESIA è stato applicato al sistema di IA FAITH-FDSS, progettato per supportare l'efficienza fiscale dei comuni, nonché a una serie di sistemi di IA, ispirati agli altri quattro casi di studio sviluppati nella tesi. La necessità di uno strumento che faciliti la redazione coordinata di DPIA e FRIA deriva dal principio secondo cui questi due documenti non devono essere considerati meri adempimenti formali da compilare esclusivamente al momento del deployment o da esibire in caso di data breach, ma strumenti operativi che accompagnano l'intero ciclo di vita del sistema [4][2][1], guidando costantemente le scelte progettuali a beneficio della qualità del prodotto finale e fondando una cultura del *know-how* responsabile.

A supporto di questa impostazione, i requisiti chiave di *compliance* sono stati valutati anche per sistemi di IA non classificati come high-risk, o in contesti in cui non risultavano esplicitamente obbligatori, in linea con un adeguamento volontario (Considerando 7 e 27, Art. 95 in [1], Art. 40 in [2]) e [3]. In tale contesto, si è reso necessario determinare, caso per caso, il *trade-off* ottimale tra conformità normativa e utilità operativa, al fine di massimizzare entrambe senza conflitti, secondo il principio di proporzionalità.

In conclusione, il presente lavoro di ricerca rappresenta una esperienza integrata e conduce allo sviluppo concreto di sistemi di IA orientati all'affidabilità, principio etico cruciale che consente ai diversi *stakeholder* di adottare con fiducia, consapevolezza e sicurezza le nuove soluzioni basate sull'IA.

Parole chiave: IA affidabile, conformità legale, Normativa EU sull'IA, DPIA, FRIA, moda, fiscalità, sanità

Contents

List of Figures

List of Tables

Abbreviations

ALTAI Assessment List for Trustworthy AI

CAM Class Activation Map

CE Cross Entropy

DP Differential Privacy

DPIA Data Protection Impact Assessment

DL Deep Learning

DSS Decision Support System

FL Federated Learning

FRI Fundamental Rights Impact Assessment

GAP Global Average Pooling

HIC Human-in-Command

HLEG High-Level Expert Group

HITL Human-in-the-Loop

LGBM Light Gradient Boosting Machine

LSTM Long Short Term Memory

MAAPE Mean Arctangent Absolute Percentage Error

MAPE Mean Absolute Percentage Error

MC Monte Carlo

ML Machine Learning

MLP Multi Layer Perceptron

MSE Mean Square Error

MMAPE Modified Mean Absolute Percentage Error

msMAPE Modified Symmetric Mean Absolute Percentage Error

PbD Privacy by Design

PFL Personalized Federated Learning

ReLU Rectified Linear Unit

RBF Radial Basis Function

RMSE Root Mean Square Error

SGDR Stochastic Gradient Descent with Warm Restarts

SHAP SHapley Additive exPlanations

SKU Stock Keeping Unit

SVR Support Vector Regression

TC Trans Cerebellar

TT Trans Thalamic

TTA Test Time Adaptation

TV Trans Ventricular

tSNE t-distributed Stochastic Neighbor Embedding

US Ultrasound

WAPE Weighted Absolute Percentage Error

WRMSPE Weighted Root Mean Square Percentage Error

XCA X-ray Coronary Angiography

List of Publications

Publications from Thesis as Primary Author

1. G. Migliorelli, L. Romeo, and E. Frontoni, "FAITH-FDSS: Fiscal AI for Trustworthy and High-compliance Federated Decision Support Systems." Manuscript under submission, 2025

2. G. Migliorelli, L. Romeo, and E. Frontoni, "COHESIA: a Cohesive Impact Assessment Framework for DPIA and FRIA with Applications to Four Case Studies." Manuscript under submission, 2025

3. M. Di Cosmo, G. Migliorelli, F. P. Villani, M. Francioni, A. Muçaj, E. Frontoni, S. Moccia, and M. C. Fiorentino, "FedStenoNet X-ray Coronary Angiography Dataset for Stenosis Detection," Oct. 2025

4. M. C. Fiorentino, G. Migliorelli, F. P. Villani, E. Frontoni, and S. Moccia, "Contrastive prototype federated learning against noisy labels in fetal standard plane detection," *International Journal of Computer Assisted Radiology and Surgery*, vol. 20, no. 7, pp. 1431–1439, 2025

5. G. Migliorelli, R. Pietrini, A. Galdelli, E. Frontoni, and M. Paolanti, "FashionDSS: A Human-in-the-Loop Decision Support System for forecasting production efficiency and predicting item defects in the luxury fashion industry." Manuscript under submission, 2025

6. G. Migliorelli, R. Pietrini, A. Galdelli, E. Frontoni, and M. Paolanti, "FashionSight: A Human-in-the-Loop Decision Support System for Sales Forecasting in the Luxury Fashion Industry." Manuscript under submission, 2025

Others

1. M. Di Cosmo, G. Migliorelli, M. Francioni, A. Muçaj, A. Maolo, A. Aprile, E. Frontoni, M. C. Fiorentino, and S. Moccia, "A federated learning framework for stenosis detection," in *Image Analysis and Processing - ICIAP 2023 Workshops* (G. L.

Foresti, A. Fusiello, and E. Hancock, eds.), (Cham), pp. 211–222, Springer Nature Switzerland, 2024

2. M. Francioni, A. Muçaj, A. Maolo, A. Aprile, R. Manfredi, M. D. Cosmo, M. C. Fiorentino, G. Migliorelli, E. Frontoni, S. Moccia, and M. Marini, "Clinical insights into coronary angiography heterogeneity: a federated learning investigation." Manuscript being revised for resubmission, 2025

Introduction

1 Motivation

After a period of relatively modest progress, the year 2012 marked a decisive turning point in the evolution of AI. Researchers could now harness the computational power of GPUs to accelerate the training of Deep Learning (DL) models at an unprecedented scale.

Over the following decade, the convergence of enhanced computing capacity, the advent of big data, the introduction of exceptionally powerful new architectures, and the large-scale parallelization of algorithms propelled AI to the forefront of technological innovation, establishing it as one of the most strategic and influential technology of the twenty-first century.

As global progress in AI accelerated substantially, massive investments began to flow from leading technology companies, driven by the recognition of AI's transformative potential and the pursuit of competitive advantage. This unparalleled wave of innovation, fuelled by substantial public and private financial resources directed toward AI research, development, and deployment, has raised profound policy and ethical questions concerning governance, accountability, and the protection of fundamental rights.

In response to these emerging challenges, since 2017 the European Union has undertaken an ambitious and comprehensive policy agenda aimed at establishing a human-centred, trustworthy AI ecosystem which effectively fosters innovation. In October 2017, the European Council affirmed that, in order to successfully build a "Digital Europe," the Union must cultivate "a sense of urgency to address emerging trends" [13] and invited the

European Commission to develop a coordinated European approach to AI. Subsequently, in June 2018, the Commission established the High-Level Expert Group on AI (AI HLEG), an independent advisory body tasked with producing two landmark documents: (1) the Ethics Guidelines for Trustworthy AI [3] and (2) the Policy and Investment Recommendations for Trustworthy AI [14], while also launching the European AI Alliance to promote multi-stakeholder dialogue. In February 2020, the White Paper on AI [15] outlined policy options designed to enable both an ecosystem of excellence and an ecosystem of trust. This was followed, in July 2020, by the AI HLEG's Final Assessment List for Trustworthy AI (ALTAI) [16], which provided practical guidance for the voluntary self-assessment of AI systems. The regulatory process reached a historic milestone in August 2024 with the entry into force of the Regulation (EU) 2024/1689 of the European Parliament and of the Council [1], establishing the first comprehensive European legal framework for AI. In February 2025, the European Commission issued the Guidelines on Prohibited AI Practices and the Guidelines on the Definition of AI Systems, offering further clarification and practical direction to ensure coherent and effective implementation of [1] across Member States.

In this highly dynamic and rapidly evolving landscape, a careful and rigorous examination of how AI applications can achieve practical compliance with emerging regulatory and ethical standards represents both a crucial opportunity and an imperative for all actors involved in the life-cycle of AI systems, from research to industry.

Such an inquiry is not only essential for ensuring that AI technologies are developed and deployed responsibly and safely, but also provides a foundation for aligning innovation with the European vision of human-centred and trustworthy AI.

Against this backdrop, the research objectives of this thesis are motivated by the need to operationalize compliance and to identify practical measures for implementing data governance, transparency, and human oversight in AI systems.

By broadening the scope of the investigation to encompass both diverse domains and AI systems exhibiting varying degrees of risk, this work further proposes original technical solutions aimed at addressing challenges that might otherwise compromise system functionality and, as a consequence, adherence to regulatory and ethical standards.

2 Research Objectives

In this research, compliance is interpreted as adherence to the principles of trustworthy AI, whose tenets, as stated in [3], are lawfulness, robustness, and ethics. Accordingly, compliance is understood not only as conformity with the legal frameworks established by [2] and [1], but also as the pursuit of high technological quality standards in AI models and alignment with ethical principles that promote responsible and human-centred AI.

Achieving genuine legal compliance cannot be separated from the technical soundness of the models and the robustness of the data-preprocessing pipeline, as demonstrated, for instance, by Artt. 10, 13, and 15 of [1]. For this reason, a substantial portion of the present research is devoted to the technical and implementation aspects that underpin regulatory adherence. While avoiding unnecessary digressions, this dissertation aims to provide a transversal perspective on the overarching theme, integrating supplementary analyses wherever they yield additional insight.

The main objectives of this research are formulated as follows:

1. **Technological optimization:** for each case study, conduct an in-depth investigation of the most effective technological solutions to address the task defined by its intended purpose.

2. **Compliance verification:** ensure that the Technological optimization comply with all the relevant legal provisions as well as with the non-binding principles of trustworthy AI.

3. **Transversality:** guarantee a transversal scope across application domains (industrial decision support, municipal tax revenue administration, and healthcare), learning tasks (regression, classification, and object detection), and data types (tabular, time-series, and imaging data), thereby promoting the generalizability and methodological coherence of the proposed approach.

4. **Trustworthiness:** apply the above criteria consistently across all risk classes in accordance with the principle of proportionality and, where appropriate, on a voluntary basis, to foster the development of genuinely trustworthy AI systems that extend beyond mere regulatory compliance.

5. **Operationalization of legal compliance:** further operationalize regulatory principles through the design and application of an integrated semi-quantitative framework for conducting DPIA and FRIA, employing this framework as a comparative and evaluative tool across diverse AI systems.

It should be noted that the activities undertaken to achieve these objectives are necessarily developed in parallel, thus implementing, in practical terms, a form of "compliance by design" which extends the "privacy by design" defined in Art.25(1) of [2].

3 Structure of the Thesis

Chapter 1 provides a concise overview of the soft law landscape, encompassing a range of ethical guidelines, policy papers, and other non-binding instruments that integrate legal, technological, economic, and theoretical perspectives. These instruments constitute the conceptual and substantive foundation upon which [1] has been developed. The chapter also reviews selected decisions of the Regional Administrative Courts (TAR) and the Council of State (CdS). These judgments illustrate an increasing judicial awareness of the legal issues posed by AI, while simultaneously highlighting the limitations of pre-existing legal frameworks in effectively addressing the challenges raised by AI. The chapter concludes with a concise overview of the sections from [1] most frequently referenced throughout this thesis, which serve as a continuous foundation for the discussion of case studies and guide the conception and implementation of the proposed compliance framework.

Chapter 2 introduces two case studies in the industrial decision support domain, motivating their classification as low-risk applications under the risk-taxonomy in [1]. The chapter explores how such systems can be aligned with the principles of Data Governance, Data Protection, Human Oversight, and Transparency, adopting a voluntary compliance approach inspired by the overarching concept of trustworthy AI.

Chapter 3 elaborates on the first case study introduced in **Chapter 2**. It provides a comprehensive description of the domain (tabular data derived from fashion manufacturing) and the specific tasks addressed, namely defect classification and production time regression. The chapter discusses the design, implementation, and evaluation of the developed AI system and concludes with the presentation of the main findings and insights.

Chapter 4 focuses on the second case study introduced in **Chapter 2**. It examines time-series data concerning the sell-out performance of a prestigious Italian company and addresses the task of sales forecasting. The chapter presents an AI-based predictive model designed to optimize performance relative to non-automated benchmarks, followed by an analysis of results and concluding remarks.

Chapter 5 introduces one case study within the domain of fiscal revenue administration, justifying its classification as moderate-risk under the risk-taxonomy in [1]. The chapter investigates how these system can ensure compliance with the principles of Data Governance, Data Protection, Human Oversight, and Transparency, following a partially voluntary approach aligned with the broader goal of promoting trustworthy AI.

Chapter 6 provides a detailed account of the case study introduced in **Chapter 5**. It describes the domain (tabular data from fiscal operations) and the specific task of payment prediction, discriminating between settled and outstanding payment notices. The design, implementation, and evaluation of the AI system are presented, followed by the results obtained and final observations.

Chapter 7 introduces two case studies within the domain of biomedical imaging, providing a rationale for their classification as high-risk applications under the risk-taxonomy in [1]. The chapter subsequently examines how these systems can be rendered compliant with the core regulatory principles of Data Governance, Data Protection, Human Oversight, and Transparency, illustrating their practical implementation.

Chapter 8 presents the detailed development of the first case study introduced in **Chapter 7**. It offers an in-depth characterization of the biomedical domain (X-ray imaging) and the specific task (object detection of stenoses). The chapter outlines the design and implementation of an AI system developed within a Federated Learning (FL)

framework, integrating an innovative approach to address non independently and identically distributed (non IID) data. The results achieved and the corresponding conclusions are subsequently discussed.

Chapter 9 details the second case study introduced in **Chapter 7**, focusing on the biomedical ultrasoud (US) imaging domain and the task of fetal standard plane classification. It presents the design of an AI system developed under a FL framework, introducing an original methodology to mitigate noise in the training data. The results and conclusions are then reported and critically analyzed.

Chapter 10 introduces COHESIA, an integrated methodological framework that consolidates the DPIA and the FRIA into a unified semi-quantitative model. In this process, well-established tools are employed for conducting the DPIA, while more experimental and still-evolving instruments are used for the FRIA, drawing upon recent literature on the practical implementation of the [1]. COHESIA is then applied to four distinct AI systems within the high-risk class, each exhibiting different levels of impact, legal compliance, and technical robustness. These four case studies are conceived as thought experiments to illustrate how the framework can facilitate comparative analysis while deriving meaningful insights from the AI systems effectively implemented throughout this research. The first case study is a modified version of the one presented in **Chapter 6**, in which the addition of a third, higher-risk task leads to its reclassification into the high-risk category. The second represents a further variation of the first, intentionally designed to weaken its technical soundness. The third is inspired by the case studies discussed in **Chapters 8** and **Chapters 9**, while the fourth, is drawn from the field of legal process automation and is characterized by high impact and limited technical soundness, a condition which makes its lawfulness highly questionable.

4 Contibutions

This thesis provides transversal insight into the different risk classes of AI systems as established by [1]. Drawing on six case studies, it examines the low, moderate, and high-risk

categories, and proposes original technical solutions tailored to their specific scope. Particular attention is paid to the solution that implements an operational framework to apply DPIA and FRIA documents in a consistent and cohesive manner. For the case studies, publications have been produced, either accepted or under submission, which concretely anchor the research. The following sections discuss these contributions, highlighting their significance and value.

The first contribution [9] introduces FashionDSS, an innovative Human-in-the-Loop Decision Support System (DSS) that integrates human expertise with advanced AI to enhance quality assurance and product manufacturing. Critical tasks such as early prediction of product defects and regression over manufacturing times are essential in the luxury fashion industry to maintain brand reputation, ensure customer satisfaction, and enable informed adjustments to materials, designs, or production processes before production begins. The results demonstrate that FashionDSS significantly improves operational efficiency, reduces defect rates, and enhances customer satisfaction by supporting proactive quality management. FashionDSS was specifically designed to comply seamlessly with the requirements of transparency, performance, and human oversight.

The second contribution [10] introduces FashionSight, a modular, human-in-the-loop DSS for sales forecasting at the Stock Keeping Unit (SKU) level in the luxury fashion industry. In the rapidly changing and highly seasonal world of luxury fashion, accurately forecasting sales at both product and market levels remains a critical and complex challenge. Brands must manage global inventories across diverse regions, cope with short and intermittent time series, and address the growing demand for transparency and data privacy. Traditional forecasting methods often prove inadequate in this context due to their reliance on global models or complex architectures that lack interpretability. To overcome these limitations, FashionSight employs a per-series ensemble architecture that combines weak learners, including Support Vector Regression (SVR), Light Gradient Boosting Machine (LGBM), and linear models, through a weighted aggregation strategy based on in-sample error performance. This design ensures high adaptability to data heterogeneity and facilitates business-driven interpretation and validation. Empirical evaluation using real-world sales data from an important Italian fashion brand, covering 14 international markets, demonstrates that FashionSight consistently outperforms an internal benchmark

and a range of alternative approaches, supporting data-driven decision-making while maintaining trust and regulatory compliance.

In public administration, the systematic adoption of new technologies reflects the recognition that efficiency is not merely an operational goal but a public responsibility. In the specific context of tax revenue administration, enhanced efficiency enables both the optimization of public budgets and the delivery of improved services through effective and accurate revenue management. In this regard, the third contribution [5] introduces OptiRevDSS, a Machine Learning (ML) DSS designed for municipal revenue officers. The system aims to improve efficiency in tax revenue administration by providing interpretable predictions accompanied by uncertainty assessments, thereby supporting informed and accountable decision-making. Implemented as a shallow Multi-Layer Perceptron (MLP) architecture, the framework reinforces data protection through the integration of FL and (ε, δ)-Differential Privacy (DP). Transparency is enhanced through the analysis of feature importance, while trustworthiness is strengthened through uncertainty estimation based on Monte Carlo dropout (MC dropout). Particular attention is devoted to ensuring compliance with [**?**] and [1], as well as alignment with the broader European ethical framework outlined in [3].

Following the entry into force of [1] on 1 August 2024, the requirement to conduct a FRIA for high-risk AI systems, as mandated by Art. 27 of the Regulation, will become mandatory as of August 2026. The preparation of the FRIA will complement the preparation of a DPIA, as required under Art. 35 of [2] for the processing of personal data that poses significant risks. In addition, beyond the contexts in which DPIAs and FRIAs are strictly mandatory, these assessments may also be conducted voluntarily, in line with [3]. In the fourth contribution [6], we propose COHESIA, a unified framework that facilitates the drafting and analysis of both documents, integrating quantitative and qualitative elements. COHESIA provides convenient indicators for both diachronic comparison of multiple document versions, since both DPIAs and FRIAs are to be maintained throughout the life-cycle of the AI application, and synchronic comparison of documents produced for different AI systems of similar nature, or, when DPIAs and FRIAs are conducted independently, by different actors. To illustrate its benefits, we apply COHESIA to the tax revenue AI system implemented in [5], as well as to three additional AI systems,

included purely as a thought experiment but inspired by the other case studies presented in this research.

The fifth contribution [7] presents FedStenoNet, a novel Personalized FL (PFL) framework for stenosis detection in X-ray Coronary Angiography (XCA) images. This work introduces the first PFL framework for stenosis detection in XCA that jointly addresses both inter- and intra-client variability across three real-world, non-identical, independently distributed (non-IID) datasets. An extensive and uniform XCA dataset has been curated and will be publicly released upon publication to support reproducible research and promote further advances in federated medical image analysis. To support both the ethical and regulatory framework, as well as to enhance the performance and trustworthiness of the system, FedStenoNet integrates Histogram Matching (HM) as a privacy-preserving inter-client adaptation strategy to harmonize image appearance across institutions. Additionally, a novel Test Time Adaptation (TTA) mechanism is employed for intra-client adaptation, enabling the update of client-specific weights and improving generalization during inference without requiring any data other than the single sample itself. In terms of transparency, an explainability technique, particularly well-suited for unstructured data, is incorporated to visually highlight the regions of an input image that are the most influential in the model's decision.

The sixth contribution [8] addresses the problem of federated fetal standard plane detection in the presence of noisy labels, using a large dataset of 5187 images from Europe and a smaller dataset of 450 images from Africa. Within a FL framework, the largest and presumably most representative client is leveraged to extract robust embeddings via contrastive learning. These embeddings capture the most significant geometrical features inherent to each standard plane (i.e., brain, femur, abdomen, thorax) and are used to refine noisy image labels for the same client. Prototypes computed from the noise-free embeddings, along with the backbone trained through contrastive learning, are then shared with the smaller client to enable robust labeling and guide the FL process. Experimental results demonstrate that this approach mitigates the impact of noisy labels across clients of different sizes, improving the overall performance of standard plane detection while supporting accuracy and reliability as mandated by regulatory frameworks for AI systems. The same explainability technique used in [7], is employed to visually highlight the regions

of an input image that most strongly influence the model's predictions, thereby enhancing interpretability. In line with the vision of AI systems that promote human wellness advocated in [3], this work investigates the development of an application that leverages technology in scenarios where biomedical devices operate with low-resolution capabilities.

Chapter 1

Tracing the Roots of the AI Act's Risk-Based Approach: Insights from Soft Law and Pre-Existing Italian Administrative Case Law

1.1 Soft Law and the AI Act: Historical Roots and Enduring Interaction

The substance of [1] cannot be fully understood without taking into account the pre-existing soft-law framework, which encompasses non-binding instruments such as ethical guidelines, white papers, opinions, and official communications. This complex ecosystem reflects the European vision of AI and its intended applications, and has significantly informed AI governance practices since 2018.

In this section, following the timeline presented in Fig. 1.1, we provide an analysis of the significance and scope of selected soft-law milestones.

In June 2018, the Commission established the High-Level Expert Group on AI (HLEG), an independent advisory body tasked with producing two landmark documents:

1. the Ethics Guidelines for Trustworthy AI [3] and

2. the Policy and Investment Recommendations for Trustworthy AI [14]

In parallel, a piloting phase led to an initial version of the Assessment List for Trustworthy AI (ALTAI) [16], which was subsequently revised and finalized in July 2020, with the aim of providing all relevant stakeholders with a practical tool to facilitate the implementation of trustworthy AI principles.

Despite the non-binding nature of the HLEG deliverables, stakeholders are strongly encouraged to take the guidelines into account and to voluntarily align their practices with them, in order to foster trust in AI and promote a culture of good practices. As will be discussed in Chapters 2, 3, and 4, even AI systems classified as low-risk are advised to adhere to these principles, while avoiding unnecessary burdens, with the aim of enabling end-users to assess their reliability and, as far as possible, gain insight into the underlying decision-making processes, thereby enhancing trust in AI-based tools. As shown in Fig. 1.2, with reference to the relationship between [3] and [1], the three tenets of trustworthy AI (lawful, ethical, and robust) deserve careful consideration. A form of complementarity emerges, whereby [3] primarily address the ethical and robust dimensions, while [1] explicitly focuses on the lawful dimension. The connection between the two documents is particularly evident in the following excerpts:

- In the Executive Summary of [3], where it is stated that: "The framework does not explicitly deal with trustworthy AI's first component (lawful AI). Instead, it aims to offer guidance on the second and third components, ethical and robust AI, by fostering and securing ethical and robust practices."

- In Recitals (7) and (27) of the [1], where the influence of [3] is clearly reflected; in particular, Recital (27) expressly refers to the seven key requirements identified by [3] and affirms that: "The application of those principles should be translated, when possible, in the design and use of AI models. They should in any case serve as a basis for the drafting of codes of conduct under this Regulation. All stakeholders, including industry, academia, civil society and standardisation organisations, are encouraged

> to take into account, as appropriate, the ethical principles for the development of voluntary best practices and standards."

After having thoroughly examined the philosophy and practice laying the foundations of trustworthy AI, in [14] the HLEG proposes 33 recommendations, underpinned by the principle of proportionality, aimed at fostering sustainable, inclusive, and competitive trustworthy AI, while safeguarding and empowering human beings. The recommendations target a broad spectrum and are inherently inclusive, addressing all areas in which the HLEG considers trustworthy AI to have the greatest potential for positive impact:

- society and individuals
- the private sector
- the public sector
- Europe's research and academic communities

In parallel, the HLEG identifies key enablers to support these impacts, including:

- data and infrastructure availability
- skills and education
- governance and regulation
- funding and investment

Trustworthy AI, envisaged as the only recognized and endorsed form of AI in Europe, should serve as the hallmark of the European AI strategy, with the implementation of its principles promoting both individual and societal well-being.

In February 2020, the European Commission published the White Paper on Artificial Intelligence [15], outlining policy options aimed at enabling the trustworthy and secure development of AI in Europe, in full respect of the values and rights of EU citizens. Once again, the objective is twofold: on the one hand, "to achieve an 'ecosystem of excellence' along the entire value chain, leveraging Europe's research excellence"; on the other,

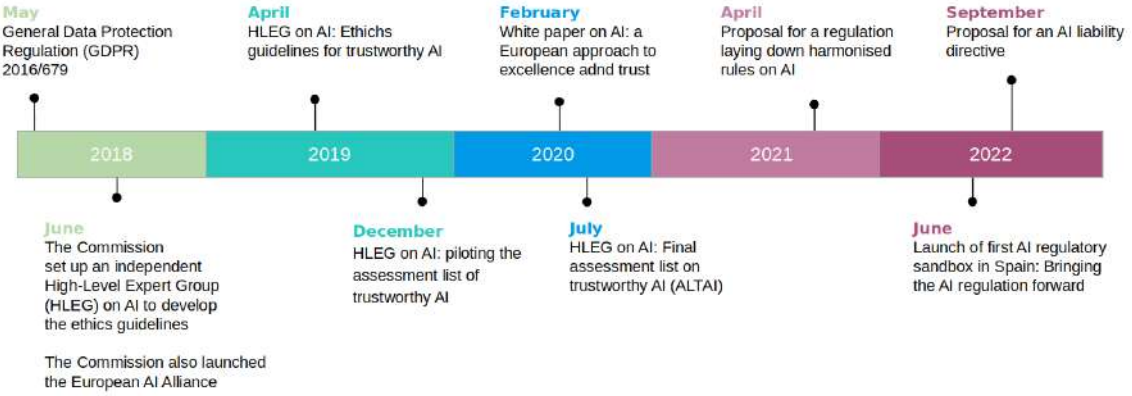


FIGURE 1.1: From Ethics to Regulation: Milestones in the European Governance of AI

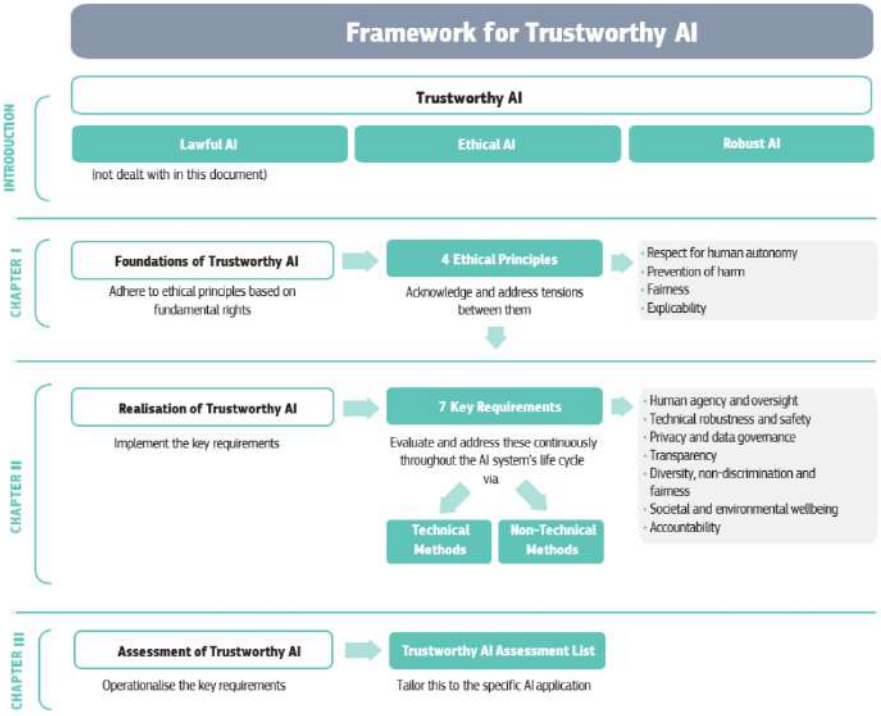


FIGURE 1.2: The HLEG Guidelines as a framework for Trustworthy AI

"to promote a future regulatory framework for AI in Europe that will create a unique 'ecosystem of trust'." Sections 2 and 4, respectively titled "Capitalising on Strengths in Industrial and Professional Markets" and "An Ecosystem of Excellence", discuss strategies from two different yet interconnected perspectives. Section 3, titled "Seizing the Opportunities Ahead: The Next Data Wave", adopts a forward-looking and concrete approach towards the economic potential of data-driven innovation. Section 5 delineates the ethical dimension, referring to the seven key principles identified in the Guidelines, and lays the foundations for what would later evolve into the Regulatory Framework for AI, subsequently embodied in the [1].

1.2 The Italian Context Before the AI Act: Selected Administrative Case Law

To gain a comprehensive understanding of the national normative and regulatory trajectory preceding the entry into force of [1], we have examined four court[1] decisions issued between 2018 and 2021 that are particularly relevant to the subject.

Depending on the court, we may observe varying degrees of openness and foresight with regard to the phenomenon of AI. Irrespective of such differences, the judgments collectively contribute to shaping a legal culture and sensitivity toward emerging technological issues, driven by the urgency of public debate and the pressing need to confront a reality that can no longer be ignored.

1.2.1 Judgment No. 2270/2019 (CdS): Algorithmic Scoring in Public Recruitment

The legal dispute concerns the use of non-transparent algorithmic decision-making processes in the management of career progression within the public administration. In Section 7 of the judgment, the appellants contend that their relocations have been determined

[1]The *TAR* (*Tribunale Amministrativo Regionale*, Regional Administrative Court) is the first level of administrative justice in Italy, whereas the *CdS* (*Consiglio di Stato*, Council of State) represents the second and highest level.

through a fully automated procedure, governed by an opaque algorithm that disregards the candidates' stated preferences and fails to provide any reasoning or justification for the decisions rendered. In addition, no human authority is identified as responsible for the decision or exercising oversight over the decision-making process.

In Section 8.3 of the judgment, invoking Art. 97 of the Italian Constitution, the Court acknowledges not only the legitimacy but also the advisability of employing algorithms for tasks such as the classification of a large number of candidates according to predefined rules that duly reflect the customary administrative procedures. That being said, in Section 8.4, the Court underscores the imperative of full transparency in the algorithmic decision-making process, including the disclosure of the algorithm's underlying logic, design methodology, evaluative criteria, and pertinent data inputs. Such transparency is not only for the benefit of the addressees of the decisions but, importantly, enables the algorithmic process to be subject to judicial review by the administrative courts and thereby ensures compliance with procedural fairness and constitutional safeguards, as set forth in Law No. 241 of 7 August 1990.

In the document, several pressing core issues can already be identified, which are subsequently addressed by [1]: first, the evident tension between fostering technological innovation and its dissemination, on the one hand, and safeguarding fundamental rights, on the other; second, the imperative need for transparency and trustworthy human oversight. Finally, it is worth noting that the algorithm under consideration represents a relatively simple AI tool compared to current advancements in the field.

1.2.2 Judgment No. 10964/2019 (TAR Lazio): Access to Algorithmic Decision-Making Systems

In this case, for a similar scenario, the Court holds that no administrative task, regardless of the number of cases to be handled, can justify delegating the procedure entirely to an impersonal algorithmic process, thereby excluding human involvement. This principle applies all the more when the decision-making procedures have an impact on individuals' rights.

The Court considers the absence of a reasoned decision to be unacceptable, as such a deficiency undermines procedural guarantees, in violation of Art. 24 of the Italian Constitution. In these circumstances, both the recipients of the decision and the judiciary are prevented from understanding the rationale underlying the administrative act, thereby compromising the right to defense and effective judicial review.

The Court ultimately, invoking respect for Arts 3,24, and 97 of the Italian Constitution, as well as Art. 6 of the European Convention on Human Rights, affirms the absolute necessity of human oversight and redefines the role of automated processes as subordinate and ancillary.

1.2.3 Judgment No. 8472/2019 (CdS): Human Oversight and Algorithmic Decision-Making

Although the CdS ultimately upholds the illegitimacy of the contested administrative measure in the specific case, thus confirming the first-instance judgment of the TAR, it adopts a more balanced tone and presents its reasoning in a different way. As a preliminary observation at Section 9.1, and in line with Judgment No. 2270/2019 (CdS), the Council reaffirms that technology constitutes an indispensable support tool and a valuable opportunity to enhance administrative efficiency in the interest of all stakeholders involved in the decision-making process. It then sets out three fundamental and inescapable requirements that any algorithmic decision-making process must respect in order to avoid infringing upon the rights of the individuals concerned:

- Transparency. The decision-making process must be fully knowable to the data subject, and the resulting decisions must be intelligible and clearly reasoned. This requirement is grounded in Recitals 39 and 58 and Art. 5(1)(a) of [2], which establish transparency as a core principle of lawful data processing. It is further supported by Art. 41 of the Charter of Fundamental Rights of the European Union, which affirms each individual's right "to have his or her affairs handled impartially, fairly and within a reasonable time by the institutions, bodies, offices and agencies of the

Union." Without adequate reasoning, these rights become ineffective or unenforceable. It is also emphasized that individuals must be properly informed, whenever automated decision-making procedures are applied to them.

- Human Oversight. The need for meaningful human involvement is underlined in Recital 71, Art. 22(1) of [2], which grant individuals the right not to be subject to decisions based solely on automated processing, including profiling, that produce legal effects or similarly significantly affect them

- Prevention of Discriminatory Effects. Recital 71 of [2] also establishes the duty of data controllers to prevent discriminatory outcomes resulting from automated decision-making. This entails the use of appropriate mathematical or statistical procedures, as well as the implementation of adequate technical and organizational measures to correct inaccuracies in personal data and minimize the risk of error. These safeguards are intended to prevent, *inter alia*, algorithmic discrimination, and they anticipate broader obligations now developed under [1] in relation to fairness, bias mitigation, and data governance.

The following discussion highlights the forward-looking approach of the CdS and its careful attention to the challenges posed by emerging technologies. As a preliminary point, the Court acknowledges the challenge posed by transparency in an interdisciplinary context, where actors with profoundly different backgrounds must be capable of mutual understanding. It affirms that such complexity cannot exempt algorithmic systems from the requirement of being fully knowable and intelligible to all stakeholders involved in the decisional process. Differently from Judgment No. 10964/2019 (TAR Lazio), the judgment also underscores the quantitative dimension of data, whose ever-increasing volume, though subject to legal safeguards and individual rights, necessitates the adoption of new technologies as an essential and unavoidable resource for public administration.

1.2.4 Judgment No. 7891/2021 (CdS): Algorithmic Profiling and Automated Administrative Decisions

This judgment illustrates the inherent difficulty in formulating a definition of automated procedures or algorithms that is both sufficiently precise and legally actionable. Notably, Art. 3 of [1] introduces a definition of artificial intelligence which, reflecting the complexity of the subject, underwent several revisions during the legislative process before ultimately being adopted. By way of note, the final definition of an AI system, as provided in Art. 3 of [1], reads as follows: "a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers from the input it receives how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments." This definition represents a revision of the initially overly technical formulation, aligning it with the general and flexible approach advocated by the Organisation for Economic Cooperation and Development (OECD), which accommodates the continuously evolving nature of technological advancements in the field of artificial intelligence. As observed by the European Law Institute (ELI), even an algorithm with no degree of automation may fall within the definition, including any algorithm, whether or not it exhibits adaptiveness.

In the specific case, the CdS is able to anticipate certain critical issues and nuances that would later resurfacedin the broader regulatory debate. First, the CdS emphasizes that, in light of current technological advancements, an algorithm inherently involves some degree of automation, while remaining distinct from AI and provides a definition of algorithm. Second, it establishes a clear distinction between a traditional algorithm and a machine learning-based algorithm, as follows: "In this case, the algorithm incorporates machine learning mechanisms, yielding a system that does not merely apply pre-programmed rules and parameters, as a "traditional" algorithm would, but continuously generates new inference criteria from the data and allows these criteria to guide its decisions, in accordance with an automated learning process." It further acknowledges that an algorithm exhibiting a higher level of automation, though fundamentally still an algorithm, should

be classified within a higher functional category and, as such, merits commensurate regulatory consideration.

1.3 Selected Provisions of the EU AI Act Implementing the Risk-Based Approach

In the following, we provide a concise overview of the sections of the [1] most frequently used in this research, which serve as a constant reference in the discussion of our case studies and have guided the development of our compliance framework.

Art. 5 is fundamental for determining the admissibility of an AI system and, consequently, for assessing its feasibility. It prohibits all AI systems that pose a serious threat to fundamental human rights, in particular, including those that employ subliminal techniques, exploit the vulnerabilities of specific groups of individuals, engage in social scoring practices, or are specifically designed to infer emotions or biometric data to classify individuals.

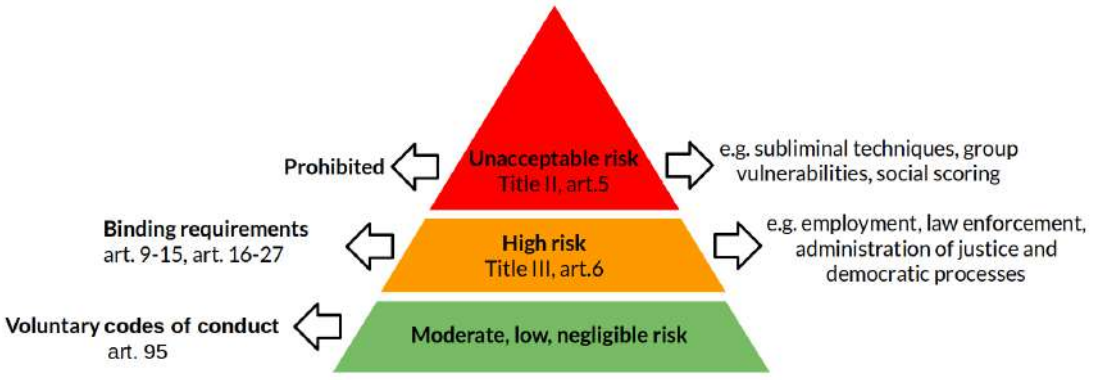


FIGURE 1.3: Risk-based classification of AI systems in the AI Act

Art. 6 addresses the identification of high-risk AI systems. These include AI systems that constitute a safety component of a product or products listed in Annex I, as well as AI systems listed in Annex III. In the latter case, an exception applies to systems that do not pose a significant risk of harm to the health, safety, or fundamental rights of natural persons, including instances where they do not materially influence the outcome

of decision-making. Detailed provisions specify the conditions under which this exception applies.

Chap.3, Sec. 2 and Sec. 3 are dedicated to high-risk AI systems and enumerate both the requirements these systems must meet (Chapt. 3, Sec. 2) and the obligations of their providers and deployers (Chapt. 3, Sec. 3). These sections are relevant not only in cases where a high risk has been established, but also more generally, as they provide guidelines for good practice that, to varying degrees, should be followed by all AI systems, regardless of the assigned risk category. In doing so, they align with many of the key requirements identified by the HLEG, although the legal and ethical frameworks operate with different emphases and nuances.

Among the articles of particular relevance in Chapt. 3, Sec. 2 are: Art. 9 on the risk management system, Art. 10 on data and data governance, Art. 13 on transparency, Art. 14 on human oversight, and Art. 15 on accuracy, robustness, and cybersecurity. Within Chapt. 3, Sec. 3, the most relevant provision for our research is Art. 27, which concerns the FRIA assessment for high-risk AI systems, a fundamental tool that, regardless of the assigned risk category, is worth applying on a voluntary basis in proportion to the characteristics of the specific AI system.

Another article frequently referenced in this study is Art. 95, which once again underscores the importance of voluntary adherence.

Chapter 2

Operationalizing Compliance for Low-Risk AI Systems

In the following, we provide a well-founded rationale for the inclusion of low-risk AI systems in our research.

Despite the non-binding nature of [3], all stakeholders committed to promoting trustworthy AI are encouraged to voluntarily implement the guidelines in [3] throughout the entire lifecycle of their AI systems.

Trustworthy AI, as defined in [3], is founded on three pillars: (1) lawfulness, (2) ethics, and (3) robustness. However, as explicitly noted in the Executive Summary, the guidelines do not directly address the first pillar, which falls within the remit of [1]. Complementarily, [1] refers to [3], specifically to the seven key requirements of trustworthy AI, and explicitly invites stakeholders "to take into account, as appropriate, the ethical principles for the development of voluntary best practices and standards."

As explained in [3], under certain circumstances, an AI system that fully complies with legal requirements may nonetheless fall short of fully meeting ethical principles, due to minor or temporary discrepancies between the regulatory framework and ethical standards. Likewise, an AI system that complies with both legal and ethical requirements may not

be entirely robust and cause harm or disadvantage, potentially undermining user trust in the system.

The significance of voluntary adherence to standards promoting ethical and robust AI is evidenced by initiatives aimed at introducing voluntary certification schemes, such as the Voluntary Labelling for Non-High-Risk AI Applications promoted in [15], and, notably, by Art. 95 of [1] and Art. 40 of [2].

We may also note that, in order to facilitate experimentation and ensure greater flexibility in relation to compliance with [2] and [1], datasets from the fashion domain are used. These datasets consist exclusively of non-sensitive and non-personal data, as defined in Art. 4(1) of [2]. This choice allows for testing under reduced legal constraints while remaining fully consistent with applicable data protection standards.

In the following sections, we examine the degree of compliance of the low-risk AI systems presented in Chapter 3 and Chapter 4 with respect to selected key requirements.

2.1 Determination of the Risk Class

The case studies presented in Chapter 3 and Chapter 4 are designed to support the optimization of production and marketing workflows within companies operating in the fashion sector. These activities are intended solely for process and marketing efficiency improvements and do not involve purposes or domains identified as high-risk in Annex III of Art. 6(2) of [1]. Accordingly, the systems and procedures described in these case studies cannot be classified as high-risk under the regulation.

2.2 Addressing Data Protection Requirements

Since the data at issue are non-personal, their regulation falls outside the scope of the [2], potentially falling instead under Regulation (EU) 2018/1807 [17], which, however, does not prescribe any specific compliance obligations. That said, an appropriate degree of confidentiality remains necessary to prevent large-scale data leakage, as the data in

question may contain potentially sensitive information regarding companies' production or marketing activities.

2.3 Addressing Data Governance Requirements

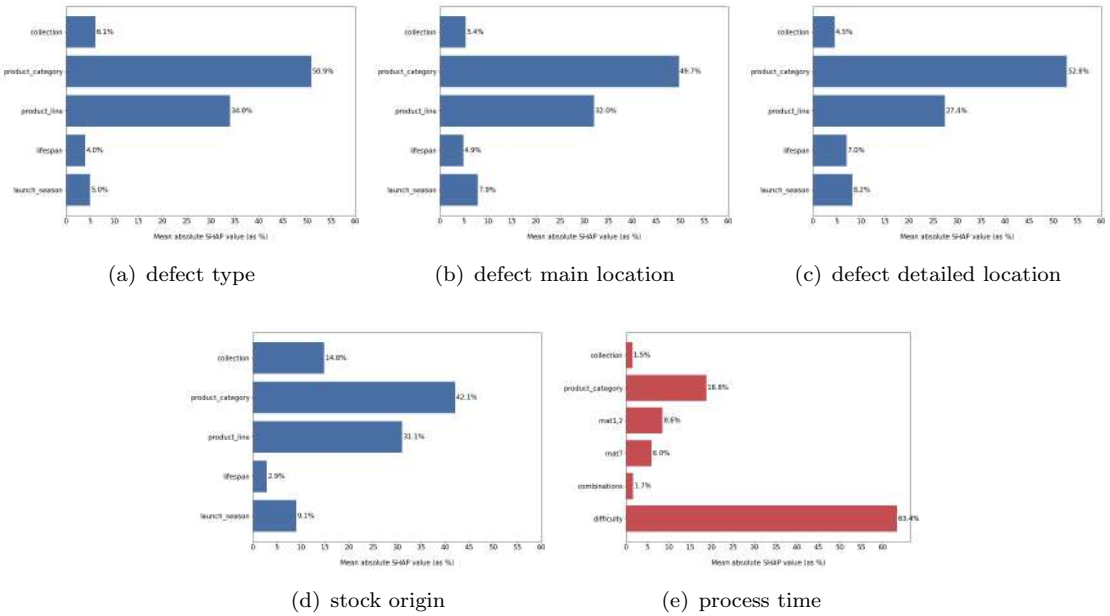
In the context of low-risk AI systems, no specific provision of the [1] directly applies. Moreover, the data employed in the case studies presented in Chapter 3 and Chapter 4 pertain to products rather than individuals, thereby excluding the applicability of the [2].

That said, the principles of Data Quantity, Representativeness, and Balance, as set out in Arts. 10(2), 10(2)(e) of the [1], and the requirements of Bias Detection, Prevention, and Mitigation, as set out in Art. 10(2)(f)–(g) of [1], remain pertinent and relevant, albeit motivated by factors other than regulatory or ethical requirements. For example, while Art. 10 of [1] primarily aims to prevent discrimination, in this context the emphasis is on constructing valid datasets for training and testing, ensuring that the model is properly trained and that the assessment of its performance accurately reflects its true capabilities, with a particular focus on performance.

2.4 Addressing Transparency Requirements

To investigate the model's global decision-making process, we employ SHapley Additive exPlanations (SHAP) [18], which allows the identification of features with the greatest relevance to the output, thereby providing an average explanation of how the model arrives at its predictions.

With reference to Fig. 2.1, the task related to defect prediction is represented in blue, while the task concerning manufacturing time regression is shown in red. A comparison among Fig. 2.1 (a), (b), and (c) reveals that all physical characteristics associated with defects exhibit a strong dependence on product type and product line, displaying similar patterns of influence. Fig. 2.1 (d) shows that stock origin deviates somewhat from this trend, as both the collection and launch season features acquire greater relevance. This



(a) defect type (b) defect main location (c) defect detailed location

(d) stock origin (e) process time

FIGURE 2.1: SHAP values illustrating the importance of features across different targets and models

finding is reasonable, given that the type of defect is expected to be closely linked to product type and product line, whereas for stock origin, indicating whether a claim was raised by a retailer or by an end customer, other features, albeit secondary, may also affect the outcome. Based on the distribution of feature importance, the model's decision-making process, although not inherently transparent, may nonetheless be regarded as logically consistent.

Fig. 2.1 (e) reports the feature importance for the regression on manufacturing time. Interestingly, two features stand out for their relevance: difficulty, a subjectively assessed variable that encapsulates multiple contributing factors and reflects the manufacturer's expertise and product category.

2.5 Addressing Human Oversight Requirements

In the context of low-risk AI systems, no specific provision of the [1] directly applies. Likewise, Art. 22(1) of the [2], as anticipated by Recital (71), appears inapplicable, since no

individuals are directly affected by the outcomes of the decision-making process. Nevertheless, if the system were allowed full autonomy in decision-making, the outcomes could either degrade or impair production, potentially causing economic harm to the company. Therefore, ensuring adequate human oversight remains a priority to prevent adverse effects and maintain alignment with the guidelines for trustworthy AI.

Below, we present the questionnaire proposed by [16] for self-assessing required oversight measures in AI applications, along with the corresponding responses for the low-risk use case from Chapter 3, which also represent the case study in Chapter 4.

Q1 Please determine whether the AI system (choose as many as appropriate):

- Is a self-learning or autonomous system;
- Is overseen by a Human-in-the-Loop (HITL);
- Is overseen by a Human-on-the-Loop (HOTL);
- Is overseen by a Human-in-Command (HIC);

The AI system incorporates a Human-in-the-Loop (HITL) approach, as human personnel are actively involved at each stage of the application life-cycle: given a new product with specific characteristics, the predictive quality analyst forecasts potential defects supported by a forecasting AI application trained on past products and their claims history and provides recommendations to enhance the product's quality; The manufacturer can use the forecast as an indication of the production timeline for a new product, starting from raw materials, while still relying on their personal expertise to identify adjustments that could shorten manufacturing time.

Q2 Have the humans (human-in-the-loop, human-on-the-loop, human-in-command) been given specific training on how to exercise oversight?

Both the predictive quality analyst and the manufacturer can already rely on their expertise.

Q3 Did you establish any detection and response mechanisms for undesirable adverse effects of the AI system for the end-user or subject?

No specific adverse effect is anticipated. In the worst-case scenario, a defect that would have manifested anyway will occur, meaning that the software would have been ineffective but not harmful.

Q4 Did you ensure a 'stop button' or procedure to safely abort an operation when needed?

A dedicated 'stop button' is unnecessary for this AI systems.

Q5 Did you take any specific oversight and control measures to reflect the self-learning or autonomous nature of the AI system?

Not yet.

2.6 Addressing Accuracy and Robustness Requirements

A thorough analysis of the strategies implemented to optimize the performance of the low-risk case studies is provided in Chapter 3 and Chapter 4. In both instances, a rigorous comparison across alternative models is conducted, and, where available, established benchmarks are employed to substantiate the superiority of the automated approaches over human decision-making. In doing so, particular attention is paid to paragraphs 2 and 3 of Art. 15 in [1].

Chapter 3

Low-Risk AI Case Study I: FashionDSS

The fashion industry, which includes clothing, footwear, make-up and other accessories, is estimated to be worth more than 3 trillion US dollars[1]. This colossal industry provides the diverse apparel and accessory needs of the world's population. The digitization of the fashion retail supply chain, coupled with emerging trends in consumer behaviour, has given a significant impetus to fashion e-commerce [19]. The e-commerce fashion industry is expected to experience a compound annual growth rate (CAGR) of 14.2% from 2017 to 2025, reaching a valuation of $1 trillion by 2024 [20]. Sales of apparel, footwear, and accessories are steadily increasing, with the U.S. market alone reaching $204.9 billion. This figure is projected to grow by 13% this year, with consumers expected to spend $204.9 billion on online fashion purchases[2]. The luxury fashion industry has always been at the forefront of defining trends, expanding frontiers and creating desirable products with a careful combination of time-honoured tradition and innovation [21]. Maintaining the highest standards of quality is a non-negotiable aspect in this sector, where consumer perception and experience have a direct impact on a brand's reputation and commercial success [22]. However, even within this sphere of excellence in quality and care, product defects can occasionally occur, leading to customer complaints and dissatisfaction [23].

[1] https://fashionunited.com/global-fashion-industry-statistics/
[2] https://www.shopify.com/enterprise/ecommerce-fashion-industry

Recognising and predicting these potential defects has traditionally relied on meticulous craftsmanship and rigorous quality checks [24]. However, the rise of big data and predictive analytics offers an unprecedented opportunity to enhance these processes, shifting quality management from a reactive to a more proactive mode [25], [26]. A defect in a fashion item typically refers to any irregularity, fault or flaw that deviates from the expected quality, design or performance of the product [27]. These defects can include stitching errors, faulty zips, inconsistent colouring, incorrect sizing, material discrepancies or any other imperfection that adversely affects the overall appearance, functionality or durability of the item. Importantly, the identification of these defects can occur either at the point of sale or after purchase, by either the seller or the customer respectively. At the point of sale, the seller could identify the defect during quality control, while preparing the product for display, or even during the transaction process. For example, an employee may notice a torn seam while folding a garment, or a mislabelled size while restocking.

Despite the availability of various technologies and methodologies to improve manufacturing processes and quality control, it remains a significant gap in effectively predicting and mitigating potential product defects before they impact customer satisfaction. Most existing systems focus on reactive measures, addressing problems after they have occurred, rather than proactively identifying and resolving potential problems during the design and manufacturing stages [21]. To address the challenges of the luxury fashion industry, DSS can benefit greatly from the use of high-quality information encapsulated in customer models (CMs) [28]. These models, created through the application of AI techniques, provide a rich foundation for making informed decisions [29], [30], [31]. They enable the personalisation and adaptation of products and services to closely match the customer's unique needs and preferences, thereby improving the overall customer experience and satisfaction [32]. A data-driven approach can be effectively implemented in the early stages of the product life-cycle, particularly during the design phase, to improve quality management [33], [34]. Using historical data from similar products makes a predictive quality management system feasible, offering significant advances in proactive defect prevention [35], [36]. This type of system, when integrated during the preliminary design phase of new fashion items, allows the analysis of the BOM and the list of required operations, and uses this data to calculate the probability of certain defects. As a result, fashion designers can identify

potential weaknesses in material selection or manufacturing processes before these issues appear in the final product, as already done in the automotive sector [37], [38]. Such preventive analysis not only improves product reliability but also reduces waste, rework and associated costs by allowing informed changes to be made early in the design process.

In this paper, FashionDSS is designed to bridge this gap in the fashion domain by providing a comprehensive approach to predicting manufacturing process times and potential product defects in the luxury fashion industry. The objective of FashionDSS is twofold: to improve operational efficiency by accurately predicting the time required to manufacture prototypes (Task T1), and to improve product quality and customer satisfaction by predicting the most likely defects in new products (Task T2). Using a collected dataset of customer complaint history, this research aims to highlight the unexploited value of such feedback, not just as an expression of dissatisfaction, but as a critical, data-rich source of tangible insights into product performance, consumer expectations and latent product defects. This dataset is unique in being a comprehensive aggregation of three key sources of information: the Product Registry, the BOM and the Claim Record. The acquisition of such high-quality dataset in the luxury fashion industry presents significant challenges due to the exclusive and highly confidential nature of the sector. The dataset used in this study, although relatively small, including 600 products and 1800 defect claims, provides an exceptional depth of information that is rarely accessible in this sector. It integrates detailed product specifications, comprehensive material and manufacturing process data, and nuanced defect reports. This level of granularity reflects the complexity of luxury fashion production, where limited product quantities, customized designs and sensitive data collection practices inherently limit the size of the dataset. Despite these limitations, the richness and complexity of the data provide a robust foundation for developing advanced predictive models that enable proactive quality management and production optimization. By incorporating data-driven insights into material selection and process planning, FashionDSS facilitates the creation of fashion items that meet higher quality standards, leading to greater consumer satisfaction and a more sustainable manufacturing cycle. This approach represents a significant shift from reactive to proactive quality management and highlights the importance of predictive analytics in modern product design and development. In addition, FashionDSS can estimate production time for previously unproduced

items, taking into account factors such as material properties, quantities, operations and the complexity of context switching between different production tasks, as well as other less quantifiable aspects. This approach aims not only to improve product quality and reduce the incidence of defects but also to improve customer satisfaction by addressing and mitigating problems before they affect the end consumer. In doing so, the paper aims to contribute a novel perspective to the literature on quality assurance and customer feedback management in the luxury fashion sector, with practical implications for brands seeking to achieve excellence in product quality and customer experience.

While our research is primarily focused on the luxury fashion industry, the methodologies and models developed herein have broader applicability across various retail sectors. This general applicability stems from the common characteristics that are prevalent in all retail domains, such as the availability of historical data to support forecasting accuracy, the need to predict outcomes for newly introduced products with no prior time series data, the maintenance of comprehensive registries detailing the attributes of both past and future items, and the clear separation of products across seasons with no overlap. The adaptability of our proposed approach therefore lies in its ability to exploit these similarities, enabling it to provide valuable insights and predictive capabilities not only in the luxury fashion domain but also in other retail contexts where similar challenges are encountered. This cross-industry applicability underlines the potential of our work to significantly impact a wide range of retail operations, providing a versatile tool for improving product quality, anticipating consumer needs and optimizing inventory management.

The main contributions of this paper are manifold and reflect significant advances in the application of data analytics to the luxury fashion industry. These contributions are not only central to the field of fashion retailing but also have broader implications for data-driven decision-making in retail and product quality management. Key contributions include: i) the introduction of an innovative approach to transforming customer complaint data into a proactive tool for product improvement; ii) the development and the validation of a comprehensive analytical framework that integrates product, material and defect data for predictive analysis; iii) the implementation of advanced AI-based models for predicting production efficiency (Task T1) and potential product defects (Task T2) before production starts iv) potential generalizability of the framework to other retail domain, beyond the

luxury fashion sector; v) actionable strategies for improving product quality, reducing defect rates and increasing customer satisfaction through proactive decision making; vi) the introduction of a unique and high-value dataset providing detailed product, material and defect information for the luxury fashion industry. By demonstrating how detailed analysis of historical data can inform future product development and quality assurance strategies, the paper encourages retailers to adopt a more data-centric approach to decision-making.

The paper is structured as follows: Section 8.1 provides a comprehensive overview of the research design, data processing and analysis techniques that are the basis of our study, ensuring replicability and transparency. In Section 3.2, we present the results of our research and demonstrate the predictive accuracy of our models in terms of process time estimation and defect prediction. Finally, Section 3.3 concludes by summarizing the key contributions and findings, underlining the importance of our innovative approach to using customer complaint data for predictive analytics. This section also outlines directions for future research, suggesting how the methods and findings of this study can be further refined and applied to other areas of the retail industry. Potential areas for further development, including the application of more advanced analytical techniques and the exploration of additional data sources, are identified to encourage continued innovation and research in this area.

3.1 Materials and Methods

In this section, we outline the basic elements and analytical techniques that constitute our FashionDSS. Central to our investigation are two tasks, an extensive data set, and the predictive models we employed to forecast manufacturing times and anticipate product defects in the luxury fashion sector. Each component is critical to the understanding the the novel approach our study introduces with the use of customer feedback and manufacturing data for accomplishing quality assurance and improve operational efficiency in fashion retailing. FashionDSS is being deployed in an Italian fashion company of a prestigious luxury fashion brand, embodying a state-of-the-art DSS designed to significantly improve the brand's product development and quality assurance processes. FashionDSS

addresses two tasks and we refer to the first task as **T1** and the second task as **T2**. Here is a detailed description for each of them:

T1: For a prototype, i.e. a product that has never been manufactured before, the framework is asked to predict the process time. The process time, a sub-interval of the manufacturing time, is the time required to manufacture the final product from the raw materials.

T2: At some point during its life, any product may prove to be or become defective. This event may occur earlier, after the product has been distributed but before it has been retailed, or later, after the product has been retailed, in which case it is usually the customer who has purchased the product who makes a claim. Here, the framework is asked to predict the most likely defect for a given new product.

The importance of **T1** and **T2** cannot be overstated, as an estimate of the process time can help the company to allocate resources carefully, or even to decide whether a product should be adapted to reduce the process time. On the other hand, any foreknowledge of possible defects can help to take prompt corrective actions to improve customer satisfaction.

The details of FashionDSS components are explained in the following subsections to ensure a thorough understanding of our methodology and its application. Figure 3.1 schematically depicts the workflow of our FashionDSS.

3.1.1 Dataset

As shown in Figure 3.1(a), the dataset encompasses three main sources of information: the Product registry, which contains the general characteristics of the product, such as colour, product type, product line; the BOM, which contains a detailed list of all the materials used to manufacture the final product and, for each material, the exact quantity required, the function of the material (inner, outer, reinforcement), and also the list of all the operations performed during the manufacturing process along with the time required to complete each step; the Record of Claims provides detailed information about

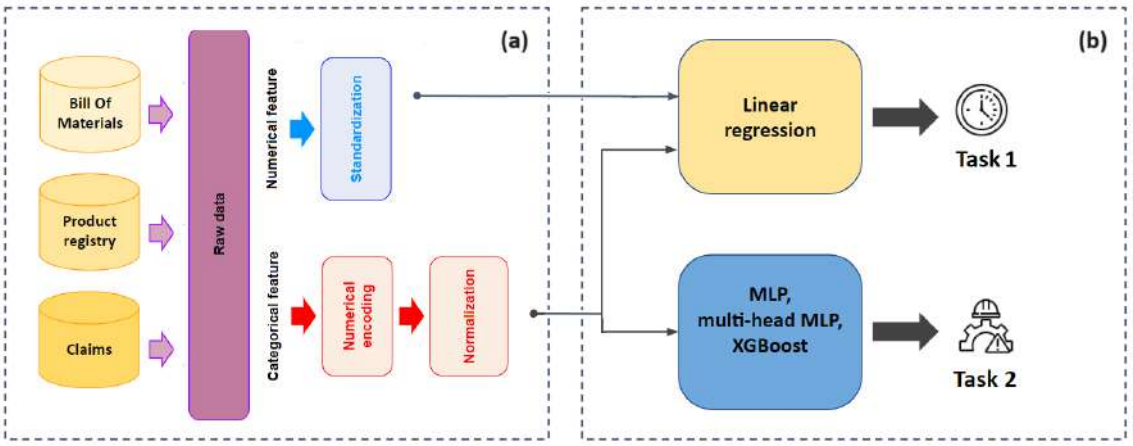


FIGURE 3.1: FashionDSS workflow. a) Data pre-processing and feature extraction b) Proposed models to solve task **T1** and task **T2**. All models are trained by exploiting data collected for products already manufactured and commercialised, for which both the complete BOM and the claims record are available

any product claim, including details of the customer who filed the claim if the defect was discovered post-purchase, or the retailers (or their representative) if the defect was identified pre-retail. Importantly, it also includes comprehensive information about the type of defect discovered. The number of products in the dataset is 600, while the number of claims is 1800. For training, **T1** relies mainly on information from the BOM available for each of the 600 products, and **T2** relies mainly on information from the Record of claims for each of the 1800 claims. Although the dataset comprises 600 products and 1800 defect claims, its size is inherently limited by the unique and highly specialised content of the information it contains. In the luxury fashion industry, access to such detailed and sensitive data is extremely rare. Luxury brands work in competitive environments where manufacturing processes and defect data are closely protected due to concerns over intellectual property and brand reputation. Gathering this level of granular data - including material specifications, manufacturing workflows and detailed defect reports - is both costly and demanding logistically. Luxury fashion production typically involves limited product quantities, reflecting a business model based on exclusivity and craftsmanship rather than mass production. As a result, production and defect data in this sector is naturally smaller in scale, but much richer in detail. In addition, the complexity of standardising and systematising defect reporting across tailored products further limits the scalability of such datasets. Each product may have unique designs, materials and craftsmanship, making

defect classification and reporting inherently complex. However, the depth of information compensates for the limited size of the dataset, providing a high-quality, multi-dimensional resource that enables sophisticated analysis of production processes and product performance.

3.1.2 Data privacy concerns

The dataset used in the present work does not contain any sensitive data, such as customers' personal data, which were initially filtered out because considered of negligible contribution. Complaints themselves are not formulated as opinions expressed by any subject, rather as objective simple pieces of information (e.g. location of the defect, material affected by the defect).

3.1.3 Dataset split during training and evaluation

For **T1**, dataset was split in training, validation and test sets in the ratio 80:10:10, whereas for **T2** a more complex approach is applied, as shown in Figure 3.2. In this case, nested cross-validation is used, which has a twofold advantage: on the one hand, the final evaluation by the 5-fold outer splits helps to reduce the risk of overestimating performance, and on the other hand, the 4-fold inner loops allow a grid search through the hyperparameter space. The split approach is stratified in that, for each class, it retains the same proportion of samples found in the original dataset. For the sake of reproducibility and fair comparison between different models, both outer and inner splits are computed and stored beforehand.

3.1.4 Data pre-processing, feature selection and target definition

With reference to Figure 3.1, data is parsed from the product registry, the claims record, and the BOM. Zero-variance data is filtered out, as well as any data that lacks essential information (e.g. the name of the customer who made the claim, the progress of the claim). Numerical features are standardised, while categorical features are first numerically encoded and then normalised in the range $[0, 1]$. Numerical encoding is then applied. Indeed,

although one-hot encoding is usually considered the method of choice for neural networks, using it for features with several categories (refer to Table 3.1) introduces sparsity. In particular, the numerical encoding we use consists in frequency encoding, also referred to as count encoding, which encodes categorical features by assigning numerical values based on how frequently each category appears in the dataset. After pre-processing, the most relevant subset of features is defined on the base of experience in a round of discussions with our technology partners which allowed us to leverage their domain expertise.

In order to train **T1**, *collection* and *product category*, two general features from the Product registry, are used in combination with more specific features from the BOM: $matx, y$ represents the total number of different materials used in the manufacturing process, where $x, y \in \{1, 2, 7\}$ are indices identifying the function of the material (1=$inner$, 2=$outer$, 7=$reinforcement$). Optionally, it is possible to use the total amount of materials used for each macro category instead of the number of different materials from it. *Combination* is a technical feature that takes into account combinations of materials and shapes. Finally, *Difficulty* is a feature whose value has been subjectively estimated for each product by experts in the manufacturing process. It varies in the range $[3, 9]$ and is very informative as it strongly correlates with the complexity of a given product and thus with the process time. Although available, low-level details about each operation have been deliberately omitted to avoid the risk of sparse data due to the limited number of products/samples in the dataset. For **T1** the numerical target is time.

In order to train **T2**, features relating to the characteristics of the product that turned out to be defective are taken from the Record of claims. **T2** performs the prediction for four different targets: *stock origin*, which returns the identity of the subject who filed the claim, either the retailer or the customer, and helps to identify the point in time when a defect tends to occur; *defect main location* (e.g. *main body*, *handle*, *lining*) reports the exact location of the defect in the product; *defect detailed location* (e.g. *fabric*, *leather*) refines the information and reports the defective material in the product; *defect type* (e.g. *broken*, *stained*, *stitched*, *unstitched*) provides an accurate description of the type of defect detected. For each target, those categories referring to similar defects or defect location are aggregated to support the decision systems with additional information. A few different types of aggregation are then experimented.

TABLE 3.1: For each task, it is given the list of features, either (n)umerical or (c)ategorical, including the number of categories for the latter. In the last column, the targets for each task are given.

Task	Features	#categories	Targets
T1	collection (c)	2	process time (n)
	product category (c)	10	
	mat1, 2 (n)	-	
	mat7 (n)	-	
	combinations (n)	-	
	difficulty (n)	-	
T2	collection (c)	2	stock origin (c)
	product category (c)	10	defect main location (c)
	product line(c)	80	defect detailed location (c)
	lifespan(c)	3	defect type (c)
	launch season(c)	16	

3.1.5 Predictive Models

To tackle **T1**, a linear regression approach is implemented in the form of a shallow neural network (NN), with no hidden layer, to allow for easy experimentation with different configurations, whereas a larger variety of models are experimented for **T2**, including an ensemble of Multi Layer Perceptron (MLP) models, working independently, one for each target; a unique multi head MLP (mhMLP) model, simultaneously making predictions over all targets; a set of ensembled binary XGBoost models, each ensemble separately trained for a specific target according to One vs All (OvA) strategy and finally providing a different binary classifier for each category of the given target. In order to rule out any misunderstanding, we point out that, although XGBoost is an ensembled method, here the term *ensemble* refer to the use of many binary classifiers in lieu of a unique multi-class classifier. During hyper-parameter grid search, the shallow architecture at the core of the the MLP/mhMLP is occasionally transformed into a deeper architecture to increase complexity, by adding either one or three hidden layers. In the following section, details are provided for each of the model.

3.1.5.1 Linear regression

This model, in charge for the forecasting of process times, defined as task **T1**, implements linear regression, relying on a shallow NN. More formally, we want to learn a linear regression function f defined as follows:

$$f: \mathbb{R} \cup \{0,1\}^n \rightarrow \mathbb{R} \tag{3.1}$$

where the target represents the predicted process time and features can be either numerical or categorical. Mean Squared Error (MSE) is used for the loss:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2 \tag{3.2}$$

where N gives the number of samples/products in the dataset, x_i represents the predicted process time for sample i and y_i the actual process time recorded during the manufacturing of sample i.

In most experiments, a shallow/linear NN is used, in which case a linear activation function is added at the top of the network. In some experiments, during an initial exploratory phase, when non-linearity was of interest, hidden layers were added to make the shallow NN deeper, and an activation function of the following form was added between each hidden layer and the one immediately preceding it in the network:

$$Relu(z) = max(0, z) \tag{3.3}$$

3.1.5.2 Supervised multiclass classification via multiple single head MLP

In this first approach, task **T2** is solved relying on multiple single-head MLP models, whose architecture allows for a number of hidden layers which can be equal to 0, 1 or 3. In **T2**, four different targets undergo prediction, as described in 3.1.4, and for each of them a different training of the MLP is performed. The loss consists in Cross Entropy (CE), weighted according to the class imbalance present in the training set and SoftMax is used as the final activation function.

3.1.5.3 Supervised multiclass classification via a single multi head MLP

In this second approach, task **T2** is solved relying on a unique MLP model having four head/outputs. As a consequence, the four different targets undergo prediction simultaneously and only one training process is performed. Also in this case, the number of hidden layers can be equal to 0, 1 or 3, the loss consists in weighted CE and SoftMax is used as the final activation function.

3.1.5.4 Supervised multiclass classification via XGBoost

In this third approach, task **T2** is solved via XGBoost [39]. For each target, the classification problem is binarized and one binary model is trained for each single category of the target. Training data are rearranged according to OvR strategy.

3.1.5.5 Training phase

With reference to task **T2**, for all models, the training procedure is repeated for each inner split returned by nested cross validation and the balanced accuracy of the splits is then averaged to identify the best performing hyper parameters. Hyperparameters for the MLP/mhMLP include number of *epochs* (200 or 300), initial *learning rate* for Adam optimizer (0.001 or 0.0001) and number of *hidden layers* (no hidden layers, 1 hidden layer with 16 units or 3 hidden layers with 16, 32 and 64 units respectively). In all experiments, the batch size was 32. Hyperparameters for the XGBoost include *learning rate* (0.01, 0.05, 0.1, 0.15, 0.2, 0.25, 0.300000012, 0.310000012, 0.320000012), *gamma* (0, 0.05, 0.1, 0.15, 0.2), *max depth* (3, 5, 6, 8, 10) and *min child weight* (1, 3, 5).

With reference to task **T1**, our experiments use a batch size of 32 and the model is trained over 1,000 epochs. Adam optimiser is used with a learning rate of 0.1. It is important to note that process times of less than 20 minutes were excluded from our analysis because likely erroneous.

3.1.5.6 Evaluation phase

For task **T2**, after the training phase, for each model category, the model whose hyperparameters performed best through inner cross validation is retrained. Then, for each outer fold, the retained model is retrained through the complete outer train set and evaluated through the (outer) test set. Results for all the 5 outer fold are finally presented via box plots.

For task **T1**, evaluation is performed on the test set after standard training.

3.1.5.7 Evaluation metrics

For regression task **T1**, performance is evaluated by means of Equation (8.6), as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2} \tag{3.4}$$

where the argument of the square is the MSE defined in 8.5

For the multi-class classification task **T2**, the performance of different models is evaluated by means of balanced accuracy (b-acc), f1-score, precision (prec) and recall, whose general formulas follow:

$$prec = \frac{TP}{TP + FP} \tag{3.5}$$

$$recall = \frac{TP}{TP + FN} \tag{3.6}$$

$$f1\text{-}score = 2 \cdot \frac{prec \cdot recall}{prec + recall} \tag{3.7}$$

$$b\text{-}acc(y, \hat{y}, w) = \frac{1}{\sum \hat{w}_i} \cdot \sum_i 1(y_i = \hat{y}_i) \cdot \hat{w}_i \tag{3.8}$$

$$\text{where } \hat{w}_i = \frac{w_i}{\sum_j 1(y_j = y_i) \cdot \hat{w}_j}$$

In particular, balanced accuracy is used during hyperparameter search to pick the best configuration. Although the splits of the dataset are stratified, considering the imbalance all targets (apart from *stock origin*) are affected by, with some categories much larger

than other, balanced accuracy seems a reasonable metric to study performance. In addition to the original data imbalance, when multi-class problems are binarized in the case of XGBoost, further imbalance can be introduced as an artifact of binarization, supporting the use of balanced accuracy. F1-score, precision and recall are to be intended as *micro*, that is computed globally through all available classes. In any case, the behaviour of each model is finally discussed in view of all proposed metrics. For each of them, metrics are computed for all the 5 outer folds of nested cross validation and reported via box-plots in Figure 3.3. Table 3.1 numerically sums up the same results via the Median and the Interquartile Range (IQR).

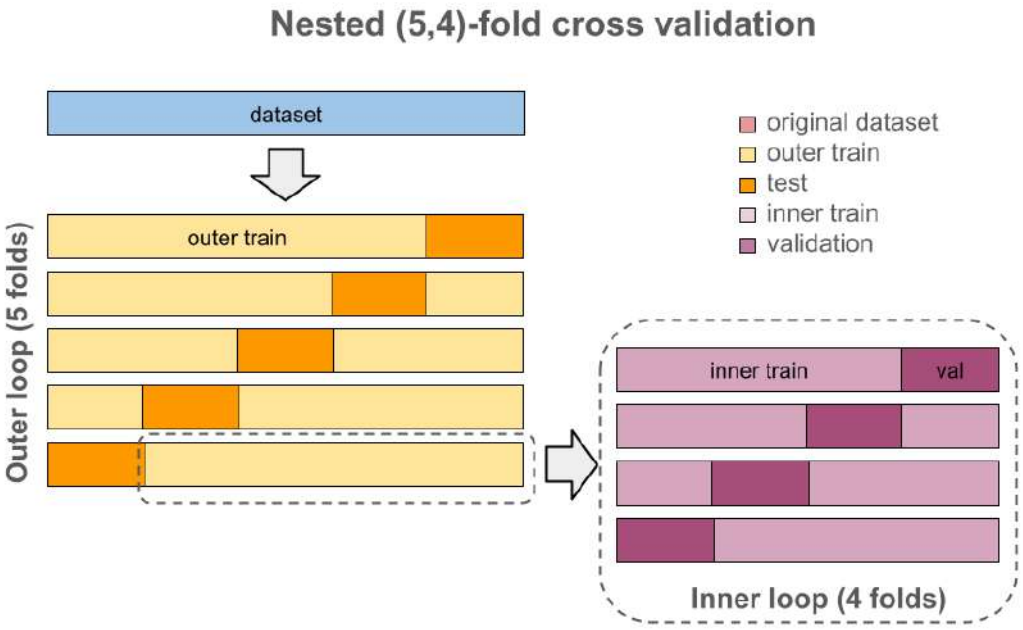


FIGURE 3.2: Nested cross-validation is applied with 5 outer folders and 4 inner folders respectively. The inner loop is responsible for hyper-parameter tuning whereas the outer loop is used to evaluate each model and allows to mitigate the risk of overconfidence

3.2 Results and Discussions

In this section, we present the results of our study. On the base of them, we define our proposed workflow FashionDSS and discuss its applicability to real world scenarios.

As far as results relating task **T2** are concerned, with reference to Figure 3.3, all metrics support the superiority of XGBoost. Table 4.2 further backs this condition

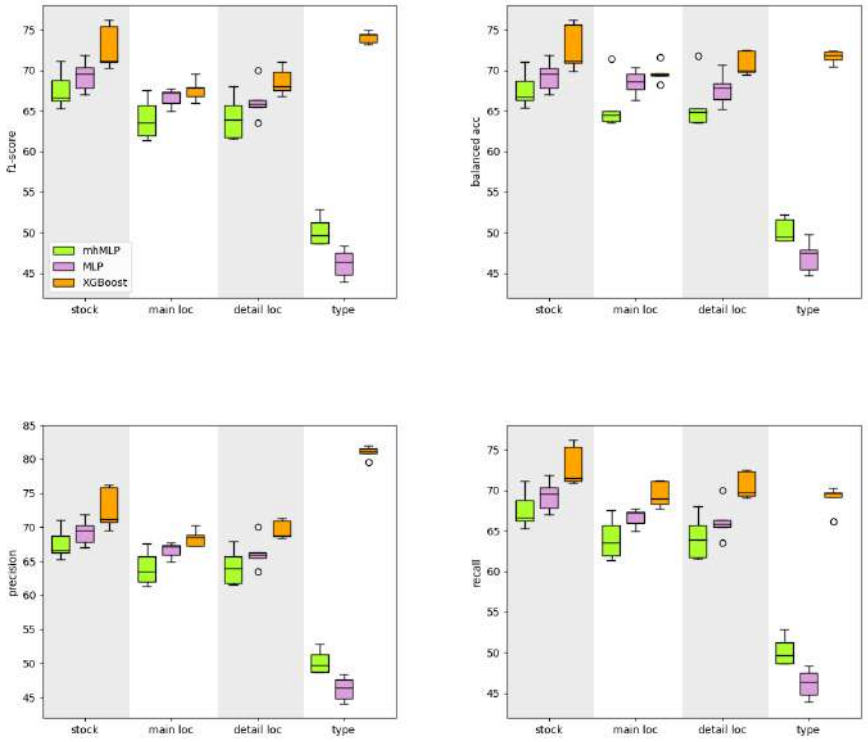


FIGURE 3.3: Box-plots for Balanced accuracy, F1-score. Precision and Recall obtained when **mhMLP** (green), **MLP** (purple) or **XGBoost** (orange) is applied.

up, showing that XGBoost routinely overcomes both MLP methods in the Median and almost always displays the narrowest IQR, which is indicative of greater stability in results through different outer folds. Also precision and recall are most times above 70% for XGBoost. From Table 4.2, a change in the aggregations of classes for target *detailed location* and target *type* clearly benefits XGBoost suggesting that a search in the domain of different aggregations may lead to much better results for this method. With reference to Figure 3.4, we tried two different aggregations for target *defect detailed location* and target *type*. The former, target *defect detailed location*, includes five categories/classes initially arranged into two groups as follows: ¡*Fabric*, *Leather*, *Other Materials*¿ and ¡*Functional Accessories*, *Ornamental Accessories*¿. As an alternative aggregation, we tried to further split the first group into ¡*Fabric*, *Leather*¿ and ¡*Other Materials*¿. Similarly we reasoned for target *type*, investigating a different and more informative aggregation, which proved to be advantageous.

TABLE 3.2: Median with IQR in brackets for all metrics and targets, inclusive of alternative aggregation modes. With reference to each column, the highest median and the lowest IQR (best results) are highlighted in red whereas the lowest median and highest IQR (worst results) are highlighted in blue

		stock	main loc	detail loc (A)	detail loc (B)	type (A)	type (B)
f1-score	**MLP**	69.51 (3.67)	67.12 (2.06)	65.89 (3.63)	58.48 (4.9)	46.35 (3.57)	47.82 (2.96)
	mhMLP	66.6 (4.17)	63.5 (4.95)	63.88 (5.15)	47.57 (10.03)	49.71 (8.69)	41.55 (2.98)
	XGBoost	71.04 (5.15)	67.78 (2.33)	67.99 (3.25)	77.47 (2.9)	74.39 (1.4)	79.81 (1.05)
b-acc	**MLP**	69.5 (3.65)	68.53 (2.96)	67.83 (3.7)	60.35 (4.78)	47.49 (3.72)	52.35 (4.62)
	mhMLP	66.66 (3.93)	64.48 (4.51)	64.85 (4.95)	51.61 (6.24)	49.49 (8.77)	47.35 4.31
	XGBoost	71.07 (5.48)	69.4 (1.84)	69.93 (2.82)	71.45 (4.55)	71.87 (1.49)	73.23 (0.38)
prec	**MLP**	69.51 (3.67)	67.12 (2.06)	65.89 (3.63)	58.48 (4.9)	46.35 (3.57)	47.82 (2.96)
	mhMLP	66.6 (4.17)	63.5 (4.95)	63.88 (5.15)	47.57 (10.03)	49.71 (8.69)	41.55 (2.98)
	XGBoost	71.15 (5.91)	68.43 (2.35)	68.75 (2.55)	81.51 (2.74)	81.04 (1.51)	87.97 (0.88)
recall	**MLP**	69.51 (3.67)	67.12 (2.06)	65.89 (3.63)	58.48 (4.9)	46.35 (3.57)	47.82 (2.96)
	mhMLP	66.6 (4.17)	63.5 (4.95)	63.88 (5.15)	47.57 (10.03)	49.71 (8.69)	41.55 (2.98)
	XGBoost	71.46 (4.76)	68.9 (3.1)	69.64 (3.21)	74.51 (4.04)	69.55 (2.36)	73.53 (1.72)

TABLE 3.3: Results for the forecast of process time in **T1**

Target range in minutes	RMSE on test set
[20, 250]	27.72
[20, 270]	27.83
[20, 280]	38.90
[20, 300]	47.60
[20, 325]	41.98
[20, 350]	51.38
[20, 375]	45.62
[20, 400]	42.81

Compared to XGBoost, MLP-based models lag behind for all targets, with a particularly dramatic drop for target *type*, the one with more categories. A comparison of the multiple single-head MLP models *versus* the multi-head single MLP model shows that simultaneous prediction reduces performance.

Our attention back to **T1**, with reference to Table 3.3, we see that the prediction of the process time is characterised by an $RMSE$ that depends on the range considered. On the one hand, up to 250 minutes, the error seems to be reasonable, although we could aim for better results. On the other hand, if we consider longer processing times, the

performance deteriorates. This could be a consequence of the imbalance in the dataset, where the longer process times are less frequent than the shorter ones. We could not do much about this kind of imbalance, as the number of samples with longer process times was very small.

Considering the results we have collected, we can define our proposed workflow, FashionDSS, as a combination of a linear regression component, implemented in the form of a shallow NN, to handle task **T1**, and an ensemble of carefully tuned XGBoost models to handle task **T2**. The results of our experiments demonstrate how FashionDSS has the potential to significantly impact decision-making processes in luxury fashion manufacturing and quality control. Finally, given that a minimal set of features has been exploited, we expect that relevant improvements can be achieved by introducing new informative variables.

3.3 Conclusions and future developments

This paper presents an innovative approach to using customer complaint data to anticipate potential defects in luxury fashion products, marking a significant shift in the way such data is perceived and used in the industry. Through the development and implementation of a DSS, called FashionDSS, we demonstrated the feasibility of predicting the location and severity of potential defects, ranging from those that are immediately apparent at the point of sale to those that only surface after a period of use by the customer. Our findings show that by leveraging the power of AI and ML techniques, coupled with a comprehensive dataset comprising the Product Registry, BOM and the Record of claims, it is possible to significantly improve prediction accuracy. This not only enables early identification and mitigation of potential product defects but also facilitates a more personalized and proactive approach to product design and manufacturing. The practical implications of this research are profound, providing luxury fashion brands with a novel tool to improve product quality, reduce the incidence of defects and thereby increase customer satisfaction and loyalty. In addition, our work contributes to the broader discourse on the application of data analytics to improve operational efficiency and decision-making in the retail sector.

Additional research directions have been identified. First, the potential for applying FashionDSS predictive models across different segments of the fashion industry and beyond suggests the need for broader validation studies. Exploring the applicability and effectiveness of the system in different retail contexts could provide deeper insights into its versatility and adaptability. Second, the integration of real-time data analytics represents an exciting frontier. By incorporating live customer feedback and production data, FashionDSS could offer even more dynamic and responsive predictive capabilities, further reducing the time between defect detection and corrective action. Third, further refinement of the AI and ML models used in FashionDSS could improve predictive accuracy and efficiency. Exploring advanced algorithms and incorporating additional data sources, such as social media sentiment analysis or global supply chain information, could provide more nuanced and comprehensive insights into potential product defects. Finally, a deeper exploration of the ethical considerations and consumer privacy implications associated with the use of customer data in predictive modelling is warranted. As AI and data analytics play an increasingly important role in retail decision-making, it will be crucial to ensure that these technologies are used responsibly and transparently.

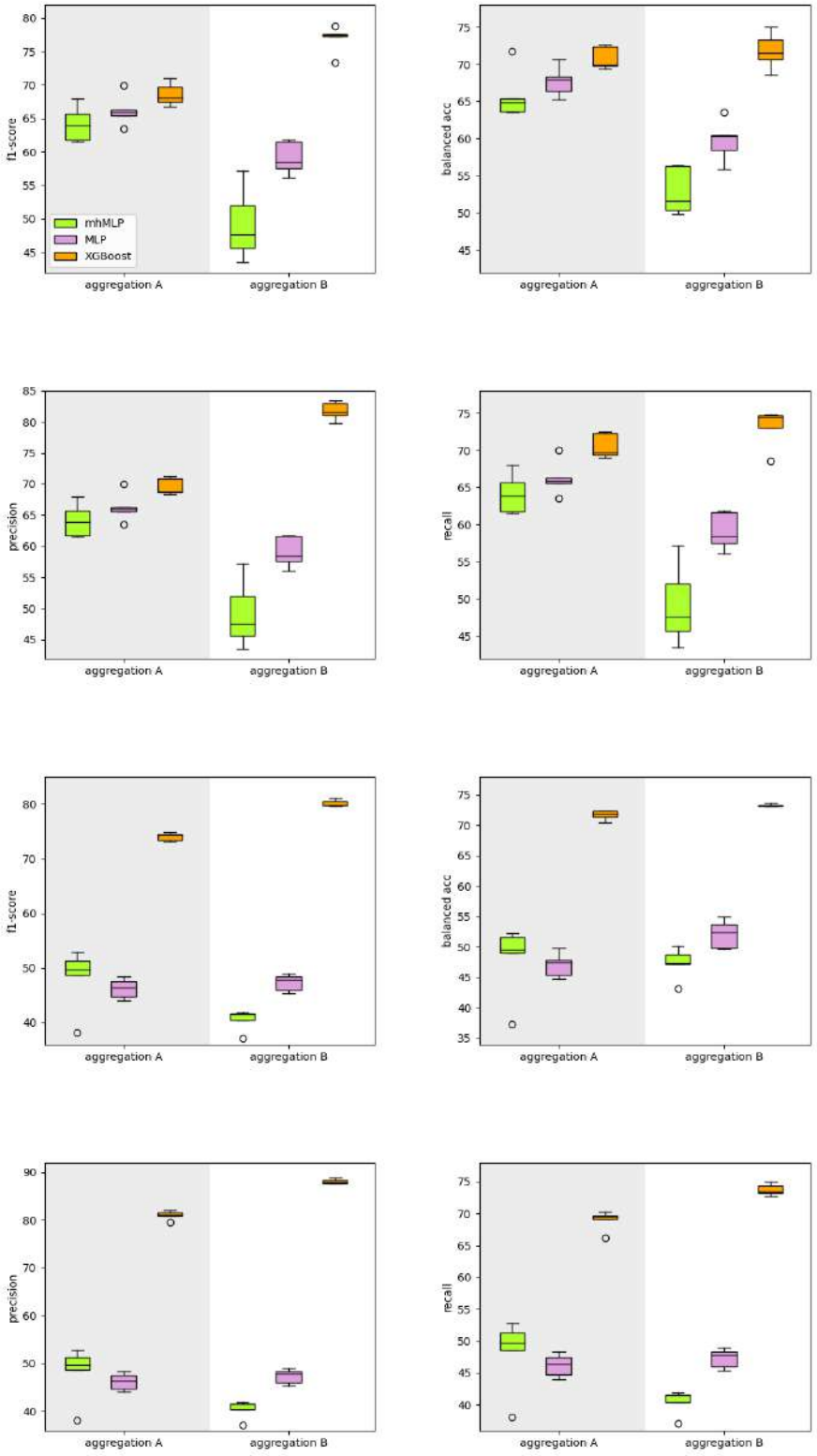


FIGURE 3.4: Box-plots for two different aggregations A and B

Chapter 4

Low-Risk AI Case Study II: FashionSight

In the fashion industry, sales forecasting is a highly important task, as decisions regarding procurement, distribution and merchandising depend heavily on the ability to anticipate consumer demand [40, 41, 42, 43]. In the luxury fashion sector, the consequences of these decisions are even more significant [44, 45]. Products tend to be expensive, seasonal and produced in limited quantities. Consumer behaviour is influenced by not only style trends, but also tourism flows, localised events and socio-economic dynamics. Furthermore, product performance is often market-specific, with different regions exhibiting different sales profiles for the same item [46, 47, 48]. Despite advancements in forecasting methodologies, the unique properties of luxury fashion sales, such as intermittent demand, limited historical data, and market heterogeneity, make this a conventional Artificial Intelligence (AI) models [49]. At the same time, human planners and analysts remain deeply involved in forecasting processes, relying on a combination of experience, intuition and contextual information. Therefore, forecasting systems must do more than just automate predictions; they must also support decision-making under uncertainty in collaboration with human stakeholders.

Many current forecasting tools operate as 'black boxes', often requiring centralised

data and offering little interpretability or transparency. In highly regulated or privacy-sensitive environments, such as those of luxury fashion brands, this limits their usability. Furthermore, excluding human expertise from the modelling process often results in poor adoption by business users, who may struggle to trust or act upon the system's output. However, human expertise in luxury fashion remains irreplaceable. For example, a merchandiser may know that a product sells better in summer due to cultural preferences in a particular region or that a sudden spike in sales of a low-performing Stock Keeping Unit (SKU) is due to a one-off marketing campaign. This type of knowledge is local and contextual and cannot always be inferred from data alone. Therefore, forecasting systems must support expert judgment rather than replace it, offering a collaborative framework in which data and human insight can work together. This requires a HITL approach to sales forecasting, where forecasting models are transparent and configurable and can provide human analysts with meaningful explanations. Such systems are particularly valuable when data alone is insufficient to capture demand dynamics or domain knowledge is essential to ensure relevance and accountability.

Most state-of-the-art forecasting methods focus primarily on predictive performance at the expense of usability, or require access to large, stable and richly annotated datasets [40]. Deep learning architectures, such as Long Short-Term Memory (LSTM) networks [50] and transformers [51], often struggle to generalise when faced with the short, noisy and irregular time series typical of the luxury fashion industry. Foundation models such as TimeGPT offer new capabilities [52], but they raise concerns about explainability, complexity and data security, particularly when external models are used on proprietary datasets. Furthermore, ensemble methods that could enhance robustness are rarely adapted to facilitate human transparency or localised model control. Few existing systems enable domain experts to understand why a forecast was made, what data it was based on or how recent history affects the prediction. In practice, luxury fashion brands often rely on expert-curated spreadsheets or manual overrides, which limits the scalability and reproducibility of forecasting workflows. Furthermore, little research has been conducted into how humans can remain involved in forecasting tasks, particularly in domains such as fashion, where expert insight is essential. There is a significant gap in the design of forecasting systems that are both methodologically sound and usable in practice.

In this light, this paper introduces FashionSight (Forecasting Analytics with Semantic Human Interaction for Optimized Navigation in Sales Insight Generation with Human-in-the-loop Transparency), a Human-in-the-loop (HITL) DSS for sales forecasting in the luxury fashion industry. FashionSight is designed to integrate domain-specific constraints—such as market heterogeneity, privacy, and short historical windows, while enabling transparent and expert-aware forecasting. FashionSight combines interpretable ensemble models with a weighting mechanism based on recent in-sample performance, allowing the model to dynamically adapt while remaining understandable to business users. Rather than aiming to replace planners or category managers, FashionSight is designed to amplify their decision-making with traceable, contextual predictions and error-aware output. By emphasizing transparency, modularity, and privacy, FashionSight addresses a critical need in luxury fashion: a forecasting system that aligns with both business workflows and data science best practices, while keeping humans at the center of the analytical process.

The key contributions of this work are as follows: i) a HITL forecasting system specifically designed for sales forecasting in the luxury fashion industry. The system is built with human-in-the-loop principles to ensure interpretability, configurability and business alignment. ii) Market- and SKU-specific modelling: FashionSight handles each time series independently to avoid global model pooling. This approach safeguards data privacy and enables the generation of highly granular, domain-specific forecasts. iii) Transparent Ensemble Methodology: We design a weighted ensemble of machine learning models using in-sample error smoothing to guide the aggregation strategy. iv) The system is tested using data from the 14 international markets of a global luxury fashion brand with a forecast horizon of six months.

The paper is organized as follows: Section 4.1 reviews relevant literature in the fields of time series forecasting and retail analytics, with a focus on luxury fashion. Section 4.2 presents the materials and methods, including a description of the dataset, the business benchmark, the forecasting architecture, and privacy considerations. Section 4.3 discusses the experimental results and ablation studies. Finally, Section 4.4 concludes the paper and outlines directions for future work.

4.1 State of the Art

In this section, we provide an overview of recent developments in sales forecasting, focusing particularly on machine learning-based approaches applied in various fields such as retail, manufacturing and fashion. Although there has been significant progress in shifting from traditional univariate forecasting methods to more complex, data-driven techniques, many existing solutions are still disconnected from the practical constraints and organisational realities of high-end fashion retail. We examine contributions exploring feature selection, hybrid modelling strategies, similarity-based learning and deep neural networks, highlighting their strengths and limitations when applied to contexts characterised by data sparsity, product heterogeneity and the need for human oversight.

Significant progress has been made in the field of sales forecasting, particularly in the transition from traditional univariate techniques to more sophisticated approaches [53]. For instance, Nguyen et al. [54] proposed a model for predicting the sales of remanufactured products, using two feature selection strategies: Boruta and Recursive Feature Elimination. The researchers then trained several models, including regression tree algorithms (CART, M5 and RF), linear regression and an artificial neural network. In another study, Dai and Huang [55] focused on sales prediction in the automotive industry. Rather than identifying the most relevant input variables, they concentrated on finding products similar to a new one based on Euclidean similarity. The products were ranked and a subset (k) was selected based on a metric called sales consistency. Sales were estimated through a weighted average of the historical sales of the selected products. This method proved to be more effective than traditional regression techniques. Furthermore, Miguéis et al. [56] used a range of machine learning and statistical models, including LSTM networks, feed-forward neural networks, support vector regression, random forests and the Holt–Winters model, to predict the sales of a particular type of fish in the food retail industry. The aim was to identify the most accurate predictive approach.

There is relatively little research focusing on sales prediction for new products. In [32], the authors developed a machine learning-based approach to estimating the sales of new items using data from previous product collections. They used a brute-force grid search to identify key variables, applying both shallow regression algorithms and deep

learning techniques. To validate their results and minimise overfitting, they performed a bootstrapping evaluation, finding that Random Forest and Deep Neural Networks yielded the most promising results. Giri et al. [57] introduced a novel method that uses the Inception V3 deep neural network [58] to extract visual features from images of women's apparel. These features were combined with past sales records and used to train a non-linear neural network regression model for forecasting future sales. Furthermore, in [59], the authors proposed a hybrid model incorporating K-means clustering [60], Random Forest [61] and Quantile Regression Forest [62] to estimate overall sales and their progression over time. They applied their method to five distinct datasets from various industries, each containing unique features. Rather than selecting variables, the model used all available inputs. Clustering was employed to group similar demand patterns, followed by classification to assign categories and regression to forecast cumulative sales and associated uncertainty.

In a more recent contribution that focused specifically on the fashion retail sector [42], a two-stage methodology was introduced to address the issue of censored demand, which is often overlooked in traditional models. Four different approaches were initially tested to reconstruct demand from the historical sales data of previously marketed products: three were based on sales-weighted averages, and one used an Expectation-Maximisation (EM) algorithm that considered substitution effects among products. The EM-based method yielded the most accurate estimation of primary demand. In the subsequent stage, this refined demand data was used as input for predictive modelling with Random Forest, Deep Neural Networks and Support Vector Regression. Additionally, a similarity-weighted method was implemented where demand from previous collections was used to forecast sales of upcoming items based on their similarity. Recent efforts to address the limitations of deterministic models in fast fashion forecasting have led to the adoption of generative methods. In [63], a novel approach called MDiFF has been proposed. This method uses the capabilities of diffusion models through a two-step, multimodal pipeline specifically designed for new Fashion Product Performance Forecasting (NFPPF). First, a score-based diffusion model is used to generate multiple future sales trajectories over time. These predictions are then refined using a lightweight multi-layer perceptron (MLP) to produce the final forecast. By combining the generative strengths of diffusion processes

with the efficiency of traditional neural networks, MDiFF achieves state-of-the-art performance in forecasting new product sales in the fast fashion industry, effectively addressing uncertainty and generalisation across novel items.

Considering this scenario, our paper proposes a different yet complementary direction by focusing on the practical and logistical challenges of sales forecasting in the luxury fashion industry, where issues such as data sparsity, intermittence, and confidentiality play a major role. While previous work has advanced the use of machine learning for sales prediction in various sectors, including automotive, food retail and general fashion, many of these approaches rely on centralised, high-volume datasets and prioritise model performance over interpretability and human usability. In contrast, we introduce FashionSight: a human-in-the-loop decision support system enabling localised, SKU-level forecasting via a modular ensemble of interpretable models. Rather than relying on pooled training data or global feature sets, FashionSight generates forecasts independently for each product-market combination, making it suitable for highly heterogeneous retail contexts. Furthermore, unlike prior work that treats human expertise as external to the modelling process, our system is explicitly designed to incorporate domain professionals, aligning with the growing need for transparency, adaptability and trust in real-world decision-making. This makes our contribution particularly relevant for luxury fashion brands, where decisions are often made amid significant uncertainty and require both data-driven support and contextual insight.

4.2 Materials and Methods

This section provides a comprehensive overview of the data, evaluation baseline, and methodological framework underlying our study. At the core of our study is a proprietary dataset provided by a global luxury fashion brand, which includes historical monthly sales for a wide range of products across 14 international markets. These markets differ significantly in volume, product mix, and demand patterns, requiring a modeling approach that is both flexible and localized. To assess the effectiveness of our system, we compare it against an internal benchmark that reflects business-as-usual forecasting practices, incorporating manual corrections and expert oversight. The forecasting engine at the heart of

FashionSight is a modular, interpretable ensemble model designed to produce SKU- and market-specific forecasts under real-world constraints such as data sparsity and privacy. In designing the system, we placed a strong emphasis on transparency and traceability, ensuring that forecasts can be understood, validated, and refined by human experts. Furthermore, given the sensitivity of business data and regulatory constraints (such as [2]), our approach is tailored to operate without centralized model sharing or global learning across time series. The following subsections describe each component of the system in detail: the dataset and its preprocessing pipeline, the business benchmark used for evaluation, the structure of the ensemble forecasting model, and the specific privacy-aware choices that guided our design. A schematic representation of these phases and their interactions is provided in Figure 4.1, which offers a high-level view of how data flows through the system from raw input to forecast output.

4.2.1 Problem Formulation

The core forecasting task addressed by FashionSight involves generating monthly sales predictions for individual product-market pairs based on univariate time series. The data consists of historical sales collected from 14 international markets, where each market is denoted as a dataset D_i, with $i = 1, \ldots, N$, and each product within that market is uniquely identified by a SKU. Each time series $y_j^i \in D_i$ corresponds to the monthly unit sales of product j in market i. To mitigate irregularities due to varying month lengths, the number of units sold is normalized by the number of retail weeks in each month. Furthermore, all time series are aligned to the **4-5-4 retail calendar**, a standard in the fashion industry that segments the year into 12 months of either 4 or 5 weeks. This calendar ensures consistent alignment of weekends and improves comparability of sales trends across years.

A univariate time series is represented as:

$$y_j^i = \{y_{j,1}^i, y_{j,2}^i, \ldots, y_{j,T+H}^i\}$$

where:

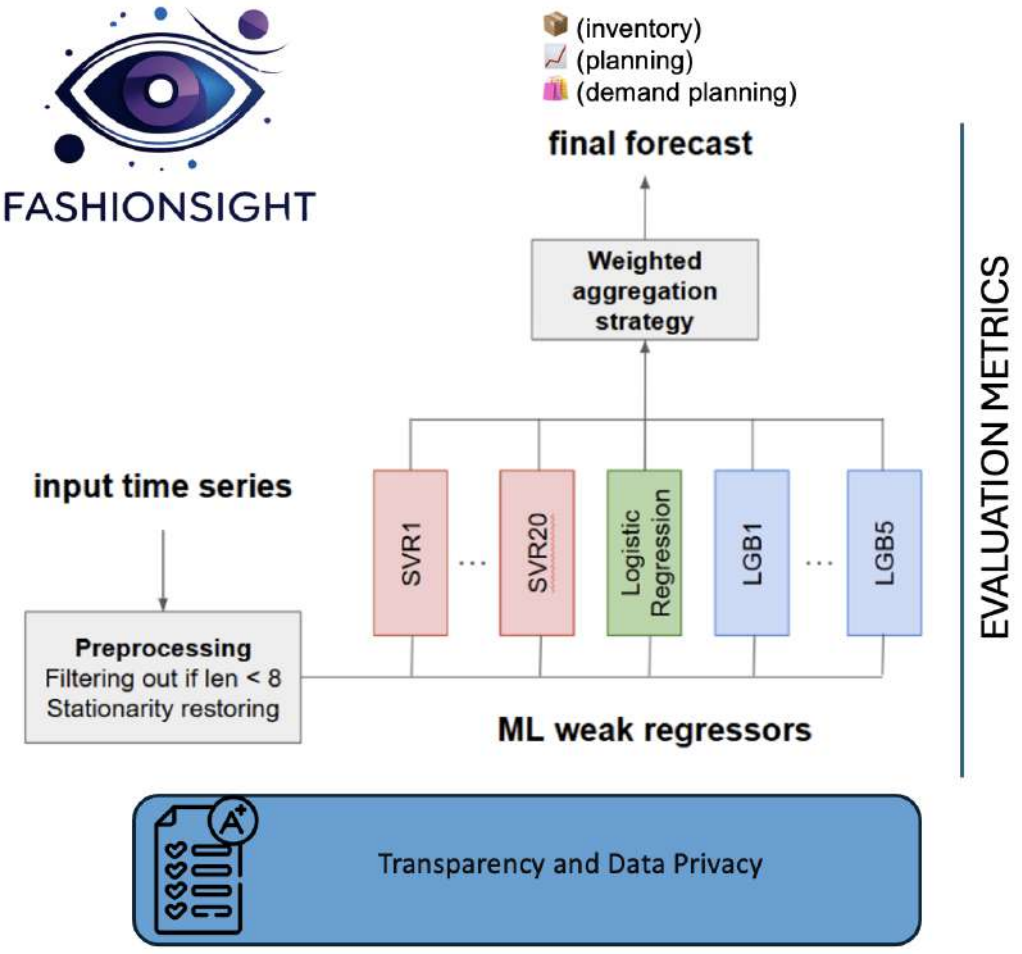


FIGURE 4.1: Monthly sales per market, normalized by the number of weeks per month, averaged over the entire dataset. Markets are grouped by volume.

- T is the number of historical data points used for training;
- H is the forecast horizon, corresponding to the number of future months to be predicted;
- $T + H$ is the total length of the time series.

The final point in the training sequence, $y^i_{j,T}$, is referred to as the *forecast origin*. The forecasting objective is to learn a model $M(\theta)$, parameterized by θ, that takes as

input the first T values and predicts the next H future sales values:

$$\hat{y}^i_{j,t} = M(y^i_{j,1}, \ldots, y^i_{j,T}) \quad \text{for } t = T+1, \ldots, T+H$$

Each model $M(\theta)$ is trained independently for each time series. This per-series modeling strategy is chosen for three main reasons:

1. It enables adaptation to the unique sales dynamics of each SKU-market pair, which can vary significantly across regions;
2. It circumvents the risk of information leakage and supports privacy constraints by avoiding pooled data;
3. It improves transparency and explainability — key features for a human-in-the-loop forecasting system.

In this study, the parameters are defined as follows:

- $N = 14$: the number of distinct international markets;
- $H = 6$: the forecast horizon in months, covering the first semester of 2023;
- $T = L^i_j - H$: the number of training observations, where L^i_j is the total length of the time series for SKU j in market i.

Each product-market time series is treated independently, even when the same product is sold in multiple countries. This allows the model to capture local trends and idiosyncrasies such as demand spikes, cultural preferences, or market-specific seasonality. The resulting problem formulation defines a flexible and scalable forecasting framework tailored to the operational complexity of luxury fashion retail. It lays the foundation for FashionSight's modular system design, which integrates predictive modeling with privacy-aware data handling and interpretable ensemble forecasting.

4.2.2 Dataset Overview and Preprocessing Pipeline

The dataset used to develop and evaluate FashionSight was sourced from the internal systems of a global luxury fashion company. It includes historical monthly sales data from 14 international markets, spanning key product categories such as apparel, footwear, and accessories. Each item is identified by a unique Stock Keeping Unit (SKU), and occurrences of the same SKU across different markets are treated as independent product-market pairs. This enables the model to capture market-specific demand behavior and avoid confounding global patterns with local dynamics. All sales data are normalized by the number of retail weeks per month, in accordance with the **4-5-4 retail calendar** an industry standard in fashion that divides the year into months of either four or five weeks. This adjustment ensures uniform alignment of weekdays across years and improves comparability of sales trends over time. Data extraction was performed via SQL queries on the company's relational database, with preprocessing steps guided by both internal business logic and modeling constraints. Only **full-price** sales (i.e., transactions with zero discount) were retained to avoid introducing noise from promotional or clearance activity. This choice limits the scope of the model to regular pricing scenarios, which are more stable and operationally relevant for strategic planning. Sales were captured across multiple retail channels. For all markets except China (MCN), both primary (brick-and-mortar) and e-commerce sales were included. In MCN, however, only primary channel sales were used, as per stakeholder policy. This filtering step ensures that the data respects market-specific strategies and avoids inconsistent channel attribution.

To reflect the heterogeneity in demand volume across regions, markets were categorized into three groups based on average monthly sales volume, as shown in Figure 4.5:

- **High-volume:** MEU (Europe), MCN (China), MUS (USA)
- **Mid-volume:** MLA (Latin America), MFJ (Japan), MFK (Korea)
- **Low-volume:** MCA (Canada), MSG (Singapore), MAU (Australia), MTH (Thailand), MMA (Malaysia), MHK (Hong Kong), MMC (Macau), MTW (Taiwan)

Time series with fewer than 8 monthly observations were excluded from the analysis. This cutoff reflects the minimum number of lags used by both machine learning and deep learning models in the system. While a lag length of 12 would better capture seasonality, enforcing that requirement would substantially reduce the dataset's coverage. As shown in Figure 4.3, most SKU-level series satisfy the 8-observation minimum, making this threshold a practical compromise.

Although contemporary forecasting models do not strictly require stationary inputs, improving stationarity can enhance model interpretability and error diagnostics. A two-step transformation pipeline was applied to promote stationarity. First, the Augmented Dickey-Fuller (ADF) test, whose null hypothesis asserts the presence of a unit root, and then then the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test, whose null hypothesis asserts stationary, were applied sequentially to each series. If either test rejected stationarity, a first-order differencing was applied. The differenced series was then retested; if still non-stationary, a second-order differencing was used. No further transformations were applied beyond this step. The procedure is visualized in Figure 4.4 (left), and its effectiveness is quantified in Figure 4.4 (right), which reports significant improvements in the percentage of stationary series across all markets.

While the 4-5-4 calendar mitigates most seasonal misalignments, real-world disruptions—such as macroeconomic shifts or tourism-driven surges—can distort expected demand patterns. For example, seasonal peaks in MEU often correlate with vacation travel. Additionally, although we model all available markets, primary analytical emphasis is placed on high-volume regions due to their strategic relevance.

We also acknowledge the issue of sales intermittency, which is further analyzed in Section 4.2.7 (e.g., Figure 4.6). However, no specialized modeling techniques (e.g., Croston variants) were employed to address this challenge. Lastly, while the COVID-19 pandemic had a substantial impact on retail sales, its effects were not modeled explicitly. Data retention policies limit historical scope to five years, and more recent data are expected to exhibit diminishing pandemic-related distortions.

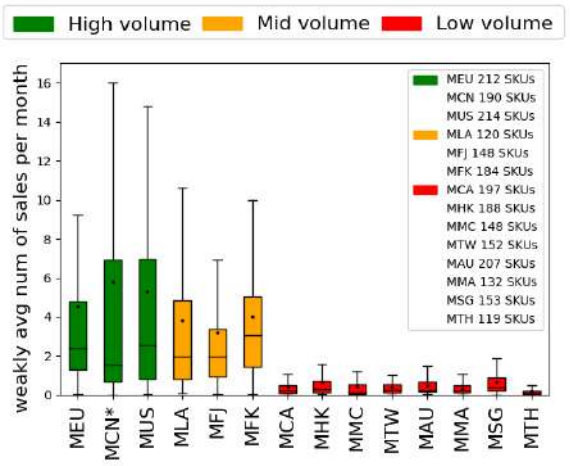


FIGURE 4.2: Monthly sales per market, normalized by the number of weeks per month, averaged over the entire dataset. Markets are grouped by volume.

4.2.3 Business Forecast Benchmark

In order to evaluate the practicality of FashionSight in the real world, we compared its forecasts with a benchmark produced by our internal business teams. This baseline forecast is used for operational planning and decision-making, so it represents a practical standard against which the system must demonstrate its value. The benchmark consists of variable-length time series from the same 14 international markets, spanning a six-month forecasting horizon, the first half of 2023. It is worth noting that the benchmark forecasts were generated using a rolling approach: after each month's actual sales were recorded, the next month's predictions were updated accordingly. Consequently, while the benchmark nominally covers six months, it is more accurately interpreted as six individual single-step forecasts rather than one multi-step prediction. Another important distinction lies in the temporal granularity. While our modelling pipeline aligns sales data to the 4-5-4 retail calendar (normalising the number of weekdays and weekends), the benchmark expresses sales in solar calendar months. To enable a fair comparison, we applied interpolation techniques to align the benchmark predictions with the retail calendar dates used in our system. Initial performance evaluations of the benchmark relied exclusively on Mean Absolute Percentage Error (MAPE). However, given MAPE's limitations in scenarios involving low-volume or intermittent sales, we aimed to investigate alternative error metrics that would provide a more robust and interpretable evaluation across diverse market contexts.

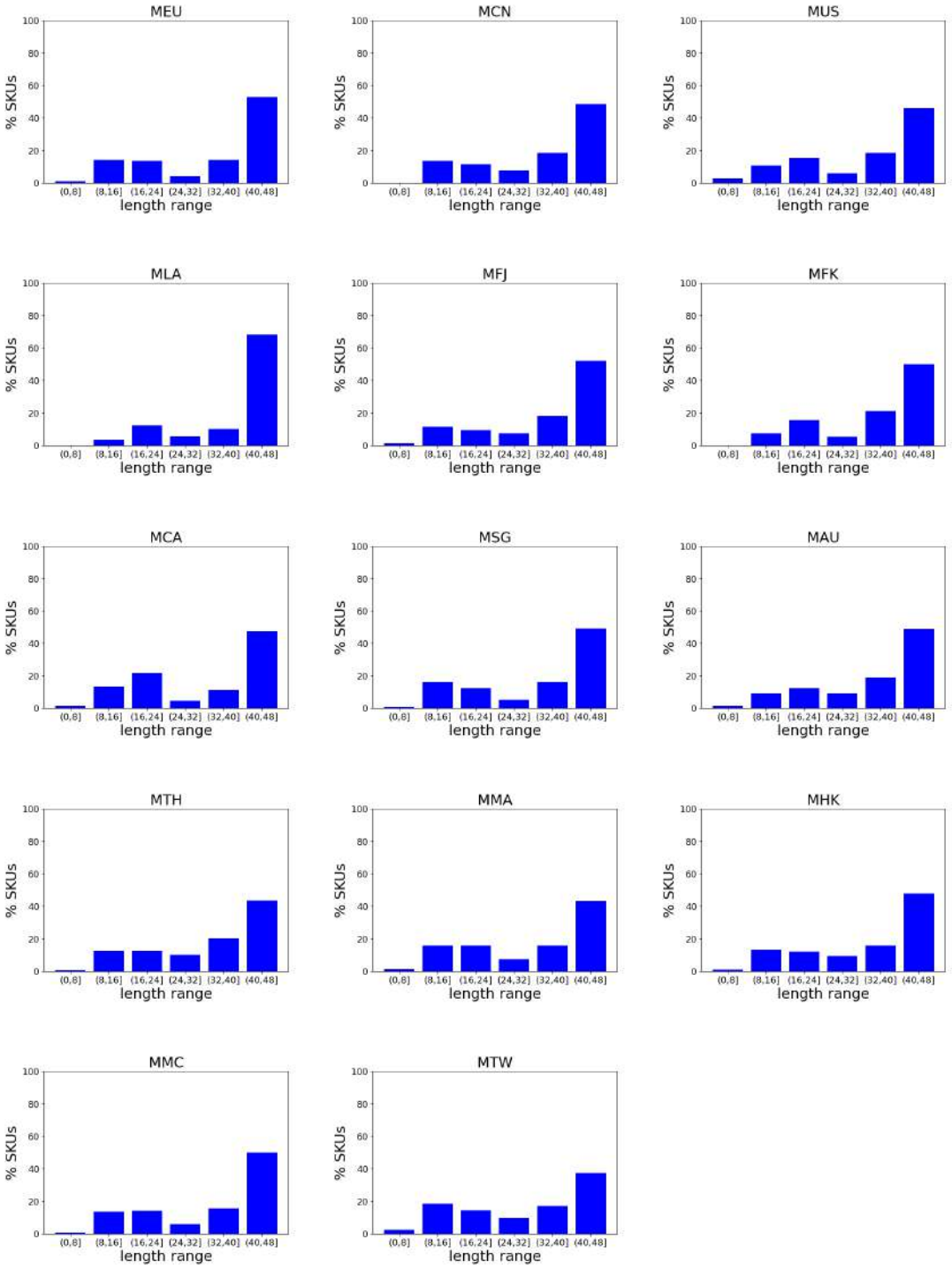
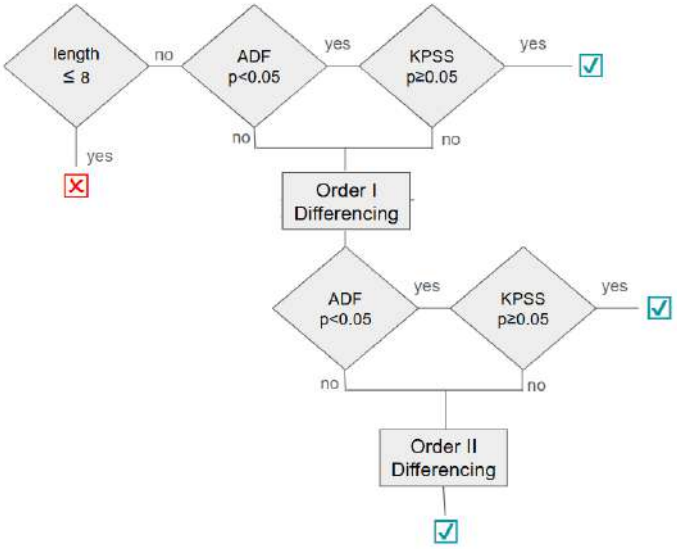


FIGURE 4.3: Distribution of SKU lengths for different markets



Market	% stationary SKUs	
	Original	Differenced
MEU	14.62	85.85
MUS	42.06	86.45
MCA	53.33	94.87
MLA	55.83	93.33
MFJ	45.27	93.92
MFK	28.8	94.02
MCN	33.16	90.0
MHK	50.0	94.62
MMC	60.81	93.92
MTW	60.93	92.72
MAU	52.66	94.69
MMA	62.12	95.45
MSG	43.42	92.76
MTH	61.54	96.58

FIGURE 4.4: Left: Stationarity testing and differencing workflow. Right: Percentage of stationary time series before and after transformation, by market.

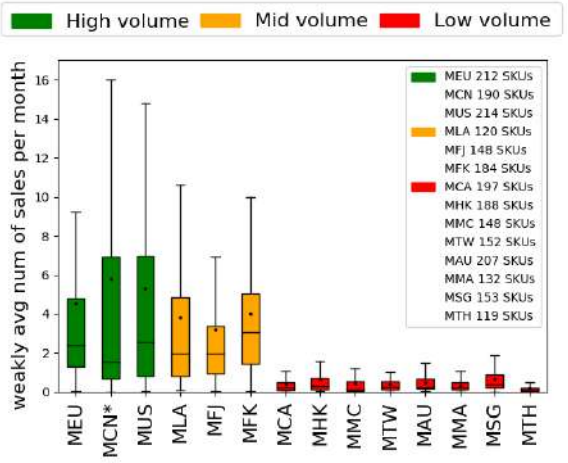


FIGURE 4.5: The number of sales for retail month, averaged by the number of weeks per month, is reported for each market in the dataset.

4.2.4 Forecasting Architecture

The FashionSight system's core engine is responsible for forecasting retail sales over the horizon defined in Section 4.2.2. This component is designed to be modular, interpretable and highly adaptable to the heterogeneity of SKU-level time series in the luxury fashion domain. Rather than relying on a single global model, FashionSight trains a distinct ensemble model for each time series. This avoids pooling and enables the forecasts to remain market- and product-specific. At the heart of this architecture is an ensemble of

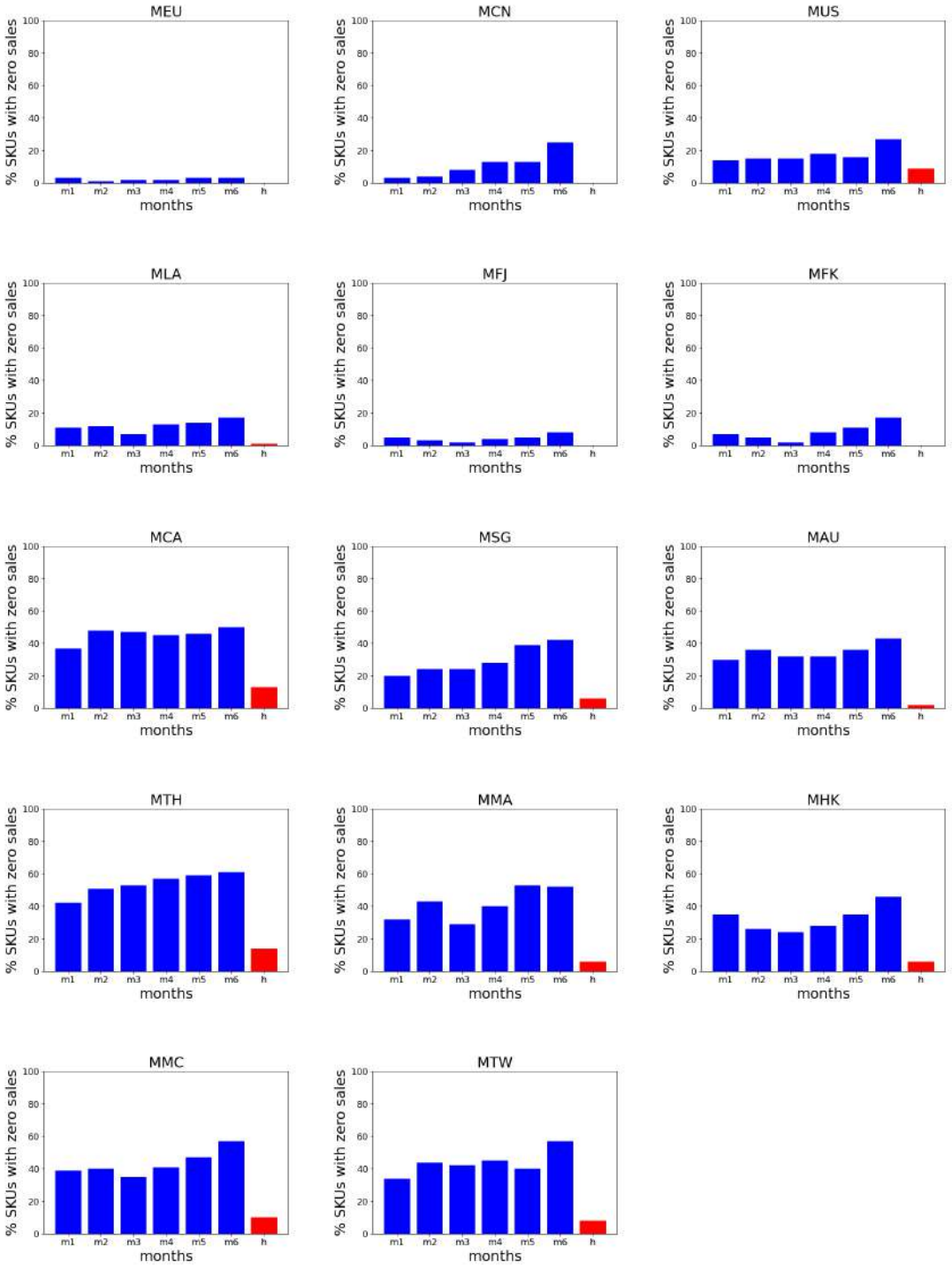


FIGURE 4.6: Distribution of SKU intermittency for different markets

$K = 26$ weak regressors. These are trained independently on the available historical data

and include a mix of algorithmic families: one linear regression model, five Light Gradient Boosting Machines (LGBM) with varied hyperparameter configurations, and twenty Support Vector Regressors (SVR) using different kernel parameters. The SVRs rely on the radial basis function (RBF) kernel and differ in the values of C and γ. Due to the limited length of many series, which precludes splitting the data into stable training and validation sets, we deliberately avoid using grid search or Bayesian hyperparameter optimisation. Instead, we exploit ensemble diversity by using predefined hyperparameter settings to capture different patterns in the data. Each weak learner k is trained independently on the training portion of the time series y_j^i from market i and product j. Let L_j^i be the total number of available observations, and H the forecasting horizon. We define the training set as the first $T = L_j^i - H$ points. The regressor k is then used to produce in-sample predictions $\hat{y}_{j,t}^{i,k}$ over this training period, and its in-sample Root Mean Square Error (RMSE) is recorded:

$$err_{j,t}^{i,k} = RMSE(y_{j,t}^i, \hat{y}_{j,t}^{i,k}) \tag{4.1}$$

To emphasize recent performance while smoothing over noise, we compute a weighted summary of the in-sample error over a fixed-length window, using exponential smoothing. We define a window of length $L_{win} = \min(12, T)$ (typically corresponding to one year), and compute smoothed errors as follows:

$$\begin{cases} s_{T-L_{win}}^k = err_{j,T-L_{win}}^{i,k} \\ s_{T-L_{win}+t}^k = \alpha \cdot err_{j,t}^{i,k} + (1-\alpha) \cdot s_{t-1}^k \qquad \alpha \in (0,1) \end{cases} . \tag{4.2}$$

The final weight $w_j^{i,k}$ assigned to each weak regressor is computed as the inverse of the smoothed in-sample error:

$$w_j^{i,k} = \frac{1}{s_T^k} \tag{4.3}$$

In this way, models with better recent in-sample performance contribute more heavily to the final prediction. The aggregated forecast for each future time step $t \in [T+1, T+H]$ is then computed as a weighted sum over the K weak predictions:

$$\hat{y}^{i}_{j,T+t} = \sum_{k=1}^{K} w^{i,k}_{j} \cdot \hat{y}^{i,k}_{j,T+t} \tag{4.4}$$

This ensemble strategy combines diversity in model behavior with a transparent and interpretable weighting scheme, allowing FashionSight to produce robust and adaptive forecasts while remaining explainable to business users.

4.2.5 Model Components

FashionSight integrates a diverse set of forecasting models, each contributing complementary strengths to address the heterogeneity of SKU-level time series in retail sales. Support Vector Regression (SVR) [64] is a kernel-based supervised learning method that projects input data into higher-dimensional spaces to enable linear regression, with the Radial Basis Function (RBF) kernel commonly used to capture non-linear relationships. Its core hyperparameters—C and γ—balance model complexity and generalization. Light Gradient Boosting Machine (LGBM) [65] is a gradient-boosted ensemble method based on decision trees, which iteratively focuses on correcting residual errors and is particularly efficient for large or sparse datasets; key hyperparameters include the number of estimators (nITE) and the learning rate (LR). LSTM networks [66, 50], a type of recurrent neural network, are equipped with memory cells and gating mechanisms that enable them to model long-term dependencies and mitigate vanishing gradient issues, making them highly suitable for time series forecasting tasks. TimeGPT[1], a transformer-based foundation model pre-trained on a large corpus of public time series data, is also included in our study to evaluate the potential of general-purpose temporal models; however, its complexity and opacity may limit its applicability in sensitive business contexts, as discussed in Section 4.2.6. Finally, the forecasting engine leverages ensemble learning principles [67], combining multiple weak

[1] `https://www.nixtla.io/`

learners in parallel. The FashionSight ensemble avoids hyperparameter tuning via resampling and instead exploits diversity across fixed model configurations, using a weighted aggregation approach based on recent in-sample performance.

4.2.6 Transparency and Data Privacy

Two key principles must be observed when designing algorithms that operate on proprietary business data: transparency and privacy. Transparency is fundamental to fostering human-in-the-loop collaboration, allowing domain experts to comprehend, interpret and potentially override model outputs based on business context. For this reason, we favour relatively low-complexity, interpretable models, avoiding overly opaque solutions. Privacy, on the other hand, is both a technical constraint and a legal obligation, especially given regulations such as [2] and [1]. These frameworks require strict control over how data is processed and shared, even within the same organisation. To align with these constraints, we adopt a fully localised modelling strategy: each time series is handled independently and no cross-series information or pooling is used during training. This design choice ensures that no sensitive data is leaked across markets, products or business units, thereby supporting both legal compliance and internal governance policies.

4.2.7 Evaluation Error Metrics

Given the varying scales and intermittent sales patterns across markets (see Figures 4.2 and 4.6), selecting appropriate error metrics is crucial. An effective metric must be scale-independent, defined even when actual sales are zero, and interpretable in a business context.

MAPE (Mean Absolute Percentage Error) is widely used for its intuitive interpretation:

$$MAPE = \frac{100}{N} \sum_{i=1}^{N} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \tag{4.5}$$

However, it is undefined for $y_i = 0$ and over-penalizes errors when y_i is small, making it unsuitable for intermittent time series.

To address this, we consider two modifications. The first is a **Modified MAPE (MMAPE)**, which smooths the denominator for zero actuals:

$$MMAPE = \begin{cases} \frac{100}{N} \sum_{i=1}^{N} \left| \frac{1+y_i-\hat{y}_i}{1+y_i} \right| & \text{if } y_i = 0 \\ MAPE(\hat{y}_i, y_i) & \text{otherwise} \end{cases}$$

While not theoretically grounded, MMAPE retains MAPE's interpretability and offers practical advantages in low-volume markets.

We also adopt the **Modified Symmetric MAPE (msMAPE)** [68]:

$$msMAPE = \frac{200}{N} \sum_{i=1}^{N} \frac{|y_i - \hat{y}_i|}{\max(|\hat{y}_i| + |y_i| + \epsilon, 0.5 + \epsilon)} \tag{4.6}$$

This formulation avoids undefined errors at zero actuals but introduces a scale-dependent threshold ϵ and lacks theoretical rigor.

To better handle intermittency, we include **MAAPE** (Mean Arctangent Absolute Percentage Error) [69]:

$$MAAPE = \frac{1}{N} \sum_{i=1}^{N} \arctan \left| \frac{y_i - \hat{y}_i}{y_i} \right| \tag{4.7}$$

It is always defined but penalizes zero-sales errors heavily, assigning them the maximum error ($\pi/2$). MAPE and MAAPE are thus used complementarily: MAPE for high-volume, non-intermittent products, MAAPE for capturing behavior under sparsity.

Finally, we report **WAPE** and **WRMSPE**, which aggregate errors over the forecast horizon and are rarely undefined:

$$WAPE = \frac{\sum_{i=T+1}^{T+H} |y_i - \hat{y}_i|}{\sum_{i=T+1}^{T+H} |y_i|}, \quad WRMSPE = \frac{\sqrt{\frac{1}{H} \sum_{i=T+1}^{T+H} (y_i - \hat{y}_i)^2}}{\frac{1}{H} \sum_{i=T+1}^{T+H} |y_i|} \tag{4.8}$$

WAPE offers a reliable alternative to MAPE in the presence of zeros, while WRMSPE further emphasizes large deviations.

Each metric is used selectively in Section 4.3 to reflect the characteristics of specific markets and ensure robust, fair model evaluation.

4.3 Results and Discussions

In this section, we report and discuss the empirical performance of FashionSight across different markets and error metrics. We first present a comparative evaluation against benchmarks and alternative models, and then perform ablation studies to understand the contribution of its components.

4.3.1 Performance assessment and comparison

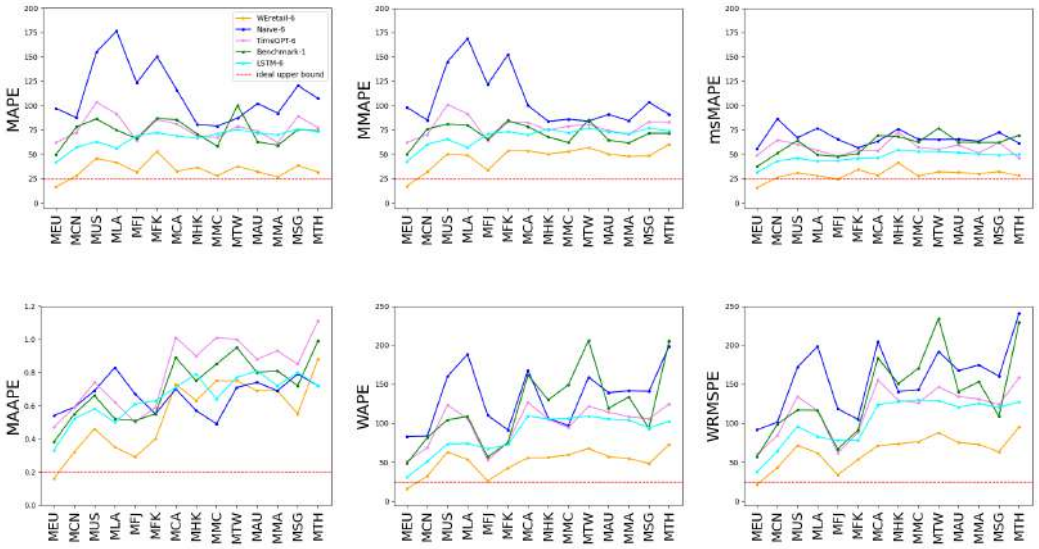


FIGURE 4.7: Results in terms of performance is reported for all markets and all error metrics investigated to allow comparison among different methods

TABLE 4.1: For the three high volume markets, the percentage of SKUs suffering from over-estimation is compared with the percentage of SKUs suffering from under-estimation, according to MAPE error metric.

Market	**%SKUs ↑**	**%SKUs ↓**	**MAPE ↑**	**MAPE ↓**
MEU	44.19	55.81	20.82	12.92
MCN	49.68	50.32	34.1	20.12
MUS	68.3	31.81	61.81	10.76

For different error metrics, Figure 4.7 reports a comparison among our method FashionSight, the benchmark, a naive forecaster which exploits the forecast origin as constant prediction throughout the entire horizon, a DL method based on LSTM, a FM

method based on TimeGpt. The method exploiting LSTM relies on 8 different configurations, whose settings have been defined on the base of our experience: in order to limit the size of the network, the encoder is given either 1 or 2 layers and both the encoder and decoder hidden size is set to 4. The number of iterations is set either to 200 or 300 for smaller learning rates. The number of lags is always 8, for all methods tested, as discussed in Section 4.2.2. The learning rate is chosen in the set [0.001,0.01,0.1]. The best performing LSTM is always chosen to make the comparison more challenging.

The mean of the metric over the entire horizon is always considered.

For what concerns the most relevant high-volume markets, MEU, MCN and MUS, it is apparent that the proposed method outperforms the benchmark as well as all proposed alternatives for any metric. MEU and MCN perform well also absolutely. The spread between MAPE and MMAPE is small because intermittence is a minor phenomenon for high-volume markets. For the same reason MAAPE is reasonably good. For MEU and MCN both WAPE and WRMSPE are good.

For what concerns mid-volume markets, MFJ performs rather well almost in line with high-volume MEU and MCN, whereas HMK lags behind and behaves more similarly to MUS, the worst performing high-volume market.

For what concerns low-volume markets, it becomes clear that both MAPE and MMAPE are inadequate metrics, even though the latter is a little more reliable. We observe that the decrease in MAPE is an artifact of the metric which completely disregards null sales, a common occurrence in highly intermittent time series. For these markets, MAAPE increases markedly due to intermittence. According to this metric, our method even struggles with the Naive/6. We can reason that the Naive-6 may take advantage of highly intermittent horizons for which a series of all zeros may be a good prediction. In addition, we know that MAAPE disproportionately penalizes those forecast which are not null when the actual sales are. WAPE and WRMSPE increase as expected with the latter being more sensitive to large errors, hence confirming their greater robustness under intermittence.

To gain insights in an aspect which is crucial for business, we investigated the degree of over-estimation compared to under-estimation. With reference to three the most relevant high-volume markets, Table 4.1 reports the percentage of SKUs either overestimated or underestimated and the corresponding MAPE. Results show that over-estimation and under-estimation occur almost equally, but with under-estimation displaying a much lower MAPE. This means that, although we have roughly the same percentage of over and under-estimation, under-estimation holds minor impact. As explained at the beginning of the chapter, this is relevant from a business point of view for which under-estimation is much more detrimental than over-estimation.

4.3.2 Ablation studies

In order to study how the inclusion of different sets of weak regressors affects the final performance of the proposed method, we evaluated its performance under the following scenarios (Table 4.2 reports results for the three high-volume markets):

- the model performing best over one market is chosen among candidate weak regressors
- a set of 16 SVR (SVR16) weak regressors (RBF kernel, $C \in [0.1, 1, 10, 100], \gamma \in [0.0001, 0.001, 0.01, 0.1]$) and our weighted aggregation strategy is applied
- a set of 20 SVR (SVR20) weak regressors (RBF kernel, $C \in [0.1, 1, 10, 100, 1000], \gamma \in [0.0001, 0.001, 0.01, 0.1]$) and our weighted aggregation strategy is applied
- SVR16 and linear regression and our weighted aggregation strategy is applied
- SVR20 and linear regression and our weighted aggregation strategy is applied
- SVR16, linear regression and a set of 5 LGB (LGB5) weak regressors ((LR=0.001, nITE=5000), (LR=0.01, nITE=3000), (LR=0.1, nITE=800), (LR=0.2, nITE=800), (LR=0.3, nITE=500)) and our weighted aggregation strategy is applied
- SVR20, linear regression and LGB5 and our weighted aggregation strategy is applied

TABLE 4.2: Results are here reported to support the ablation studies.

	Weak regressors	Aggregation	MAPE	MMAPE	MAAPE	WAPE	WRMSPE
MEU	Best single model	none	40.67	40.98	43.06	0.36	40.03
	SVM16	weighted	21.08	21.92	0.2	20.22	27.12
	SVM20	weighted	18.09	18.99	0.18	17.42	23.63
	LinearReg+SVM16	weighted	19.59	20.5	0.19	18.66	25.16
	LinearReg+SVM20	weighted	17.21	18.16	0.17	16.49	22.51
	LinearReg+LGB5+SVM16	weighted	18.68	19.61	0.18	17.63	23.97
	LR+LGB5+SVM20	**weighted**	**16.5**	**17.48**	**0.16**	**15.66**	**21.58**
	LR+LGB5+SVM20	mean	42.65	42.96	0.39	61.02	63.72
	LR+LGB5+SVM20	median	41.71	41.98	0.37	64.8	68.27
	LR+LGB5+SVM20	truncated Q1Q3	43.01	43.16	0.38	62.24	65.26
	LR+LGB5+SVM20	winsorized Q1Q3	46.12	46.13	0.41	58.41	60.9
MCN	Best single model	none	65.36	65.12	59.15	0.55	68.3
	SVM16	weighted	36.31	39.35	0.38	40.11	51.35
	SVM20	weighted	29.9	33.89	0.34	34.73	45.63
	LinearReg+SVM16	weighted	33.22	36.5	0.35	37.28	48.22
	LinearReg+SVM20	weighted	28.63	32.71	0.33	33.39	44.04
	LinearReg+LGB5+SVM16	weighted	32.44	35.77	0.35	36.34	47.04
	LR+LGB5+SVM20	**weighted**	**27.92**	**32.05**	**0.32**	**32.54**	**42.95**
	LR+LGB5+SVM20	mean	60.95	60.74	0.52	74.54	80.61
	LR+LGB5+SVM20	median	72.98	71.58	0.57	78.26	85.92
	LR+LGB5+SVM20	truncated Q1Q3	65.4	64.63	0.54	78.19	84.73
	LR+LGB5+SVM20	winsorized Q1Q3	62.08	61.56	0.53	76.87	82.84
MUS	Best single model	none	76.17	74.24	53.48	0.61	91.81
	SVM16	weighted	53.05	56.22	0.5	69.73	78.4
	SVM20	weighted	48.25	52.3	0.47	65.56	74.22
	LinearReg+SVM16	weighted	51.69	55.18	0.49	68.23	76.64
	LinearReg+SVM20	weighted	47.45	51.72	0.47	64.66	73.21
	LinearReg+LGB5+SVM16	weighted	50.02	53.77	0.48	66.39	74.63
	LR+LGB5+SVM20	**weighted**	**45.88**	**50.4**	**0.46**	**62.93**	**71.31**
	LR+LGB5+SVM20	mean	65.33	65.07	0.58	101.72	110.37
	LR+LGB5+SVM20	median	78.06	75.7	0.62	108.96	119.01
	LR+LGB5+SVM20	truncated Q1Q3	73.58	71.92	0.6	108.92	118.5
	LR+LGB5+SVM20	winsorized Q1Q3	67.82	67.06	0.59	105.49	113.85

- SVR20, linear regression and LGB5 and the mean of all predictions is used as aggregation strategy

- SVR20, linear regression and LGB5 and the median of all predictions is used as aggregation strategy
- SVR20, linear regression and LGB5 and the mean over predictions within the truncated inter-quartile range between Q1 and Q3 is used as aggregation strategy
- SVR20, linear regression and LGB5 and the mean over predictions within the winsorized inter-quartile range between Q1 and Q3 (the minimum and the maximum of the replace those values which are below or above the range respectively, without truncation of data) is used as aggregation strategy

The proposed model FashionSight (LR+LGB+SVM20) performs always better, proving that all week regressors in the set play a role in enhancing performance. In particular, we notice that a single model is unable to provide good performance whereas all ensemble approaches relying on our weighted aggregation strategy obtain good results according to all metrics, with a performance which increases as the set of week regressors is enriched. We also notice how the weighted aggregation definitely outperforms alternative aggregations such as mean, median, truncation and winsorizing.

4.4 Conclusions and Future Works

This paper presents FashionSight, a modular and interpretable forecasting system tailored to luxury retail sales at the SKU-market level. FashionSight demonstrates that lightweight, modular and transparent solutions can rival—and even surpass—state-of-the-art methods in real-world retail forecasting, providing a valuable tool for data-driven decision-making in fashion and beyond. The proposed solution uses a per-series ensemble learning strategy that combines a variety of weak regressors, including SVRs, LGBMs and linear models, into a transparent aggregation mechanism guided by in-sample performance. This design enables FashionSight to adapt to the heterogeneous and often intermittent nature of sales data in global fashion markets. Experimental results from 14 international markets show that FashionSight outperforms the business benchmark and other modern forecasting approaches, such as LSTM-based deep learning and TimeGPT, by a significant margin. Notably, the system achieves robust performance even under severe data constraints, such as

short time series, high intermittency and limited visibility into promotional effects. Ablation studies confirm the importance of ensemble diversity and error-weighted aggregation. In addition to accuracy, the system was designed with a focus on interpretability and data privacy. By avoiding global pooling and model sharing across series, FashionSight remains fully compliant with privacy regulations such as [2], offering traceable and justifiable outputs for business users. Future work will extend this research in several directions. Firstly, we intend to explore adaptive weighting schemes that evolve over time in order to capture temporal changes in performance more effectively. Secondly, integrating external features, such as macroeconomic indicators, holiday calendars or marketing events, could enrich the input space to improve context awareness, particularly in high-volume regions. Thirdly, we plan to investigate hybrid approaches combining local modelling with light global priors to retain privacy while leveraging structural regularities across markets. Finally, more advanced intermittency handling, including probabilistic models or zero-inflated regressors, could further improve performance in low-volume markets.

4.5 Acknowledgement

This work is funded by Salvatore Ferregamo an Italian luxury fashion company `www.ferragamo.com/`.

Chapter 5

Operationalizing Compliance for Moderate-Risk AI Systems

The AI system under examination is designed to support public administration in addressing the following institutional objectives:

T1 estimation of compliant tax revenue: providing assistance to the competent municipal authority in determining the amount of tax liabilities anticipated to be duly paid, for the purpose of projecting the municipality's annual fiscal receipts;

T2 forecasting recoverable tax arrears: assisting the municipal administration in identifying taxpayers predicted, on the basis of risk analysis, to be at high probability of default, and in pursuing recovery through the issuance of formal payment notices;

5.1 Determination of the Risk Class

In this section, the risk classification of the proposed application is examined with reference to the provisions in [1]. An initial and indispensable step in establishing its legal admissibility consists in verifying that the system does not fall within any category deemed to pose an unacceptable risk. To this end, Art. 5 of [1] is considered, as it delineates the

specific categories of AI systems whose use is explicitly prohibited within the European Union.

Among the prohibitions enumerated in this provision, only Art. 5(1)(c) appears potentially relevant to the system under analysis. This clause proscribes the deployment of AI systems designed to evaluate or classify individuals (so called "social scoring"), in circumstances where such evaluations give rise to one or both of the following consequences:

i) detrimental or unfavourable treatment of certain natural persons or groups of persons in social contexts that are unrelated to the contexts in which the data was originally generated or collected

ii) detrimental or unfavourable treatment of certain natural persons or groups of persons that is unjustified or disproportionate to their social behaviour or its gravity

Task **T1** does not involve the assessment of individual persons; rather, it concerns the aggregate total of amounts classified by the system as "NOT PAID." Task **T2**, by contrast, introduces a slightly more individualised dimension, as it supports the revenue officer in assessing whether the issuance of a payment reminder would be appropriate. As discussed in Chapter 6, the system generates its output in the form of a confidence interval. Consequently, when a subject remains in default and the interval is particularly narrow, the system indicates that sending a reminder would likely be ineffective. Nevertheless, the ultimate decision as to whether to issue the reminder remains entirely within the discretion of the revenue officer.

With reference to criterion (i), the operational context exactly corresponds to that for which the AI application was originally conceived, and it is not applied in any domain other than its intended one. Regarding criterion (ii), the potential consequence of omitting to issue a reminder cannot be considered detrimental, unjustified, or disproportionate. It is not detrimental, since the payer has already been duly notified of their default, with the omission of an additional notice resulting at most in minimal inconvenience or lack of further support for the taxpayer; nor is it disproportionate, as the revenue officer may provide a rationale for the decision not to have sent the reminder.

Having established that the application does not fall within the category of AI systems presenting an unacceptable risk, the analysis now turns to its potential classification as high-risk under Art. 6 of the [1].

The assessment of whether the application constitutes a high-risk AI system is governed by Art. 6 of the [1]. Paragraph 1 designates an AI system as high-risk if two cumulative conditions are met: (a) the system falls within the categories listed in Annex I, and (b) a conformity assessment is required as specified therein. Annex I primarily addresses AI systems that constitute safety components or otherwise pose significant risks, which is not the case for the application under consideration.

Paragraph 2 refers to Annex III, subject to certain exceptions. Among the eight categories listed in Annex III, the one potentially relevant to the application is category 5, which concerns the allocation of services. In this context, it is necessary to assess whether the omission of a payment reminder could be considered a disadvantageous condition for the taxpayer, as it may potentially deprive them of a reminder that others would receive. However, since the taxpayer has already received the necessary reminders, the practice of issuing additional reminders only to selected taxpayers constitutes an administrative efficiency measure, and does not amount to the denial of a service or right. Based on this reasoning, Tasks **T1** and **T2** cannot be classified as high-risk, and, at most, may be assigned a precautionary designation of moderate risk.

5.2 Addressing Data Protection Requirements

Unlike the case studies introduced in Chapter 2, since this application deals with personal data, we evaluate it in the context of [2]. This analysis is conducted on a voluntary basis, as the risks posed by the system are minimal: indeed, unlike [1], the [2] delegates the assessment of risk to the responsibility and discretion of the data controller (supported by the data processor, as required under Art. 28(3)(f) of [2]), without prescribing a formal risk classification to be adhered to. In the following, we provide an evaluation of the AI system with regard to the articles deemed most relevant.

1. Art.14 [2]: this article is applicable because the personal data are already held by the municipality for the purpose of conducting standard local tax collection procedures and have not been obtained through direct requests to the data subject. The processing in the context of tasks **T1** and **T2** of the application under consideration is lawfully grounded in Art. 97 of the Italian Constitution, which affirms the duty for administrative procedures to be efficient within the limits of the current state of technological development (see Chapter 1 for a detailed discussion). The processing does not involve profiling, as explained in Section 5.1, and the human oversight required by Art. 22(1) [2] is fully ensured, as detailed in Section 5.5. The data subject will be informed of the processing via a note added to the payment notice, indicating that the notice has been issued by a revenue officer, potentially assisted by an automated AI process;

2. Art.25(1) [2] (privacy by design): to ensure pseudonymization as defined in Art. 4(5) [2], the system removes directly identifying data and incorporates two security layers. A Federated Learning (FL) framework which enables training across multiple datasets without sharing local data and a Differential Privacy (DP) layer which prevents inferring whether a specific individual belongs to a dataset from the final model, thereby providing robust privacy guarantees. Data minimization is achieved by using a minimal feature set, explicitly excluding any information that could potentially identify individuals, even indirectly.

3. Art.25(2) [2] (privacy by default): with respect to our AI system, personal data selected exclusively for the explicitly declared intended purpose are retained for a maximum period of four years to allow the construction of training and testing datasets, after which only the outputs of the trained model are preserved. The data are processed solely for the purpose of training and testing the model, and access is restricted to the data controller or processor.

Further insights into some of these aspects are provided in Chapter 10 and Appendix A, where a DPIA document is prepared and presented for the AI system under consideration.

5.3 Addressing Data Governance Requirements

Although, as argued in Section 5.1, the system does not fall into the category of high-risk AI systems, in alignment with the good practices of trustworthy AI [3], and as an academic exercise, this section examines the extent to which the application complies with Art. 10 in [1], concerning data governance. Each paragraph of Art. 5 is examined systematically, with those deemed inapplicable omitted from the analysis:

In compliance with Art.10(2)(e), an assessment is carried out regarding the availability, quantity and suitability of the data sets in use.

In compliance with Art.10(2)(h), dataset D5, the only one with insufficient data and an unacceptable minority-to-majority class imbalance, is included in the analysis but flagged as an outlier.

In compliance with Art.10(2)(f), biases related to age, gender, or ethnicity are not present because these variables are not included among the features. The only potential source of bias could be the region of residence: although this information is not explicitly included as a feature, the system is trained on region-specific datasets. However, global high performance during model testing across regions indicates that the system is not biased by this information.

In compliance with Art.10(3), as three additional datasets (D2, D3 and D4), alongside D5, present class imbalance, during preprocessing, downsampling is applied, and class imbalance is addressed during training by (1) weighting the loss function and (2) using balanced batches.

In compliance with Art.10(4), no differences in model performance were observed across regions due to geographical characteristics. An exception is dataset D5, where limited data availability accounts for the observed deviation. As already explained it is marked as outlier.

Further insights into some of these aspects are provided in Chapter 10 and Appendix B, where a FRIA document is prepared for the AI system under consideration.

5.4 Addressing Transparency Requirements

To investigate the model's global decision-making process, we employ SHAP [18], which allows the identification of features with the greatest relevance to the output, thereby providing an average explanation of how the model arrives at its predictions. Fig. 5.1 presents the SHAP values associated with the input features. After normalization, these values represent the relative contribution of each feature to the average deviation of individual predictions from the mean prediction across all samples in the test set. In other words, SHAP values quantify the importance of each feature in driving the model's decisions.

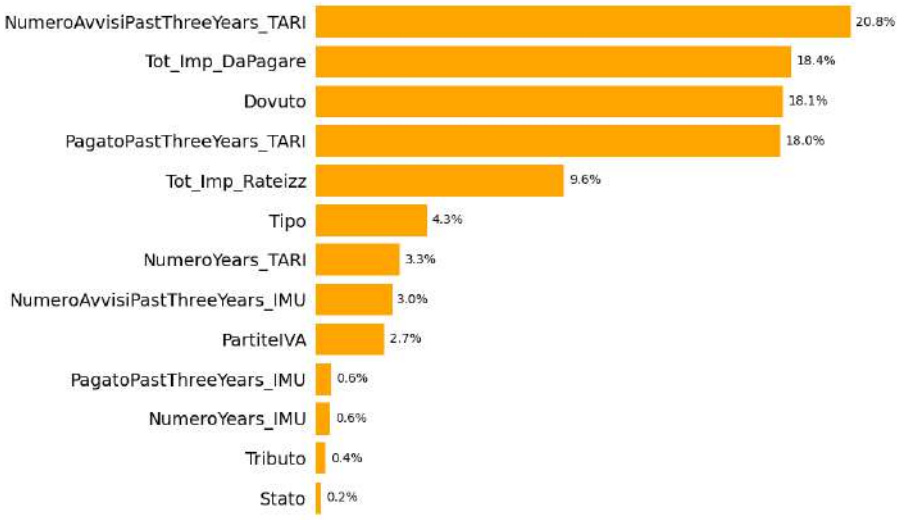


FIGURE 5.1: SHAP values illustrating the importance of features across different targets and models

This form of explanation enhances Explainability by revealing which factors most strongly influenced the predictions and enabling an assessment of whether these contributions are reasonable. If the model relies counterintuitively on a trivial feature, it warrants careful reconsideration. Similarly, if the model is found to rely predominantly on a sensitive or bias-inducing feature, its trustworthiness and fairness may reasonably be called into question.

5.5 Addressing Human Oversight Requirements

In the context of moderate-risk AI systems, no specific provision of [1] applies directly; however, it is evident that, in several instances, [1], either explicitly or implicitly, defers to the [2], thereby establishing a form of regulatory complementarity. That said, Art. 22(1) of [2], as anticipated by Recital (71), establishes that: "automated individual decision-making, including profiling, should not be based solely on automated processing". Similarly, within the pre-existing soft-law framework, *Respect for Human Autonomy* is identified as one of the four ethical principles underpinning trustworthy AI [3]. This principle is subsequently translated into one of the seven key requirements, namely *Human Agency and Oversight*, which operationalizes the ethical commitment to ensuring meaningful human control over AI systems [3]. Below, we present the questionnaire proposed in [16] as a tool for self-assessing the necessary oversight measures in AI applications, along with the corresponding responses for the moderate-risk use case described in Chapter 6.

Q1 Please determine whether the AI system (choose as many as appropriate):

- Is a self-learning or autonomous system;
- Is overseen by a Human-in-the-Loop (HITL);
- Is overseen by a Human-on-the-Loop (HOTL);
- Is overseen by a Human-in-Command (HIC);

> *The AI system incorporates a Human-in-the-Loop (HITL) approach, as human personnel are actively involved at each stage of the application lifecycle: the revenue officer may rely on the AI system solely to receive recommendations regarding the likelihood that a tax payment notice will either be disregarded by the taxpayer or duly settled.*

Q2 Have the humans (human-in-the-loop, human-on-the-loop, human-in-command) been given specific training on how to exercise oversight?

> *The revenue officer may consider the recommendation generated by the AI system as a basis for conducting further inquiry into a specific case or for*

issuing an additional notification aimed at drawing the taxpayer's attention to the pending tax liability.

Q3 Did you establish any detection and response mechanisms for undesirable adverse effects of the AI system for the end-user or subject?

No specific adverse effect is anticipated. In the worst-case scenario, the taxpayer may simply not receive any additional reminders.

Q4 Did you ensure a 'stop button' or procedure to safely abort an operation when needed?

A dedicated 'stop button' is unnecessary for this AI systems.

Q5 Did you take any specific oversight and control measures to reflect the self-learning or autonomous nature of the AI system?

Both systems support horizontal scalability through the FL framework: as new datasets become available, they can be incorporated into the training process, which may be periodically repeated. The performance metrics established during system development can be continuously updated as the systems are applied to real-world cases, enabling ongoing monitoring of their generalization capabilities.

5.6 Addressing Accuracy and Robustness Requirements

Also for the moderate-risk case study, thorough analysis of the strategies implemented to optimize its performance is provided a dedicated chapter (Chapter 6). Initially, a model is selected that achieves robust performance with respect to the chosen accuracy metric. Subsequently, a dedicated mechanism is introduced to promote reliability by confining trust to outcomes with a low variance on repeated non deterministic evaluations, thus mitigating the inherent overconfidence of the model.

Chapter 6

Moderate-Risk AI Case Study: FAITH-FDSS

The aim of this study is to provide decision support to revenue officers during the issuance of payment notifications. The application, operating under human supervision with the final decision remaining with the officer, is designed to deliver insights into the payer's likelihood of settling outstanding obligations. Consequently, it informs the appropriateness and potential utility of issuing a reminder, with the goal of enhancing administrative efficiency by both avoiding unnecessary notifications and promoting timely payment.

In particular, our novel AI-based Decision Support System (DSS), FAITH-FDSS, assists municipal revenue officiers in the decision of whether to issue a payment notification, thereby fulfilling the efficiency requirements mandated for public administration (see for example Art. 97 of the Italian Constitution). It also provides projections of the amount of taxes likely to be collected, enabling municipalities to optimally plan the provision of public services. Its life-cycle is informed by the European regulatory framework since its earliest stages, ensuring lawfulness, robustness, and trustworthiness in compliance with [1], [2], and [3].

FAITH-FDSS is trained on historical data regarding the issuance of payment notifications and notices, to perform a binary classification task with outcomes defined by

whether the payer subsequently settled the outstanding liabilities. To operationalize compliance by design, it is engineered to reinforce computational privacy through the integration of FL and output privacy through the integration of (ε, δ)-DP, to enhance transparency via feature importance analysis, and to support trustworthiness through uncertainty estimation using MC dropout.

In addition, the DSS is intended solely as a decision-support tool for the revenue officer, without prejudice to their discretionary authority, in accordance with the principles of human oversight and the human-in-the-loop paradigm.

The contributions of this paper can be summarized as follows:

1. design and implementation of a DSS that assists revenue officers by predicting, with high reliability, whether a payment notice is likely to be paid or remain outstanding;
2. integration of the model within a FL framework, enabling comparison between centralized, and distributed approaches to enhance inter-dataset privacy preservation;
3. reinforcement of privacy protections through the implementation of (ε, δ)-DP, including an assessment of the trade-off between gradient clipping and Gaussian noise injection against model performance and utility, thereby promoting data subject privacy both within and across datasets in compliance with [2] and [1] requirements;
4. incorporation of components to enhance the transparency and trustworthiness;
5. An interdisciplinary, by-design approach that simultaneously advances both the technological implementation and the regulatory and ethical dimensions since the earliest stages of the AI system life-cycle.

While both technical [70], [71], [72], [73], [74] and legal–ethical [75], [76], [77] scholarship have extensively examined issues surrounding AI governance, few studies, to our knowledge, have effectively bridged these two perspectives. In particular, there remains a lack of research demonstrating how the normative framework can actively inform and guide technical design decisions from the very outset of an AI system's life-cycle.

6.1 Related work

The strategic relevance of tax administration, and of tax revenue as a specific use case, has been explored from an algorithmic and technical standpoint and, with growing attention, from legal and ethical perspectives. Contributing to the former, several studies have proposed data-driven and Machine Learning (ML) approaches aimed at improving fiscal efficiency, detecting tax evasion, and optimizing collection processes. [70] have investigated the most effective techniques for forecasting sales tax revenue, highlighting the pivotal role of data pre-processing, which in certain contexts may outweigh the impact of the forecasting model itself; though valuable, this research addresses only a specific aspect of AI systems for tax revenue administration. In our case, data pre-processing is a relatively straightforward task, given the limited scope and content of the datasets. [71] addresses the prediction of national tax revenue ratios using an optimized Kernel Extreme Learning Machine (KELM) model which incorporates a vector-weighted averaging algorithm. In our case, the primary focus of the research is not the development of a highly optimized model. The use of a MLP provides strong predictive performance while enabling straightforward and natural integration within the FL framework, as well as seamless incorporation of uncertainty assessment by means of MC dropout. [72] capitalizes on the emergence of big data to perform multi-source data fusion across business, corporate, and personal income taxes, aiming to enhance tax forecasting at an aggregate level. In our research, data protection is prioritized, and the fusion of miscellaneous large-scale datasets is not considered a viable option. [73] examines the effectiveness of "deep" machine learning techniques (neural networks, random forests, and decision trees) in forecasting quarterly corporate tax payments, contributing to improved corporate tax revenue prediction. Since the integration of XGBoost or similar algorithms within a FL framework, as well as the application of DP to them, requires dedicated approaches, we did not consider them a convenient option for this study.

On the regulatory and ethical side, European institutions and academic commentators have underscored the need for trustworthy and transparent AI systems in the public sector, particularly within fiscal administration. These contributions emphasize compliance with [2] and [1] as fundamental to ensuring the lawful, fair, and accountable use of

AI technologies in tax governance: [75] explores how AI influences taxpayer compliance decisions, highlighting the ethical challenges inherent in deploying AI within the highly sensitive context of tax administration; [76] examines the effects of digital transformation in tax administration on tax-payers' trust in the fiscal system, highlighting its positive influence on voluntary compliance; [77] examines the application of NLP and machine learning in tax administration, emphasizing that transparency, ethical adherence, and fairness in AI systems not only ensure regulatory compliance and protect institutional reputation but also enhance system trustworthiness, thereby fostering greater taxpayer confidence and improving compliance.

In this context, only a limited number of contributions have attempted to bridge these two dimensions by introducing privacy-preserving techniques, such as FL and DP, to enable collaborative model training across institutions without compromising taxpayer confidentiality, thereby addressing both technical performance and regulatory compliance. For example, [74] introduces a three-layered security framework combining FL, DP, and Secure Multi-party Computation (SMC). The primary advantage of SMC lies in its ability to enhance computational privacy, ensuring data protection during computation, without any degradation in model performance. This property is particularly valuable when the central server cannot be fully trusted or when it is desirable to avoid the performance penalties associated with noise injection. Although this approach is thorough and conceptually robust, in the present work DP is strictly required to ensure output privacy guarantees. In addition, experimental results indicate that a modest level of noise is sufficient to meet the intended privacy objectives. Meanwhile, computational protection is already reasonably ensured through the combined use of FL and DP, as the server only receives noisy model updates. Furthermore, in contrast to [74], our approach adopts a more comprehensive perspective by incorporating additional compliance dimensions such as transparency, robustness, and trustworthiness, by design.

Despite the growing literature on AI applications in tax administration and the concurrent legal and regulatory scholarship emphasizing integration, no prior work, to our knowledge, has proposed a unified expert system framework that simultaneously addresses the core requirements of trustworthy AI [3]. Specifically, our approach accomplishes these requirements within a real-world tax revenue use case, prioritizing them through a

proportionality-based lens: Transparency (via explainability), Privacy and Data Governance (via DP and FL), Technical Robustness and Safety (through resilient, regularized modeling), and Human Agency and Oversight (ensuring human-in-the-loop decision authority). Ultimately, this paper fills that gap by offering a deployable expert system architecture that balances predictive performance with interpretability, privacy, and regulatory compliance.

6.2 Materials

To ensure pseudonymization, as defined in Article 4(5) and mandated under Article 27(1) of [2], the pre-processing phase removes directly identifying data and incorporates two security layers. Complementarily, data minimisation, as defined in Article 5(c) and mandated under Art. 27(2) of [2], is achieved by using a minimal feature set, explicitly excluding any information that could potentially identify individuals, even indirectly. In addition, data imbalances that could potentially give rise to biases, contrary to the principles of sound data governance outlined in Art. 10 of [1], have been identified and mitigated through down-sampling.

6.2.1 Datasets

The data used in this study comprise nine datasets (D1–D9), each corresponding to one of nine Italian regions (Campania, Marche, Toscana, Lazio, Basilicata, Abruzzo, Calabria, Lombardia, Puglia), as shown in Table 6.1. The data collection period spans from 2019 to 2024; however, the most recent year (2024) is excluded due to incompleteness at the time the data are provided.

Each sample in a dataset represents a payment notice issued to a defaulting taxpayer on a specific date and is characterized by 8 continuous, 3 integer and 3 categorical fields described in Section 6.2.2 and one binary field, the target, which indicates whether the due amount has been paid in response to the notice. Due to the limited, sparse, and irregular nature of the available samples, constructing time series is not feasible. Nevertheless, aggregated features are incorporated to capture information from the preceding three

TABLE 6.1: Samples per dataset, split into train and test set with corresponding class imbalance before and after downsampling

	train set				test set	
	original		down-sampled			
	paid	not paid	paid	not paid	paid	not paid
D1	53876	83395	-	-	34819	46323
D2	17111	55219	17111	34222	15132	55384
D3	3902	16305	3902	7804	1261	7426
D4	2419	8775	2419	4838	2073	7599
D5	784	169	-	-	776	442
D6	15262	28852	-	-	16220	30010
D7	8586	9202	-	-	5643	14194
D8	4412	4552	-	-	5725	6464
D9	22933	60847	-	-	26434	64941

years as explained in Section 6.2.2. The 2022 datasets are used for training, enriched with aggregated features derived from 2019, 2020, and 2021, while the 2023 datasets serve as the test set.

The first four columns of Table 6.1 report, for each training set, the initial number of samples per class and the corresponding imbalance ratio. To mitigate class imbalance, which may otherwise bias the model, the majority class is downsampled whenever the imbalance ratio exceeds 2:1. This threshold represents a suitable trade-off between restoring an admissible balance and preserving dataset size.

6.2.2 Features

Our feature set is composed of 9 continuous features, 3 integer features, and 3 categorical features, as shown in Table 6.2. Categorical features are encoded using embeddings of dimensionality 2, 2, and 4, respectively. Embeddings are adopted instead of one-hot encoding to maintain a low-dimensional representation and avoid the creation of additional sparse vectors. The encoding is applied to the input categorical variables through an embedding layer placed immediately after the input layer in the MLP, as shown in Fig. 6.1 a) shows. Although capturing latent relationships among categorical values is not the primary motivation for their use, this property may nonetheless offer an additional advantage. 4 out 9 continuous and 2 out of 3 integer features are aggregate features added

to capture the temporal dimension: aggregate features represent averages, in particular, they capture the average number of notices received by a taxpayer over the past three years, as well as the total amount paid to settle these notices during the same period. Features potentially leading to bias or discriminatory outcomes (e.g., age) are excluded from the dataset.

TABLE 6.2: The role of a feature can be either feature (F) or aggregated feature (AF) over the past three years; its type can be categorical encoded through embeddings of size n (cat$-n$), integer (int) or float (float). The target is denoted as T

name	**role**	**type**	**description**
Municipal Tax	F	cat-2	type of tax
Type	F	cat-2	type of notice
State	F	cat-4	issuance status
VAT Number	F	int	number of VAT numbers registered to the tax payer
Num Years IMU	AF	int	num of years with at least one notice for IMU
Num Years TARI	AF	int	num of years with at least one notice for TARI
Total Due	F	float	total taxable amount due
Total Installment-Based	F	float	total taxable amount under installment plan
Due	F	float	outstanding amount
Num Notices Past Three Years IMU	AF	float	num of notices received during the past three years for IMU
Num Notices Past Three Years TARI	AF	float	num of notices received during the past three years for TARI
Paid Past Three Years IMU	AF	float	overall amount paid during the last three years for IMU
Paid Past Three Years TARI	AF	float	overall amount paid during the last three years for TARI
Payment Status	T	float	whether the notice has been duly settled

6.3 Methods

In this section, we provide a detailed description of the technical components underlying FAITH-FDSS. The system architecture is designed to ensure both predictive accuracy and compliance with European regulatory requirements on data protection and AI ethics, with each component serving this objective. In Section 6.3.1 we present the MLP architecture, which forms the computational basis of FAITH-FDSS. In Section 6.3.2, we describe the federated learning framework adopted to enable collaborative model training across multiple municipal datasets, while preserving local data sovereignty. In Section 6.3.3, we briefly present the centralized training approach, which serves as a baseline for evaluating the federated learning methodology of FAITH-FDSS and ensuring that the distributed approach does not compromise overall model performance. In Section 6.3.4, we discuss the integration of Differential Privacy (DP) mechanisms, which offer formal privacy guarantees by limiting the risk of re-identification, thereby promoting anonymization and mitigating the exposure of sensitive financial and personal data. In Section 6.3.5, we briefly introduce the use of SHapley Additive exPlanations (SHAP) to determine feature importance. This approach provides revenue officers using FAITH-FDSS with a means to understand whether the model's suggested decisions are grounded in a logical rationale and whether such rationale is acceptable and aligned with domain knowledge, thereby enhancing the system's transparency and explainability. In Section 6.3.6, we introduce MC dropout, a strategy that applies the dropout mechanism, typically used during training, also at inference time, in order to generate a distribution of predictions for each individual sample rather than a single deterministic output. By evaluating the width of the interval defined by the minimum and maximum predicted probabilities for a given sample, it becomes possible to assess the reliability of the corresponding prediction, thereby contributing to the overall trustworthiness of the model. Together, these elements constitute a robust, privacy-preserving, and trustworthy computational foundation for FAITH-FDSS.

6.3.1 FAITH-FDSS architecture

Figure 6.1 a) illustrates the architecture of FAITH-FDSS. Categorical input features are first transformed via an embedding layer, which encodes each feature into a fixed-size vector

representation. These embeddings are then concatenated with non-categorical features before being fed into the first block of the model. The core of the model is implemented as an encoder in the form of a MLP consisting of five sequential blocks. Each block comprises a fully connected layer, followed by a dropout layer for regularization and a non-linear ReLU activation function. The output of the fifth block is passed to a classification layer with two nodes, corresponding to the number of target classes, which produces the model's final prediction.

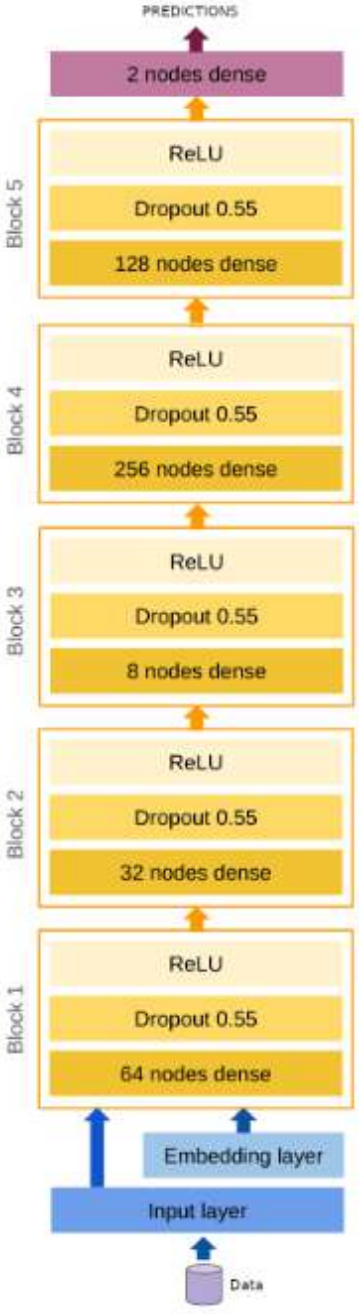
6.3.2 Distributed training

As shown in Fig. 6.1 b), the federated approach involves training models locally on datasets distributed across multiple clients, followed by aggregation of these local models into a single global model via a defined aggregation strategy. The procedure is as follows: **(1)** the server sends the current global model to all clients, **(2)** each client performs local training on its own data, **(3)** at the end of local training, each client sends its updated model to the server, and **(4)** the server aggregates the updates received from all clients into a new global model, using a specific aggregation strategy. The procedure defined by Steps **(1)**–**(4)** constitutes one federated round. Once Step **(4)** is completed, the process loops back to Step **(1)** to execute the next round.

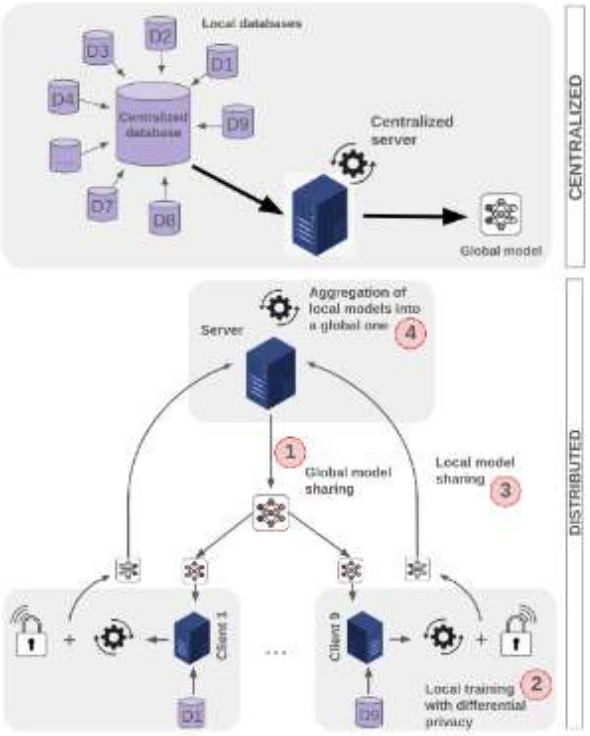
The aggregation strategy defines the behavior of the federated approach, and several strategies exist. In vanilla federated learning [78], at round t, each client k ($k = 1..K$) computes an updated model ($\mathbf{w}_k^{t+1}$) based on its local data. The server then combines these updates into a new global model ($\mathbf{w}^{t+1}$) according to the following weighted sum (FedAvg):

$$\mathbf{w}^{t+1} = \sum_{k=1}^{K} \frac{n_k}{n} \mathbf{w}_k^{t+1} \tag{6.1}$$

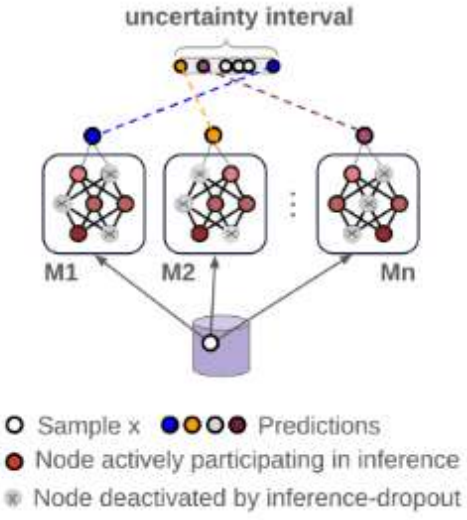
where n_k is the number of samples in dataset k, $n = \sum_{k=1}^{K} n_k$ is the total number of samples across all clients. Eq. 6.1 ensures that clients with larger datasets contribute proportionally more to the updated global model. An alternative to FedAvg is FedProx [79], which extends the former by incorporating a proximal term into the local loss function



a) The architecture of FAITH-FDSS, based on a 5-block MLP.



b) Different training approaches compared in this research work



c) MC dropout at inference time to conduct uncertainty assessment

FIGURE 6.1

minimization, as follows:

$$\min_{\mathbf{w}_k}(F_{CE}^k(\mathbf{w}_k) + \frac{\mu}{2}\|\mathbf{w}_k - \mathbf{w}^t\|^2) \tag{6.2}$$

where $F_{CE}^k(\mathbf{w}_k)$ is the cross-entropy (CE) loss at client k, $\mathbf{w}^t$ is the global model from the previous communication round, $\mathbf{w}_k$ is the local model being updated at client k. Aggregation at the server is not modified and is the same as in vanilla federated learning. The purpose of the proximal term is to regulate local updates by penalizing deviations from the global model, thereby ensuring consistency across clients during training. The strength of the penalization is controlled by the hyperparameter μ; higher values of μ correspond to stronger penalization, whereas $\mu = 0$ causes the proximal regularization term to vanish, thereby reducing FedProx to standard FedAvg. FedProx is specifically designed for scenarios with heterogeneous non Independently and Identically Distributed (non-IID) data [79], For the purposes of this case study, we employed FedProx to investigate its potential advantages in addressing variations and heterogeneity among datasets D1–D9. This characteristic closely matches our application scenario, where each dataset (D1–D9) corresponds to a distinct Italian region with its own socio-economic composition, fiscal regulations, and administrative practices. Such regional diversity introduces significant variability in feature distributions and class proportions, which can cause standard FedAvg to diverge or converge toward suboptimal local minima. By introducing a proximal term that regularizes local updates around the global model, FedProx mitigates this issue, ensuring more stable convergence and balanced knowledge aggregation across clients. In the context of our expert system, this leads to a more robust and equitable decision-support framework, capable of providing consistent recommendations even when local fiscal behaviors differ substantially among municipalities.

6.3.3 Centralized training

The upper panel of Fig. 6.1 b), illustrates the centralized approach, which trains the model on a single, aggregated dataset combining data from all sources (D1-D9). During training, a global model, applicable across all datasets, is selected at the epoch corresponding to the minimum validation loss observed before early stopping is triggered.

6.3.4 $(\varepsilon, \delta)-$ Differential Privacy

If an algorithm satisfies DP, the inclusion or exclusion of a single individual's record in the dataset induces only negligible changes in the probability distribution of its outputs, thereby formally bounding the contribution of any single sample to the overall result. More rigorously, a randomized algorithm [80] $\mathcal{A} : \mathcal{D} \to \mathcal{R}$ supports $(\varepsilon, \delta)-$DP if for any pair of datasets $(D_i,\ D_j)$ differing by at most one sample and for all sets of outputs $(S) \subseteq \text{Range}(\mathcal{A})$, the following holds:

$$\Pr[\mathcal{A}(D_i) \in S] \leq e^{\varepsilon} \Pr[\mathcal{A}(D_j) \in S] + \delta \tag{6.3}$$

where ε denotes the privacy budget, a parameter that bounds the maximum admissible privacy loss under a given setting. Privacy guarantees are inversely proportional to the privacy budget ε: larger ε implies weaker privacy protection. In the context of FL, $\mathcal{A}$ represents the training procedure, $\mathcal{A}(\mathcal{D})$ the trained weights, ε the cumulative privacy loss accumulated across all training rounds. The parameter δ denotes the probability that the privacy guarantee specified by ε fails to hold. The type of DP considered here is approximate, meaning δ is always strictly positive.

A common mechanism for implementing (ε, δ)-Differentially Private Stochastic Gradient Descent (DP-SGD), thereby enforcing DP at the level of gradient updates, proceeds in two main steps:

1. gradient clipping

$$g_i = \nabla_W \mathcal{L}_i(W), \quad i = 1, \ldots, B \tag{6.4}$$

$$\bar{g}_i = g_i \cdot \min\left(1, \frac{C}{\|g_i\|_2}\right) \tag{6.5}$$

2. Gaussian noise injection:

$$\tilde{g} = \frac{1}{B} \sum_{i=1}^{B} \bar{g}_i + \mathcal{N}\left(0, \sigma^2 C^2 I\right) \tag{6.6}$$

$$W \leftarrow W - \eta \tilde{g} \tag{6.7}$$

where C is the gradient clipping or sensitivity, σ the noise multiplier, η the learning rate (LR), g_i the gradient computed over the i^{th} mini-batch, $\bar{g}_i$ the clipped gradient, $\tilde{g}$ the perturbed gradient, B the number of mini-batches, W the model weights, I the identity matrix. By constraining gradients to a maximum norm, clipping prevents any single gradient from dominating the update, allowing smaller Gaussian noise to sufficiently mask the identity of all samples.

Inspired by the Dynamic Clipping-SGD-Percentile approach [81], we determine the clipping threshold C using a simple adaptive mechanism: C is set to a chosen percentile of the gradient norms in each mini-batch to account for the variable magnitude of learning across early and later training epochs. This approach ensures that most gradients are scaled appropriately while allowing occasional larger gradients to be clipped.

The pseudo-code for the FAITH-FDSS $(\varepsilon, \delta)-$DP-SGD FL core is described in Algorithm 1

Algorithm 1: Local DP pseudo-code

Input : initial global model w_1, number of rounds T, number of local epochs in a round E, number of clients K, dataloaders $\mathcal{D}^k_{loader}$, noise multiplier σ, initial norm clipping threshold C

Output: final global model w_T

for $t \leftarrow 1$ **to** T **do**
  $C_0^k \leftarrow C$
  for $k \leftarrow 1$ **to** K **do**
    $w_t^k \leftarrow w_t$
    for $e \leftarrow 1$ **to** E **do**
      foreach $\mathcal{B} \in \mathcal{D}^k_{loader}$ **do**
        $g_k \leftarrow \text{compute}(\ell_{\text{CE}}, w_t^k, \mathcal{B})$
        $C_{t+1}^k \leftarrow P_{90}(\{\|g\|_2 : g \in g_k\})$
        $\bar{g}_k \leftarrow \text{clip}(g_k, C_t^k)$
        $\tilde{g}_k \leftarrow \bar{g}_k + \mathcal{N}(0, (\sigma C_t^k)^2)$
        $w_t^k \leftarrow \text{update}(w_t^k, \tilde{g}_k)$
      end
    end
  end
  $w_{t+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{N} w_t^k$
end
return w_T

6.3.5 Feature importance

We conduct a feature importance analysis using SHapley Additive exPlanations (SHAP) [82] and we use the outcome to provide guidelines for the revenue officer supervising the application. As shown in Fig. 6.8, the most important feature is the number of notices received by a payer in the past three years, followed by the total amount owed and the amount already paid. TARI-tax related features carry greater weight due to the larger number of samples reporting TARI-tax compared to IMU-tax. In practice, predictions that a payment will remain unpaid are primarily based on the payer's history of multiple notices and the total or partially paid tax amount.

6.3.6 Monte Carlo Dropout

MC dropout [83] quantifies the uncertainty associated with a sample (x), performing T multiple stochastic forward passes as follows:

$$\{\hat{y}^{(t)}\}_{t=1}^{T} = \text{MC-Dropout}(x) \tag{6.8}$$

where MC-Dropout(x) denotes the stochastic forward pass over x with dropout layers kept active, in contrast to standard deterministic inference, in which they are systematically turned off, and $\hat{y}^{(t)}$ denotes the resulting prediction. Fig. 6.1 c) graphically illustrates this procedure, depicting the multiple stochastic models derived from the trained global model through the application of dropout during the inference phase.

Equation 6.8 defines an empirical distribution of predictions from which the predictive mean and variance can be estimated:

$$\mathbb{E}[y|\mathbf{x}] \approx \frac{1}{T}\sum_{t=1}^{T}\hat{y}_t \tag{6.9}$$

$$\text{Var}[y|\mathbf{x}] \approx \frac{1}{T}\sum_{t=1}^{T}\hat{y}_t^2 - \left(\frac{1}{T}\sum_{t=1}^{T}\hat{y}_t\right)^2 \tag{6.10}$$

Equation 6.10 quantifies the model's predictive uncertainty: a higher variance across stochastic inferences indicates lower confidence in the corresponding prediction. This provides a practical means to assess the reliability of model outputs and to calibrate the level of human oversight accordingly.

MC dropout is particularly well suited to the architecture underlying FAITH-FDSS, which includes one dropout layer per block, as illustrated in Fig. 6.1 a). In addition to the hyperparameters already defined by the architecture, MC dropout requires specifying the number of stochastic forward passes, which we set to 100, following [84].

6.4 Experimental settings and procedure

In this section, we describe all aspects related to the training of model used in our experiments. As described in Section 6.4.1, hyperparameter tuning is first performed in the centralized setting, which represents our baseline, to identify the optimal configuration of architectural and training hyperparameters. The resulting configuration is then transferred to the distributed setting to ensure comparability between centralized and federated experiments. Section 6.4.2 outlines the practical implementation of the training procedure for the centralized approach. Since the hyperparameters obtained from the centralized training are also adopted in the distributed setting, Section 6.4.3 introduces the single hyperparameter specific to the federated configuration that requires separate tuning, highlighting its role. Section 6.4.4 presents the practical execution of the training procedure for the federated approach. Finally, Section 6.4.5 provides the implementation details, including the development framework, auxiliary libraries, and computational resources utilized in the experiments.

6.4.1 Hyperparameter tuning for the baseline

The optimal hyperparameter set is determined by monitoring the training and validation losses, throughout the optimization process and selecting the configuration corresponding to the minimum validation loss. The early convergence criterion allows the training to terminate once performance gains plateau, thereby reducing the total number of epochs

TABLE 6.3: Summary of hyperparameter tuning for centralized and distributed training. The search space and optimal configuration identified in the centralized setup are reported for each hyperparameter. This configuration is subsequently transferred to the federated learning (FL) experiments to ensure methodological comparability. In the FL setting, the number of training epochs is replaced by the number of communication rounds, an additional hyperparameter controlling the global convergence process, while the number of local epochs per round is fixed at 10, consistent with the optimal value found in centralized training.

hyperparameter	**search space**	**optimal configuration**
nodes per layer	{(256,128,64,32,8), (128,64,32,8)}	(256,128,64,32,8)
dropout	{.40,0.45,0.50,.55,.60}	0.50
embedding dims	{(2,2,2), (2,2,4)}	(2,2,4)
batch size	{128, 256, 512}	256
initial LR (SGDR)	[0.1, 0.01]	0.01
minimum LR (SGDR)	[0.001, 0.0001]	0.0001
num of epochs	[1, 200]	84
num of rounds	[1, 50]	15

while maintaining comparable performance. In Table 6.3, the list of hyperparameters, corresponding search space and best candidate.

6.4.2 Centralized training of the baseline

From each initial training set, a validation set is subsequently obtained using an 80:20 split, maintaining the same minority-to-majority class proportion of the training set of origin. The application of cross-validation is not considered necessary, given the substantial size of the datasets, even after down-sampling. Training is performed with a batch size of 256, with each batch balanced to maintain the minority-to-majority class proportion and mitigate the effect of class imbalance on learning. Stochastic Gradient Descent with Warm Restarts (SGDR) [85] serves as the optimizer, using an initial LR of 0.01, a minimum LR of 0.0001, and 10 epochs per cycle. The loss is defined as weighted Cross-Entropy (CE) to further account for potential class imbalance. Regularization is accomplished by the Dropout layers.

6.4.3 Hyperparameter tuning for FAITH-FDSS

As shown the bottom panel of Fig. 6.1 b), in the FL setting, the training process unfolds over a sequence of communication rounds, each comprising multiple local epochs executed independently by the clients (step 2) before model aggregation is performed at the server (step 4). The number of communication rounds becomes an additional hyperparameter to be tuned in the FL experiments, replacing the notion of total training epochs in the centralized setup. In our experiments, each round consists of 10 local epochs, the minimum required for the adopted LR scheduler, Cosine Annealing with Warm Restarts, to complete a full cycle. The number of local epochs is not increased beyond this threshold to avoid potential uncontrolled divergence among local models, as reported in [79]. To tune the number of rounds, two criteria are applied: (1) the average loss across all client validation sets is evaluated at the end of each round using the current global model; and (2) when, beyond a certain point, the loss exhibits only marginal improvements across successive rounds, the configuration corresponding to the near-minimum average loss is selected. Criterion (2) is particularly relevant for the subsequent DP phase, as in (ε, δ)-DP-SGD FL, a larger number of communication rounds leads to a higher cumulative privacy loss ε, which in turn necessitates injecting larger amounts noise to maintain the same privacy guarantee.

6.4.4 Federated training for FAITH-FDSS

Local training at each client follows the same procedure as in the centralized setting, with the sole difference that it is limited to 10 epochs per communication round. FAITH-FDSS performs local training for a total of 15 rounds. In experiments employing the FedProx algorithm, the proximal term serves as an additional regularization component.

6.4.5 Implementation details and computational resources

The entire framework is built in Python v3.8.10 [86] and Pytorch v2.0.0 [87]. Flower Framework v1.22.0 [88] is exploited for the federated learning whereas Opacus v1.5.4 [89]

for local DP. Shap v0.49.1 [18] is used to analyse feature importance of the model. Computations are distributed on a node containing 4 NVIDIA A100 GPUs, each equipped with 64GB VRAM, on a system with 512GB RAM.

6.5 Results and Discussions

In this section, we summarize the main experiments conducted and the corresponding results obtained to assess the validity of FAITH-FDSS in terms of both compliance and utility. Specifically, Section 6.5.1 introduces and justifies the primary metric adopted to evaluate model performance. Section 6.5.2 presents the results of the comparison between FAITH-FDSS, the centralized baseline, and selected federated alternatives, in support of performance assessment. Section 6.5.3 reports the results concerning the privacy–performance trade-off, obtained by progressively introducing increasing amounts of noise to enhance privacy protection. Section 6.5.4 discusses the outcomes of the feature importance analysis, in support of transparency, while Section 6.5.5 presents the results of the MC dropout experiment, as evidence supporting the trustworthiness of the system.

6.5.1 Metrics

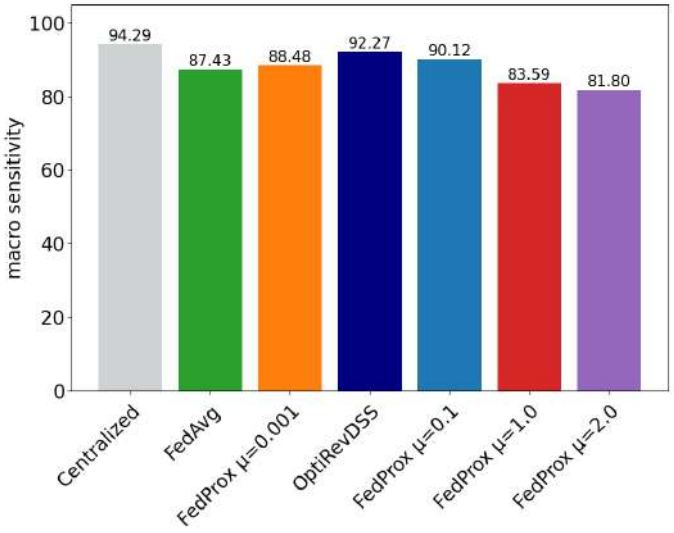
Macro-sensitivity serves as the primary metric for evaluating model performance. It represents the average sensitivity across all classes, aligning with the study's emphasis on sensitivity as a key performance dimension. In binary classification, the sensitivity of one class corresponds to the specificity of the other, and vice versa; therefore, in our research, macro-sensitivity provides balanced control over both sensitivity and specificity. The use of symmetric metrics is preferred, as they ensure an equitable assessment across classes and prevent class imbalance from disproportionately influencing the overall evaluation.

6.5.2 Performance Comparison between FAITH-FDSS, the Centralized Baseline, and Other Federated Learning Approaches

FAITH-FDSS is trained using the FedProx federated learning approach with the hyper-parameter $\mu = 0.01$. This configuration is compared with both the centralized baseline and the same FL approach under different μ values, with the aim of demonstrating that (1) the performance loss with respect to the centralized baseline is negligible, and (2) the performance of the alternative federated configurations is broadly aligned with that of FAITH-FDSS, which achieves the highest overall performance. As illustrated in Figure 6.2, FAITH-FDSS exhibits only a marginal decrease in performance compared to the centralized model. Among the federated approaches, it achieves the highest overall performance, benefiting from the mild regularization introduced by its proximal term. Moderate regularization ($\mu = 0.1$ or 0.01) enhances performance, whereas stronger regularization ($\mu = 1$ or 2) results in a more pronounced decline. The FedAvg baseline ($\mu = 0$) yields intermediate results, positioned between these two groups.

The learning dynamics shown in Figure 6.3 further support these findings. All federated configurations exhibit a rapid performance increase during the first communication rounds, followed by gradual stabilization. FAITH-FDSS, trained with FedProx ($\mu = 0.01$), consistently maintains the highest trajectory, achieving faster convergence and higher final macro-sensitivity compared with alternative setups. These trends confirm that incorporating a mild proximal constraint improves both convergence stability and predictive robustness in a potentially non-IID federated setting, making FAITH-FDSS particularly suitable for our expert system operating on regionally distributed fiscal data.

As shown in Fig. 6.4, breaking down the results by individual datasets, the performance of FAITH-FDSS, is generally consistent across all datasets except for **D5**, which can be considered an outlier. As detailed in Table 6.1, this dataset is considerably smaller and exhibits a pronounced class imbalance, which jointly limit the representativeness of its local model updates during aggregation. The convergence dynamics depicted in Fig. 6.5 confirms this behavior. Each box plot summarizes the distribution of macro-sensitivity across datasets for every communication round, when FAITH-FDSS is evaluated. The



The histogram compares the performance of the best-performing centralized model, used as the baseline, with various federated learning approaches, identifying FAITH-FDSS as the top-performing distributed configuration. Federated performance is reported at the 15th communication round

FIGURE 6.2: Performance comparison between baseline and federated

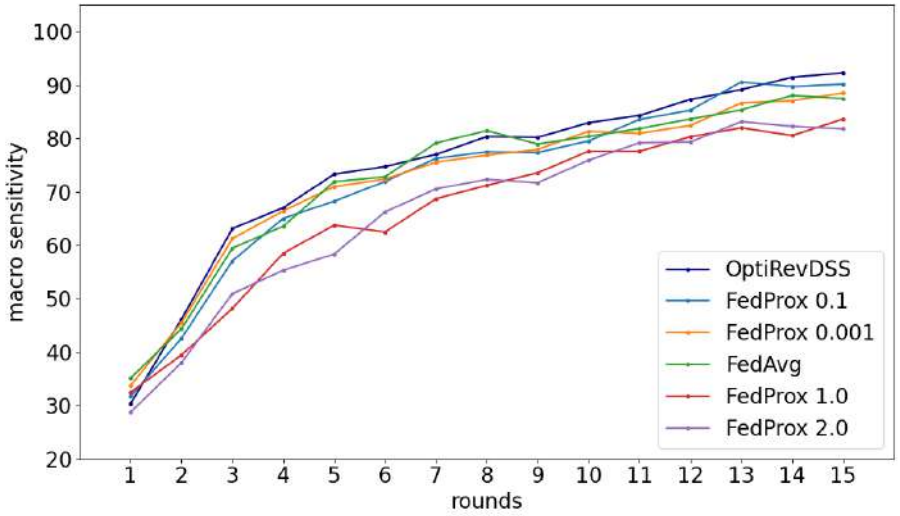


FIGURE 6.3: The plot reports the performance of different federated learning approaches, identifying FAITH-FDSS as the best-performing configuration

red line traces the trajectory of D5, showing consistently lower and more variable performance, while the blue line represents the overall mean across all datasets (D1–D9). The violet line, which excludes D5, provides a clearer view of the general trend and demonstrates steady improvement and stabilization as training progresses. The widening gap

between D5 and the remaining datasets highlights how small and imbalanced data partitions can slow convergence and introduce instability in the global model. Nevertheless, the aggregated performance, dominated by larger and more balanced clients, remains robust, confirming that FAITH-FDSS achieves reliable convergence even under realistic non-IID and unbalanced conditions, typical of fiscal data environments.

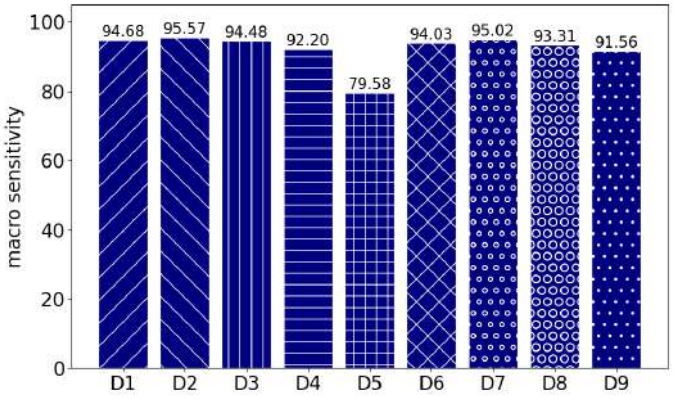


FIGURE 6.4: The histogram compares the performance of FAITH-FDSS, the best-performing federated configuration, at round 15, for different datasets

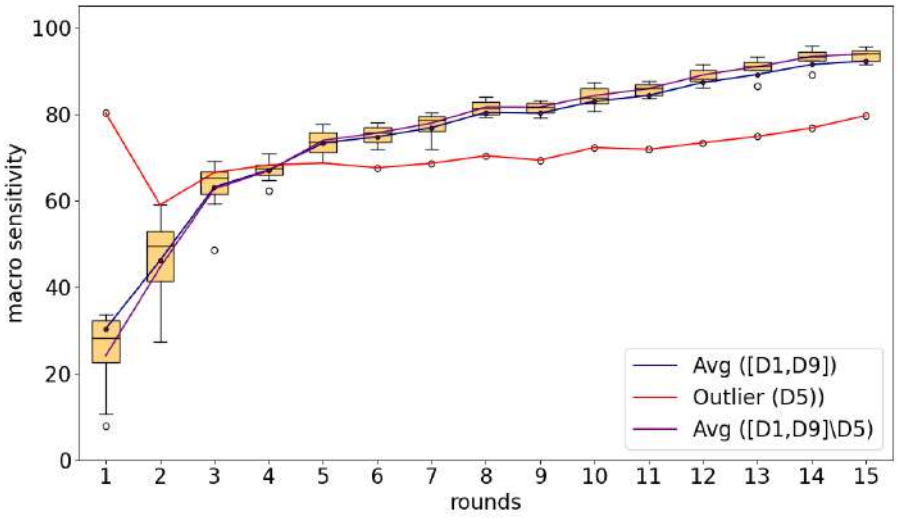


FIGURE 6.5: For each round, the box plot illustrates the macro-sensitivity achieved by the best performing federated learning approach (FedProx $\mu = 0.01$) across the datasets. The red line highlights D5, identified as an outlier due to its comparatively poor performance. The blue line traces the mean macro-sensitivity per round across all datasets (D1–D9), whereas the violet line represents the mean excluding D5, thereby accounting for the outlier's effect.

6.5.3 Experiments to define noise

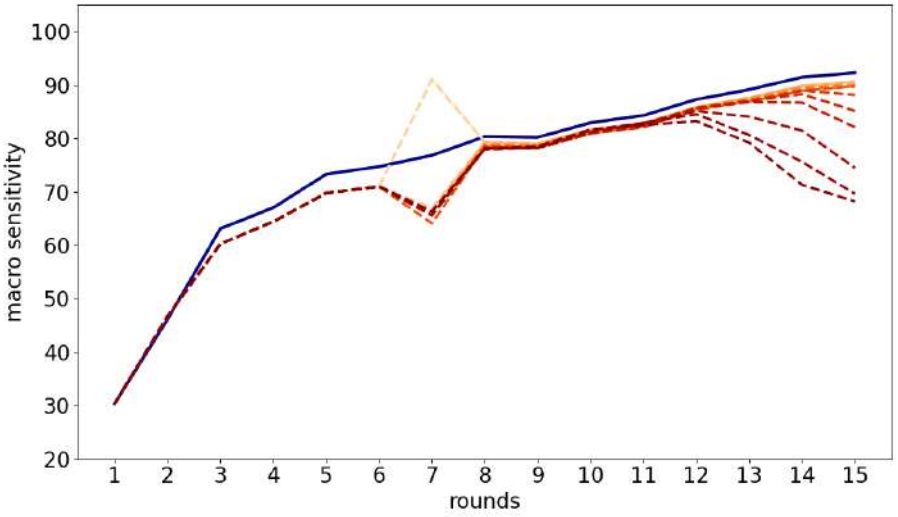
To assess the robustness of FAITH-FDSS against noise injection, with reference to the procedure explained in Eq. 6.6 and Eq. 6.7, we vary the intensity of the injected Gaussian noise by adjusting the multiplier σ within a typical range of [0.1, 2.0] [89]. For each value of σ, we compute the cumulative privacy budget ε using the Rényi DP accountant implemented in Opacus, a relaxation that provides a tighter bound than standard composition methods and offers a more accurate estimation of the effective privacy achieved [90]. In this way, the magnitude of the injected noise is calibrated to maintain a low privacy loss while preserving model accuracy.

As shown in Figure 6.6, the introduction of DP through Gaussian noise progressively affects the performance of FAITH-FDSS. As the noise multiplier (σ) increases, the magnitude of injected noise grows, producing a gradual degradation in macro-sensitivity. This behavior reflects the inherent trade-off between privacy and utility: higher σ values provide stronger privacy guarantees but reduce model accuracy. Nevertheless, when σ is maintained within the range $1.0 \leq \sigma \leq 1.4$, the decrease in performance remains moderate, indicating a practical operating region in which an adequate trade-off of privacy can be achieved without substantially compromising predictive reliability.

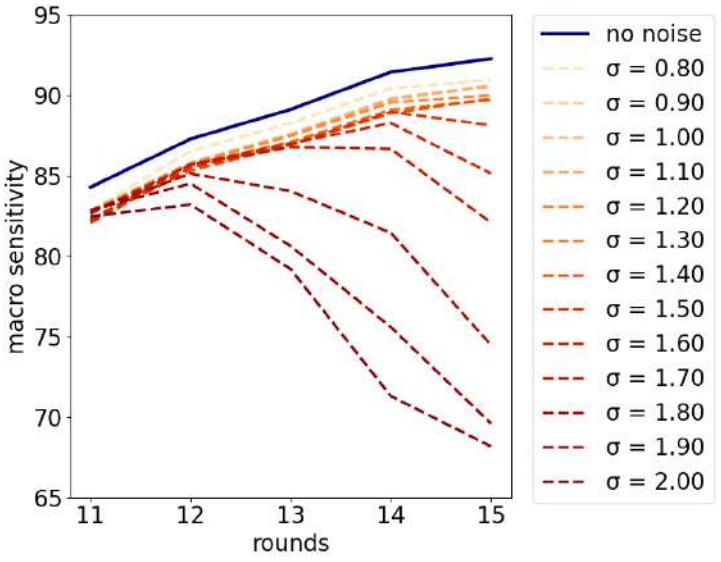
In addition, Fig. 6.7 shows that for noise values around 1.3, the cumulative ε remains reasonably low, thereby providing an adequate privacy guarantee. Tighter clipping thresholds, such as the 80th percentile, constrain gradient magnitudes more effectively and further reduce the accumulation of privacy loss across training rounds. In our experiments, a 90th percentile threshold (p) is adopted as a balanced configuration to preserve model sensitivity while maintaining sufficient privacy protection. This choice prevents clipping from being either excessively aggressive, which could dampen useful gradient information, or too soft, which would weaken privacy control.

6.5.4 Feature importance and transparency

We conduct a feature importance analysis using SHapley Additive exPlanations (SHAP) [82] and we use the outcome to provide guidelines for the revenue officer supervising the



(a) Effect of noise injection on FAITH-FDSS performance



(b) Zoom of FAITH-FDSS performance over the last 5 rounds

FIGURE 6.6: The plot illustrates how FAITH-FDSS performance progressively degrades with increasing values of the noise multiplier, which correspond to higher magnitudes of injected noise

application. As shown in Fig. 6.8, the most important feature is the number of notices received by a payer in the past three years, followed by the total amount owed and the amount already paid. TARI-tax related features carry greater weight due to the larger number of samples reporting TARI-tax compared to IMU-tax. In practice, predictions that a payment will remain unpaid are primarily based on the payer's history of multiple notices and the total or partially paid tax amount.

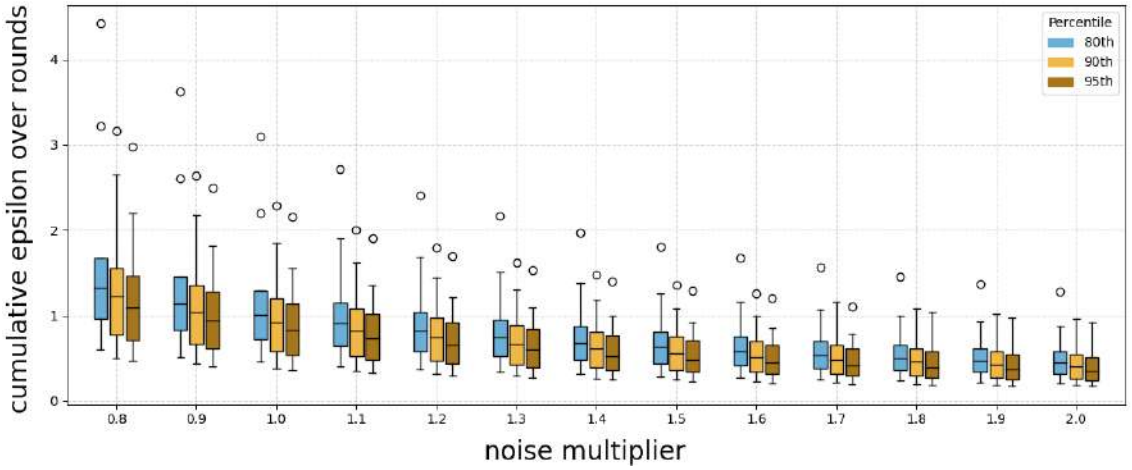


FIGURE 6.7: The plot illustrates how the cumulative ε varies with different magnitudes of the noise multiplier and varying strengths of gradient clipping, expressed as the percentile of gradient norms within each mini-batch

A minor consideration concerns the presence of two different taxes, TARI and IMU, with a clear predominance of the former. This imbalance is not explicitly accounted for, as it is deemed of limited relevance. However, it emerges in the feature importance analysis, and in that context, awareness of it allows for a proper interpretation of the results, as discussed in Section 6.5.4.

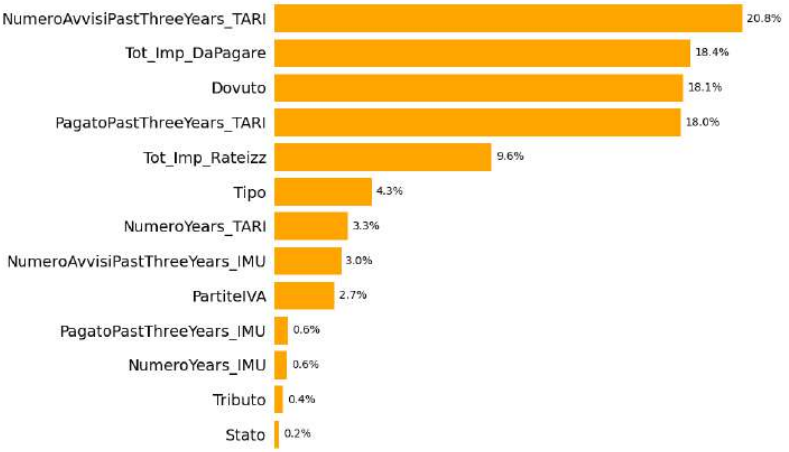


FIGURE 6.8: Mean absolute SHAP value (as %)

6.5.5 MC Dropout

Fig. 6.9 illustrates the results of the MC Dropout analysis applied to the final global model produced by FAITH-FDSS. We set a threshold of 0.71 on the x-axis and 0.32 on the y-axis, such that predictions with a posterior mean greater than 0.71 and a maximum variance of 0.32 are considered reliable (gray square). Most correct predictions, shown as blue dots, fall within this quadrant, while incorrect predictions are mostly outside it. This suggests that the chosen criterion for determining the reliability of a prediction, although not perfectly separating correct from incorrect predictions, is adequate.

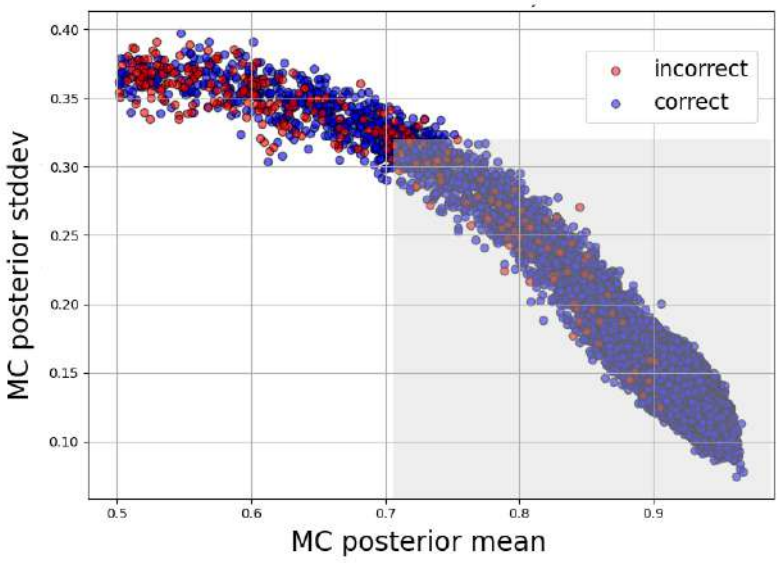


FIGURE 6.9: MC dropout results for the 'not paid' class. Red dots indicate false negatives, blue dots indicate true positives. Based on the thresholds for posterior mean and posterior standard deviation, the gray square highlights the region of high-confidence predictions

6.6 Conclusions

FAITH-FDSS is the outcome of interdisciplinary research that, by embracing and extending the principles underpinning Privacy by Design, advances concurrently on both the technological implementation and the regulatory and ethical dimensions since the earliest stages of the AI system life-cycle. The key requirements of the European regulatory and ethical framework, considered most relevant for the present case study, have been

addressed through the implementation of specific components or strategies within FAITH-FDSS. Relating [2], they can be summarized as follows:

1. **Data Anonymization (Art. 25(1)):** All direct or indirect references that could lead to the identification of any data subject have been removed; DP has been integrated to prevent re-identification of individuals within the dataset.
2. **Data Minimization (Art. 25(1)):** The feature set has been reduced to the minimum necessary to ensure adequate model performance.
3. **Computational and Output Privacy (Art.25):** Federated learning combined with DP enhances data protection to an appropriate level based on the characteristics of the datasets, while imposing minimal performance penalties.

Relating [1], they can be summarized as follows:

1. **Technical Robustness and Safety (Art. 15):** The model ensures robust performance through the design and optimization of a five-layer MLP architecture. During inference, MC-Dropout further validates robustness and allows users to accept or reject outputs based on the associated uncertainty, thereby promoting best practices in trustworthy AI.
2. **Data Governance (Art. 10):** Dataset characteristics are carefully considered to mitigate potential biases, including class imbalance or features inherently susceptible to discrimination (e.g., gender, age).
3. **Transparency (Art. 13):** Feature importance analysis enables users to verify that the decision-making process is based on a logical and interpretable rationale.
4. **Human Agency and Oversight (Art. 14):** The framework provides mechanisms to adjust the level of human oversight, which remains mandatory irrespective of the confidence associated with automated decisions.

6.7 Future developments

First and foremost, the current version of FAITH-FDSS should be complemented with comprehensive documentation, pursuant to Art. 11 in [1] on Technical Documentation, and automatic event logging, pursuant to Art. 12 in [1] on Record-Keeping, to ensure full compliance with [1] and readiness to manage tasks with higher risk levels. This approach will guarantee reliable system performance under real-world conditions while maintaining both regulatory and ethical standards. After a fully developed version of FAITH-FDSS is obtained, it should undergo intensive operational validation to test and confirm its maturity in terms of both utility and compliance. Once it is confirmed that FAITH-FDSS provides a robust foundation, a range of potential enhancements can be considered, including:

- **scalability**: expanding the federated learning setup to encompass additional regional offices or departments, ensuring efficient and secure model updates;
- **enhanced privacy mechanisms** (as required for more sensitive tasks): integrating DP with secure multi-party computation techniques to further mitigate information leakage;
- **higher-risk task integration**: gradually incorporating tasks classified as higher risk to test the robustness of FAITH-FDSS under increasingly challenging scenarios;
- **integration with complementary systems**: enabling interoperability with other government or financial systems to provide a holistic view of taxpayer data while adhering to strict data governance policies and carefully avoiding any profiling risks.

These future directions aim to make FAITH-FDSS not only fully compliant and reliable but also more versatile, scalable, and capable of supporting informed, trustworthy decisions across multiple operational contexts.

Chapter 7

Operationalizing Compliance for High-Risk AI Systems

In this chapter, we examine how the principles of Data Protection, Data Governance, Transparency, Human oversight, and Accuracy and Robustness have been implemented across two use cases involving highly sensitive data and critical tasks in the field of biomedical imaging. We first discuss the strategies adopted for data management, in accordance with the requirements set out in Art. 10(1–6) of [1], and then focus on the role of Federated Learning (FL) as a risk mitigation strategy for high-risk AI systems, designed to satisfy the obligations established under Art. 25(a) of [2]. We further analyse the challenges posed by the DL models employed, particularly with respect to Art. 13 of the [1] on transparency, and conclude by succinctly illustrating the potential integration of these tools into clinical practice, confirming their role as systems operating solely in a decision-support capacity, as envisaged under Art. 14 of [1].

7.1 Determination of the Risk Class

Under Art. 6(2) of [1], an AI system is classified as high-risk when its intended purpose falls within one of the areas referred to in Annex III, which explicitly includes systems intended for medical diagnosis. Accordingly, a stenosis detection system, designed to assist

clinicians in identifying vascular or cardiac stenoses, generally falls within this high-risk category.

Art. 2(3) provides a conditional derogation, stating that an AI system referred to in Annex III shall not be considered high-risk if it does not pose a significant risk of harm to health, safety, or fundamental rights, including by not materially influencing the outcome of decision-making. Although this could theoretically exempt some systems, in practice even diagnostic AI that merely provides information or recommendations typically guides clinical decisions. For instance, presenting a stenosis probability, heatmap, or diagnostic suggestion can influence a clinician's judgment, even though the final decision remains theirs. Thus, the criterion of material influence is generally met.

In light of these considerations, both the stenosis detection system and the fetal plane classification system are classified as high-risk. Adopting this precautionary stance ensures inclusivity and aligns with a conservative interpretation of the regulatory framework, emphasizing patient safety and compliance with [1].

7.2 Addressing Data Protection Requirements

Art. 4(15) of [2] defines "data concerning health" as "personal data related to the physical or mental health of a natural person, including the provision of health care services, which reveal information about his or her health status." Art. 9(1) establishes a general prohibition on the processing of clinical or health-related data, while Art. 9(2) specifies a set of exceptions under which such processing may lawfully take place. In the context of the health-related data employed in the two use cases presented here, and analyzed in detail in Chapter 8 and Chapter 9, the relevant exceptions are (a), (h), (i), and (j).

From a soft-law perspective, Section C(1) of [3] explicitly identifies health and well-being as a key domain for trustworthy AI, encompassing strategies aimed at safeguarding and promoting individual health, as well as supporting research in this field.

In this regulatory and soft-law context, FL is a learning paradigm, applicable to both ML and DL models, that enables model training through decentralized collaborative

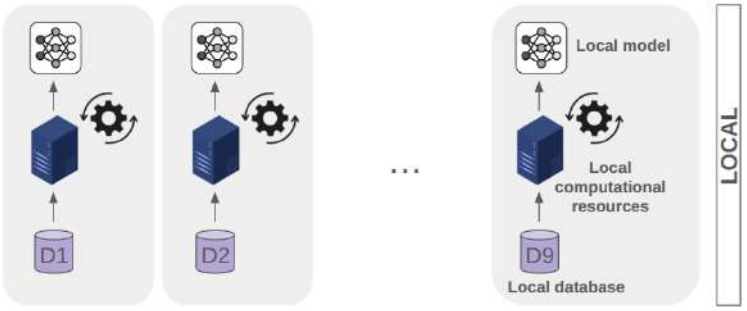
learning. Multiple isolated datasets are processed locally, without transferring raw data to a centralized server. Instead, only the gradients or weight updates of the local models are periodically transmitted to a coordinating server, where they are aggregated into a global model to synchronize the training process. This approach allows participants to benefit from the advantages of shared data, such as improved model generalization, while preserving data privacy and reducing the risks associated with centralization [91].

Fig. 7.1 compares three common approaches to model training: local, centralized, and federated. In a local setting, each user trains a model solely on their own dataset, which may limit the model's generalization capability depending on the characteristics and size of the data. A centralized model enables data sharing among multiple users, leveraging a larger and more diverse global dataset; however, this comes at the cost of exposing sensitive user data to a central server, which must be fully trusted. Furthermore, beyond privacy concerns for data subjects, it should not be overlooked that, in the context of big data, datasets may possess significant economic and strategic value (the V of Value being one of the "7 V's" [92]) and their sharing is consequently not straightforward..

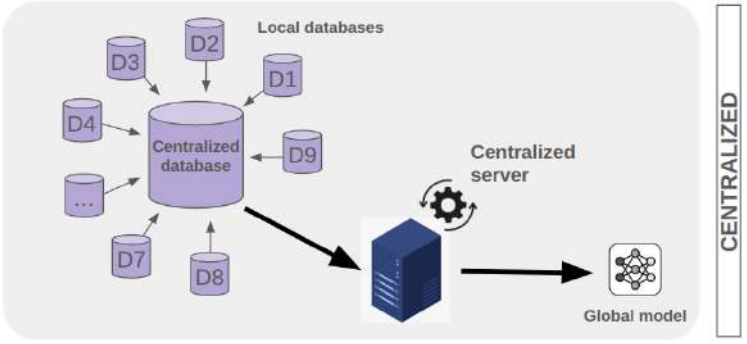
In greater detail, Fig. 7.1 c) illustrates the training process under FL. Initially (Step 1), the server distributes the same, potentially pretrained, model to all users (hereafter referred to as clients [91]). Each client then trains the received model locally using its own dataset for a specified number of local epochs (Step 2). Upon completion, each client transmits its model updates back to the server (Step 3), where the updates are aggregated (Step 4) and the resulting global model is redistributed back to the clients (Step 1). Steps 1–4 are collectively referred to as a training round.

FL provides an initial layer of protection, offering only partial security against attacks on individual data samples. Indeed, although raw data is never directly exchanged, the transmitted gradients or weight updates may, in principle, permit reconstruction of the original data under certain adversarial conditions.

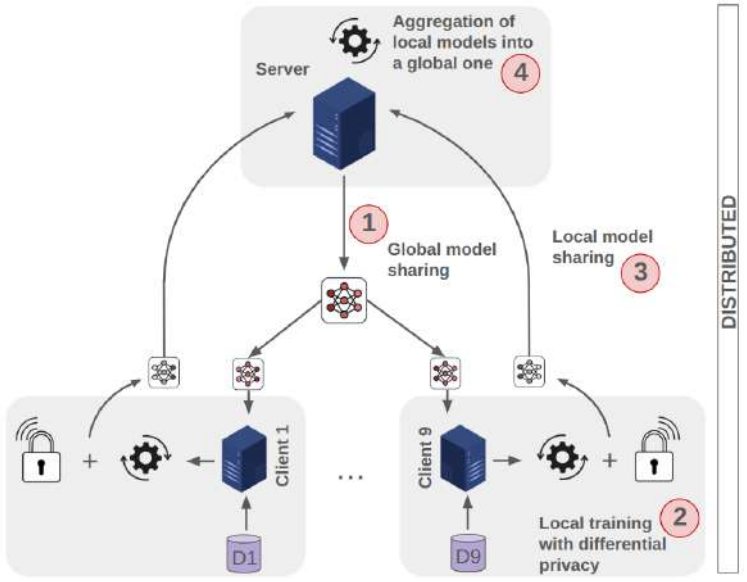
FL can be applied to any type of data, whether structured or unstructured. Chapter 8 and Chapter 9 present two relevant use cases in the domain of biomedical imaging, where privacy concerns are particularly significant. Both use cases involve unstructured



(a) **Local training**



(b) **Centralized training**



(c) **Federated training with differential privacy**

FIGURE 7.1: Different training methods (a) Local training without any data sharing (b) Centralized training with complete sharing of data and no protection c) Federated training with distributed sharing of local model weights updates and additional differential privacy protection layer

data in the form of X-ray and US images, respectively, and the models employed are typically Deep Learning architectures. The aim of this research is to broaden the scope of application domains, enabling comprehensive cross-domain investigations while introducing novel contributions within the specific biomedical context. In Chapter 6, we examined how security can be enhanced through DP, a convenient strategy, which can be applied to any data types, including images.

7.3 Addressing Data Governance Requirements

In compliance with Art. 10(2)(b) of [1], the acquisition of data and the corresponding consent procedures for the three datasets are documented in [93], [94], and [12]. The stated purpose of these activities is to obtain and make available biomedical imaging data to support the development and evaluation of DL algorithms, given the current paucity of such datasets in the biomedical domain.

In compliance with Art. 10(2) of [1], all images derived from publicly available datasets were manually annotated and curated by expert personnel, ensuring clinical accuracy and label consistency to the extent permitted by image quality. When image quality was suboptimal, as discussed in Chapter 9, specific strategies were adopted to restore an acceptable level of annotation reliability. In line with best practices for object detection, the images were then cleaned to remove potential artifacts that could mislead the neural network

In compliance with Art. 10(2)(d) of [1], Labelling and Task Definition, the images concern the identification of stenotic pathologies. Bounding boxes (BBox) provide the spatial localisation of stenotic regions, thereby defining the learning objective of the model in accordance with medical annotation practice.

In compliance with Art. 10(2) and Art. 10(2)(e) of [1], Data Quantity, Representativeness, and Balance, given the typically limited size of biomedical datasets and the scarce public availability of such resources, we address data paucity through augmentation techniques applied on the fly at each training iteration to maximise variability. A careful analysis of class imbalance is conducted to ensure fair sampling during the creation of

validation datasets. In cases where datasets are already partitioned by their authors, the balance is verified; where clear disproportions between classes is identified, the training, validation, and test sets are redefined on a patient-based split, ensuring both class balance and statistical representativeness.

In compliance with Art. 10(2)(f)–(g) of [1], Bias Detection, Prevention, and Mitigation, an examination of potential sources of bias is conducted, followed by the adoption of appropriate measures to detect, prevent, and mitigate such biases. The training, validation, and test datasets include real cases of stenotic conditions, yet true representativeness remains difficult to achieve due to the limited availability of biomedical images and the ethical approvals required for their use by institutional review boards. In the second use case, errors are artificially injected into the data to study their impact and introduce a dedicated denoising algorithm, mitigating the effect of low image quality on model performance. This approach ensures that model training remains reliable, even under conditions that would otherwise compromise diagnostic accuracy.

7.4 Addressing Transparency Requirements

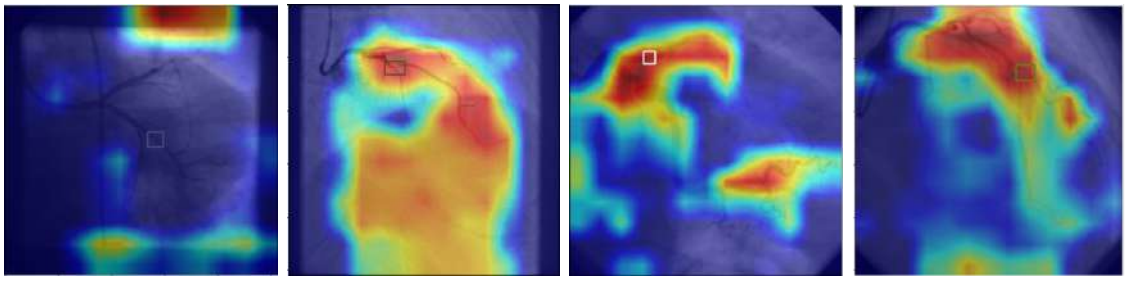
As discussed in Chapter 1, the need for a decision-making process that is comprehensible even to non-technical stakeholders predates Art. 13 of [1]: Recitals 39 and 58 and Art. 5(1)(a) of [2]. This principle also constitutes a core tenet in support of trustworthy AI, with multiple cross-references to soft-law instruments, within the preexisting non binding soft-law framework [3], where the ethical principle of Explicability translates into the practical guideline of Transparency, which includes Explainability as a key component aimed at making AI systems' behavior and reasoning understandable to humans.

Unfortunately, as models increase in complexity, they also tend to become more and more opaque, making their Explainability increasingly elusive. For images, one possible way to provide at least a partial account of the rationale underlying automated decisions is through the application of interpretation and visualization techniques, which contribute to enhancing transparency and explainability.

In our research, we specifically employ two such techniques, which are briefly described below, together with the results of their application to use cases examined in Chapter 8 and Chapter 9.

7.4.1 Gradient-weighted Class Activation Mapping (Grad-CAM)

Grad-CAM [95] is Class Activation Map (CAM) explainability technique particularly well-suited for unstructured data. It generates class-discriminative localization maps that visually highlight the regions of an input image most influential in a CNN-based model's decision. This is achieved by propagating the gradients of a target concept into either the final convolutional layer or one of the immediately preceding layers, thereby capturing high-level features. In the resulting Grad-CAM visualizations, color encoding conveys the relative importance of image regions in the model's decision-making process: red areas indicate maximum relevance, whereas orange, yellow, and blue correspond to progressively lower levels of attention. These heatmaps thus provide an intuitive representation of the model's internal reasoning, allowing for the identification of salient visual patterns underlying each prediction.



(a) No stenosis on focus (b) Diffused focusing (c) Misleading focusing (d) Precise focusing

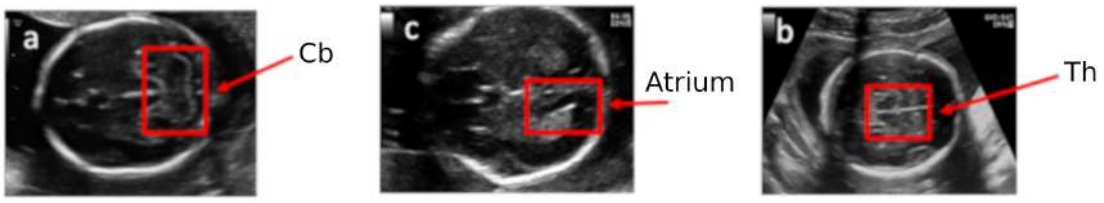
FIGURE 7.2: Grad-CAM heatmaps revealing where in the image the model focuses to provide a decision in the case study from Chapter 8

Fig. 7.2 presents four examples of Grad-CAM visual explanations applied to object detection from use case in Chapter 8 where a stenosis is defined as an abnormal narrowing in the lumen of some coronary segment and its ground-truth annotations in each sample image is identified by a rectangles mark. In panel (a), the model appears to focus predominantly on the image boundaries, attending to non-meaningful areas that may contain shortcuts, that is spurious visual cues the network erroneously associates with a class and

it completely disregards the stenosis. In (b), although the model correctly attends to the region containing the target object, it also diffusely activates across irrelevant portions of the image, suggesting some degree of uncertainty. In (c), the model concentrates on two distinct areas, of which only one corresponds to a stenotic lesion, being the second a false positive. Finally, in (d), the model confidently focuses on the clinically relevant region, correctly identifying the stenosis.

We now display Grad-CAM visual explanations applied to the classification of fetal standard planes (brain, femur, cervix, and abdomen) from the use case to be detailed in Chapter 9. Specifically considering the Brain class, this category can be further subdivided into three subtypes characterized by distinctive anatomical features (Fig. 7.3):

1. Trans-Ventricular (TV) – atrium
2. Trans-Thalamic (TT) – thalami
3. Trans-Cerebellar (TC) – cerebellum

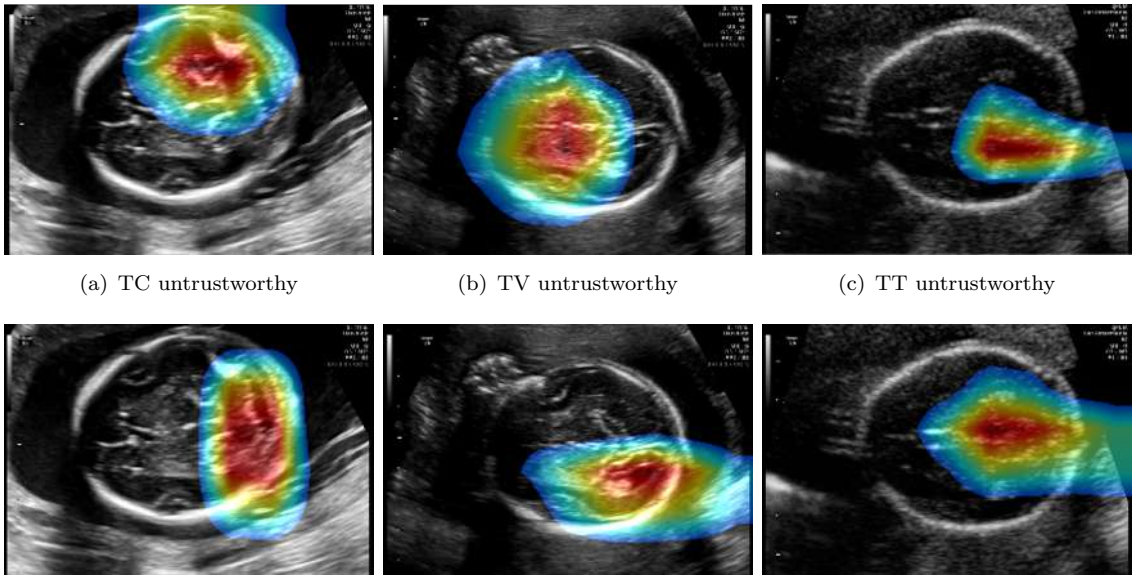


(a) Trans-cerebellar (TC) (b) Trans-ventricular (TV) (c) Trans-thalamic (TT)

FIGURE 7.3: Anatomical features of the fetal brain that differentiate the principal standard planes

In Fig. 7.4. the analysis evaluates the model's ability to make decisions grounded in these anatomically meaningful regions, thereby demonstrating genuine learning rather than reliance on superficial cues. This step is crucial to exclude the possibility that the network's decisions are guided merely by global shapes (e.g., the elliptical contour of the skull) or other shortcut features, instead of the relevant anatomical structures. In Fig. 7.4, panels (a) through (f) illustrate representative cases:

a) the cerebellum is only partially localized;



(a) TC untrustworthy (b) TV untrustworthy (c) TT untrustworthy

(d) TC trustworthy (e) TV trustworthy (f) TT trustworthy

FIGURE 7.4: Grad-CAM heatmap depicting the contribution of distinct areas of the input image to the model's decision-making process in the case study from Chapter 9

b) the cerebellum is correctly and fully highlighted;

c) the atrium is entirely missed;

d) the atrium is correctly identified;

e) the thalami are completely missed;

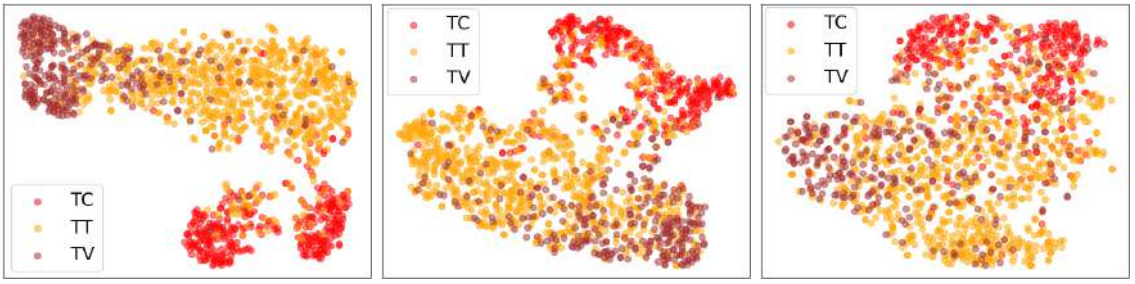
f) although thalami are not entirely delineated, the model correctly focuses along the midline.

Overall, Grad-CAM provides valuable insights into the degree of anatomical awareness acquired by the model during training, allowing an assessment of whether its internal representations are truly aligned with clinically relevant visual patterns.

7.4.2 t-distributed Stochastic Neighbor Embedding (t-SNE)

t-SNE [96] is a nonlinear dimensionality reduction technique commonly used to visualize and interpret how a model assimilates and distinguishes data. It provides insights into the

internal logic of the model and allows for an assessment of the quality of its learned representations. For instance, if data samples that are similar in the real world appear close to one another in the t-SNE projection, while clearly separated from clusters representing dissimilar samples, this suggests that the model has effectively captured meaningful similarities among data points. Conversely, the appearance of small, isolated clusters detached from the main group may indicate the presence of outliers. A key advantage of this technique lies in its ability to map high-dimensional feature spaces, often difficult to interpret due to their size, into a two or three-dimensional representation that is more intuitive for human analysis. The technique is applied to the use case detailed in Chapter 9. With



(a) Fetal brain (b) Fetal thorax (c) Fetal thorax

FIGURE 7.5: t-SNE plots related to fine grained brain classification of standard planes

reference to Fig. 7.5, each plot represents the subclasses of the Brain category (as introduced in Fig. 7.3) using distinct colors. Panel (a) illustrates a classification, where the model successfully distinguishes among classes, with only limited overlap at the boundaries. Panel (b) depicts a more ambiguous scenario, in which the Trans-Thalamic (TT) and Trans-Ventricular (TV) classes exhibit partial overlap. Panel (c) shows an unacceptable degree of confusion, suggesting that the model fails to properly separate the classes and has not developed an adequate internal representation of their differences.

7.5 Addressing Human Oversight Requirements

In the context of high-risk AI systems, *human oversight* is explicitly mandated by Art. 14 of [1]. Within the pre-existing regulatory framework, however, Art. 22(1) of the [2], as anticipated by Recital (71), had already established that *automated individual decision-making*,including profiling, *should not be based solely on automated processing*. Similarly,

within the pre-existing soft-law framework, *Respect for Human Autonomy* is identified as one of the four *ethical principles* underpinning trustworthy AI [3]. This principle is subsequently translated into one of the *seven key requirements*, namely *Human Agency and Oversight*, which operationalizes the ethical commitment to ensuring meaningful human control over AI systems.

All the use cases examined in this research assume the implementation of human oversight through Human-in-the-Loop and Human-in-Command governance mechanisms, irrespective of their risk classification.

Below, we present the questionnaire proposed in [16] as a tool for self-assessing the necessary oversight measures in AI applications, along with the corresponding responses for the high-risk use cases described in Chapter 8 and Chapter 9.

Q1 Please determine whether the AI system (choose as many as appropriate):

- Is a self-learning or autonomous system;
- Is overseen by a Human-in-the-Loop (HITL);
- Is overseen by a Human-on-the-Loop (HOTL);
- Is overseen by a Human-in-Command (HIC);

> *The AI system incorporates a Human-in-the-Loop (HITL) approach, as human personnel are actively involved at each stage of the application lifecycle: clinicians participate during the annotation process and subsequently when the tool is applied in real-world scenarios. They retain ultimate authority over individual decisions, given that the application is intended solely to support human decision-making in the review of specific cases. The system also incorporates a Human-in-Command (HIC) framework: in the healthcare domain, the Ethics Committee is responsible for defining the rules, boundaries, and approval criteria for AI deployment. This includes specifying operational constraints, conducting risk assessments, establishing protocols, and determining which data may be used.*

Q2 Have the humans (human-in-the-loop, human-on-the-loop, human-in-command) been given specific training on how to exercise oversight?

> *Clinicians, as end users of the AI applications, are expected to have advanced expertise and substantial experience in interpreting US or X-ray images within their specific medical domain.*

Q3 Did you establish any detection and response mechanisms for undesirable adverse effects of the AI system for the end-user or subject?

> *End users were equipped with visual inspection tools to understand the rationale behind individual decisions, with the goal of preventing potential adverse effects.*

Q4 Did you ensure a 'stop button' or procedure to safely abort an operation when needed?

> *A dedicated 'stop button' is unnecessary for this AI systems.*

Q5 Did you take any specific oversight and control measures to reflect the self-learning or autonomous nature of the AI system?

> *Both systems support horizontal scalability through the FL framework: as new datasets become available, they can be incorporated into the training process, which may be periodically repeated. The performance metrics established during system development can be continuously updated as the systems are applied to real-world cases, enabling ongoing monitoring of their generalization capabilities.*

7.6 Addressing Accuracy and Robustness Requirements

Also for high-risk case studies, thorough analysis of the strategies implemented to optimize their performance is provided in dedicated chapters (Chapter 8 and Chapter 9). Here, we anticipate a brief discussion on how compliance with Art. 15 of [1] is pursued.

In the first case study, which focuses on the detection of stenoses in radiographic images, particular attention is devoted to enhancing the robustness of the proposed FL framework both within individual clients (intra-client) and across different clients (inter-client). To this end, we first adopt a well-regularized FL approach, and subsequently introduce strategies for information sharing that preserve data privacy, such as the exchange of intensity histograms. In addition, a test-time mechanism specifically designed for this scenario is incorporated to further improve accuracy, relying solely on the data from the individual test sample.

In the second case study, the system's robustness is assessed by progressively introducing label noise, to be intended as the incorrect annotation of US images. To address this challenge, we propose a mechanism capable of substantially mitigating the impact of noise and almost completely recovering the correct labels. This approach significantly enhances the robustness and resilience of the system, making it particularly suitable for scenarios in which only low-quality images are available.

Chapter 8

High-Risk AI Case Study I: FedStenoNet

Coronary heart disease refers to the formation of plaque in the heart arteries that could lead to ischemic stroke. Current research in automatic artery stenosis detection mainly focuses on the analysis of computed tomography angiography [97, 98], while niche attention is given to X-ray Coronary Angiography (XCA), which remains the current gold standard in clinical practice [99]. This may be explained considering that XCA images present numerous challenges that can lead to stenosis misdetection, including uneven illumination, low contrast, motion artifacts, shadowing, and the overlay of structures such as body tissues or catheters [100]. Additionally, the variability in stenosis (in terms of shape, pattern, and severity, differences in imaging equipment across hospitals, and varying levels of contrast agent depending on the patient's characteristics and clinician's evaluation [101] contribute to domain-shift issues, which should be addressed by algorithms for effective stenosis detection.

DL approaches have monopolized the attention of the medical-image-analysis community that works in the field, demonstrating promising potential [102, 103, 104, 105]. Recently, the interest in developing DL approaches for XCA analysis has been growing,

confirmed also by the recent release of a public dataset within the Arcade grand challenge [94]. This research challenge aims to provide a standardized benchmark for evaluating DL models in XCA, facilitating comparative assessments, and promoting the broader applicability of these models.

Despite these advancements, most literature relies on centralized, often private, datasets for training DL models [106]. This dependency on dataset-specific characteristics, such as imaging equipment, acquisition protocols, and patient variability, results in models that may not generalize well across different clinical settings. Consequently, these models are often difficult to compare and have limited broader applicability. Additionally, using centralized datasets poses issues relevant to privacy concerns [107], which limit data sharing and integration across different institutions.

Federated learning (FL) [108] is a promising approach for training models across multiple clinical institutions while preserving patient privacy and hospital governance. One of the primary challenges in developing models based on multicentric datasets is managing inter-dataset domain-shift issues, which further exacerbate the intrinsic complexity of XCA and its inherent intra-dataset shifts [109]. Strategies to partially mitigate domain shifts are grouped in federated domain generalization (FedDG) and personalized FL (PFL) [110, 111], which address the issues from complementary perspectives [112]. For our task, as shown by preliminary work in the literature [113, 114], PFL appears particularly promising. While FedDG tends to favor a generalizable global model at the expense of individual client performance [112], thus not focusing on ensuring accurate performance in stenosis detection for any dataset, PFL customizes the model to each client's unique data distribution, enhancing relevance to specific clinical environments and patient populations [110]. However, a PFL approach alone may not be sufficient to identify common features across various datasets and generalize well due to the high variability in multicentric data. This issue remains an open problem in the literature, with the development of domain adaptation approaches aimed at mitigating the large domain gap between different hospital datasets. Domain adaptation methods include invariant feature learning [115, 116, 117], pseudo-label based self-training [118, 119], image translation [120, 121, 122, 123], domain randomization [124], and mean-teacher training

[125, 126]. Many of these domain adaptation approaches, such as invariant feature learning and pseudo-label-based self-training, typically require access to the source and target datasets. This requirement poses a significant challenge in an FL setting where direct access to other datasets is not feasible due to privacy constraints.

One promising approach to address this limitation of PFL is histogram matching (HM) [127], a method that involves sharing and aligning histogram statistics across datasets to standardize the distribution of pixel intensities and reduce variability. As proposed in [128], this technique enables the alignment of feature distributions across different hospitals while still preserving patient privacy. This makes it suitable for integration within an FL framework, offering a practical solution for effective domain adaptation in a privacy-preserving manner.

During the training and testing phases, PFL faces challenges not only with inter-center variability, but also with domain shifts within the same dataset. To address this, we integrate a test-time adaptation (TTA) strategy can be integrated, allowing models to adapt dynamically to new, unseen data distributions during deployment without requiring ground-truth labels or full retraining [129]. TTA adjusts model weights during inference using unsupervised consistency losses computed on augmented views of each test image, thus operating on a per-image basis. This approach can be particularly valuable for improving the detection of stenosis in XCA images, effectively enhancing performance regardless of the inherent variability in these images. The detection of stenosis could benefit from TTA as it could help models remain robust despite these variations. Notably, no current research has proposed a TTA approach specifically designed for stenosis detection.

In this study, we present FedStenoNet, a novel PFL framework for stenosis detection in XCA images. Its main contributions are:

- FedStenoNet is the first PFL framework for stenosis detection in XCA that jointly addresses both inter- and intra- client variability across three real-world, non-identical and independently distributed (non-IID) datasets. The framework integrates HM as a privacy-preserving inter-client adaptation strategy to harmonize image appearance across institutions, and a novel TTA mechanism as intra-client adaptation to update client-specific weights and improve generalization during inference.

TABLE 8.1: Overview of key characteristics of each client dataset (A, B, and C), emphasizing their differences.

	Dataset A	**Dataset B**	**Dataset C**
Property	Released with this work [7]	Open source [94]	Available upon request [93]
Number of patients	245	1500	29
Mean age ± standard deviation, years	68.4 ± 7.5	45.8	60.3 ± 13.8
Gender distribution (male/female)	148/96	855/645	-/-
Number of images	1565	1500	2483
Number of stenosis per image	from 1 up to 4	from 1 up to 8	1
Scanner	Azurion Clarity IQ (Philips) / Allura Xer FD10 (Philips)	Azurion 3 (Philips) / Artis Zee (Siemens)	Coroscop (Siemens) / Innova (GE Healthcare)
Annotators	3	10	1
Frame selection	Multiple frames selected by high-contrast dye, varied viewpoints, diastolic phase of heart cycle, stenosis visibility	One frame selected by optimal contrast, minimal blurriness, stenosis visibility	Multiple sequential frames selected based on stenosis visibility
Stenosis definition			
	– Stenosis diameter $\geq 70\%$ – Stenosis diameter $\geq 60\%$ and lumen cross-sectional area $< 4.0mm^2$ – According to European Society of Cardiology Guidelines [130]	– Stenosis diameter $\geq 50\%$ and lumen thickness $\geq 1.5mm$ – According to Syntax Score [131]	– Stenosis diameter $\geq 70\%$ – Stenosis diameter $\geq 50\%$ and fractional flow reserve $\leq 80\%$ – According to 2017 US appropriate use criteria for coronary revascularization in patients with stable ischemic heart disease [132]

- A novel TTA module is introduced, specifically designed for stenosis detection. It enforces prediction consistency across augmented views of the same test image by optimizing an innovative adaptive loss that combines classification accuracy, bounding box stability, and feature-level invariance.
- We curated an extensive, uniform XCA dataset that will be publicly available upon publication, to support reproducible research and foster further advances in federated medical image analysis.

8.1 Methods

In this section, we begin with an overview of the datasets used, highlighting the challenges associated with the data (Sec. 8.1.1). This is followed by a detailed description of the

novel proposed methodology (Sec. 8.1.2, Sec. 8.1.3, Sec. 8.1.4) and the entire experimental setup, including implementation details (Sec. 8.1.5), comparison with the literature and ablation study (Sec. 8.1.6), and metrics considered to evaluate the experimental performance (Sec. 8.1.7).

8.1.1 Datasets description

To leverage PFL for XCA detection, we used three datasets: dataset A is curated by us for this work, dataset B was released within the Arcade grand challenge [94] and dataset C is also available upon request from [93].

These datasets, sourced from different medical centers using a variety of XCA scanners and different acquisition and annotation protocols, showcase a broad spectrum of patient demographics and XCA characteristics, illustrating the extensive real-world heterogeneity of XCA. Table 8.1 provides a detailed summary of each dataset.

8.1.1.1 Dataset A

The dataset A was acquired at Ospedali Riuniti (Ancona, Italy) using Azurion Clarity IQ (Philips) and Allura Xer FD10 (Philips) scanners and comprises 1565 XCA images from 244 patients, who underwent XCA procedures during the year 2020. These patients were diagnosed with "significant coronary stenosis" as indicated in their examination reports. According to European Society of Cardiology Guidelines [130], a coronary stenosis was considered "significant" if presenting an area $\geq 70\%$ measured using quantitative coronary angiography (QCA). When necessary, diagnostic confirmation was performed through intravascular US (IVUS) [133], based on which a stenosis is considered as significant if the minimum lumen area exceeds $4mm^2$. The images (varying per patient) were selected based on criteria such as high-contrast dye visibility, varied viewpoints, and the diastolic phase of the heart cycle. The data collection adhered to the Helsinki Declaration and was approved by the Local Ethics Committee (CET 59/2024), with all patients providing informed consent, and it was supervised by three expert clinicians who also provided annotations of the stenotic regions along the left coronary artery. For model training, the

dataset, partitioned based on patients involved, comprises 1106 images from 174 patients for the training set, 347 images from 45 patients for validation, and 112 images from 25 patients for the test set.

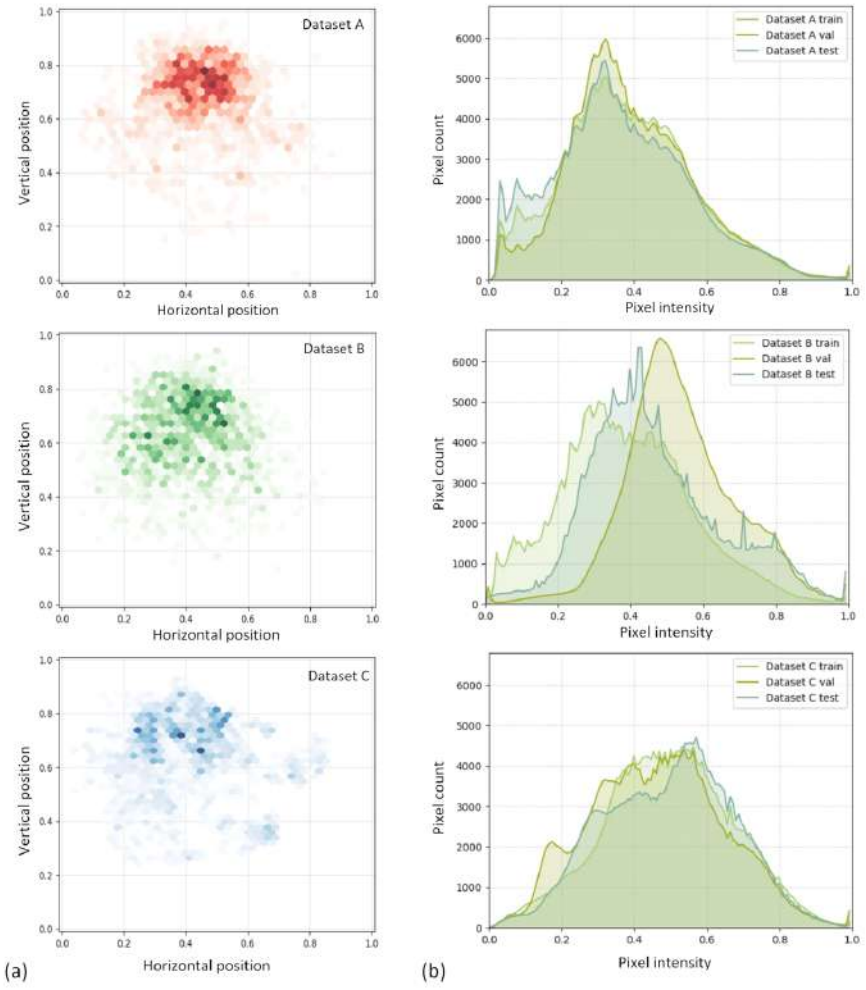
8.1.1.2 Dataset B

The dataset B was publicly released in [94], including one XCA image from each of 1500 patients acquired at the Research Institute of Cardiology and Internal Diseases (Almaty, Kazakhstan) using Azurion 3 (Philips) and Aris Zee (Siemens) angiographs. The dataset includes views of the left and right coronary arteries, and images were selected based on optimal contrast, minimal blurriness, target lesion visibility, and clinical relevance. Stenotic plaques were annotated with cross-validation by experienced cardiologists based on the SYNTAX Score definition [131], in which stenosis is defined as a coronary lesion with 50% and higher narrowing in vessels with a thickness of 1.5 mm or more. The data split into training, validation, and testing sets include 1000, 200, and 300 images in each respective partition.

8.1.1.3 Dataset C

The dataset C consists of a subset of 2483 XCA images from 30 patients, derived by Danilov et al. [93], which acquired one-vessel stenotic sequences from 100 patients using Coroscop (Siemens) and Innova (GE Healthcare) at the Research Institute for Complex Problems of Cardiovascular Diseases (Kemerovo, Russia). Manual annotation of the presence or absence of stenotic lesions for each image was performed by a single operator. All patients had angiographically and /or functionally confirmed one-vessel coronary artery disease and significant stenosis was defined as a stenosis with diameter $\geq 70\%$ by QCA or ≥ 50–69% with fractional flow reserve (FFR) $\leq 80\%$ or stress echocardiography evidence of regional ischemia, according to 2017 US appropriate use criteria for coronary revascularization in patients with stable ischemic heart disease [132]. Selecting a smaller subset from the whole dataset aims to explore the potential federation benefits for clients with limited patient data availability. This subset is further divided into training, validation, and testing sets of 1860 images from 24 patients, 295 images from 3 patients, and 328

images from 3 patients, respectively.



(a) Average intensity distribution of pixel values, and (b) average bounding box distribution](a) Average intensity distribution of pixel values across training, validation, and test subsets, and (b) average bounding box distribution, both shown for dataset A (top), dataset B (middle), and dataset C (bottom)

FIGURE 8.1: [

As shown in Fig. 8.1 a), the inherent heterogeneity among and within these datasets is also evident in the considerable variability of their intensity distributions: in particular, the average intensity distribution of dataset C exhibits a shift with the other datasets; in addition, dataset B shows internal variability, with a shift between training, validation, and test sets. The variability is not limited to intensity distributions, but it also extends to the distribution of bounding boxes delineating stenotic regions as identified by experts

within images, as illustrated in Fig. 8.1 b). Concerning how the annotation is performed, bounding box dimensions also vary from client to client, and, based on COCO standards [134], the number of small bounding boxes ($area < 32^2$) is 1140 for dataset A, 326 for dataset B, and 1626 for dataset C; the number of medium bounding boxes ($32^2 \leq area \leq 96^2$) is 761 for dataset A, 1712 for dataset B, and 858 for dataset C; and the number of large bounding boxes ($area > 96^2$) is 1 for dataset A, 379 for dataset B, and 0 for dataset C.

These characteristics of inherent heterogeneity and domain shift reflect the challenges prevalent in real-world clinical settings, where differences in acquisition protocols and annotation standards across institutions are common.

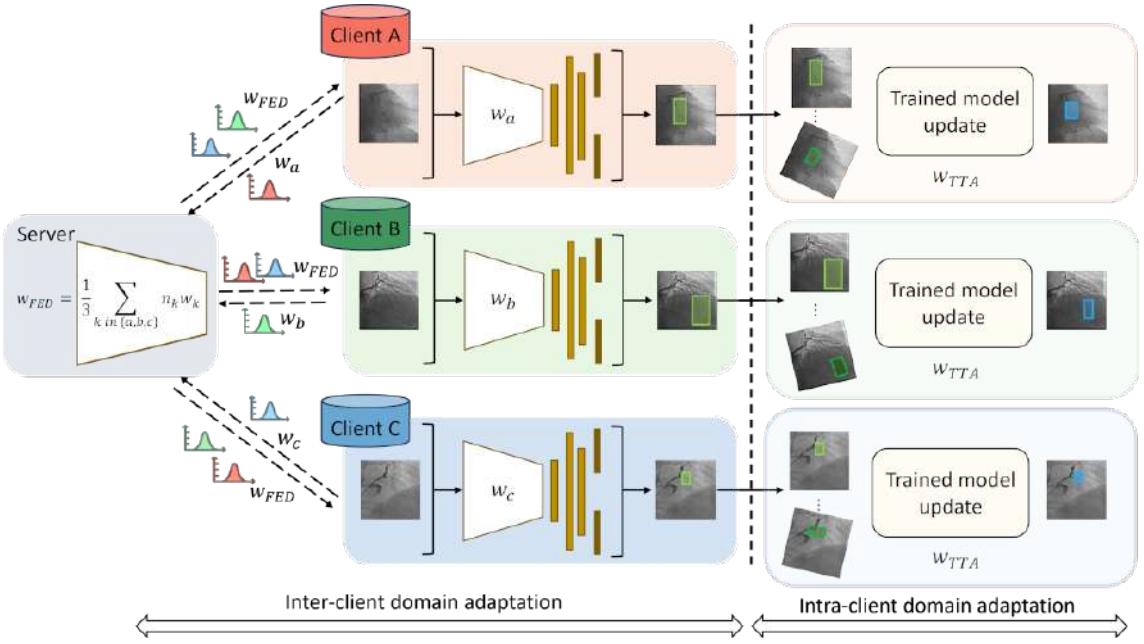


FIGURE 8.2: Overview of the FedStenoNet framework. The central server aggregates the backbone weights w_a, w_b, and w_c of a Faster R-CNN model from each client (A, B, C) and shares average intensity distributions among clients to perform inter-client domain adaptation using histogram matching (HM). The weights aggregation is performed as: $w_{FED} = \frac{1}{3} \sum_{k=A,B,C} w_k$. During test-time adaptation (TTA), the same test image at each client is transformed into multiple augmented views, and the local model generates predictions on each of these views. The predictions are then merged, with overlapping bounding boxes aggregated through non-maximum suppression (NMS), while updating the local weights (w_{TTA}) to produce a final refined prediction. The TTA module ensures that the final prediction is robust against variations introduced by augmentations, effectively handling intra-client domain shifts and improving model performance

8.1.2 Framework description

FedStenoNet, which is shown in Fig. 8.2, leverages a Faster R-CNN [135] for collaborative learning through a central server that aggregates weights of the Faster R-CNN backbone. The choice of using Faster R-CNN is due to the proven effectiveness of an R-CNN architecture with an FPN in detecting stenosis [93, 136, 137].

The aggregation strategy we used is FedProx [138], which extends the standard federated averaging (FedAvg) algorithm by adding a proximal term to address clients' heterogeneity. This is especially useful in scenarios where data distributions vary significantly across clients.

In the FedProx framework, the objective function for each client k is modified to include a proximal term as follows:

$$\min_{w} \left\{ F_k(w) = f_k(w) + \frac{\mu}{2} \|w - w_g\|^2 \right\} \tag{8.1}$$

where $f_k(w)$ is the local objective function for client k, μ is a positive constant that controls the strength of the proximal term, w is the model parameter, w_g is the global model parameter from the previous round.

The central server aggregates the updated weights from all clients using the following weighted average:

$$w_{g+1} = \frac{1}{n_s} \sum_{k=1}^{K} n_k w_k \tag{8.2}$$

where K is the total number of clients, n_k is the number of data samples on client k, n_s the total number of samples, and w_k is the updated model parameter from client k.

While FedProx introduces a mechanism to handle some degree of heterogeneity by penalizing large deviations from the global model, it cannot fully resolve the fundamental issue of non-IID data distributions in XCA, failing to mitigate inter-client data shifts. These shifts, often caused by variations in contrast and intensity across different datasets, significantly affect model performance.

8.1.3 Inter-client domain adaptation

To reduce the domain gap among clients, visible in the average intensity distribution plots shown in Fig. 8.1 a), we introduce HM as a data augmentation technique in the federated framework: HM has been recently used in domain adaptation to improve the generalization ability of convolutional neural networks by transferring the intensity distribution of the source dataset to the target dataset [127].

Given the average intensity distribution H_s of the training set of one client (the "source"), and an image x_{tg} from another client (the "target"), HM aims to shift the intensity levels of x_{tg} such that its intensity histogram H'_{tg} matches H_s. This procedure involves the computation of the cumulative distribution function CDF of the average source histogram CDF_s and target histogram CDF_{tg}, and a transformation function T that maps each intensity level i in x_{tg} to a new intensity level i' in the shifted image:

$$T(i) = \min\left\{i' : CDF_s(i') \geq CDF_{tg}(i)\right\}, \quad i \in [0, 255] \tag{8.3}$$

T is applied to each pixel in x_{tg} with intensity distribution h to obtain the image (x'_{tg}) with matched histogram:

$$x'_{tg}(w, h) = T(x_{tg}(w, h)), \quad \forall (w, h) \in x_{tg} \tag{8.4}$$

8.1.4 Intra-client domain adaptation

Each client has its unique data distribution and internal variability, which may result in discrepancies between training and testing sets, such as those visible among the subsets of dataset B in Fig. 8.3 a). Therefore, we design a TTA module for stenosis detection, inspired by a state-of-the-art approach [139].

The approach in [139] offers a lightweight TTA module that adapts models to distribution shifts using augmentation techniques, making it an excellent starting point for handling the variability in XCA images. However, while [139] is focused primarily on classification tasks, we adapt this module to tackle the unique challenges of object detection

in XCA imaging. Our approach innovatively combines multiple augmentations with three complementary loss functions, each acting on different levels to improve robustness and accuracy in detecting stenosis: specifically, global entropy loss for classification, bounding box consistency loss for localization, and feature alignment loss for semantic consistency. This multi-faceted approach ensures that the model not only adapts to new data distributions at test time but also produces a unique, robust prediction that is less sensitive to variations in the image content. Figure 8.2 provides a schematic representation of the designed TTA module.

For each client k, our TTA module optimizes the model f_θ^k, parameterized by its weights θ_k, which has been produced during the collaborative learning phase for k. To this end, TTA executes an additional training stage, during which for given client k and given an image x_0 in its test set, a set of three randomized geometric transformations including horizontal flip, rotation and translation, is applied to produce an *ensemble* of augmented views $\{x_0, x_1, x_2, x_3\}$.

At each iteration of TTA training, the union of bounding boxes returned by f_θ^k for each view is processed through non-maximum suppression (NMS) to isolate single predictions. Nevertheless and unless differently specified, all bounding boxes, no matter how overlapping the maximally scoring ones, are retained in the following computations, to avoid any information drop.

For each prediction $p = 0 \dots n$ (where n is the total number of predictions), we denote the set of bounding boxes and corresponding confidence scores as $y_p = f_\theta^k(x_k) = \{(bbox_l^k, s_l^k)\}_{l=0}^{b_k-1}$. Here, $bbox_l^k$ represents the l-th bounding box for client k and s_l^k is the associated confidence score.

After the NMS, we minimize a composite loss function across all predictions, defined as:

$$L_{TTA} = L_h + L_{bboxes} + L_{features} \tag{8.5}$$

In Eq. 8.5, L_h is the global entropy loss, adapted from [139] for object detection task replacing class scores with bounding box confidence scores. For each prediction p, we compute the entropy $\mathcal{H}$ of the sum of the maximum score from each view contributing to

the prediction, and then sum up for all p:

$$L_h = -\sum_{p=0}^{n-1} \mathcal{H}(\sum_{v=0}^{3} s^k_{J(p,v)}) \tag{8.6}$$

where $J(p,v) = \arg\max_{j,V(p,j)=v}(s^p_j)$ is the index of the bounding box with the maximum confidence score in view v contributing to prediction p, s^p_j is the confidence score of the j-th bounding box in prediction p, and $V(p,j)$ denotes the view from which bounding box j in prediction p comes.

To make the loss more specific for object detection, we introduce a bounding box consistency loss L_{bboxes}, which measures the consistency of regressed bounding boxes. This is defined for each prediction p as the cumulative squared distance between the center of each bounding box in the prediction and the global centroid $\overline{C}_p$:

$$L_{bboxes} = \sum_{p=0}^{n-1} \sum_{l=0}^{b_p-1} d^2(C_{p,l}, \overline{C}_p) \tag{8.7}$$

where $C_{p,l}$ is the center of the l-th bounding box within prediction p, and $d^2(C_{p,l}, \overline{C}_p)$ is the Euclidean distance between $C_{p,l}$ and $\overline{C}_p$. $\overline{C}_p$ is computed as the mean of the centers of all bounding boxes in the prediction, weighted by confidence scores:

$$\overline{C}_p = \frac{\sum_{l=0}^{b_p-1} s^p_l C_{p,l}}{\sum_{l=0}^{b_p-1} s^p_l} \tag{8.8}$$

where s^p_l is the confidence score of the l-th bounding box within prediction p.

To further enforce the agreement of views from a semantic point of view, we introduce a feature alignment loss component $L_{features}$, accounting for the similarity of high-level feature maps among views. After extracting feature maps from the convolutional layers of the fourth block of the backbone in f^k_θ for each view, we downsample them cross-channel-wise, using global average pooling (GAP). $L_{features}$ is then defined as the sum of pairwise distances among corresponding feature maps of different views:

$$L_{features} = \sum_{v_0=0}^{|A|} \sum_{v_1=0, v_1 \neq v_0}^{|A|} d^2(\mathrm{GAP}(f^k_\theta(x_{v_0})), \mathrm{GAP}(f^k_\theta(x_{v_1}))) \tag{8.9}$$

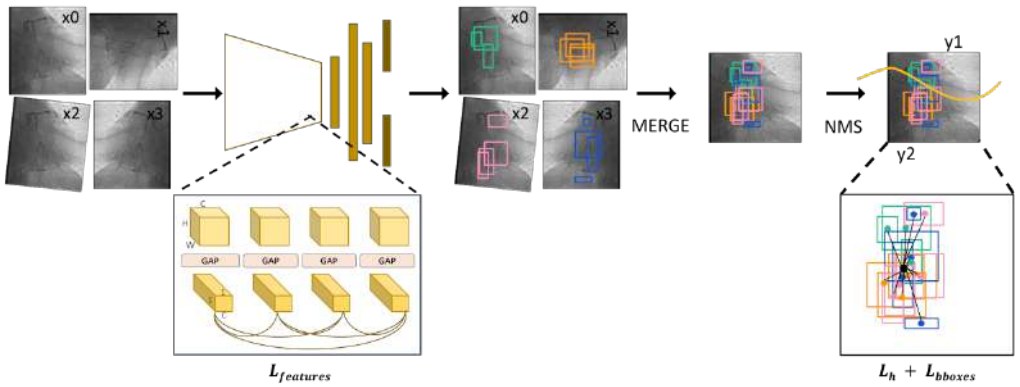


FIGURE 8.3: The image illustrates the test-time adaptation (TTA) algorithm for object detection in XCA images. Key steps include: creating multiple views x_0, x_1, x_2, x_3 from the original image x; processing each augmented image through the trained Faster R-CNN model to generate bounding box and class predictions with confidence scores for each view; merging and refining predictions from all views using non-maximum suppression (NMS); and ensuring consistency by aligning bounding boxes and feature maps through global average pooling (GAP). The three components of the TTA loss function L_{TTA} are depicted: global entropy loss L_h for classification, bounding box consistency loss L_{bboxes} for localization, and feature alignment loss $L_{features}$ for semantic consistency across views

where x_v is the image from view v with different indexes (v_0 and v_1), $|A|$ is the number of applied augmentations, d^2 is the Euclidean distance between the feature maps $f_\theta^k(x_v)$ extracted for client k and view v.

8.1.5 Implementation details

In all experiments, we used a Faster R-CNN with Resnet-50-FPN [140, 141] as the backbone, pre-trained on the COCO dataset [134], and used the Stochastic gradient descent optimizer combined with cosine annealing warm restart scheduler, and a batch size of 16. The local model was trained to minimize a multitask loss, which includes a cross-entropy loss for stenosis detection and a Smooth L1 loss for the localization of stenosis bounding boxes. The FL framework performed 20 rounds, during which local training on the Faster R-CNN was performed for 10 epochs. A warm-up strategy extended the initial round by 5 epochs to stabilize weights. After the final round, an optional 10-epoch training phase allowed clients to autonomously adopt either the final aggregated model or their best-performing local model to provide flexibility for customized refinements. Image pre-processing consisted of normalization to [0,1] and random augmentations including brightness and contrast alterations (both in the range [0.7, 13]), rotation (in the range

Algorithm 2: Pseudocode for the TTA algorithm.

Input : number of epochs E, test sample x, set of augmentations A, model f_{θ_1}
Output: optimized model f_{θ_E}, optimized predictions $\{y_k\}_{k=0}^{n-1} = f_{\theta_E}(x)$

$\{\hat{x}^i\}_{i=0}^{|A|} \leftarrow A(x)$

while $epo \leftarrow 1$ **to** E **do**

  $\{(bbox_l, s_l)\}_{l=0}^{N_k-1} \leftarrow \text{NMS}(\{f_{\theta_{epo}}(\hat{x}^i)\}_{i=0}^{|A|})$

  $J(k, v) \leftarrow \arg\max_{j, V(k,j)=v}(s_j^k)$

  $L_h \leftarrow -\sum_{k=0}^{n-1} \mathcal{H}(\sum_{v=0}^{|A|} s_{J(k,v)}^k)$

  $L_{bboxes} \leftarrow \sum_{k=0}^{n-1} \sum_{l=0}^{N_k-1} d^2(C_{k,l}, \overline{C}_k)$

  $L_{features} \leftarrow \sum_{v_0=0}^{|A|} \sum_{v_1=0, v_1 \neq v_0}^{|A|} d^2(\text{GAP}(f_\theta^k(x_{v_0})), \text{GAP}(f_\theta^k(x_{v_1})))$

  $L_{TTA} \leftarrow L_h + L_{bboxes} + L_{features}$

  $grad \leftarrow \text{compute\_grad}(L_{TTA}, f_{\theta_{epo}})$

  $\text{apply\_grad}(grad,\ f_{\theta_{epo}})$

end

return f_{θ_E}, $\{y_k\}_{k=0}^{n-1}$

[-10°, 10°]), translation (in the range [-20,20] pixels), a linear mix-up (alpha=0.5) and HM (with a probability of 0.5). Horizontal flip, rotation, and translation augmentations (as defined during training) were applied for TTA allowing the model to not rely on different intensity distributions that arise in client B and client C, acting as a regularization term. Notably, we opted for a hold-out validation setup over cross-validation to assess model performance under realistic inter- and intra-client shift conditions, better simulating real clinical scenarios. In particular, Dataset B follows the predefined split released as part of the Arcade Grand Challenge [94], enabling direct comparison with existing state-of-the-art methods. Datasets A and C were split at the patient level to prevent data leakage and their splits were specifically designed to suit this study, aiming to comprehensively analyze and manage intra- and inter- variability within datasets.

Hyperparameter selection was done through an exhaustive grid search for optimal performance (see Table 8.2). The framework was implemented in Python 3.8.10, with PyTorch v.2.0.0 and Torchvision v.0.15.1, utilizing 8 NVIDIA A100 GPUs with 64GB

VRAM and 512GB RAM.

TABLE 8.2: Summary of hyperparameters used in the experiments tuned through grid search.

Component	Value / Setting
Backbone architecture	Faster R-CNN with ResNet-50-FPN (pretrained on COCO)
Optimizer	Stochastic Gradient Descent (SGD)
Learning rate scheduler	Cosine annealing with warm restarts
Batch size	16
Loss function	Cross-entropy + Smooth L1
Federated rounds	20
Local epochs per round	10 (and 5 Warm-up epochs for round 1 only)
Post-training refinement	Optional 10 epochs per client
Data augmentation	Brightness [0.7, 1.3]; Contrast [0.7, 1.3]; rotation [-10°, 10°]; translation [-20, 20] px; mix-up (alpha=0.5); histogram matching (p=0.5); horizontal flip
Test-time augmentation	horizontal flip; rotation [-10°, 10°]; translation [-20, 20] px
Test-time loss function	$L_h + L_{bboxes} + L_{features}$ (see Eq. 8.5)

8.1.6 Ablation study

As a first ablation study, to investigate the impact of HM two strategies are compared: a dynamic strategy (HM_rand) and the proposed approach (HM_avg), which is used in FedStenoNet. In the HM_rand strategy, the intensity distribution of a client's input image is matched to the intensity distribution of a random image from another client. Images are randomly sampled with a probability of 0.5. This comparison allowed us to determine if the dynamic approach could yield better results than the average-based method. The results of this study provide insights into the potential benefits of adopting a more flexible HM strategy in FL contexts.

To assess the effectiveness of inter-client personalization in enhancing individual client performance, we also conducted an ablation study on the TTA component of the FedStenoNet framework. Specifically, we examined the influence of the combined strategy $L_{TTA} = L_h + L_{bboxes} + L_{features}$, analyzing the effects of different TTA loss functions, including L_h only (TTA1) and $L_h + L_{bboxes}$ (TTA2). We also tested L_{bbox} and $L_{features}$ individually but found they did not contribute any improvement nor provide additional insights. A comparison of these configurations was conducted to identify the most effective approach for leveraging TTA to enhance model performance during the testing phase.

TABLE 8.3: Augmentation techniques used for TTA views include horizontal flipping, mean histograms derived from the two other datasets (HM_avg), jittering which involves four variations in brightness and contrast, geometric transformations consisting of two rotations ($-12 \leq \alpha < 0$, $0 < \beta \leq 12$) and four translations ($x, y > 0$, $x, y < 0$, $x > 0, y > 0$, $x < 0, y > 0$).

	flipping	HM_avg	jittering	rotation	translation
E1 (proposed)	●	❍	❍	●	●
E2	●	●	●	❍	❍
E3	●	❍	●	●	●
E4	●	●	●	●	●
E5	●	●	❍	❍	❍
E6	●	●	❍	●	●
E7	●	❍	●	❍	❍

The augmentation techniques used for TTA views were also systematically evaluated by experimenting with different types of transformations and various combinations of them, including horizontal flipping; mean histograms (HM_avg) derived from two other datasets, which could help to standardize the intensity distributions across images; jittering, which consisted of four variations of brightness and contrast; and affine transformations, which included rotations within the ranges [-12°, 0°] and [0°, 12°], and four types of translations, combining $x, y > 0$, $x, y < 0$, $x > 0 \& y < 0$ and $x > 0 \& y < 0$ in the range [-20, 20] pixels. Each experiment applied a distinct set of augmentation techniques to systematically evaluate the impact of each one (both individually and in combination with others) on model performance across the datasets. The main experiments and their configurations, defined as E1 - E7, are briefly summarized in Table 8.3. In addition to understanding the contributions of each augmentation technique, this ablation study aims to guide the selection of optimal augmentation strategies to improve the model's performance and generalization capabilities under varying conditions.

Regarding the comparison with state of the art, despite recent literature addressing stenosis detection on XCA images [102, 105, 103, 104], a quantitative comparison with these works would be unfair due to their use of centralized data for training, which is out of the scope of this study. We benchmarked FedStenoNet against two baseline FL methodologies, FedProx [138] and FedAvg [91], used in the preliminary approaches in this field [113, 114]. Moreover, in line with common practices in FL literature, we trained a centralized model on the union of all client datasets to provide an upper-bound performance reference. The centralized model used the same architecture (i.e., Faster R-CNN) and training configuration as FedStenoNet. However, no domain adaptation techniques

(e.g., HM or TTA) were applied, as these are tailored to the federated scenario and specifically designed to address inter- and intra-client domain shifts. For a fair comparison, all experiments (both federated and centralized) were conducted using the same dataset splits, training settings, and computational hardware.

8.1.7 Performance metrics

To evaluate the performance of the proposed method, we computed the Precision ($Prec$), Recall (Rec), and F1 score ($F1$). To determine the minimum overlap required between the predicted bounding box and the ground truth for a positive detection, we set the Intersection over the Union (IoU) threshold to 0.5, ensuring a balance between $Prec$ and Rec. $Prec$ measures the accuracy of positive predictions, defined as the ratio of true positive (TP) predictions to the total number of positive predictions, which includes both TPs and false positives (FPs):

$$Prec = TP/(TP + FP) \tag{8.10}$$

Rec assesses model's ability to identify all relevant instances, calculated as the ratio of TP predictions to the total number of actual positives, which includes both TPs and false negatives (FNs):

$$Rec = TP/(TP + FN) \tag{8.11}$$

$F1$ is the harmonic mean of $Prec$ and Rec, providing a single metric that balances both:

$$F1 = 2 \cdot \frac{Prec \cdot Rec}{Prec + Rec} \tag{8.12}$$

TABLE 8.4: Comparison of centralized approach and FedStenoNet in terms of Precision ($Prec$), Recall (Rec), and F1 score ($F1$) for clients A, B, and C and on average (AVG).

	Prec				*Rec*				*F1*			
	Client A	Client B	Client C	AVG	Client A	Client B	Client C	AVG	Client A	Client B	Client C	AVG
Centralized	55.77	30.67	**45.39**	43.94	**69.05**	44.82	**42.07**	**51.98**	61.70	36.42	43.67	47.26
FedStenoNet	**64.00**	**46.44**	40.00	**50.15**	63.49	**52.33**	39.02	51.61	**63.75**	**49.21**	**39.51**	**50.82**

TABLE 8.5: Comparison of FedStenoNet with standard FL approaches (FedAvg and FedProx) and ablation studies: Histogram Matching (HM) variants applied to the FedProx baseline, including dynamic HM, HM_rand, and the proposed HM_avg; 3) TTA module variations using different loss functions: L_h only (TTA1) and combined $L_h + L_{bboxes}$ (TTA2). Performance is reported in terms of average *Prec*, *Rec*, and $F1$ for clients A, B, and C and on AVG.

	Prec				*Rec*				$F1$			
	Client A	Client B	Client C	AVG	Client A	Client B	Client C	AVG	Client A	Client B	Client C	AVG
FedAvg	**65.35**	41.14	28.14	44.88	65.87	50.52	25.30	47.23	**65.61**	45.35	26.65	45.87
FedProx	64.12	42.73	35.95	47.60	**66.67**	49.48	36.28	50.81	65.37	45.86	36.12	49.12
HM_rand	64.96	45.80	30.10	46.95	60.32	46.63	28.35	45.10	62.55	46.21	29.20	45.99
HM_avg	64.29	43.38	38.51	48.73	64.29	51.81	39.02	51.81	64.29	47.23	38.91	50.14
TTA1	64.00	44.39	40.69	49.71	57.94	50.26	35.98	48.06	60.83	47.14	38.19	48.72
TTA2	62.39	44.74	**41.92**	49.68	57.94	51.81	37.20	48.98	60.08	48.02	39.42	49.17
FedStenoNet	64.00	**46.44**	40.00	**50.15**	63.49	**52.33**	**39.33**	51.61	63.75	**49.21**	**39.51**	**50.82**

TABLE 8.6: Results of FedStenoNet performances in terms of average *Prec*, *Rec* and $F1$ for clients A, B, and C, and on AVG across multiple experiments on augmentation techniques for TTA views.

	Prec				*Rec*				$F1$			
	Client A	Client B	Client C	AVG	Client A	Client B	Client C	AVG	Client A	Client B	Client C	AVG
E1, proposed	64.00	**46.44**	40.00	50.15	63.49	52.33	39.02	**51.61**	63.75	**49.21**	39.51	**50.82**
E2	63.71	45.27	40.43	49.80	62.70	50.78	**39.94**	51.14	63.20	47.86	**40.18**	50.41
E3	65.32	45.13	39.31	49.92	**64.29**	49.22	38.11	50.54	64.08	47.09	38.72	50.19
E4	64.52	45.61	39.38	49.83	63.49	48.45	39.02	50.32	64.01	46.98	39.20	50.06
E5	**65.55**	42.48	40.45	49.49	61.90	50.52	38.72	50.38	63.67	46.15	39.56	49.79
E6	65.04	45.05	**41.75**	**50.61**	63.49	49.48	37.80	50.26	**64.26**	47.16	39.68	50.36
E7	54.55	45.54	37.69	45.92	50.01	**52.85**	37.80	46.88	52.17	48.92	37.75	46.28

8.2 Results

Tables 9.3 and 9.2 present the performance of FedStenoNet along with results from various experiments. Specifically, Table 9.3 shows the comparison between the proposed FedStenoNet and the centralized model, while Table 9.2 includes comparisons with standard FL approaches (FedAvg and FedProx), HM integrated into a FedProx baseline (as HM_rand and as HM_avg), and TTA strategies (TTA1 and TTA2) integrated into the FedProx + HM_avg pipeline.

FedStenoNet demonstrated the best overall performance in almost all metrics. It achieved the highest average $F1$ of 50.82%, with individual scores of 63.75% for client A, 49.21% for client B, and 39.51% for client C. Additionally, it achieved the highest average *Prec* of 50.15% with scores of 64.00% for client A, 46.44% for Client B and 40.00% for client C. The average *Rec* is also notable achieving 51.61%, with scores of 63.49% for client A, 52.33% for client B, and 39.02% for client C.

In comparison, other models showed varied performance across different datasets. The centralized model achieved the highest average *Rec* at 51.98% but with a pronounced disparity among clients (69.05% for client A, 44.82% for client B, and 42.07% for client C). The same uneven performance is also evident in the other metrics (particularly for Dataset C), despite the overall *Prec* and $F1$ averaging 43.94% and 47.26%, respectively. FedAvg achieved the best $F1$ performance on client A with a score of 65.61% and performed well also on client B (F1=45.35%), but it underperformed on client C with an $F1$ of 26.65%, resulting in an average $F1$ of 45.87% and similar trends for the other metrics. FedProx had an average $F1$ of 49.12%, increasing performance on client C of +10.53% in terms of $F1$, demonstrating the importance of an additional regularizing term in the training process to uniform the performance across clients.

The HM_rand and HM_avg models exhibited interesting contrasts in their performance.
HM_rand achieved an average $F1$ of 45.99%, with scores of 62.55% for client A, 46.21% for client B, and 29.20% for client C. Its average *Prec* was 46.95%, with 64.96 for client A, 45.80% for client B, and 30.10% for client C. The average *Rec* for HM_rand was 45.10%, with individual *Rec* of 60.32% for client A, 46.63% for client B, and 28.35% for client C. HM_avg showed better results than HM_rand with an average $F1$ of 50.14%, with scores of 64.29% for client A, 47.23% for client B, and 38.91% for client C. The average *Prec* for HM_avg was 48.73%, with individual scores of 64.29% for client A, 43.48% for client B, and 38.51% for client C. Its average *Rec* was notably higher at 51.81%, almost equal to the average *Rec* achieved by the centralized model (-0.17%), with values of 64.29 for client A, 51.81% for client B, and 39.33% for client C. HM_avg improvements on client C indicate that it effectively mitigates some domain shift issues, but it still falls short of FedStenoNet's balanced performance across all datasets.

Regarding TTA, the inclusion of different loss components in FedStenoNet targets various aspects of the learning problem. As evident from Table 9.3, the use of $L_{features}$, in addition to L_h and L_{bboxes}, is essential for mitigating intra-dataset shifts. For instance, relying solely on L_h, TTA1 resulted in an average $F1$ of 48.72%, with individual scores of 64.83% for client A, 47.14% for client B, and 38.19% for client C. By incorporating L_{bboxes} with L_h in TTA2, the average $F1$ slightly improved to 49.17%. This adjustment

also led to improved performance for clients experiencing significant shifts, with client B score increasing from 47.14% to 48.02% and client C score rising from 38.19% to 39.42%. When including $L_{features}$, consistency and more robust performance are visible for each client.

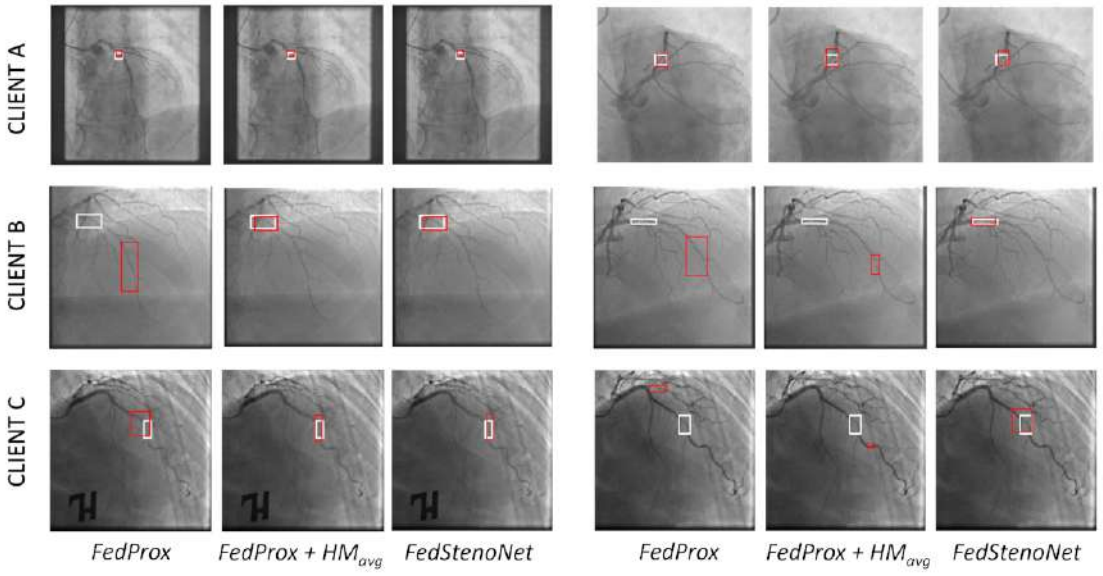


FIGURE 8.4: Visual samples of the predictions (red boxes) obtained with FedProx (1st and 4th columns), FedProx + HM_avg (2nd and 5th columns), and FedStenoNet (3rd and 6th columns) for two test images of client A, client B, and client C. Ground truth annotations (white boxes) are reported, too

The effectiveness of the TTA module is also analyzed by experimenting with different augmentation techniques, as described in Sec 8.1.6. From the comparative analysis summarized in Table 8.6, experiment E1, which combines horizontal flipping, translation, and rotation transformations, achieved the best overall performance with an average $F1$ of 50.82%, significantly enhancing the performance of client B ($F1 = 49.21\%$), which is the dataset with the highest variability among train, validation, and test subsets. In comparison, all other experiments (E2 to E7) showed less homogeneous performance across all clients, leading to imbalanced results specific to certain datasets. For instance, E4 demonstrated the highest $Prec$ on client B (49.49%) but did not perform consistently well on other metrics or clients. This indicates that while some augmentations can enhance performance on specific datasets, they may not generalize well across all datasets, underscoring the need for a balanced approach in augmentation techniques.

In addition, inference time was measured on a single GPU (NVIDIA A100, 64GB VRAM, 512GB RAM) and the average inference time per image was 0.71 s without TTA and 10.26 s with TTA, approximately increasing the time by a factor of 14.

Figure 8.4 displays visual samples of the predictions (in red) obtained with FedProx, HM_avg, and FedStenoNet. While performance is consistently strong for client A, improvements brought by HM_avg and HM_avg + TTA (i.e, FedStenoNet) are more evident for clients B and C, which present greater challenges due to internal distribution shift (client B) and the combined effect of limited data and inter-client shift (client C).

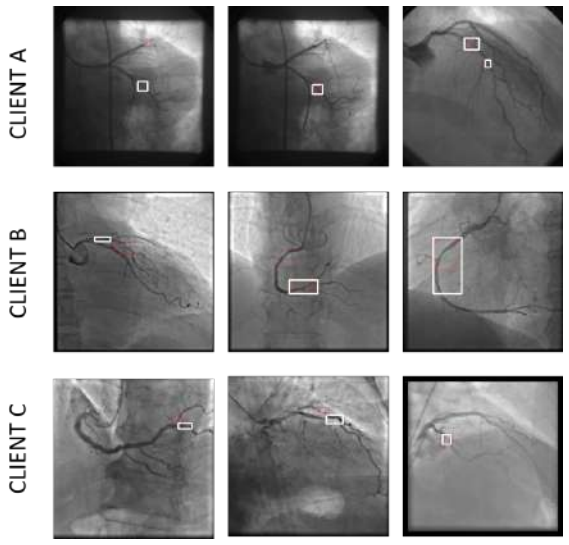


FIGURE 8.5: Visual samples of FedStenoNet predictions (red boxes) obtained for client A, client B, and client C. Ground truth annotations (white boxes) are reported, too

Figure 8.5 presents additional examples of FedStenoNet predictions, illustrating the model's strengths in clear cases and its limitations in more complex scenarios. The challenging cases reflect issues such as varying contrast levels, overlapping anatomical structures, and inconsistent annotation practices, which can compromise accurate detection.

8.3 Statistical analysis

To assess the significance of the performance differences among methods (significance level at 0.05), we conducted a Friedman test followed by post-hoc pairwise comparisons

using multiple testing corrections. The analysis was performed across the datasets of the three clients. As expected, no statistically significant differences were observed on dataset A, which is more homogeneous and less impacted by domain adaptation components. In contrast, datasets B and C, characterized by higher heterogeneity and inter- and intra- client variability, showed improvements in both $F1$ and *sensitivity* for our proposed method. In particular, Dataset B showed significant improvements for FedStenoNet compared to the centralized approach in both $F1$ (p=0.00001) and *sensitivity* (p=0.0064). Similarly, Dataset C exhibited significant gains over FedAvg with p=0.00001 for both $F1$ and *sensitivity*, and also surpassed FedProx+HM_rand with p=0.0001 across both metrics. These results confirm the added value of FedStenoNet, particularly in more challenging and variable contexts such as datasets B and C.

8.4 Discussion

Stenosis detection in XCA is a critical task that requires precision and reliability. By leveraging the collective knowledge from diverse datasets, FL models can play a pivotal role in this domain, enabling the use of data from multiple institutions to achieve robust and generalizable performance, without compromising patient privacy. However, the success of FL in this context is often challenged by domain shift issues, which can occur both inter- and intra-client. Inter-client domain shifts arise when data distributions differ significantly among institutions due to variations in imaging equipment, protocols, and patient demographics [101]. Intra-client domain shifts, on the other hand, occur within the data from a single institution over time, which are often due to changes in imaging procedures, equipment updates, and even patient population variations. In stenosis detection, the lack of standardization in imaging practices exacerbates these domain shift problems. Different institutions may use various imaging techniques, contrast agents, and resolution settings, leading to inconsistencies in the data [100]. Additionally, even within the same institution, differences in technician skills and procedural adjustments can introduce variability in the imaging data. Addressing these shifts is crucial to ensure the model's adaptability and effectiveness across different environments and conditions.

FedStenoNet aims to tackle these issues by proposing a simple, direct, and efficient HM technique to better homogenize the feature distribution across clients and a comprehensive TTA strategy to boost the feature representation further, regardless of the intra-dataset variability.

As shown in Table 9.3, the centralized approach struggles to generalize across diverse datasets, as it attempts to fit all data simultaneously, whereas FedStenoNet benefits from the aggregated learning of common features through backbone weights sharing and local specialization. Despite access to all data, the centralized model shows no clear performance advantage, while FL offers a compelling trade-off by achieving competitive performance while respecting privacy and data governance. In PFL, Table 9.2 highlights that FedStenoNet surpasses standard FL approaches like FedAvg and FedProx, resulting in only moderate improvements in $F1$ (45.87% and 45.90% on average, respectively). Specifically, FedAvg favors client A, which has the least heterogeneous dataset (according to Fig. 8.3), leading to imbalanced performance due to the absence of a regularization term, which FedProx better addresses in line with [138].

Although FedStenoNet average F1 score (50.01%) may appear modest, it aligns with the current state of the art for XCA stenosis detection, where variability in imaging protocols and annotations still limits higher performance. From a clinical perspective, the acceptability depends on the use case: in low-resource or remote settings, such results may already support screening or prioritization, while in specialized contexts, a human-in-the-loop approach could allow clinicians to tailor performance thresholds to the required accuracy. Therefore, performance evaluation should be contextualized to the clinical scenario and expected level of autonomy. Although FedStenoNet average F1 score (50.01%) may appear modest, it aligns with the current state of the art for XCA stenosis detection [94, 113, 114, 93, 142, 143], where variability in imaging protocols and annotations still limits higher performance. From a clinical perspective, performance acceptability depends on context and use-case [144, 145]: in low-resource or remote settings, such results could already provide valuable support for screening, triage, or case prioritization; in specialized environments, instead, higher accuracy might be needed, particularly if the model is intended to assist in training junior clinicians or to serve as a second-opinion decision-support tool.

Despite these limitations, overall our proposed approach shows significant improvements over traditional methods, suggesting that further refinements and the inclusion of more diverse datasets could enhance performance in future studies.

Table 9.2 also shows the results of the ablation study conducted to explore these components and demonstrates how its ability to harmonize training intensity distributions and adapt during testing provides a significant advantage over traditional methods. To harmonize intensity distributions among clients, HM is applied to a FedProx baseline, and different configurations of this technique, HM_rand and HM_avg, are explored to evaluate its effectiveness. HM_avg achieved a higher average $F1$ (50.14%) compared to HM_rand (45.99%), and a significant improvement from FedProx baseline only of +1.21 and +9.71 for client B and client C, respectively. This suggests HM_avg effectively mitigates domain shift among different clients, especially benefiting client C, which has the smallest dataset in terms of number of patients and the most evident intensity distribution shift. In contrast, HM_rand did not lead to similar improvements: randomly selecting images from other clients may introduce a degree of random variation or noise, making it harder for the model to learn generalizable representations. For very different data, like that from client C ($F1 = 31.34\%$), it may not be sufficient to bridge their differences. Hence, HM_avg might harmonize the central tendencies of the data distributions better than HM_rand, dealing more effectively with variations among clients.

To address intra-client distribution shift, we adopted a TTA strategy, whose effectiveness relies on the adaptive loss function optimized at test time. Each loss function targets different learning aspects, proving to be essential for handling complex shifts as demonstrated in Table 9.2. The primary loss function, L_h, focuses on classification accuracy, but it is insufficient alone for managing complex shifts in the input and feature spaces, as seen in TTA1 results for client B and client C. The combination $L_h + L_{bboxes}$ in TTA2 slightly improved the results in clients with evident shifts, indicating the benefit of a diversified loss strategy. However, TTA1 and TTA2 are not beneficial for client A, which presents no significant shift, suggesting potential overfitting if the training data already aligns well with the test distribution. Including $L_{features}$ in FedStenoNet is crucial for learning invariant features across different views while stabilizing the performance across different clients. This approach allows FedStenoNet to boost performances for test-shifted

clients without sacrificing the performance of client A. Nonetheless, the complexity of the loss components (e.g., quadratic terms), together with the iterative optimization at test time, introduces a notable limitation in terms of computational cost. From the inference time analysis (average 0.07 s without TTA vs. 9.35 s with TTA per image), this overhead may hinder real-time applicability and should be addressed in future work.

Finally, regarding the design of TTA views, as shown in Table 8.6, the combination of geometric augmentations (horizontal flipping, translation, and rotation) proved more effective than alternative augmentation strategies. This can be attributed to the fact that variability and shifts in terms of intensity are already extensively addressed during the federated training phase using HM, which normalizing the intensity distributions across different datasets reduced intensity-related domain shifts. Consequently, at test time, geometric transformations become crucial as they help the model achieve invariance to orientation and location, essential for accurately detecting stenosis across varying imaging conditions. Figure 8.4 underlines that incorporating diverse augmentations at both training and testing levels, as proposed in FedStenoNet, can mitigate intra- and inter-dataset shifts. The visual samples in Fig. 8.4 show that FedStenoNet's predictions are more closely aligned with the ground truth compared to FedProx and FedProx + HM_avg, suggesting its ability to effectively identify actual stenosis cases while minimizing FP predictions under varying imaging conditions and data distributions.

The importance of augmentation techniques as regularizers is evident in this study, consistent with the literature, where methods such as MixUp, Style Transfer, RandAugment, and others have been shown to enhance model robustness and accuracy by simulating real-world variations [146, 147]. This prevents the model from overfitting to specific features, promoting generalization. The lack of a unique standardized acquisition and evaluation protocol for stenosis detection further emphasizes the need for robust augmentation strategies. Each hospital or clinician may follow different procedures, resulting in discrepancies that might impact the model's performance. This real clinical scenario is reflected in the three clients and datasets used in this study. As shown in Fig. 8.5, there is a notable degree of variability in the images in terms of object shape and size, and due to the presence of interfering elements, such as catheters and overlapping structures, underscoring the challenge of generalizing across such diverse datasets. These discrepancies,

originating from acquisition procedures, stenosis definition protocols, and image selection and annotation processes, can lead to misleading FedStenoNet predictions. For example, in Fig. 8.5, it is shown that in client A similar frames with varying levels of contrast agent are chosen; while this variation may not be relevant to an expert for identifying a stenosis, an appropriate contrast level is critical for the algorithm to make accurate predictions. Additionally, the annotations can differ significantly across clients, also in terms of size. Client B, especially, often has larger annotations, as seen in Sec. 8.1.1 and in Fig. 8.1b. In complex images where overlapping structures are prominent, or in the presence of distracting objects such as catheters or sutures, stenosis detection is further complicated, but the model still tends to predict shapes that resemble actual stenoses. These limitations highlight the importance of developing models that can robustly handle ambiguity and noise in real clinical data, possibly incorporating uncertainty estimation mechanisms in future work. It is also important to note that XCA images represent 2D projections of a very complex 3D vascular structure; thus, a particular image might show the overlap of two vessels that, in reality, do not intersect. These distinctions only become apparent when analyzing the entire sequence and applying prior clinical knowledge. The 2D nature of the input also imposes structural limitations, as it may lead to ambiguous interpretations that could be addressed in the future by integrating temporal or 3D spatial information. Overall, while there is room for improvement and further refinement, the network consistently performs well when a stenosis is visible, accurately centering its predictions on the stenosis regardless of the dataset, and achieving a comparable level of accuracy across all three clients.

FedStenoNet is a PFL framework designed to specialize in local datasets, leveraging a central server to aggregate Faster R-CNN backbone weights and share clients' training average intensity distributions. This federated approach facilitates effective personalization during both the training and testing phases, thereby leading to improved performance while maintaining privacy and data governance. Notably, there have been very few studies in this domain that use an FL setup [113, 114], highlighting the novelty and potential impact of our work. Moreover, FedStenoNet not only respects privacy and data governance but also takes into account the real-world XCA variability and the risk of developing models biased towards specific dataset characteristics. The development of

DL algorithms on datasets acquired through single acquisition and annotation protocols, from single sources or specific scanners, may fail to generalize to different clinical settings [148, 149]. Such models are prone to biases that can lead to misdiagnoses and unequal treatment. FedStenoNet aims to mitigate these risks by incorporating diverse data and robust strategies to handle domain shifts, ensuring more fair, accurate, and generalizable stenosis prediction across varied clinical settings. Nonetheless, performance can still be affected by protocol differences, intra-patient variability, ambiguous 2D projections, and annotation inconsistencies, common challenges in XCA analysis that may limit generalizability.

To support future research and development in this field, we are making our dataset publicly available. This resource will facilitate the evaluation and improvement of DL-based tools for CAD diagnosis, fostering further advancements in the field. In the future, FedStenoNet should be validated on a larger and more diverse set of data institutions to ensure its generalizability and robustness further. Additionally, while our approach addresses intensity and geometric variations, other sources of variability may not be fully accounted for. Future work could explore more sophisticated domain adaptation techniques and augmentation methods to address these additional factors.

8.5 Conclusion

In conclusion, FedStenoNet tackles the challenge of domain shifts across multi-institutional XCA images, demonstrating promising results that, while confirming the complexity of direct clinical deployment, provide a foundation for further research. This work advances the applicability of DL for stenosis detection in XCA imaging, and potentially the detection of other anatomical structures in different imaging modalities [150, 151], by proposing a methodological approach that addresses multi-centric dataset variability. Moreover, by releasing a new dataset to the scientific community, we hope this work encourages further research in this field, aimed at enhancing diagnostic accuracy and fostering more equitable and effective patient care.

Its adaptable nature could enhance the detection of other anatomical structures in different imaging modalities [150, 151], offering a versatile tool for medical image analysis. We hope this work fosters future research where DL can be more readily applied to medical imaging, overcoming the challenges posed by multi-centric dataset variability. Such advancements will not only improve diagnostic accuracy but also contribute to more equitable and effective patient care.

Chapter 9

High-Risk AI Case Study II: Fetal Plane Classification of Noisy US images

Fetal standard planes, such as the abdominal, brain, and cardiac views, are key in prenatal Ultrasound (US) screening for measuring fetal biometric parameters and identifying abnormalities like growth restrictions and congenital anomalies [152, 153]. Traditionally, locating these planes has relied heavily on the skill of experienced sonographers. However, the subjective nature of manual interpretation can affect the precise identification of these planes, driving the need for deep learning (DL) methods to streamline the automatic identification of these planes and improve the accuracy of biometric measurements [154].

While DL models have shown considerable promise in automating the detection of standard planes, their implementation in clinical practice remains limited. This can be explained by the limited representativeness of training datasets, which often fail to capture the full variability encountered in real-world clinical settings [155]. Most datasets used for training come from single-center studies or involve subjects from specific demographic groups [156, 152, 157], limiting DL model ability to generalize to diverse populations. This can lead to biased predictions and potential disparities in clinical outcomes [158]. To address this limitation, there has been growing interest in multicenter studies, such

as those by [159], [160], and [158], which, however, may raise privacy concerns due to the integration of data from multiple centers. Federated Learning (FL) has emerged as a promising solution to mitigate this issue, enabling collaborative model training across multiple institutions without sharing local images [161]. This decentralized approach preserves patient privacy and facilitates the inclusion of data from multiple hospitals, potentially enhancing models' robustness and generalizability. At the same time, the involvement of multiple hospitals in an FL setting introduces new challenges, such as different dataset sizes among clients [162], potentially leading to imbalanced contributions to the global model, with biased performance favoring clients with larger datasets, and variation in data labeling quality [163]. Today, there is an established literature relevant to centralized learning with noisy labels [164] but centralized noise filtering algorithms cannot always be exploited in multicentric studies due to privacy concerns. In other fields of medical image analysis, researchers have started to explore techniques to handle noisy labels in FL [165] but the problem of different dataset sizes among clients is not considered.

In this work, we address the problem of federated fetal standard plane detection in the presence of noisy labels from a large (5187 images) [166] and a small (450 images) dataset [159] from two different continents (Europe and Africa, respectively).

Our strategy begins by leveraging the largest and, therefore, presumably the most representative client in the federation to extract robust embeddings using contrastive learning. Our idea is that these embeddings should capture the most significant geometrical features inherent to each standard plane (i.e., brain, femur, abdomen, thorax), as preliminarily shown in [152], and can be used to refine noisy image labels of the same client. Prototypes computed from the noise-free embeddings, along with the backbone trained via contrastive learning, are shared with the smallest client to label it robustly and guide the learning process in the federation. We show that we can mitigate the impact of noisy labels in clients with different data size, improving the overall performance of standard plane detection across the federation.

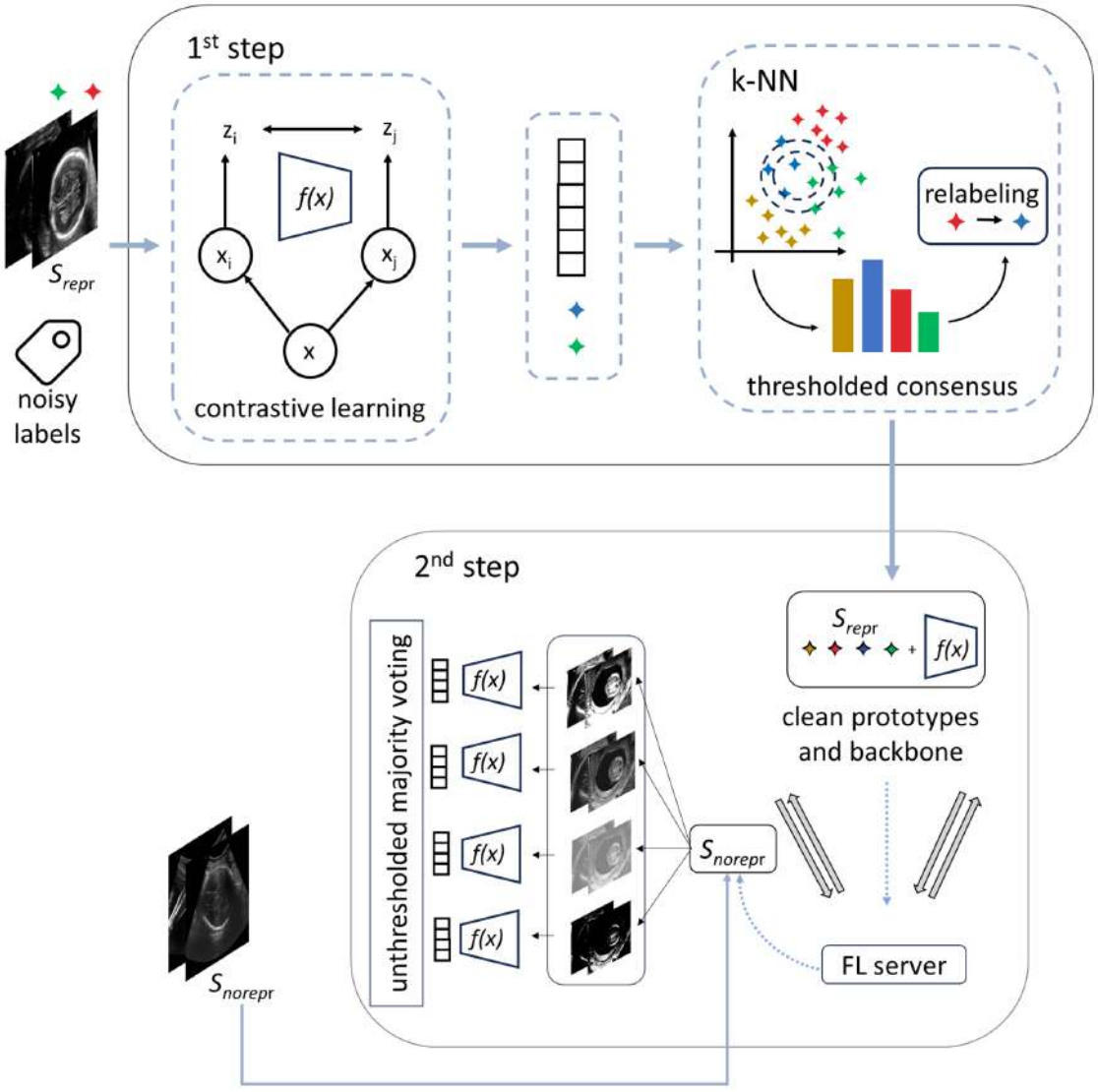


FIGURE 9.1: Overview of the proposed FL framework. In the first step, contrastive learning (SimCLR) is used on the largest dataset in the federation (S_{repr}) to produce embeddings. A k-NN-based thresholded consensus is used for relabeling S_{repr} samples in the embedding space, resulting in a noisy label-free S_{repr}. In the second step, prototypes computed from clean embeddings, along with the SimCLR backbone pretrained self-supervisely on S_{repr} ($f(x)$), are shared through the FL server with the smallest dataset (S_{norepr}). Multiple augmented views of S_{norepr} are generated, their embeddings are extracted, and unthresholded majority voting is used to assign labels to S_{norepr} based on the clean prototypes.

9.1 Method

Figure 6.1 shows an overview of the proposed framework. The first step involves applying contrastive pre-training using SimCLR [167] on the largest, most representative client (S_{repr}) to learn robust noise-free embeddings. Following SimCLR, given an input image (x), we generate two augmented views of it, denoted by x_i and x_j, by applying two random transformations ($R_i(x)$ and $R_j(x)$). Both the augmented images are fed into a backbone encoder network ($f(\cdot)$), which maps them into a latent space in the form of two embeddings ($z_i = f(x_i)$ and $z_j = f(x_j)$). $f(\cdot)$ is trained to maximize the agreement between these two embeddings in the latent space, encouraging invariant representations learning of the input, while maximizing the distance from negative pairs to prevent complete collapse. This is achieved by minimizing:

$$\mathcal{L}_{\text{SimCLR}} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{r=1}^{2N} \mathbb{1}_{[r \neq i]} \exp(\text{sim}(z_i, z_r)/\tau)} \tag{9.1}$$

where $\text{sim}(z_i, z_j)$ represents the cosine similarity between the embeddings z_i and z_j, τ is the contrastive temperature parameter, $\mathbb{1}$ is the indicator function, which is equal to 1 if $y_j = c$ and 0 otherwise, and N is the batch size. By leveraging the structure of the latent space resulting from SimCLR, we use k-nearest neighbors (k-NN) to refine noisy labels in S_{repr}. For each image embedding, we identify its k nearest neighbors based on the Euclidean distance. A threshold (th) is further set so that a sample is kept if at least the th% of its k nearest neighbors shares the same label (there is a consensus), otherwise the sample is discarded. We choose the value of th as a trade-off between minimizing the number of samples with noisy labels and maximizing the number of samples to be kept.

In the second step, we focus on the client with the smallest sample size (S_{norepr}). Instead of performing re-labeling as done for S_{repr}, we directly discard S_{norepr} labels. In fact, S_{norepr} may struggle to benefit from the self-supervised learning approach applied in the first step due to its limited size [167]. To comply with the privacy-preserving nature of FL, our strategy only shares the pre-trained backbone $f(\cdot)$ from the first step and class-specific prototypes computed from S_{repr}, where the prototype (p_c) for class c is computed

as:

$$p_c = \frac{1}{\mathcal{D}_c} \sum_{z \in \mathcal{D}_c} z \tag{9.2}$$

with $\mathcal{D}_c$ being the set of embeddings labeled as c.

These prototypes act as compact, privacy-compliant feature representations, guiding S_{norepr} to align with the knowledge captured in the initial phase. To ensure robustness to possible variability in image acquisition across centers, we apply several augmentations to each image from S_{norepr}. For an image $x \in S_{\text{norepr}}$, we generate T augmented views ($\{x_1, x_2, \ldots, x_T\}$) through operations such as random flips, rotations, and brightness changes. These transformations are denoted as:

$$\mathcal{A}(x) = \{R_1(x), R_2(x), \ldots, R_T(x)\} \tag{9.3}$$

where each $R_i(x)$ represents the transformation applied to the image x, with $R_1(x)$ being the identity transformation. Each augmented image is fed to the shared $f(\cdot)$, which extracts the corresponding embedding. This results in a set of embeddings $\{z_1, z_2, \ldots, z_T\}$. For each extracted embedding, the closest S_{repr} prototype within the latent space is identified, and its label is assigned to it. The final sample label ($\hat{y}$) is assigned based on the class that receives the majority of votes among the embeddings for that sample:

$$\hat{y} = \arg\max_c \sum_{z_j \in \mathcal{N}(z_i)} \mathbb{1}[y_j = c] \tag{9.4}$$

where $\mathcal{N}(z_i)$ denotes the neighborhood of the embedding z_i, consisting of the closest embeddings to z_i in the feature space.

9.1.1 Datasets

We here use two multi-centric datasets, which correspond to S_{repr} and S_{norepr}. Table 9.1 provides an overview of dataset size, Countries involved, number of fetal US images available for the anatomical planes (i.e. abdomen, brain, femur, and thorax), and acquisition devices used.

TABLE 9.1: Dataset information in terms of Country, device, and standard planes.

Country	Device	Abdomen	Brain	Femur	Thorax
	S_repr				
Spain	Voluson E6, Voluson S8, Voluson S10, Aloka	711	1718	1040	1718
	S_norepr				
Malawi	Mindray DC-N2	25	25	25	25
Egypt	Voluson P8	25	25	25	25
Uganda	ACUSON X600	25	25	25	0
Ghana	EDAN DUS 60	25	25	25	0
Algeria	Voluson S8	25	25	25	25

Dataset 1: This dataset acts as S_{repr} and is a publicly available dataset collected from two centers in Spain, released by [166]. Several operators with similar experience acquired fetal US images from six different US machines including three Voluson E6, one Voluson S8, one Voluson S10, and one Aloka. The images are acquired using a curved transducer with a frequency range of 3 to 7.5 MHz for abdominal US, and a 2 to 10 MHz vaginal probe for cervical US screening during the second and third trimesters.

Dataset 2: This dataset acts as S_{norepr} and is a publicly available dataset with images collected from 5 African countries (Malawi, Algeria, Uganda, Ghana, and Egypt), released by [159]. Different operators acquired the images using US scanners from various vendors, including GE Medical Systems, Siemens, Edan Instruments, Shenzhen Mindray Bio-Medical Electronics and Aloka. The acquisition was done using a curved transducer with a frequency range of 3 to 7.5 MHz, during the second and third trimester of pregnancy.

9.1.2 Experimental settings

For SimCLR, ResNet-50 is used as $f(\cdot)$. A projection head with a single non-linear layer comprising 256 nodes processes embeddings of length 2048 obtained from the final average pooling layer. The batch size N is set to 256 and the temperature τ is set to 0.5 as in [152]. For k-NN classification, the value of k is set to 50, while th is set to 40.

To set the value of th, we take into consideration values from 40 to 70 with steps of 10 and assess, at varying levels of noise (0%, 20%, 50%), the trade off between the number

of preserved samples and the percentage of residual noisy labels. We consider 40 as a minimum value to ensure enough agreement among the 4 classes within the neighborhood.

For the transformations in Eq. 9.3, we use random horizontal flips, rotations up to 15 degrees, and shifts up to 12 pixels along both the x and y axes. Brightness and contrast adjustments are performed within a range 0.7 to 1.3 to simulate diverse lighting conditions. Linear mix-up ($\alpha = 0.5$) is used for regularization. For both datasets, we follow the train-test split proposed in the original papers presenting the datasets. We further split the training set of each dataset, allocating 20% of the samples to a validation set to monitor the learning process and detect potential overfitting during training. Following [168, 169], for simulating label noise, we consider uniformly distributed noise among all classes. Model updates from the local clients are aggregated at the central server using Federated Averaging (FedAvg). Training is conducted over 5 communication rounds, each with 20 local epochs. The number of rounds is set to 5 based on the minimum validation loss observed when running the FL process up to 100 rounds. While increasing the number of epochs per round can reduce communication overhead, excessive epochs in clients with high inter-client heterogeneity may hinder convergence [138]. For this reason, we set 20 epochs per round, the minimum required to complete a full cycle of the learning rate scheduler. The optimization process uses the Stochastic Gradient Descent (SGD) optimizer with a Cosine Annealing Warm Restart scheduler. The scheduler starts with an initial learning rate of 0.05, which is reduced to 0.00001 within 20 epochs (1 round). Warm restart applies at the beginning of the next round. The batch size is set to 16, and optimization is performed using the standard Cross-Entropy loss function. The entire framework is built in Python 3.8.10, using PyTorch 2.0.0 and Torchvision 0.15.1. Computations are distributed across 4 NVIDIA A100 GPUs, each equipped with 64GB of VRAM, on a system with 512GB of RAM, ensuring efficient and scalable training.

9.1.3 Ablation study

The performance of our framework was first compared with that of traditionally (i.e. locally) trained models (Local_train) in the presence of noise, as to evaluate the potential

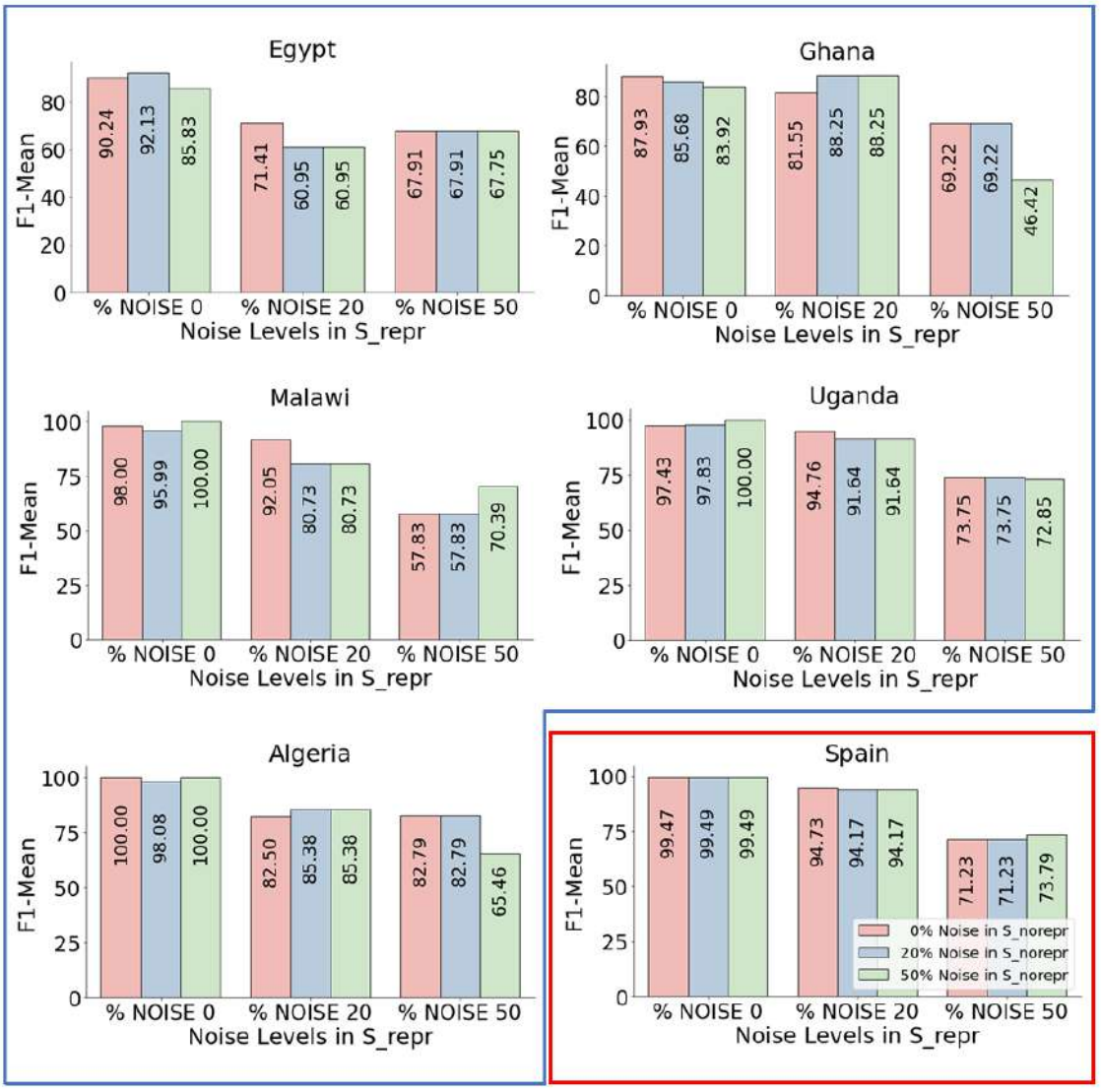


FIGURE 9.2: Impact of varying noise levels (0%, 20%, and 50%) in the S_{repr} dataset (x-axis) on the F1-mean scores across different Countries for Simple_FL. The African countries belonging to the S_{norepr} dataset (Egypt, Ghana, Uganda, Malawi, Algeria) are highlighted with a blue box, while Spain, the S_{repr} dataset is highlighted with a red box

gain of training a joint, federated model specifically designed to address noise-related challenges.

As a first experiment (Simple_FL) for FL, we analyzed the performance of FedAvg under varying noise levels (0%, 20%, 50%) in S_{repr} and S_{norepr} to assess how noise affects the detection performance independently of the proposed strategy for cleaning image labels in the first step of our framework.

TABLE 9.2: F1-score averaged over all classes for S_{repr} and S_{norepr} for Local_train.

	S_{repr}	S_{norepr}				
%noise	Spain	Algeria	Egypt	Ghana	Malawi	Uganda
0	99.40	96.15	96.15	84.56	100.0	100.0
20	93.02	86.38	94.15	86.90	82.90	100.0
50	60.48	36.15	59.45	59.68	60.83	61.00

TABLE 9.3: Effects of label denoising applied to S_{repr} with varying percentage of noise (% noise) in terms of percentage of preserved samples.

	before denoising	after denoising	
%noise	#samples	#samples	%preserved samples
0	2840	2706	95
20	2840	2646	93
50	2840	2169	76

To assess the impact of the second step of our framework, we used FedAvg, filtered the noisy labels in S_{repr} and introducing 0% (Baseline) 20% (Baseline20) and 50% (Baseline50) of label noise in S_{norepr}. Prototype sharing was evaluated in the Proto+Baseline configuration, i.e. we excluded the use of views from the proposed framework to check if this was enough to tackle possible variability among image acquisition protocols across centers.

The combination of prototypes with multiple views outside FL was explored in the experiment Proto+views, where training was performed on S_{repr} and S_{norepr} was classified based on S_{repr} prototypes using augmented views. This experiment aimed to assess whether differences between the two datasets, potentially caused by varying acquisition protocols, could be mitigated through the application of transformations and comparison against prototypes, without requiring S_{norepr} to participate in the FL process.

As a last experiment, we assessed the performance of $f(\cdot)$ trained on S_{repr} with noise free labels and tested on S_{norepr}. This experiment (Pretrained weights) assessed the potential benefits of transferring knowledge from a clean, representative dataset on a small one.

TABLE 9.4: F1-score averaged over all classes for the ablation study.

		mean-F1 over classes by country						
Method	% noise	Algeria	Egypt	Malawi	Uganda	Ghana	Mean	↑↓
Baseline	0	96.13	87.99	84.17	98.00	100.00	93.26	0
Baseline20	20	94.03	90.12	84.17	95.99	100.00	92.86	-0.40
Baseline50	50	94.03	85.81	81.85	93.89	100.00	91.12	-2.14
Proto+Baseline	no labels	93.96	81.34	88.43	98.00	98.67	92.08	-1.18
Proto+views	no labels	100.00	80.53	90.87	95.99	97.43	92.97	-0.29
Proposed framework	no labels	98.10	79.49	94.87	100.00	100.00	94.49	+1.23
Pretrained weights	-	98.00	86.15	83.49	98.00	96.10	92.35	-0.91

9.2 Results and Discussion

Results from Local_train are shown in Table 9.2. With such a simple training strategy, S_{repr} and, more evidently, S_{norepr} experienced a drastic drop in performance at high noise levels. This first result emphasizes the need of (i) guiding the training on S_{norepr} via the larger S_{repr} dataset and (ii) exploiting noise filtering.

Moving to Simple_FL, Figure 9.2 shows the F1-score obtained for S_{repr} and S_{norepr} with varying noise levels for each client. The figure presents the results obtained setting noise levels at 0%, 20%, and 50% in S_{repr}, while independently varying the noise levels in S_{norepr} across the same range. The results for S_{norepr} are reported individually for images acquired from Egypt, Ghana, Malawi, Uganda and Algeria, as they may exhibit differing image distributions The impact that the noise in S_{norepr} had on S_{repr} did not appear to be particularly significant, as the client generally maintained performance within the same order of magnitude across varying noise levels. This observation supports the initial intuitive hypothesis of this study: knowledge transfer is predominantly driven by the most representative client. Clients with smaller datasets benefit from the influence of the representative client, as their learning process is positively guided, even in the presence of noise from incorrectly labeled data.

Baseline showed to be robust to noise injected to S_{repr}. We only had a slight decline in performance when 20% of its data was mislabeled. However, performance deteriorated significantly as noise levels increased to 50%. As expected, the noise introduced in S_{norepr} did not appear to affect the performance on S_{repr}. Overall, the performance of Baseline on S_{repr} remained superior to that of Local_train (shown in Table 9.2).

(a) Thorax, Algeria (b) Thorax, Algeria (c) Abdomen, Ghana

(d) Abdomen, Ghana (e) Abdomen, Ghana (f) Femur, Ghana

FIGURE 9.3: Samples where Baseline fails, while the proposed framework provides correct results. (a) Baseline incorrectly classified the sample as abdomen, overlooking crucial features such as the ventricles; (b) Baseline erroneously predicted the sample as brain ; (c) the prediction is femur, likely influenced by the contrasted upper part of the abdomen; (d) the sample is predicted as brain by Baseline, possibly misled by its shape (e) the sample is predicted as femur as for c; (f) is completely misunderstood as thorax by FedAvg.

As shown in Table 9.3, using $th = 40$ allowed the proposed framework to obtain a clean version of S_{repr}, preserving 93% of samples with 20% noise injected and 76% with 50% noise injected. This threshold was selected because the higher tested thresholds (i.e., 50, 60, and 70), drastically reduced the percentage of preserved samples as follows: under 50% noise (36%, 3%, and 0%, respectively) and 20% noise (77%, 64%, and 46%, respectively).

As shown in Table 9.4, as the filtering is applied in S_{repr}, with Baseline we achieved a mean F1-score of 93.26% under noise-free conditions. However, as noise is introduced in S_{norepr}, the decline in performance remains relatively controlled. With Baseline20, the mean F1-score drops to 92.86%, and even Baseline50, the model maintains a mean score of 91.12%. This highlights the robustness of FedAvg, largely attributed to the presence of

a clean, representative client (S_{repr}), which plays a crucial role in mitigating the impact of noise on smaller datasets. The pivotal role of this dataset is further evidenced by Pretrained weights experiment, where, even in the absence of labels, the model maintains a relatively high performance, achieving a mean F1-score of 92.35%.

When Proto+Baseline was tested, we saw that incorporating prototypes into the FL process gave mixed results. The overall mean F1-score reached 92.08%, showing an improvement over Baseline50 but a slight decline compared to noise-free Baseline. This suggests that incorporating prototypes may not be enough to mitigate the effects of noise without additional refinement, as performance can be impacted by potential shifts introduced by different vendor machines. In fact, the Proto+views approach achieved a better mean F1-score of 92.97%. This highlights the potential performance boost that can be achieved by addressing potential distribution shifts by using augmented views.

The most promising results come from the proposed framework, which integrates prototypes, views, and FL. This approach achieved the highest mean F1-score of 94.49% (+1.23% over Baseline). The improvement stemmed from the complementary strengths of the components. The prototypes provided a stable, noise-resistant foundation by leveraging robust embeddings from the large and clean dataset, effectively transferring knowledge to the small client. The ensemble views further enhanced the performance by introducing diverse perspectives for each sample, helping to smooth out inconsistencies or shifts in data distributions. Finally, federated averaging ensured that all clients benefitted from shared global knowledge. Figure 9.3 shows that our framework enabled a robust extraction of discriminative features, being able to capture key characteristics, such as roundness of brain, the straight line of femur, and chamber shape of the thorax.

Chapter 10

COHESIA: Cohesive Impact Assessment Framework for DPIA and FRIA

In this chapter, we first present a schematic comparison between the DPIA and the FRIA, which primarily serve complementary functions. Next, we discuss the proposed framework in detail, explaining how it builds upon well-established support tools for the DPIA and integrates more experimental and still-evolving instruments for the FRIA, the latter drawing from recent literature on the practical implementation of the [1]. Finally, we simulate the drafting of both a DPIA and a FRIA for selected AI systems through the application of COHESIA, and we discuss the resulting findings.

10.1 Interplay between DPIA and FRIA

Under Art. 27(4) of [1], the FRIA limits its scope to cases where the relevant aspects have already been addressed within the DPIA, for which the latter retains primary responsibility in ensuring compliance.

At the same time, the FRIA functions as an extension of the DPIA, broadening its scope beyond data protection to encompass the full range of Fundamental Rights (FRs)

enshrined in the [170]. Conversely, unlike the DPIA, the FRIA narrows its application by focusing specifically on AI systems, although the definition of an AI system, as set out in Art. 3(1), has been formulated as broadly as possible. This focus reflects the particularly significant implications that AI technologies may have for both individual and collective rights.

That said, an area of overlap exists between the DPIA and the FRIA, allowing for the potential reuse of sections from the former when drafting the latter [171]. In this regard, a framework enabling the parallel execution of both assessments may serve as a catalyst for mutual reinforcement, enhancing the coherence, efficiency, and overall robustness of the compliance process. The following sections compare the two assessments with respect to selected points of interest.

10.1.1 Legal Basis

DPIA. The legal basis for the DPIA, in force since 25 May 2018, is established under Art. 35 of [2].

FRIA. The legal basis for the FRIA, which will take effect on 1 August 2026 following a two-year grace period from the entry into force of [1], is provided by Art. 27 of [1].

10.1.2 When Mandatory

DPIA. Pursuant to Art. 35(1) [2], DPIA must be undertaken whenever processing is likely to pose a risk to the rights and freedoms of natural persons, particularly where new technologies are employed, and with regard to the nature, scope, context, and purposes of the processing. Unlike [1], which explicitly sets out risk criteria, [2] largely entrusts the concrete determination of risk to the controller, or, where relevant, to other actors involved in the processing. Art. 35(3) [2] specifies instances in which, "in particular," a DPIA is mandatory:

(a) systematic and extensive evaluation of personal aspects of natural persons, based on automated processing (including profiling), where decisions produce legal effects concerning the individual or similarly significantly affect them;

(b) large-scale processing of special categories of data pursuant to Art. 9(1), or of personal data relating to criminal convictions and offences pursuant to Art. 10;

(c) large-scale systematic monitoring of publicly accessible areas.

In addition, non-binding guidelines, issued by the former Art. 29 Working Party [4], identify further contexts in which a DPIA is strongly recommended. These guidelines refer not only to Art. 35 but also to Recitals 71, 75, and 91 in [2], as well as to other provisions of [2] where the expression "likely to result in a high risk" is adopted. The guidelines highlight several types of processing falling into this category, including:

1. Evaluation or scoring;
2. Automated decision-making producing legal or comparably significant effects;
3. Systematic monitoring;
4. Processing of sensitive or highly personal data

As a general rule, [4] considers a DPIA necessary when at least two of these criteria are present. Nonetheless, the ultimate responsibility lies with the controller, who may decide to conduct DPIA even if only one of the criteria applies. In [172], the Italian Data Protection Authority further clarifies the categories of processing that necessitates DPIA, with particular emphasis on new technologies, including AI (point 7). Notably, in such cases, DPIA is required even if only one of the above mentioned criteria is met.

FRIA. Art 27(1) [1] mandates FRIA for any high-risk AI system referred to in Art. 6(2) or Annex III, with the exception of high-risk AI systems intended to be used in the area listed in point 2 of Annex III.

Art 27(4) [1] elucidates the complementerity between FRIA and DPIA stating that "If any of the obligations laid down in this article is already met through the data protection impact assessment conducted pursuant to Art. 35 of Regulation (EU) 2016/679 or

Art. 27 of Directive (EU) 2016/680, the FRs impact assessment referred to in paragraph 1 of this article shall complement that data protection impact assessment."

10.1.3 Execution

DPIA. Art. 35(7), anticipated by Recitals 84 and 90 [137], clarifies the minimum content that a DPIA must include:

(a) a systematic description of the envisaged processing operations and the purposes of the processing, including, where applicable, the legitimate interest pursued by the controller;

(b) an assessment of the necessity and proportionality of the processing operations in relation to the purposes;

(c) an assessment of the risks to the rights and freedoms of data subjects referred to in paragraph 1; and

(d) the measures envisaged to address the risks, including safeguards, security measures and mechanisms to ensure the protection of personal data and to demonstrate compliance with this Regulation taking into account the rights and legitimate interests of data subjects and other persons concerned.

FRIA. Art. 27(1) [1], having defined the cases in which FRIA is mandatory, subsequently specifies its required content:

(a) a description of the deployer's processes in which the high-risk AI system will be used in line with its intended purpose;

(b) a description of the period of time within which, and the frequency with which, each high-risk AI system is intended to be used;

(c) the categories of natural persons and groups likely to be affected by its use in the specific context;

(d) the specific risks of harm likely to have an impact on the categories of natural persons or groups of persons identified pursuant to point (c) of this paragraph, taking into account the information given by the provider pursuant to Art. 13;

(e) a description of the implementation of human oversight measures, according to the instructions for use;

(f) the measures to be taken in the case of the materialisation of those risks, including the arrangements for internal governance and complaint mechanisms.

10.1.4 Timeframe and Deployment

DPIA. Art. 35(1) [2] requires that DPIA be conducted "prior to the processing," a requirement already anticipated in Recitals 90 and 93, and further reinforced by the recommendations in [4]. This temporal requirement reflects the principle of "privacy by design", as enshrined in Art. 25(1) [2]. Undertaking the DPIA at the earliest possible stage, and, where appropriate, in parallel with the design and development of the processing operation, ensures that compliance considerations are embedded throughout the process, thereby securing conformity with [2] on a continuous basis.

We note that, although the publication of a DPIA is not legally mandated, it is strongly encouraged by [4], even if only in partial form. As with other non-binding best practices, such disclosure remains voluntary, yet it is recommended as a means of fostering transparency, enhancing trustworthiness, and promoting a culture of sound data protection practices. The sole circumstance under which publication becomes mandatory is established by Art. 36 of [2], which applies in cases where significant residual risks are identified, notwithstanding the fact that all protective measures foreseen in the DPIA have been designed and implemented during the development of the data processing activity.

FRIA. Relating FRIA, Art. 27 [1] establishes the time frame within which the risk assessment must be conducted and explicitly provides that it shall be carried out prior to any deployment. Art. 27(3) specifies the delivery requirements: once the assessment under Art. 27(1) has been completed, the deployer must notify the market surveillance authority of the results, using the template referred to in Art. 27(5).

10.1.5 Accountability

DPIA. Art. 24(1) of [2] designates the controller as ultimately accountable for the execution of the DPIA. In this context, the controller may rely on documentation or input provided by the data processor, where applicable. Notably, Art. 28(3)(f) [2] requires that the processor assist the controller in ensuring compliance with the obligations set out in Artt. 32, 33, 34, 35, and 36. The same recommendation is also reiterated in [4].

FRIA. With reference to FRIA, pursuant to Art. 27 [1], deployers that are bodies governed by public law, private entities providing public services, or deployers of high-risk AI systems are required to prepare a FRIA. Deployers may, however, rely on FRIAs previously prepared by providers.

10.2 Cohesive Impact Assessment framework

The entry into force of [1] marks a pivotal milestone in the European Union's effort to ensure the safe, transparent, and rights-respecting deployment of AI systems. Among the regulatory novelties introduced, Art. 27 of [1] establishes the requirement to conduct a FRIA for AI systems classified as high-risk under Art. 6 and Annex III.

This requirement complements the DPIA, mandated by Art. 35 of [2], which has long served as the principal mechanism for evaluating risks associated with personal data processing. Although the obligation to perform a FRIA will become fully enforceable in August 2026, Art. 27(5) of [1] anticipates that "the AI Office shall develop a template for a questionnaire, including through an automated tool, to facilitate deployers in complying with their obligations under this Article in a simplified manner."

In the meantime, several studies and technical proposals explore methodologies to operationalise the FRIA ahead of the official template. Some authors [171] focus on adapting or extending the DPIA methodology, given the strong conceptual and procedural affinities between the two instruments, while others [173] have emphasized the need for a systematic and quantitatively grounded approach.

Against this background, we introduce COHESIA, a cohesive impact assesment semi-quantitative methodological framework that integrates the DPIA and the FRIA into a unified structure. To the best of our knowledge, this proposal is the first to systematically consolidate these two legal instruments, highlighting their interplay in terms of complementarity, overlap, and extension.

Indeed, while the FRIA extends the DPIA beyond data protection to encompass all FRs potentially affected by AI systems, it also represents a specialisation tailored to the AI context. The overlap is particularly evident in relation to Artt. 1, 7, and 8 of [170], which concern human dignity, liberty, and respect for private life, rights that the DPIA already seeks to safeguard within the data protection domain.

Practically, COHESIA, builds on two pre-existing tools, CNIL-PIA [174] and FRI-Act [175], respectively, supporting the DPIA and FRIA processes. The framework enforces a symmetric three-layer structure across both assessments: 1) a qualitative layer, which guides the narrative and contextual analysis; 2) a semi-quantitative layer, hereafter referred to as Questionnaire, following the nomenclature introduced in [175], consisting of a structured survey assigning numerical impact scores to each question; and 3) a quantitative layer, based on the construction of a severity–likelihood matrix to assess risks in the DPIA and infringements of FRs in the FRIA, in a way that, similarly to what happens in [175], renders results directly comparable. This architecture ensures methodological consistency and facilitates comparison across assessments.

Furthermore, COHESIA introduces simple visualization instruments enabling comparative analysis, both diachronically (across successive versions of the same document) and synchronically (across multiple DPIAs or FRIAs, or between a DPIA and FRIA for the same system). These tools allow assessors to visualise convergence or divergence in evaluations, identify potential inconsistencies, and quantitatively support deliberative decision-making.

To illustrate the framework's capabilities and practical benefits, COHESIA is applied to a modified version of a FAITH-FDSS presented in Chapter 6. The analysis is complemented by three additional AI systems, included as a thought experiment, to

demonstrate how the proposed framework supports comparison across systems with differing levels of risk and compliance maturity. In doing so, the paper reflects on the evolving methodological landscape of AI compliance assessment in Europe, where both mandatory and voluntary DPIA/FRIA practices are expected to become increasingly common as part of a broader culture of trustworthy AI development.

We conclude this introduction with a brief outline of the following sections. Section 10.2.1 introduces the CNIL-PIA [174] and FRIAct [175] tools, which are respectively employed to conduct the DPIA and the FRIA and Section 3.1 describes how CNIL-PIA [174] has been revised by integrating: 1) a quantitative evaluation within the CNIL-PIA [174] questionnaire, following the same methodology as in FRIAct [175]; 2) a quantitative matrix for the numerical assessment of the three categories of risks addressed in the DPIA, Illegitimate access to data, Unwanted modification of data, and Data loss. Section 3.2 explains the overall structure of the COHESIA framework. Section 10.2.2 presents the COHESIA framework and details its three-layer architecture, while Section 10.2.3 focuses on the visualization tools introduced in COHESIA. Finally, 10.2.4 introduces the four case studies, applies COHESIA to each of them, presents and discusses the results, highlighting the methodological advantages of the proposed approach.

10.2.1 Related Tools COHESIA Builds Upon

This section presents the two methodological tools employed in our research, each corresponding to one of the two assessments integrated within the COHESIA framework.

The first tool is CNIL-PIA [174], a well-established and predominantly qualitative instrument developed by the CNIL (Commission Nationale de l'Informatique et des Libertés) institute, which supports the preparation of DPIAs.

The second tool is FRIAct [175], a more recent and semi-quantitative tool, which facilitates the execution of FRIAs. Together, they provide the conceptual and procedural foundations upon which the COHESIA framework builds to achieve an integrated and harmonized approach.

10.2.1.1 CNIL PIA Tool for DPIA

For the DPIA simulation, we employed [174], a tool provided by CNIL, which comprises a series of structured forms guiding the user through the three principal sections of a DPIA: Context, Fundamental Principles, and Risks. Beyond generating the DPIA document

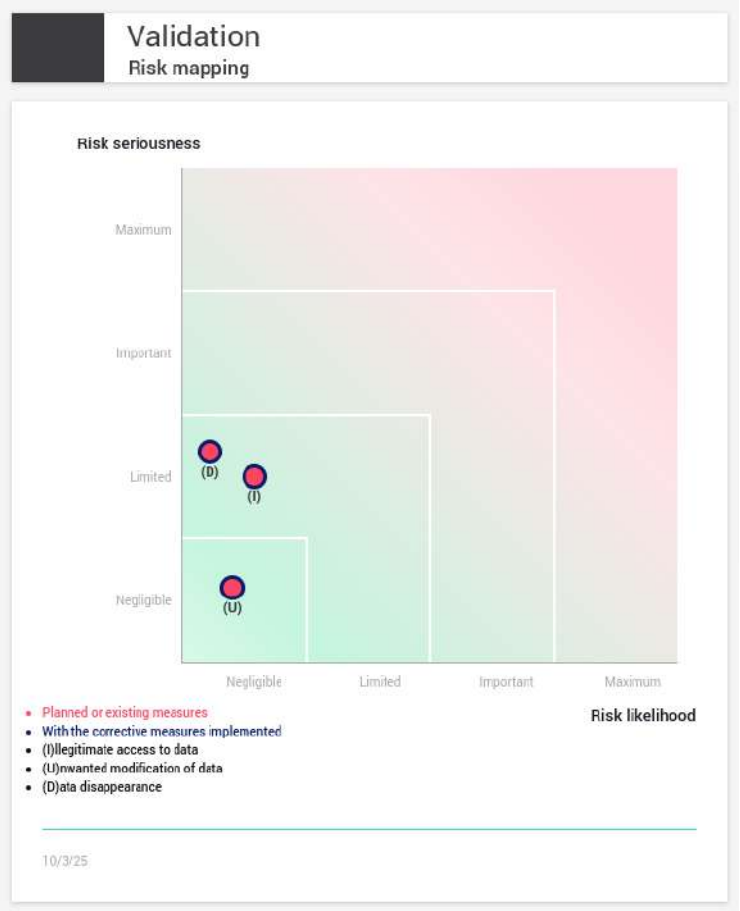


FIGURE 10.1: DPIA Risks placement in CNIL-PIA tool

itself, as shown in Fig. 10.1, the tool produces a visual summary output that provides a concise representation of the third section, Risks. Each risk category, encompassing Illegitimate access to data, Unwanted modification of data, and Data loss, is represented according to the likelihood and severity as assessed by the evaluator. These two dimensions constitute the only semi-quantitative inputs, defined on a limited scale ranging from 1 to 4.

10.2.1.2 FRIAct Tool for FRIA

FRIAct [175], developed to support a comprehensive and reproducible assessment of the impact on FRs, comprises two distinct modules: the Pre-Deployment Questionnaire and the Matrix. The first module, referred to as the Pre-Deployment Questionnaire, is divided into a qualitative introductory component and a quantitative component, collecting the following sections: Categories of natural persons and groups; Deployment process; Input data and Fairness; Transparency; Performance; Human oversight; and AI system Monitoring and Maintenance.

Each section comprises one or more questions, to which the assessor can assign a quantitative response on a scale from 1 to 10, followed by space for a qualitative and discursive justification. Fig. 10.2 illustrates one section as an example. The average of

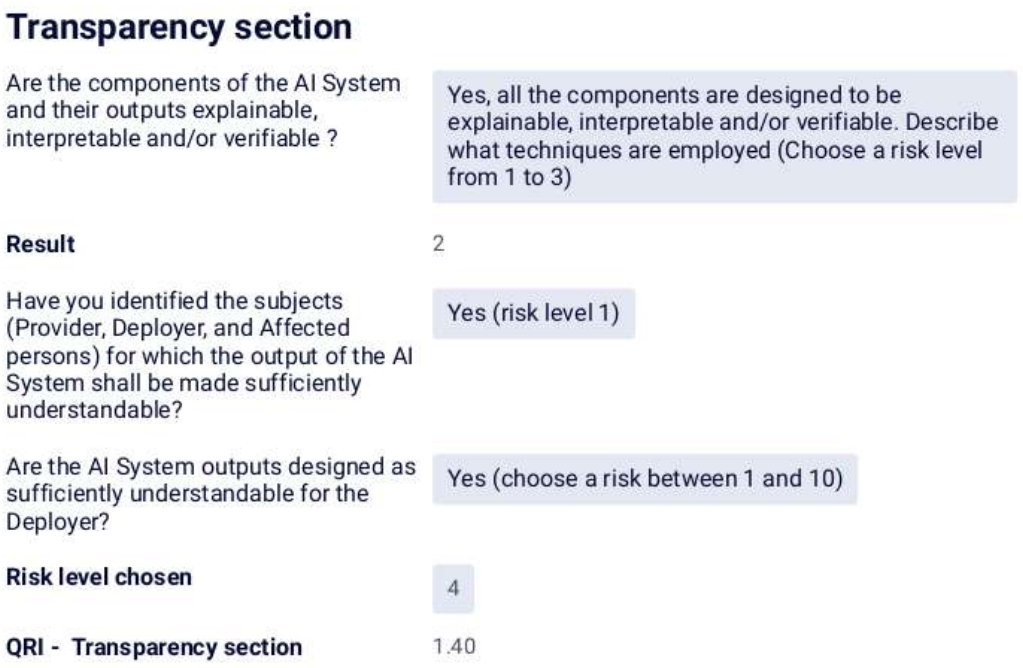


FIGURE 10.2: Transparency section from FRIAct tool

the responses for all questions within a section yields the Specific Risk Indicator (SRI) for that section. Subsequently, the average of all SRIs across the sections produces the Questionnaire Risk Indicator (QRI), which ranges from 1 (low risk) to 10 (high risk). The second module, referred to as the Matrix, is structured to mirror the categories of FRs

in [170]: Dignity (Artt. 1–5), Freedoms (Artt. 6–19), Equality (Artt. 20–26), Solidarity (Artt. 27–38), Citizens' Rights (Artt. 39–46), and Justice (Artt. 47–50).

Article 8 of the EU Charter of Fundamental Rights enshrines the right to theprotection of personal data. It ensures that personal data must be processed fairly,for specified purposes, and on the basis of the consent of the person concerned oranother legitimate basis laid down by law. Moreover, everyone has the right to accessand rectify data collected about them. This right is particularly significant in thecontext of AI, where vast amounts of personal data can be processed automatically.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Protection of personal data	3	3	3	2	2	6	13.05

FIGURE 10.3: Risk assessment in FRIAct for FR in Art.8 of CFREU

Fig. 10.3 illustrates the construction of the matrix specifically for Art. 8. For each FR, the potential infringements are evaluated along two dimensions: Severity and Likelihood. Severity is further decomposed into Intensity, capturing the magnitude of potential harm through a conservative assessment of worst-case outcomes, and Effort of Remediation, reflecting the reversibility of the harm and the difficulty of mitigation.

From these evaluations, an Impact Significance (IS) score is calculated for each FR using classical risk assessment methodology as the product of Severity and Likelihood. In FRIAct, the IS is combined with the QRI obtained from the Pre-Deployment Questionnaire to compute the FRIAct score; in COHESIA, the IS is used as-is, while the QRI is visualized separately to enable a more analytical understanding of the contributions from both the questionnaire and the matrix.

To contextualize each scale, FRIAct [175] provides refined guidance for every dimension or sub-dimension contributing to the final IS. For example, regarding the Effort of Remediation dimension, the 1-to-10 scale is defined as follows: 1–2 (Trivial) indicates minimal corrective actions, with existing resources considered sufficient; 3–4 (Modest) entails moderate effort and some coordination; 5–6 (Substantial) requires significant resource allocation, potentially involving specialized expertise; 7–8 (High) corresponds to complex,

time-intensive interventions that may disrupt standard operations; 9–10 (Nearly Impracticable) denotes extremely difficult or costly measures that could compromise feasibility.

10.2.1.3 Adapting the CNIL-PIA to Fit the COHESIA Framework

First, in order to align the CNIL-PIA [174] with the more quantitative Pre-Deployment Questionnaire proposed in FRIAct [175], we revise each open-ended question from the sections Proportionality and Necessity, and Controls to Protect the Personal Rights of Data Subjects within the DPIA module, integrating them with a quantitative evaluation. The average of the responses for all questions in a section yields the SRI for that section, following the methodology implemented in FRIAct [175].

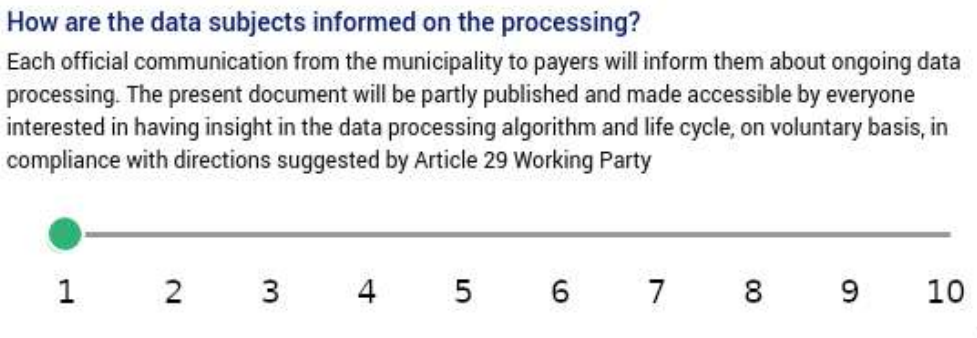


FIGURE 10.4: Open-ended question in CNIL-PIA integrated with quantitative assessment in COHESIA

For a practical illustration, Fig. 10.4 shows one open-ended question in CNIL-PIA [174] integrated with a quantitative assessment. Second, we focus on the following risk categories identified by CNIL-PIA [174], which ideally correspond to the privacy-related articles of the [170] (Art.1, Art.7, and Art. 8): Illegitimate Access to Data, Unwanted Modification of Data, and Data Disappearance. For each category, CNIL-PIA [174] provides:

- three open-ended questions addressing Potential Impacts, Threats, and Sources;
- a semi-structured selection tool labeled Measures, where the user can specify the safeguards implemented to manage potential adverse events: this includes a list of

standard measures suggested by the tool (e.g., encryption, archiving, anonymisation, minimisazion) and user-defined measures entered in the previous module;

- a severity score and a likelihood score, each on a 1 to 4 scale (Negligible, Limited, Important, Maximum).

In our approach, and to ensure methodological consistency with the FRIAct [175] scoring matrix, each individual risk category within the DPIA is considered as it were a FR. Following the way FRIAct [175] evaluates FRs, Severity is decomposed into two sub-dimensions, Intensity and Effort, each quantitatively assessed on a scale from 1 to 10. The Likelihood dimension is similarly refined using the same numerical range. Subsequently, three IS values are computed as the product of severity and likelihood which yields the risk.

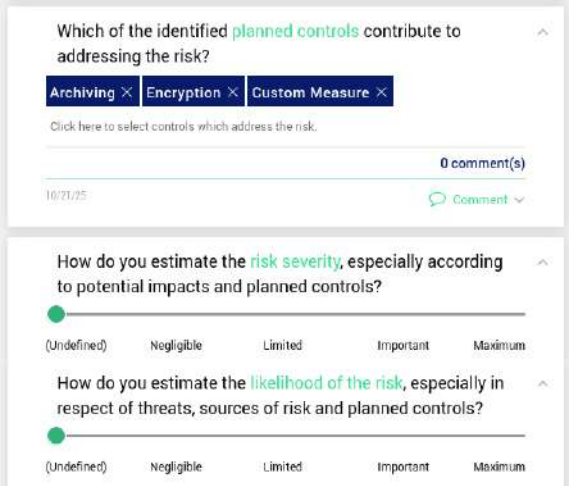


FIGURE 10.5: Original form in CNIL PIA tool

For a practical illustration, Fig. 10.5 shows the original open-ended question in CNIL-PIA [174], whereas Fig. 10.6 depicts its adapted version. In line with FRIAct [175], but introducing some novel elements, we incorporate three dimensions to estimate severity: the Effort of Planned Controls (*ex ante*), the Effort of Remedies (*ex post*), and Intensity, which captures the magnitude of potential harm through a conservative assessment of worst-case outcomes. Similarly, in analogy with FRIAct [175], the SRIs across all sections are averaged to produce an overall score for the DPIA, analogous to the QRI for the FRIA.

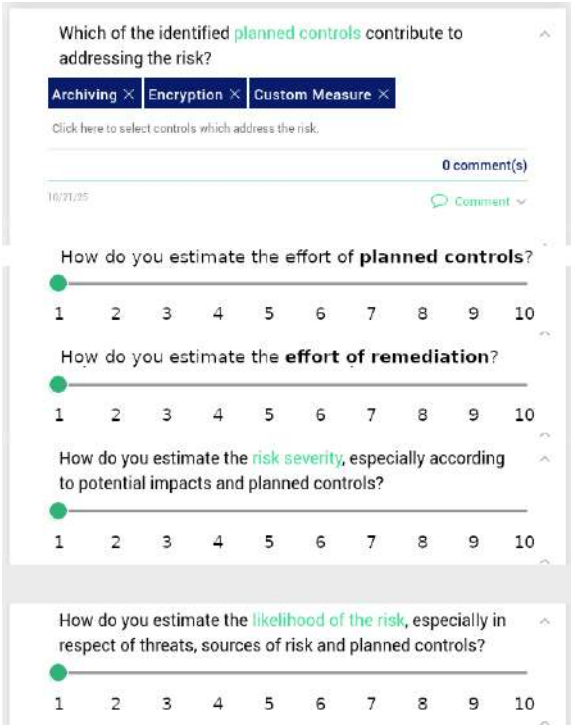


FIGURE 10.6: The same form revised in COHESIA

10.2.2 The Integrated Framework

After adapting the DPIA into a more analytical and quantitatively oriented instrument in line with the FRIAct [175] methodology, the integration of both components within the COHESIA framework yields, in addition to the respective DPIA and FRIA reports, the following outputs:

1. a histogram reporting both the DPIA QRI, which represents the average of SRIs over sections Proportionality and Necessity and Controls to Protect the Personal Rights of Data Subjects of the DPIA; and the FRIA QRI , which represents the average of SRIs over sections Categories of natural persons and groups, Deployment process, Input data and Fairness, Transparency, Performance, Human oversight, and AI system Monitoring and Maintenance of the FRIA, as calculated in FRIAct [175];

2. a scatter-plot representing the distribution of risks across the 50 FRs in [170] (FRIA) plus the three risk assessments from DPIA, reporting the IS for each of them;

3. heatmaps that decompose risk into its components, Severity and Likelihood, for the FRs and risks already reported in the scatter-plot.

10.2.3 Visualization Tools Introduced to Provide a Compact Overall Assessment

Results from COHESIA are presented through three graphical representations, designed for immediate interpretability, which facilitate comparison between different case studies, even at a glance. The first visualization component consists of histograms displaying DPIA QRI and FRIA QRI relating the same AI system as paired bars. This representation allows for an immediate understanding of the specific risk associated with data protection infringements, as well as the risk inherent to the AI system and the task it is intended to perform, highlighting the complementarity of DPIA and FRIA. The second visualization component is a simple scatter plot representing, for each case study under analysis, the IS scores as defined in FRIAct [175]. This score, calculated as the product of likelihood and severity, directly quantifies the risk of infringing a specific FR. On the x-axis of the plot, FRs are grouped according to their category in the [170]. Each point in the scatter is labeled to clearly identify the corresponding FR. The risk values range from 1 to 100. This visualization allows for an immediate understanding of which category of FRs is most affected by each AI system, and the magnitude of this impact. The third representation component consists of a series of 10×10 heatmaps, plotting likelihood against severity. Each entry of the heatmap reports the amount of FRs that exhibit a specific Likelihood and Intensity. This approach allows the two factors contributing to risk to be disentangled; given that we are, in any case, analyzing AI systems from the high-risk category, a high-risk AI system can be well-designed and compliant, minimizing likelihood, yet still be inherently "very risky" due to the nature of the task and its application domain; an AI system with a moderately significant impact but poor design may also be very risky; finally, there are the cases of well-designed systems with limited impact, which are less risky, versus those with maximal impact and poor design, representing the most hazardous class and potentially becoming unacceptable. In addition to the global heatmap encompassing all FRs, a separate heatmap has been generated for each category to enable a more in-depth analysis.

10.2.4 Case studies

To demonstrate its advantages, COHESIA is applied to four AI systems representing distinct risk classes and varying levels of compliance with the European regulatory and ethical framework. With reference to Fig. 10.9, Fig. 10.10, Fig. 10.11, and Fig. 10.12 the four AI systems are selected close to the corners of the heatmap to emphasize different possible types of behavior:

(A) low-impact, compliant, technically robust AI system, whose intended purpose relates the optimization of local tax revenue. **A** is obtained from the AI system introduced in Chapter 6 with minor adaptations. These adjustments enable the system to incorporate a higher-risk task and makes the analysis more informative and instructive. Specifically, alongside the original tasks **T1** and **T2** discussed in Chapter 6, we introduce task **T3**, which concerns the classification of taxpayer behaviour. The task facilitates the categorisation of non-compliant taxpayer conduct into three distinct typologies: (i) deliberate non-compliance, (ii) unintentional default, and (iii) default attributable to partial or temporary financial hardship, thereby supporting proportionate and differentiated administrative responses. For individuals identified under category (iii), who are presumed to experience financial hardship, the objective is to offer a tailored instalment plan that appropriately reflects the taxpayer's financial capacity and enhances the likelihood of effective debt recovery. Now **A** is classified as high-risk because, if the system fails the prediction, some taxpayers may be devoid of a privilege potentially critical for their life. According to Art 6. Annex III (system category 5) [1], this makes the AI system high risk. However, even after adding **T3**, the AI system remains admissible. The key issue is whether the application could be deemed to involve profiling in relation to **T3**, a situation which, under the relevant provisions of the [1], would classify the system as forbidden. While it is true that taxpayers are categorised in a manner that may suggest profiling, [1] stipulates that any system performing categorization falls within the forbidden category only if at least one of the following conditions is met:(1) the categorisation of individuals is used for purposes other than the system's originally intended purpose; or (2) the system could potentially cause disproportionate and detrimental harm to individuals.

Condition (1) does not apply because the categorization is strictly functional to the original intended purpose of the system. Condition (2) appears inapplicable as well, as the absence of an instalment-payment proposal does not constitute a definitively harmful situation and is readily rectifiable upon the taxpayer's request. Indeed, the individual is fully capable of being informed of this option and of asserting their entitlement to it, should they deem it applicable.

(B) low-impact, partially compliant, technically unsound AI system: **B** has the same intended purpose of **A** but differently from **A** lacks technical soundness and full compliance;

(C) high-impact, compliant, technically robust system: inspired by our experience in biomedical images, we conceived **C** as an AI system which supports a clinician in deciding a treatment for a patient, identifying in X-Ray biomedical images features of clinical relevance. According to Annex III [1], AI systems concerning the health of individuals are always to be considered high-risk. **C** has been designed to be compliant with: human oversight (Art. 14 [1]), technical robustness and safety (Art. 15 [1]), data governance (Art. 10 [1]), transparency (Art. 13 [1]), privacy by design (Art. 25(1) [2]) and by default (Art. 25(2) [2]). This careful design makes **C** a high risk fully compliant AI system which minimizes risks, in spite of the remarkable impact associated with its task;

(D) high-impact, not compliant, technically unsound AI system: this hypothetical AI system is designed for partially automating decision-making in legal proceeding . It is assumed to exhibit certain deficiencies in data protection, data governance, and human oversight, which altogether render it unacceptable in real-world deployment due to the magnitude of its potential impact. Of particular concern is the inadequate handling of bias, which may occasionally lead to discriminatory decisions and to arbitrary correlations among unrelated personal attributes, ultimately resulting in unjustified outcomes and severe violations of FRs.

The application of COHESIA to the four case studies, along with the corresponding DPIA and FRIA documentation, also produces the visual representations shown in Figs. 10.7–10.12, which are discussed in the following paragraph. The histograms in

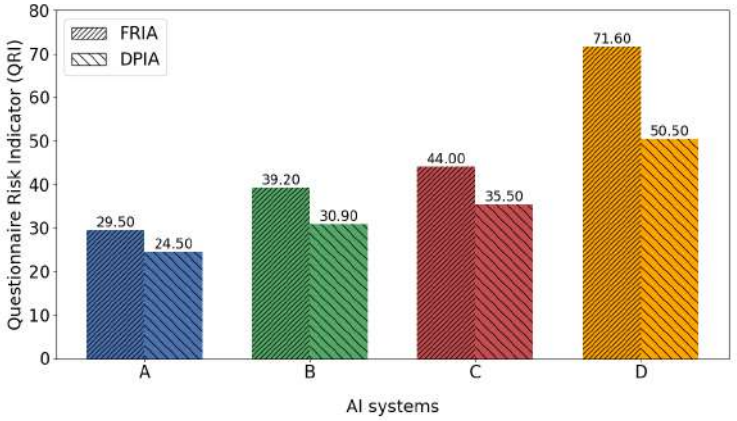


FIGURE 10.7: Histogram reporting FRIA and DPIA QRI for AI systems A to B

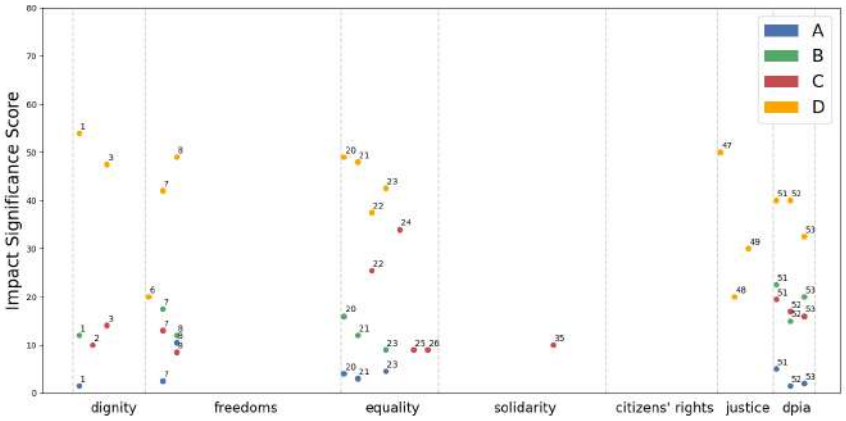
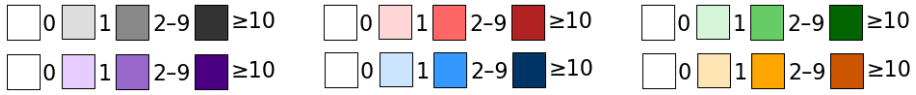
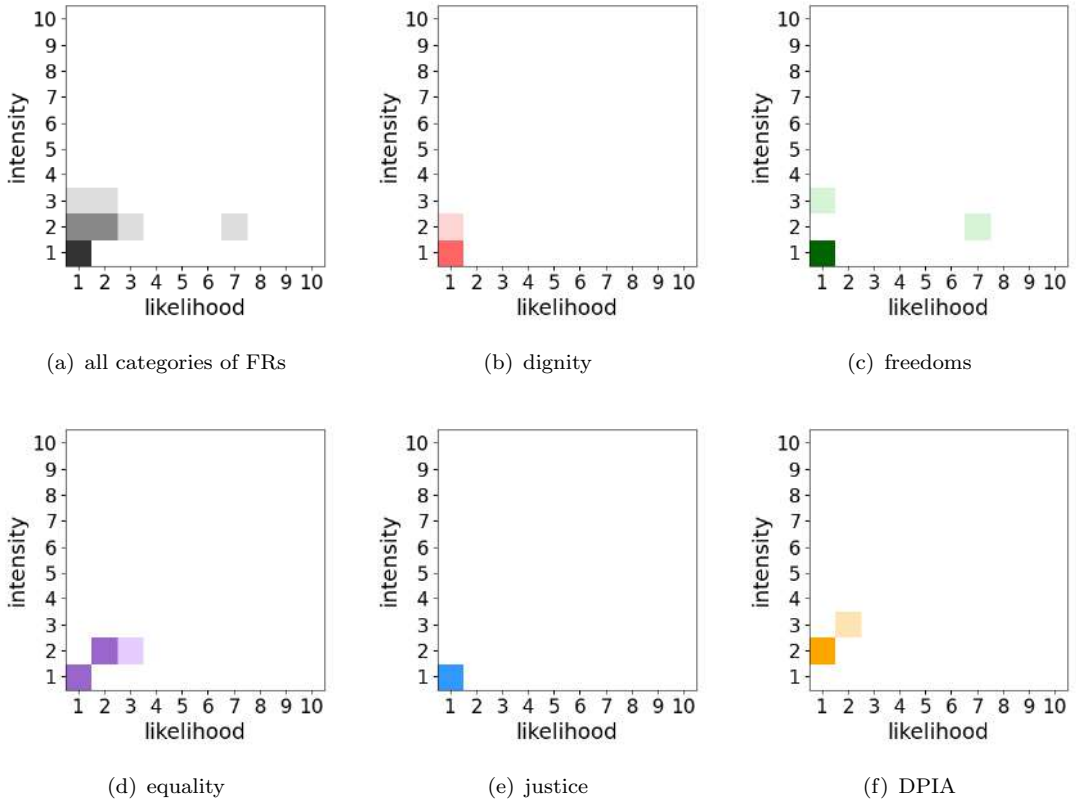


FIGURE 10.8: Plot reporting the risk of infringing any FR in CFREU and for the three risk categories (Illegitimate access to data, Unwanted modification of data, and Data loss) analyzed in DPIA relating , for AI systems A to B

Fig. 10.7 are based on the quantitative components of the questionnaires related to the DPIA and the FRIA, as indicated in the legend. They represent and compare the risk levels across different AI systems though the QRI: the FRIA column accounts for aspects such as technical robustness, human oversight, transparency, and data governance, whereas the DPIA column primarily reflects data protection considerations. Systems **A** and **B** are not markedly different, as the impact of both is modest; even though **B** is poorly designed, its overall risk remains limited. Figure 10.8 presents the infringement risk analysis for all FRs, highlighting the systems associated with the highest infringement risk. Along the X-axis, the FRs are grouped according to the CFREU classification, allowing for the immediate identification of the most affected categories. In Figs. 10.9–10.12, the infringement risk analyzed in the previous plot is decomposed into its two components:



Legend: color intensity reflects the number of FRs mapped to each (likelihood, intensity) pair

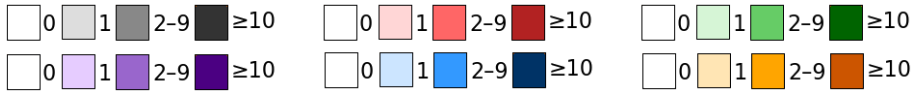


(a) all categories of FRs (b) dignity (c) freedoms

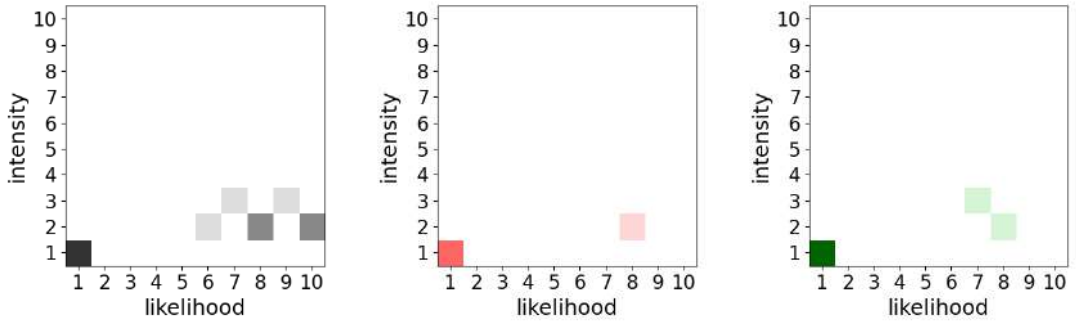
(d) equality (e) justice (f) DPIA

FIGURE 10.9: For AI system **A**, the first heatmap decomposes risk into likelihood and severity, with color indicating the number of FRs corresponding to each likelihood–severity combination. The subsequent heatmaps present the same information distributed across categories of FRs. In all cases, **A** occupies the left bottom quadrant, displaying low intensity and likelihood.

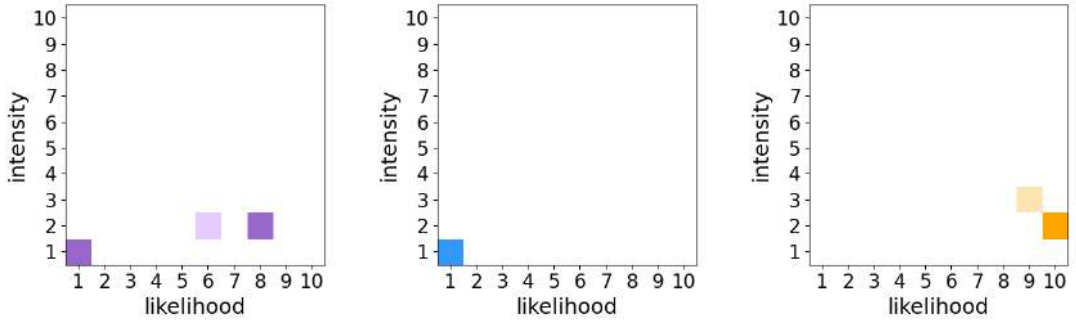
likelihood and severity. By examining them, the differences among the various systems become immediately apparent. System **A** (Fig. 10.9) lies in the lower-left quadrant, indicating that, despite belonging to the high-risk class, its potential impact is modest. Its likelihood of infringement is also low, making the system overall relatively safe. System **B** (Fig. 10.10), positioned in the lower-right quadrant, shows that although its impact is comparable to that of **A**, it lacks adequate technical and legal soundness, resulting in poor compliance. As shown in Fig. 10.8, **B** poses threats to FRs in the categories of Freedoms and Equality, as well as risks related to data protection (DPIA category). System **C** (Fig. 10.11), located in the upper-left quadrant, confirms its high potential impact but



Legend: color intensity reflects the number of FRs mapped to each (likelihood, intensity) pair



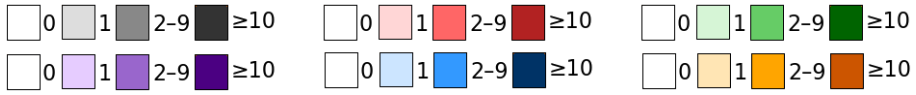
(a) all categories of FRs (b) dignity (c) freedoms



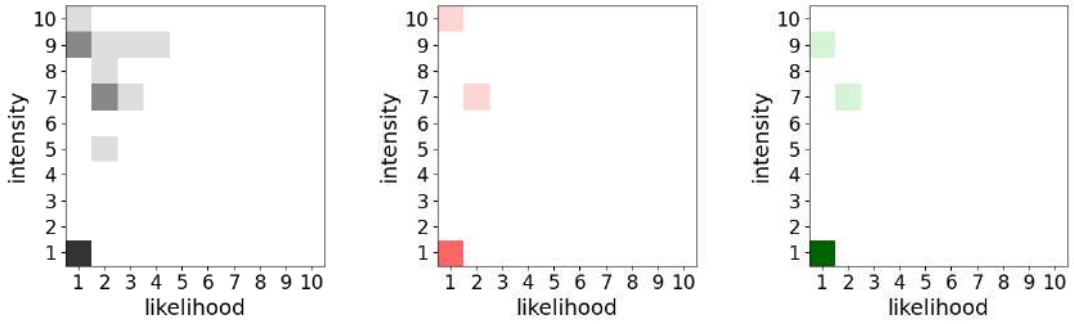
(d) equality (e) justice (f) DPIA

FIGURE 10.10: For AI system **B**, the first heatmap decomposes risk into likelihood and severity, with color indicating the number of FRs corresponding to each likelihood–severity combination. The subsequent heatmaps present the same information distributed across categories of FRs. The most affected categories occupy the bottom right quadrant indicating a high likelihood of occurrence combined with moderate impact.

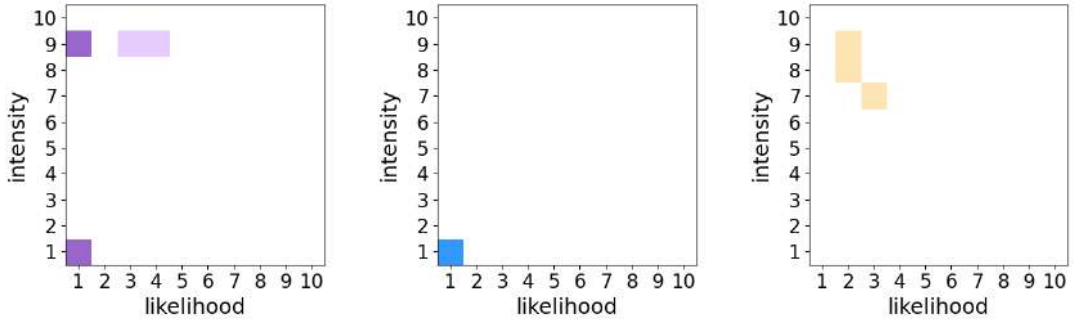
also demonstrates proper technical and legal design and development. Ultimately, the heatmaps provide a clear and concise means to assess AI system quality. Systems located in the bottom-left quadrant are low-impact and well-designed, while those in the top-left quadrant are high-impact and well-designed. Systems in the bottom-right quadrant are low-impact but poorly designed, whereas those in the top-right quadrant combine high impact with poor design. This classification suggests that systems on the left represent good quality, whereas those on the right reflect poor quality. For low-impact systems, poor quality may lead to unwanted outcomes and inconveniences for individuals, whereas high-impact, poorly designed systems should be rejected. The plot in Fig. 10.8, combined with



Legend: color intensity reflects the number of FRs mapped to each (likelihood, intensity) pair



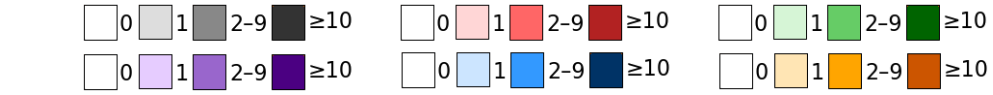
(a) all categories of FRs (b) dignity (c) freedoms



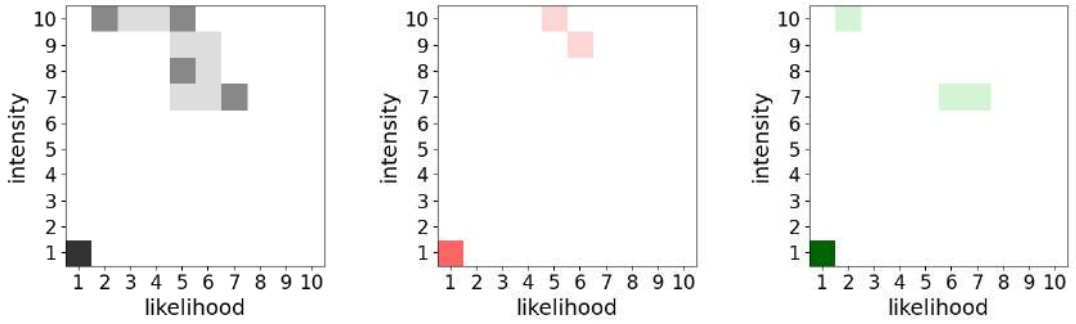
(d) equality (e) justice (f) DPIA

FIGURE 10.11: For AI system **C**, the first heatmap decomposes risk into likelihood and severity, with color indicating the number of FRs corresponding to each likelihood–severity combination. The subsequent heatmaps present the same information distributed across categories of FRs. The most affected categories occupy the top right quadrant indicating that, although **C** has a high impact, a careful design allowed to keep the likelihood low.

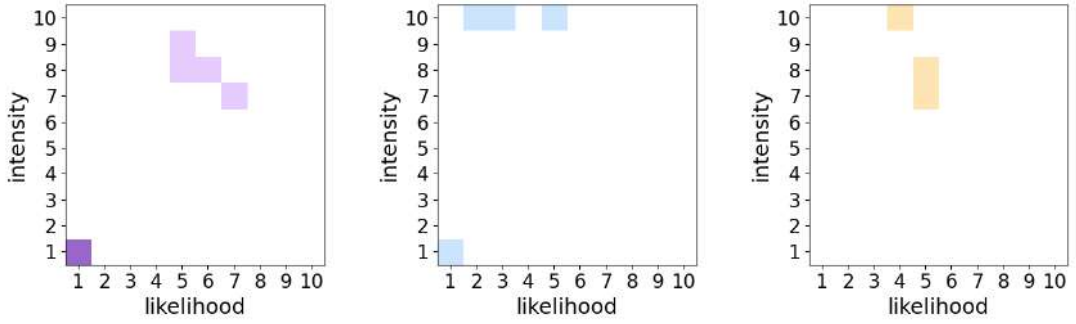
the heatmaps that decompose individual categories, supports the identification of which FRs are most at risk and highlights the components of the AI system that may require correction. This analysis is valuable throughout the entire life cycle of the AI system, providing guidance for timely interventions and preventing design errors from accumulating and becoming increasingly difficult to address. It is also useful whenever the AI system is updated, enabling the monitoring of the impact of new features on technical robustness and regulatory compliance.



Legend: color intensity reflects the number of FRs mapped to each (likelihood, intensity) pair



(a) all categories of FRs (b) dignity (c) freedoms



(d) equality (e) justice (f) DPIA

FIGURE 10.12: For AI system **D**, the first heatmap decomposes risk into likelihood and severity, with color indicating the number of FRs corresponding to each likelihood–severity combination. The subsequent heatmaps present the same information distributed across categories of FRs. The most affected categories occupy the upper central region of the matrix, shifting rightward relative to system **C**. This indicates that system **D** exhibits higher severity and likelihood, thereby rendering its overall trustworthiness questionable.

10.3 Conclusions

The proposed framework offers several notable advantages for the assessment and integration of DPIA and FRIA in AI systems:

1. **DPIA QRI** integration: the open-ended questions in the main sections of the DPIA are enriched with a quantitative dimension, synthesizing responses into numerical values, following the methodology implemented in FRIAct [175]. These quantitative

responses are averaged within each section and subsequently across all sections to produce a Questionnaire Risk Indicator (QRI), providing a concise measure of overall risk.

2. **Enhanced granularity in risk quantification**: by introducing a quantitative assessment of risk in CNIL-PIA [174] based on four dimensions—effort of planned controls, effort of remedies, intensity of harm under a worst-case scenario, and likelihood, COHESIA enables fine-grained estimates of DPIA risks. This approach provides more precise and informative outputs than the standard CNIL-PIA [174] tool while aligning with the FRIAct [175] methodology. These four dimension are relevant to tell apart the *ex ante* impact and *ex post* for risks analysed in DPIA; this finer decomposition allows the assessor to better calibrate the contribution of each factor;

3. **Comprehensive overview of impact assessments**: the framework offers a holistic perspective on both DPIA and FRIA outcomes, leveraging multiple complementary visualization tools. Histograms compare the QRI, a single, intuitive indicator that summarizes the sections of DPIA and FRIA. Heatmaps allow for a more detailed analysis of the two risk dimensions compared to the CNIL DPIA, as they incorporate two additional quantitative dimensions and align with FRIAct [175], thereby enabling the integrated assessment of the three DPIA risk categories alongside the FRs considered in [170]. The scatter plot provides an overall view of risk while simultaneously allowing at-a-glance comparison across categories, highlighting which FRs are most affected by any AI system, both qualitatively and quantitatively. Additionally, the distribution of IS scores across DPIA risk categories and the 50 FRs is examined. Histograms are employed to facilitate an intuitive, simultaneous interpretation different AI systems.

4. **Improved coordination among contributors**: the framework fosters effective communication between contributors responsible for DPIA and FRIA preparation, providing simple, interpretable quantitative and semi-quantitative indicators. This is especially valuable when different actors are involved in drafting the two assessments, ensuring alignment while maintaining accountability for subjective risk judgments.

5. **Diachronic assessment and life-cycle support**: COHESIA supports the longitudinal evaluation of DPIA and FRIA scores, recognizing that both instruments are dynamic and must evolve alongside the AI system throughout its life-cycle. The framework further allows comparison of assessments produced by different individuals or across similar AI applications prepared by different assessors. While quantitative scoring provides structure, COHESIA preserves a component of expert judgment, reflecting the responsibility and accountability of the assessor in motivating risk evaluations in accordance with the principle of proportionality.

In summary, COHESIA provides a structured, reproducible, and interpretable approach to integrating DPIA and FRIA assessments, bridging quantitative rigor with qualitative reasoning. By facilitating comparability, transparency, and accountability, the framework contributes to the broader adoption of trustworthy AI practices and supports compliance with both the [2] and the [1] within a life-cycle oriented perspective which expands Art. 25(1) and makes compliance "by design" not only privacy.

Conclusions

This doctoral research has demonstrated that AI systems can be conceived, designed, and implemented responsibly. To achieve this objective, the principles of lawfulness, robustness, and ethics are embedded in the AI system "by design" from the very outset of its life cycle. This coordinated approach, spanning both regulatory and technological dimensions, yields tangible benefits for the overall quality of the AI system, demonstrating that compliance and performance can not only coexist, but also mutually reinforce each other in a virtuous cycle.

Through six case studies in the taxation, fashion, and healthcare sectors, the research explored the full spectrum of AI risk levels, establishing that trustworthiness can be applied consistently and advantageously across domains, tasks, and data types. The investigation of different risk classes also aligns with the European approach to proportionality and voluntary adherence to ethical standards, promoting both technical excellence and social responsibility, while helping to bridge potential gaps between what is legally required and what is ethically recommended.

Five of the six contributions involve the practical development of AI systems, each addressing the scientific-technical and regulatory challenges specific to its domain. The sixth contribution, serving as a methodological cornerstone, introduces the COHESIA framework, an integrated model that unifies DPIA and FRIA into a single semi-quantitative compliance methodology. Remarkably, this work includes a practical application of the COHESIA framework to four high-risk AI systems, inspired by the other five contributions of the thesis and exhibiting distinct nuances of risk within that class. It demonstrates how an integrated, operational analysis of compliance within the framework,

complemented by graphical visualizations of the results, enables stakeholders to assess the limitations and potential improvements of their AI systems. This approach positions the drafting and continuous updating of DPIA and FRIA documents as a genuine driver of technical robustness, transparency, and accountability.

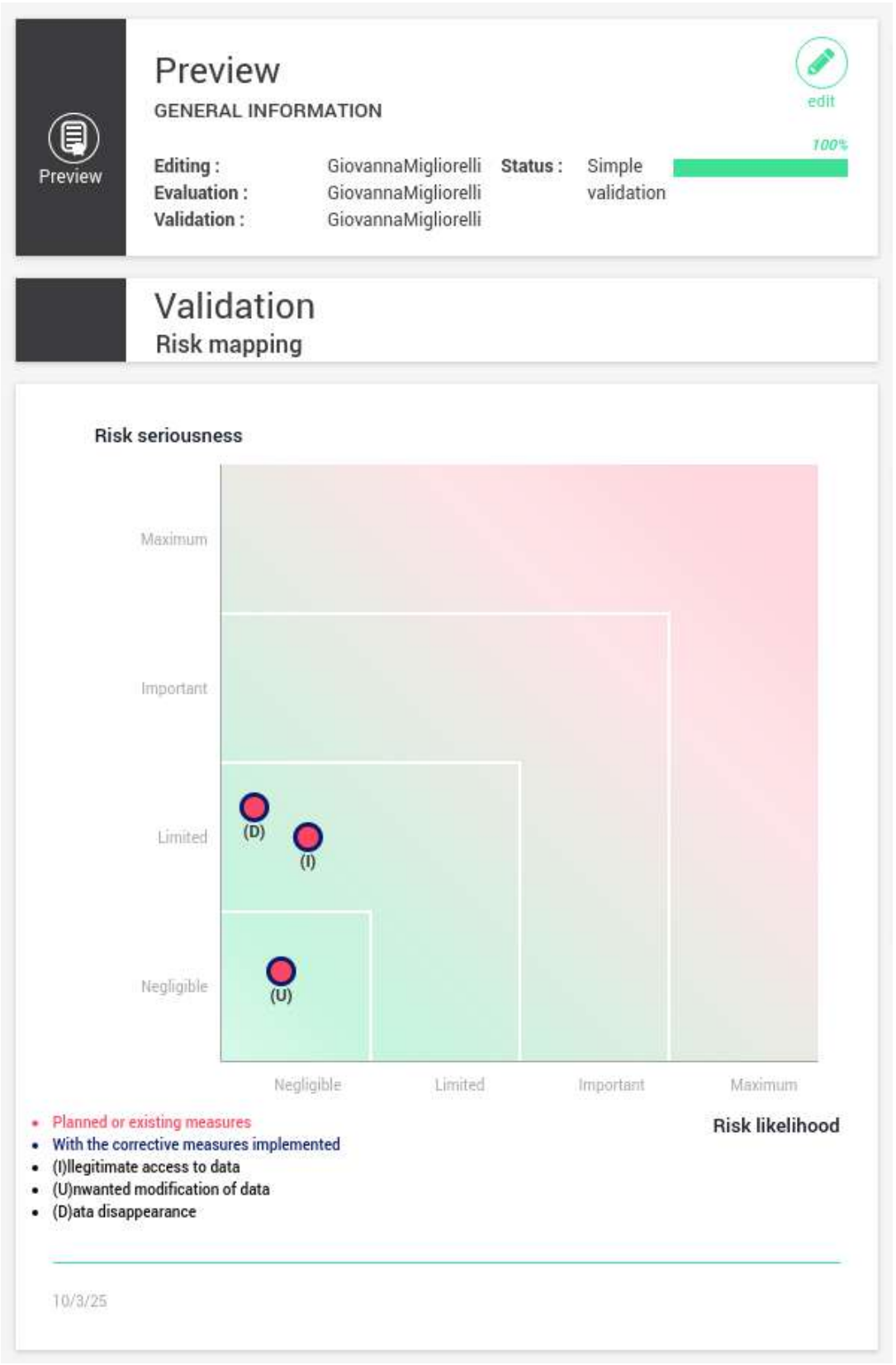
Ultimately, this research affirms that the future of Artificial Intelligence depends not solely on the increasing capabilities of systems, but on the responsibility with which they are designed and implemented, fostering transparency, fairness, and trustworthiness.

Appendix A

Data Protection Impact Assessment (DPIA) Generated Using CNIL Tool

In this chapter, we present the DPIA document for the case study discussed in Chapter 6.

The DPIA document is produced with the support of [174]. The plots summarizing risks are ported first a they are returned by [174]. The more discursive part follows thereafter.


Preview
Preview
GENERAL INFORMATION
edit
100%
Editing : GiovannaMigliorelli
Evaluation : GiovannaMigliorelli
Validation : GiovannaMigliorelli
Status : Simple validation
Validation
Risk mapping
Risk seriousness
Maximum
Important
Limited
Negligible
(D)
(I)
(U)
Negligible
Limited
Important
Maximum
Risk likelihood
Planned or existing measures
With the corrective measures implemented
(I)llegitimate access to data
(U)nwanted modification of data
(D)ata disappearance
10/3/25

Validation

Action plan

Overview

Fundamental principles

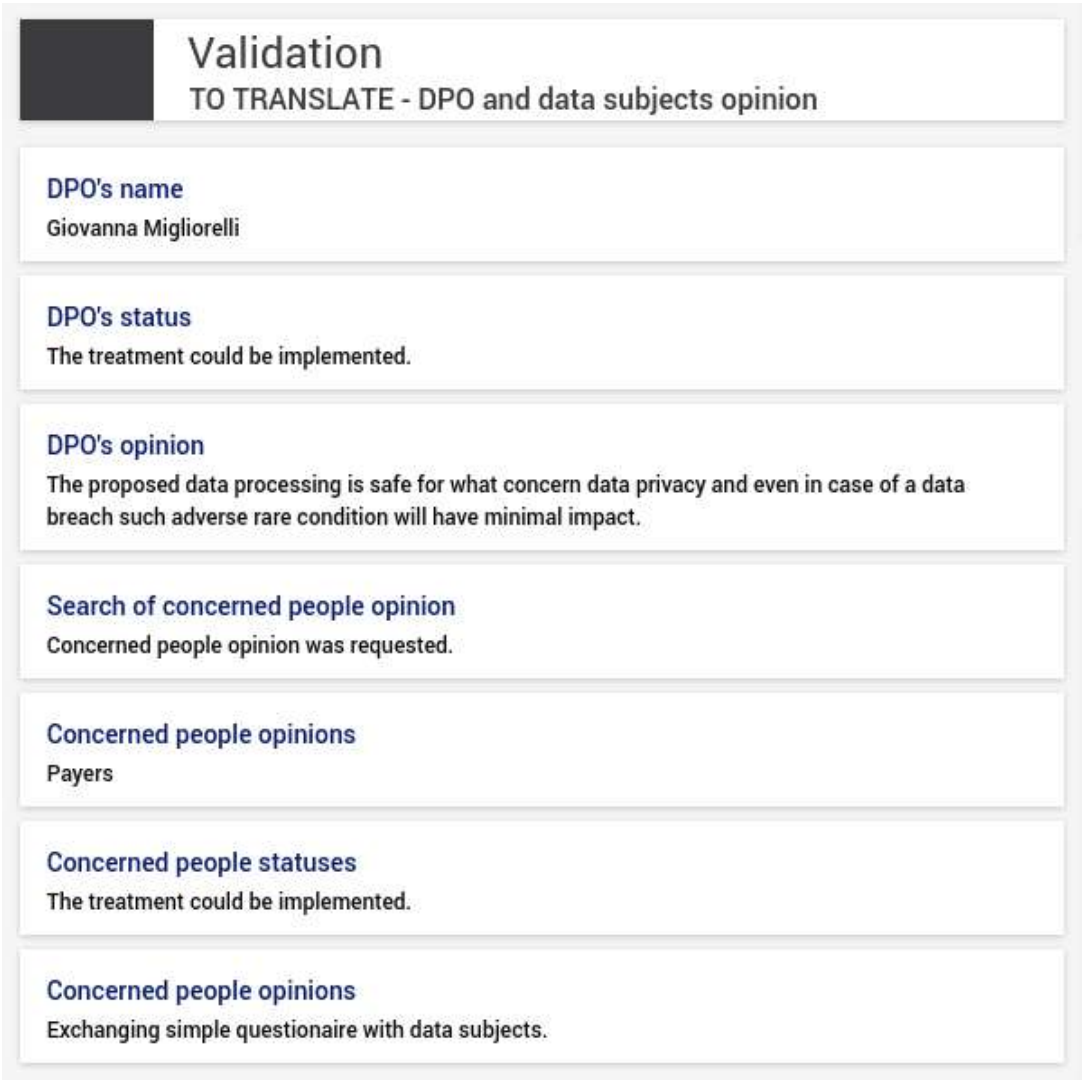
No action plan recorded.

Existing or planned measures

No action plan recorded.

Risks

No action plan recorded.


Validation
TO TRANSLATE - DPO and data subjects opinion
DPO's name
Giovanna Migliorelli
DPO's status
The treatment could be implemented.
DPO's opinion
The proposed data processing is safe for what concern data privacy and even in case of a data breach such adverse rare condition will have minimal impact.
Search of concerned people opinion
Concerned people opinion was requested.
Concerned people opinions
Payers
Concerned people statuses
The treatment could be implemented.
Concerned people opinions
Exchanging simple questionaire with data subjects.

Context

Overview

What is the processing under consideration?

This document sets out the Data Protection Impact Assessment (DPIA) prepared pursuant to Article 35 of Regulation (EU) 2016/679 of the European Parliament and of the Council (hereinafter "GDPR"), concerning the envisaged processing of personal data aimed at predicting whether a payment notice will be paid or not.

The processing operation involves the use of historical data relating to the status of previous payment notices (PAID/NOT PAID), collected from 89 Italian municipalities. The intended purposes of the processing are as follows:

1. to forecast the shortfall that a given municipality may record in the forthcoming fiscal year in relation to amounts due, thereby supporting budgetary planning;
2. to assess whether the transmission of a reminder notice is appropriate when the prediction indicates that the original notice is unlikely to be paid, with the aim of increasing compliance.

The processing is justified under Article 6(1)(e) GDPR, insofar as it is necessary for the performance of a task carried out in the public interest, namely municipal revenue collection and the efficient management of local taxation (IMU, TARI). The necessity of the processing derives from the municipalities' statutory duty to ensure the collection of local taxes and to prevent or reduce fiscal shortfalls.

To achieve the purposes outlined above, the following methodologies are employed exclusively on the data made available by Andreani S.r.l.:

- **Machine Learning (ML):** predictive modelling using supervised algorithms trained on data limited to the relevant payment notice;
- **Federated Learning (FL):** a decentralized training approach enabling the sharing of model parameters without the transfer of raw data, thereby implementing data protection by design (Art. 25 GDPR);
- **Differential Privacy (DP):** an additional privacy-preserving technique integrated into FL, enhancing resilience against inference and re-identification risks in case of targeted cyberattacks.

Each phase of the methodology involves the intervention of one or more human operators. Consequently, the process does not entail decision-making based solely on automated processing, within the meaning of Art. 22(1) GDPR. The system functions exclusively as a decision-support tool, providing municipalities with estimations of the tax gap and indications of when a reminder notice may be appropriate. The final decision rests with the competent municipal authority.

The organizational process under examination is structured into the following phases:

1. **Data quality assessment**;
2. **Construction of the Analysis Dataset**, derived from the raw data provided by Andreani Solution S.r.l.;
3. **Construction of the Control Dataset**;
4. **Selection and implementation of the ML model** for binary classification (PAID/NOT PAID);

The processor as defined in Art. 3(8) GDPR is the person who has developed the AI application implementing the directions and on behalf of the controller as defined in Art 3(7) GDPR and in this case identified in Adreani s.r.l.

What are the responsibilities linked to the processing?

The processor identidied in the person who has developed the AI application applied according to the "privacy by design" principle (Art. 25 GDPR) made it possible to apply the federated framework since the very beginning, hence allowing the datasets not to be ever merged. The processor also made it available as soon as possible and as a priority a Differential Privacy layer, in order to mitigate the system vulnerabilities. The processor was also responible for implementing the Machine Learning model for classification to ensure an optimal performance and avoid misclassification. Each classification response was enriched with a list of importance for each feature utilized in the prediction and furthermore an evaluation of the confidence was also provided to allow the final user to understand the rational behind a decision and assess how confident be in that result respectively.

The controller safegarded sensitive data during the collection phase, applied pseudonimization of data. With regard to the pseudonymisation of data, this is provided for as an option under Article 1, paragraph 682, of Law No. 160 of 27 December 2019, in line with the principle of accountability that underpins the current framework on the processing of personal data. In order to comply with the relevant guidelines and to foster user trust, pseudonymisation is systematically applied.

In this respect, it should be recalled that the concept of pseudonymisation is defined in Art. 4(5) GDPR as *"the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person."*

Both the processor and the controller also applied "privacy by default" (GDPR art 25) limiting the amount of collected data to data collected to the extent strictly necessarily for training and testing the classifier. The processor processed data strictly only for the purpose of classification.

Are there standards applicable to the processing?

Not any standard yet

Evaluation : Acceptable

Context

Data, processes and supporting assets

What are the data processed?

Data processed consist in a collection of CSV files each containing a table with data from some municipals from a given Italian region. Altogether 9 shuch datasets were processed separately under the federated framework. Each table row refers to a tax payment notice sent to some payer. Each row contains 15 fields, 14 features and 1 target.
The 14 features represent inopt data used by the ML classifier to infer the target.

The features are:

- 3 categorical:
  - Tributo
  - Tipo
  - Stato
- 3 integer type
  - number of PartiteIVA registered
  - numer of Years at least one notice has been received for IMU
  - numer of Years at least one notice has been received for TARI
- 4 continuos
  - Tot_Imp_DaPagare_Prima_emission
  - Tot_Imp_DaPagare
  - Tot_Imp_Rateizz
  - Dovuto
- 4 continuous aggregated
  - NumeroAvvisiPastTwoYears_IMU average number of notices received in the past two years for IMU
  - NumeroAvvisiPastThreeYears_TARI average number of notices received in the past two years for TARI
  - PagatoPastThreeYears_IMU average number of paid notices received in the past two years for IMU
  - PagatoPastThreeYears_TARI average number of paid notices received in the past two years for TARI

The target is a binary 0/1 value corresponding to PAID/NOT PAID and telling whether the specific tax payment notice has been paid or not. Aggregated features are an attempt to include time in the prediction. Since aggregation has a lookbehind equal to two years only, at most we need to look into data collected in the two years, before the current one limiting the storing of data for our purposes at a minimum.

In order to foster transparency at least in the rational behind the decision taken by the classification model, a feature importance study has been conducted. The importance detected for each feature applyng SHAP library is reported in attachment shap_plot.png

How does the life cycle of data and processes work?

The prediction of the outcome of a tax payment notice is carried out according to the following criteria:

- maximization of predictive performance;
- sharing of information across multiple datasets while preserving their separation;
- establishment of an acceptable level of security;
- explainability of the classification outcome;
- assessment of the reliability of the classification outcome.

The lifecycle of data processing can be described as follows:

Development phase

Data Quality Analysis
Raw data are filtered on the basis of accuracy and completeness of the values recorded for each field. Redundant or non-relevant fields are removed. The volume and temporal distribution of data are assessed along the relevant timeline, in order to verify consistency and identify sparse or irregular data flows.

Dataset Construction
Raw data are first grouped at the municipal level and subsequently at the regional level, with the region chosen as the unit of granularity. Datasets deemed excessively small in size or statistically negligible in comparison with outliers are excluded. For each dataset, data are aggregated for training purposes, with the addition of four aggregated features: (i) the number of years, within the preceding two years, in which at least one payment notice was received, and (ii) the average number of payment notices received per year. Furthermore, the level of imbalance between the two classes (PAID/NOT PAID) is documented.

Model training and testing
A model chosed on the base of the processor's experience has been implemnted and inscribed in the Federated Learning framework in order to ensure a first line safety level. As soon as possible a Differential Privacy layer has been added to add an additional safety layer, in accordance with "privacy by design" principle as stated by Art. 25 GDPR. The model hyperparameters have been carefully tuned and also the noise injection evaluated in order to grant the best trade off between safety and performance. A privacy budget epsilon computed by means of Renyil accounting has been provided to assess the level of safety obtained. Although the basic model is not straightforward to explain and many steps are need to achieve the final version of the framework, by documenting each step the processor intended to ensure an acceptable level of transparency and reproducibility. Level training has been carried out using datasets provided by the controller and nothing else. The imbalance of classes PAID/NOT PAID has been taken into account under both scenarios:

- during training in order to mitigate the bias possibly due to imbalance;
- during testing in order to provide a reliable evaluation of the performance.

Production phase

Inference
In this phase the framework can be used to safely process new samples from any municipal interested in the data processing. Features are submitted by the municipal to the model which then applies inference and returns the predicted target (PAID/NOT PAID) about the submitted tax payment notice.

Explainability
In oredr to support explainability which is mandated by Art 15(a,h) GDPR, the feature importance has been assessed to allow a clear understanding of which variables most influence the decision taken by the classifier.

Confidence
It is well known that predictions can diaplay an high rate of correctness despite the model being unstable. In that, meaning the confidence we can trust the prediction with is low. In other terms, the model provided the right answer but it was not confident about its decision. In order to provide a higher level of confidence and make the model highly confident when it provides the right decision, little confident otherwise, we provided a confidence interval for each prediction, that is an interval around the probability with which the prediction was given, whose width is inversely proportional to confidence: the narrower the interval, the lowest the confidence.

What are the data supporting assets?

Data are stored in simple csv files on hard disks without any need to trasfer them to a DBMS neither locally not remotely.

Fundamental principles

Proportionality and necessity

Are the processing purposes specified, explicit and legitimate?

The intended purposes are clearly defined as follows:

1. to forecast the shortfall that a given municipality may record in the forthcoming fiscal year in relation to amounts due, thereby supporting budgetary planning;
2. to assess whether the transmission of a reminder notice is appropriate when the prediction indicates that the original notice is unlikely to be paid, with the aim of increasing compliance.

It is well known that efficiency is not a mere occasion but a strong recommendation at the national and European level. Enhancing some aspects in the collection of taxes can greatly affect the policy of municipalities supporting them in delivering efficient and adequate services to all individual people.

Evaluation : Acceptable

What are the legal basis making the processing lawful?

Processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller Art 6 (e) GDPR.

Evaluation : Acceptable

Are the data collected adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed ('data minimisation')?

Data collected are all common data and lies in the tax category. For each data subject, they include all the tax payment notices received in a given year and relating IMU/TARI, with information whether each was duly paid or not. Each tax notice contains data relating the amount of tax due, the type of tax, the type of notice, the number of Partite IVA belonging to the data subject, etc as already described. No name, surname, age, fiscal code are collected. Separate storage of identifying data is applied in an encrypted data base, never exposed to the processor or to the processing algorithm.

Evaluation : Acceptable

Are the data accurate and kept up to date?

Regular checks of the accuracy of the data subject's personal data (e.g. number of partite IVA registered).

Evaluation : Acceptable

What are the storage duration of the data?

Common data are stored for two years.

Evaluation : Acceptable

Fundamental principles

Controls to protect the personal rights of data subjects

How are the data subjects informed on the processing?

Each official communication from the municipality to payers will inform them about ongoing data processing. The present document will be partly published and made accessible by everyone interested in having insight in the data processing algorithm and life cycle, on voluntary basis, in compliance with directions suggested by Article 29 Working Party

Evaluation : Acceptable

If applicable, how is the consent of data subjects obtained?

Pursuant to Art 6(1)(e) GDPR, data processing will be lawful if "it is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller"; in this case the controller is exempted from collecting explicit consent.

Evaluation : Acceptable

How can data subjects exercise their rights of access and to data portability?

Data collected can be required by any data subject in the form of a history of payments. The right to data portability does not apply.

Evaluation : Acceptable

How can data subjects exercise their rights to rectification and erasure?

Simply by contacting the municipality the data subject can exercise her/his right to rectification. The right to erasure does not apply (Art. 6(1)(e) GDPR)

Evaluation : Acceptable

How can data subjects exercise their rights to restriction and to object?

Since the intended purpose is based on a valid legal ground pursuant to Art. 6(1)(e) GDPR and is carried out in the public interest on behalf of the community, the rights to object (Art. 21 GDPR) and to restrict processing (Art. 18 GDPR) are not applicable.

Evaluation : Acceptable

Are the obligations of the processors clearly identified and governed by a contract?

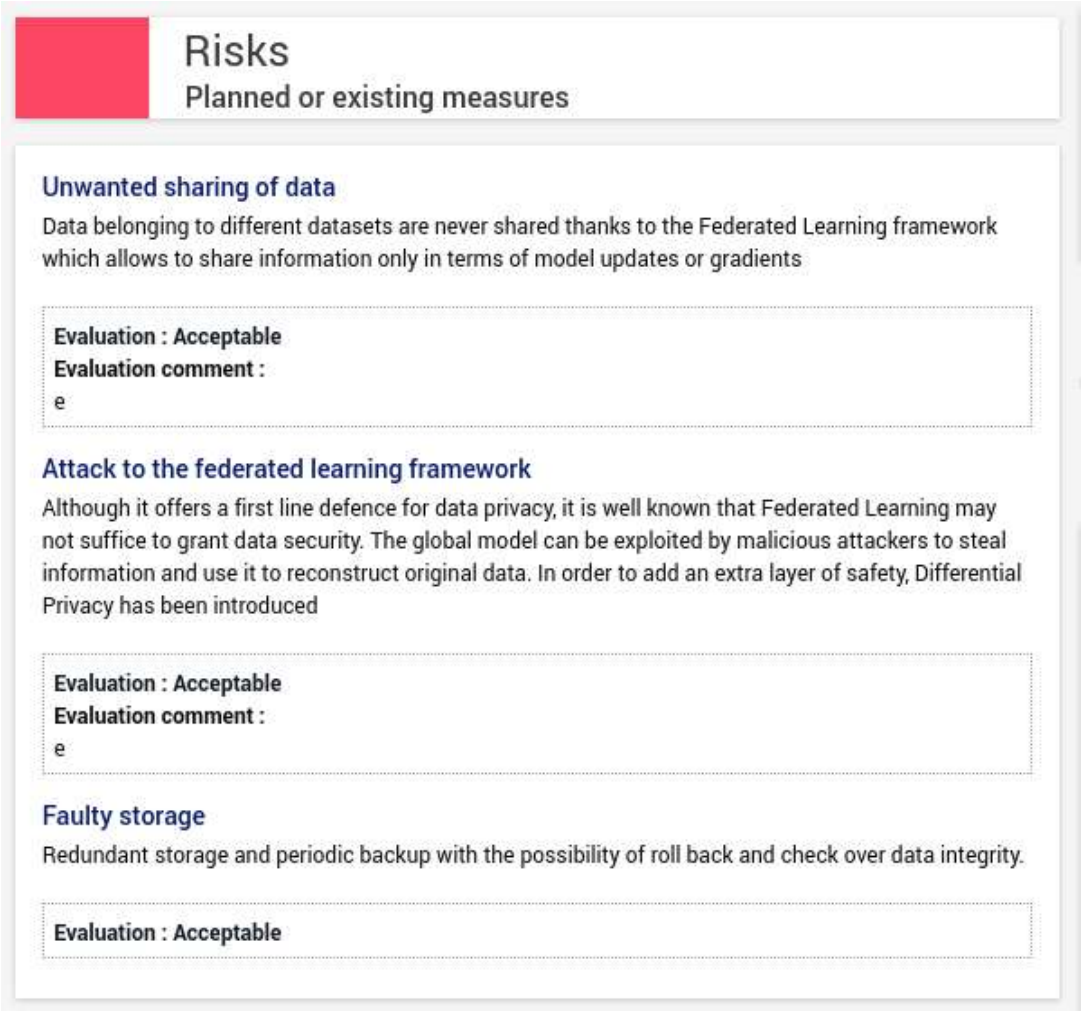
The contract between the controller and the processor commits the processor to the development of an AI application for data processing, in compliance with the principle of "privacy by design".

Evaluation : Acceptable

In the case of data transfer outside the European Union, are the data adequately protected?

Such a scenario is not envisaged.

Evaluation : Acceptable



Risks

Planned or existing measures

Unwanted sharing of data

Data belonging to different datasets are never shared thanks to the Federated Learning framework which allows to share information only in terms of model updates or gradients

Evaluation : Acceptable
Evaluation comment :
e

Attack to the federated learning framework

Although it offers a first line defence for data privacy, it is well known that Federated Learning may not suffice to grant data security. The global model can be exploited by malicious attackers to steal information and use it to reconstruct original data. In order to add an extra layer of safety, Differential Privacy has been introduced

Evaluation : Acceptable
Evaluation comment :
e

Faulty storage

Redundant storage and periodic backup with the possibility of roll back and check over data integrity.

Evaluation : Acceptable

Risks
Illegitimate access to data

What could be the main impacts on the data subjects if the risk were to occur?

Indeed the identity of the data subject is kept separately and never exposed. Tracing back to a data subject would be very difficult. The impact is mild and relates only the breach into information relating taxes paid/not paied

What are the main threats that could lead to the risk?

vulnerability in the framework which processes data

What are the risk sources?

Hacker targeting data subjects

Which of the identified planned controls contribute to addressing the risk?

Attack to the federated learning framework, Unwanted sharing of data

How do you estimate the risk severity, especially according to potential impacts and planned controls?

Limited, Limited because at worst information about which tax payment notice has been paid/not paid will be stolen.

How do you estimate the likelihood of the risk, especially in respect of threats, sources of risk and planned controls?

Negligible, Federated Learning combined with Differential Privacy is already a rather good protection strategy.

Evaluation : Acceptable

Risks
Unwanted modification of data

What could be the main impacts on the data subjects if the risk were to occur?

The data subject may receive additional tax payment notices or on the contrary receive none

What are the main threats that could lead to the risk?

IO error while transferring data

What are the risk sources?

IO hardware fault

Which of the identified controls contribute to addressing the risk?

Faulty storage

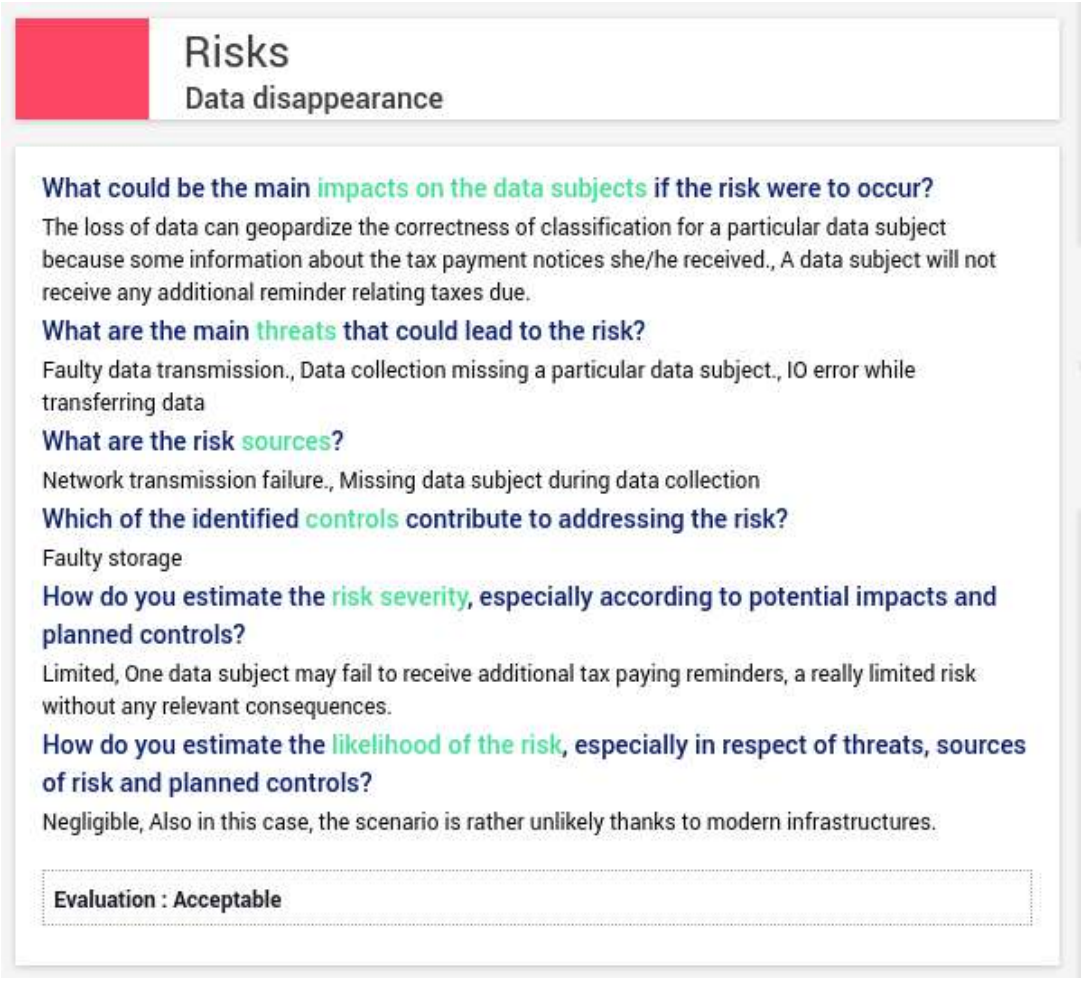
How do you estimate the risk severity, especially according to potential impacts and planned controls?

Negligible, Wrong data for a data subject may at worst prevent her/him from receiving any additional payment nitice.

How do you estimate the likelihood of the risk, especially in respect of threats, sources of risk and planned controls?

Negligible, Negligible because the amount of data is not so large and rather easy to manage also applying integrity checks.

Evaluation : Acceptable



Risks

Data disappearance

What could be the main impacts on the data subjects if the risk were to occur?

The loss of data can geopardize the correctness of classification for a particular data subject because some information about the tax payment notices she/he received., A data subject will not receive any additional reminder relating taxes due.

What are the main threats that could lead to the risk?

Faulty data transmission., Data collection missing a particular data subject., IO error while transferring data

What are the risk sources?

Network transmission failure., Missing data subject during data collection

Which of the identified controls contribute to addressing the risk?

Faulty storage

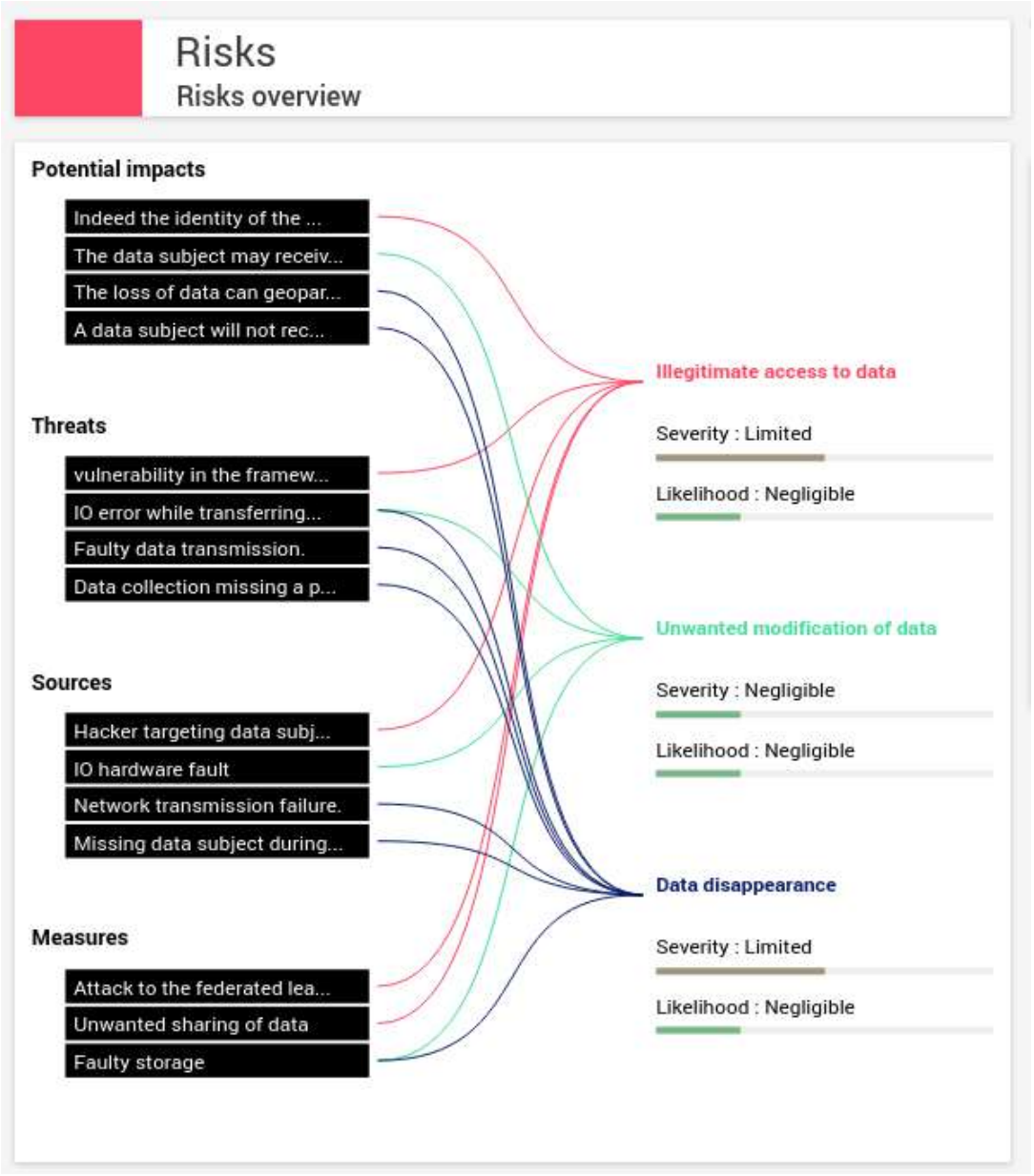
How do you estimate the risk severity, especially according to potential impacts and planned controls?

Limited, One data subject may fail to receive additional tax paying reminders, a really limited risk without any relevant consequences.

How do you estimate the likelihood of the risk, especially in respect of threats, sources of risk and planned controls?

Negligible, Also in this case, the scenario is rather unlikely thanks to modern infrastructures.

Evaluation : Acceptable


Risks
Risks overview
Potential impacts
Indeed the identity of the ...
The data subject may receiv...
The loss of data can geopar...
A data subject will not rec...
Threats
vulnerability in the framew...
IO error while transferring...
Faulty data transmission.
Data collection missing a p...
Sources
Hacker targeting data subj...
IO hardware fault
Network transmission failure.
Missing data subject during...
Measures
Attack to the federated lea...
Unwanted sharing of data
Faulty storage
Illegitimate access to data
Severity : Limited
Likelihood : Negligible
Unwanted modification of data
Severity : Negligible
Likelihood : Negligible
Data disappearance
Severity : Limited
Likelihood : Negligible

Appendix B

Fundamental Rights Impact Assessment (FRIA) Generated Using FRIAct Tool

In this chapter, we present the FRIA document for the case study discussed in Chapter 6.

The FRIA document is produced with the support of [175]. The questionnaire which represents the more discursive part is presented first. The more quantitative matrix follows thereafter quantifying the risk for each fundamental right of the [170].

Wednesday, October 15, 2025

FRIAct

Pre-deployment Questionnaire

Categories of natural persons and groups

Describe the context and the specific purpose of the AI system

The application under examination is designed to support public administration in addressing the following institutional objectives:
- T1 estimation of compliant tax revenue: providing assistance to the competent municipal authority in determining the amount of tax liabilities anticipated to be duly paid, for the purpose of projecting the municipality's annual fiscal receipts;
- T2 forecasting recoverable tax arrears: assisting the municipal administration in identifying taxpayers predicted—on the basis of risk analysis—to be at high probability of default, and in pursuing recovery through the issuance of formal payment notices;

When (dd/mm/yyyy) will the AI system begin operation?

01/01/2026

Specify the end date if known or estimate the duration in months/years

10 years

Usage Frequency: How many times per day/week/month will the AI system be used? Provide an exact count.

Ideally, the AI system will be applied monthly to analyze individual instances of non-payment and annually to generate forecasts of uncollected tax revenue resulting from non-compliant taxpayers.

What are the expected outputs of the AI system and how will they be used in decisionmaking processes?

For each reminder issued in relation to a pending payment, the AI system will generate a prediction of the likelihood that the notice will be settled by the specified due date. This information will support the municipality's budget planning activities, which will, in any case, remain under the supervision of municipal officers. Furthermore, it may assist officers in deciding whether to issue additional reminders to known defaulting taxpayers, thereby enhancing the overall efficiency of the administrative process.

Is the AI system classified as High-risk under article 6 of the AI Act and why?

The assessment of whether the application constitutes a high-risk AI system is governed by Article 6 of the AI Act. Paragraph 1 designates an AI system as high-risk if two cumulative conditions are met: (a) the system falls within the categories listed in Annex I, and (b) a conformity assessment is required as specified therein. Annex I primarily addresses AI systems that constitute safety components or otherwise pose significant risks, which is not the case for the application under consideration.
Paragraph 2 refers to Annex III, subject to certain exceptions. Among the eight categories listed in Annex III, the one potentially relevant to the application is category 5, which concerns the allocation of services. In this context, it is necessary to assess whether the omission of a payment reminder could be considered a disadvantageous condition for the taxpayer, as it may potentially deprive them of a reminder that others would receive. However, since the taxpayer has already received the necessary reminders, the practice of issuing additional reminders only to selected taxpayers constitutes an administrative efficiency measure, and does not amount to the denial of a service or right. Based on this reasoning, Tasks T1 and T2 cannot be classified as high-risk, and, at most, may be

assigned a precautionary designation of moderate risk.

Does the AI system have a GenAI component?

No, it doesn't

Does the GenAI component present systemic risks as defined by the AI Act, and if so, what are the specific factors contributing to this assessment?

The AI system does not present any specific risk.

Who are the affected persons?

The relevant population comprises taxpayers who have received payment reminders as a result of payment default.

Are there any categories of natural persons that can be considered as more vulnerable groups? What are their specific risks?

No. there aren't

Who will directly interact with the AI system?

Municipal officers.

Deployment process section

What type of algorithm(s) or model will be used? Please provide detailed examples and context for each risk level.	A System that is entirely or partially based on a self-learning algorithm, where the machine itself is finding patterns in the data (choose a risk level from 4 to 10)
Risk level chosen	6

Provide more details

The AI system is based on a shallow neural network and operates within a federated learning framework, the security of which is reinforced through the application of differential privacy techniques.

Among different alternatives suitable to achieve similar goals and performances, do you plan to use a simpler and more explainable algorithm?	No. Please explain your choice (choose a risk level between 6 and 10)
Risk level chosen	6

Provide more details

It is our assessment that the associated risk is negligible. In light of the existing human oversight and supplementary validation measures ensuring the accuracy of the model's results, the implementation of simpler models is not deemed necessary to guarantee safeguards that are already effectively ensured.

QRI - Deployment process section	6

Input data and Fairness section

Are all input data (from internal and external sources), including third-party training data and data added by the Deployer, governed by data governance and data quality processes?	Yes, all, in compliance with GDPR, AIA and IP requirements (risk level 1)
Risk level chosen	1

Provide more details

A Data Protection Impact Assessment (DPIA) was conducted to verify the absence of material risks and to evaluate the system's compliance with the provisions of the GDPR.

If the AI system includes a GenAI component and it uses personal data to produce the output, could this lead to the generation of unexpected and/or undesired content? Do not count this question for the Risk Indicator calculation in case the System has no GenAI components or if it does not use personal data to produce the output.	No (risk level 1)
Have you or your Provider analyzed the input data to assess possible biases that could lead to a negative impact on fundamental rights? Please specify your choice and provide detailed examples.	Yes, please specify (e.g. in Fairness Assessment or data exploration, documentation provided by the Provider) (choose a risk level between 1 and 10).
Risk level chosen	3

Provide more details

In general, the data used by the AI system do not contain attributes that could constitute a potential source of bias (e.g., age or gender). In particular, given the minimal risk associated with the AI system, the assessment focused solely on the primary potential bias related to the taxpayer's region of origin. Evaluation of the model's predictive accuracy did not reveal any significant imbalances indicative of bias.

Have you or your Provider collected all the relevant and available data needed to assess if the algorithm may discriminate against or disadvantage specific population subgroups (e.g., nationality, gender, age, etc.)? Please specify your choice and provide detailed examples.	Yes, please specify (e.g. in Fairness Assessment, data exploration, or in order to do this assessment; choose a risk level between 1 and 10)

Risk level chosen	3

Provide more details

As noted above, data do not include sensitive attributes that could potentially introduce bias or constitute a source of discrimination (e.g., age, gender, or ethnicity).

Do you or your provider plan to monitor the trend over time of fairness metrics and/or AI System performance referring to specific population subgroups? Please specify your choice and provide detailed examples.	No (choose a risk level between 6 and 10)
Risk level chosen	6

Provide more details

The AI system does not operate on, or make decisions concerning, particular population subgroups.

QRI - Input data and Fairness section	2.80

Transparency section

Are the components of the AI System and their outputs explainable, interpretable and/or verifiable ?	Yes, all the components are designed to be explainable, interpretable and/or verifiable. Describe what techniques are employed (Choose a risk level from 1 to 3)
Result	2
Have you identified the subjects (Provider, Deployer, and Affected persons) for which the output of the AI System shall be made sufficiently understandable?	Yes (risk level 1)
Are the AI System outputs designed as sufficiently understandable for the Deployer?	Yes (choose a risk between 1 and 10)
Risk level chosen	4
QRI - Transparency section	1.40

Performance section

Have you planned to test the AI system performance and proper functioning?	Yes (choose a risk level between 1 and 10)
Risk level chosen	2

Human Oversight section

What is the degree of automated decision-making in the AI system?	The decision is taken by a human being and the AI System provides only an additional layer of information (Choose a risk level between 1 and 2)
Risk level chosen	1
Can the operation of the AI system be interrupted through a 'stop' button or similar procedure?	Yes, intervention possible (choose a risk level between 1 and 10)
Risk level chosen	1

Provide more details

The AI system under consideration does not necessitate the implementation of safety mechanisms, including an emergency stop function.

Is it ensured that the staff and other persons dealing with the operation and use of AI systems on behalf of the Deployer and the Provider have the means to interact with the AI system and use it in an informed and conscious manner?	Yes, those subjects have received a specific training and they can consult technical guidelines or support. Please specify. (choose a risk level from 1 to 4)
Risk level chosen	1

Provide more details

The AI system's users are fully aware of the significance of the single output it provides. These users are municipal officers who retain full decision-making authority and utilize the AI system solely as an expert aid to support their judgments and enhance administrative efficiency.

Are there support mechanisms for the subjects mentioned above designed to address issues or concerns with AI system operation?	Yes (risk level 1)

QRI - Human Oversight section	1

AI system Monitoring and Maintenance

If unfair behavior emerges, will it be possible to intervene to correct it?	Yes, it will be possible to intervene (risk level 1)
Do you plan to adopt monitoring techniques and related countermeasures in order to intercept anomalous behavior or performance deterioration?	Yes (risk level between 1 and 10)
Risk level chosen	2
How do you plan to update the AI System (retrain, tune, knowledge source update, etc)?	Manual or automatic update planned (e.g., using new data, re-estimation of the hyperparameters, vector store update in RAG applications.) (choose a risk level between 1 and 9)
Risk level chosen	2
Have you planned procedures for emergency/urgent updates?	No (choose a risk level between 6 and 10)
Risk level chosen	6

Provide more details

Indeed there was no need for such procedures.

Have you planned any processes in order to ensure the ongoing and continued availability of data, coherently to the business needs?	No (choose a risk level between 6 and 10)
Risk level chosen	6
Do you plan to oversee the stability of input data/features used by the AI system?	No action planned (risk level 10)
Risk level chosen	10
QRI - AI system Monitoring and Maintenance	4.50

QRI	**2.95**

Pre-deployment Fundamental Rights Matrix

CHAPTER 1: DIGNITY

Human dignity, as enshrined in Article 1 of the EU Charter of Fundamental Rights,forms the basis of fundamental rights and is integral to the European Union's values,including freedom, equality, and solidarity. It influences various other rights, impacting everything from privacy and equality to workers' rights and social security.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Human dignity	1	1	1	1	1	1	9.55

The right to life, articulated in Article 2 of the EU Charter of Fundamental Rights,involves complex legal and ethical dimensions, particularly as they relate to newtechnologies and medical practices. The right not only prohibits arbitrary deprivation of life but also influences EU and MS laws and policies concerning healthcare access,emergency medical interventions, and ethical issues like euthanasia and assistedreproductive technologies.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Right to life	1	1	1	1	1	1	9.55

Article 3 of the EU Charter of Fundamental Rights focuses on the right to physicaland mental integrity, specifically emphasizing protections in the fields of medicineand biology.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Right to the integrity of the person	1	1	1	1	1	1	9.55

Article 4 of the EU Charter of Fundamental Rights prohibits torture and inhuman ordegrading treatment or punishment. This is particularly relevant in the fields ofjustice, home affairs, and external policies where the EU and MS's influence issignificant.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Prohibition of torture and inhuman or degrading treatment or punishment	1	1	1	1	1	1	9.55

Article 5 of the EU Charter of Fundamental Rights specifically addresses theprohibition of slavery, servitude, forced labour, and human trafficking, highlightingthe EU and MS's commitment to combatting these severe human rights violations.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Prohibition of slavery and forced labour	1	1	1	1	1	1	9.55

CHAPTER 2: FREEDOMS

Article 6 of the EU Charter of Fundamental Rights ensures the right to liberty andsecurity, which becomes particularly relevant in the domains of immigration,asylum, and criminal justice under the EU's Area of Freedom, Security and Justice.This article shapes the standards and practices concerning the detention andtreatment of individuals within these areas, ensuring that actions like arrests,detentions, or deportations comply with EU laws designed to protect personalfreedom and security.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Right to liberty and security	2	1	1.50	1	1	1.50	9.90

Article 7 of the EU Charter of Fundamental Rights emphasizes the right to respect forprivate and family life. It acknowledges the broad and dynamic nature of family lifewithin the EU, influenced by various factors including legislation on free movement,social security, and gender equality. The right to privacy under Article 7 ensures theprotection of personal and family records against unauthorized or unnecessarycollection, use, or disclosure of such information. This is particularly critical incontexts affecting children, gender equality, and migrants, where privacy and dataprotection are paramount.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Respect for private and family life	2	1	1.50	1	1	1.50	9.90

Article 8 of the EU Charter of Fundamental Rights enshrines the right to theprotection of personal data. It ensures that personal data must be processed fairly,for specified purposes, and on the basis of the consent of the person concerned oranother legitimate basis laid down by law. Moreover, everyone has the right to accessand rectify data collected about them. This right is particularly significant in thecontext of AI, where vast amounts of personal data can be processed automatically.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Protection of personal data	3	3	3	2	2	6	13.05

Article 9 of the EU Charter of Fundamental Rights addresses the right to marry andfound a family. This right involves areas related to non-discrimination, cross-borderrecognition of family statuses, and public health services affecting family life.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Right to marry and right to found a family	1	1	1	1	1	1	9.55

Article 10 of the Charter of Fundamental Rights of the European Union guaranteesthe right to freedom of thought, conscience, and religion. This includes the right tohold beliefs, change religion or belief, and manifest one's religion or beliefs inworship, teaching, practice, and observance, either alone or in community withothers. Areas of application include employment, public expression, education,healthcare.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Freedom of thought, conscience and religion	1	1	1	1	1	1	9.55

Freedom of expression constitutes a cornerstone of democratic societies. It is at thecentre of individual autonomy, public discourse, and the functioning of the media.This right encompasses the freedom to hold opinions and to receive and impartinformation and ideas without undue interference. It also ensures the freedom andpluralism of the media which is at the core of a diverse and informed public sphere.This right impact areas such as: the right to hold opinions; the right to impartinformation and ideas; and the right to receive information and ideas.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Freedom of expression and information	1	1	1	1	1	1	9.55

The right to assembly and association guarantees individuals the freedom to gatherpeacefully and to form or join groups, including political parties, trade unions, and civic organisations. This right is essential for political participation, social solidarity, and theexpression of collective interests. In this case, AI systems could limit the possibility ofindividuals

assembling, associating, and participating in political and social life, forinstance, by profiling individuals.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Freedom of assembly and of association	1	1	1	1	1	1	9.55

Article 13 of the CFEU deals with the freedom of the arts and sciences and aims to fosterknowledge, culture, and innovation. It ensures that artistic expression and scientificresearch are free from undue restrictions. Academic freedom is also a key component,safeguarding the independence of educational institutions and the development ofknowledge. AI systems raise primary questions about the creation of knowledge and thedevelopment of innovation.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Freedom of the arts and sciences	1	1	1	1	1	1	9.55

Article 14 of the CFEU is about the protection of the right to education. It ensures accessto quality education for everyone, regardless of their nationality or status. It includes theright to vocational training, non-discrimination in educational opportunities, and accessto education for EU nationals and third-country nationals under certain conditions. Theright to education also intersects with freedom of movement within the EU and theability to pursue studies across member states.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Right to education	1	1	1	1	1	1	9.55

Article 15 of the EU Charter protects the right to engage in work and pursue a freelychosen occupation. This right encompasses equal treatment for EU citizens acrossMember States regarding employment opportunities, social, and tax advantages. Forthird-country nationals, once authorized to work, they are entitled to equal treatment inworking conditions.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Freedom to choose an occupation and right to engage in work	1	1	1	1	1	1	9.55

Article 16 of the EU Charter protects the freedom to conduct business, including theright to start and manage economic activity. It is closely linked to the rights toproperty (Article 17) and work (Article 15).

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Freedom to conduct a business	1	1	1	1	1	1	9.55

Article 17(1) of the Charter protects the right to property, ensuring that individuals,companies, and entities are entitled to own, use, and dispose of their possessions.The right extends to both tangible and intangible assets and guarantees that propertycannot be taken away except in cases of public interest and under fair compensation,which must be established by law. Article 17(2) specifically highlights the growingimportance of intellectual property (IP) rights, including copyrights, patents, andtrademarks. Intellectual property extends the scope of the right to property byensuring that the creations of the mind, such as inventions, literary and artisticworks, and symbols, are recognized and protected. Given the increasing reliance onknowledge-based economies, protecting intellectual property rights is crucial tofostering innovation, creativity, and economic competitiveness in the EU. In thiscontext, EU law seeks to strike a balance between protecting IP rights and promotingthe public interest, ensuring that the use of intellectual property does not undulyrestrict innovation or fair competition.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Right to property	1	1	1	1	1	1	9.55

Article 18 of the Charter of Fundamental Rights of the European Union guarantees theright to asylum, framing it within the existing EU asylum system and protocols. Thisright ensures that individuals seeking international protection within the EU have accessto a fair and functioning asylum process. The field of application of Article 18 isprimarily centered on third-country nationals seeking protection within the EU.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Right to asylum	1	1	1	1	1	1	9.55

Article 19 of the EU Charter prohibits collective expulsion and protects individuals frombeing sent to a country where they may face torture, inhuman, or degrading treatment.It has two main components: Article 19(1) prevents the collective expulsion of EU andthird-country nationals, protecting the right to individual consideration of cases; Article19(2) prohibits extradition or removal of individuals to countries where they may facehuman rights abuses.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Protection in the event of removal, expulsion or extradition	1	1	1	1	1	1	9.55

CHAPTER 3: EQUALITY

Article 20 of the EU Charter guarantees that "everyone is equal before the law." This fundamental right is central to EU law, ensuring that all individuals receive fair and equal treatment in all situations where EU law applies. It plays a broad role in numerous areas, including employment, social welfare, taxation, and public services, where differences in treatment can arise.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Equality before the law	3	1	2	1	1	2	10.25

Article 21 of the EU Charter prohibits any discrimination based on nationality or other personal characteristics such as race, gender, religion, sexual orientation, and more. It is divided into two parts: one focusing on nationality discrimination (Art. 21(2)) and the other addressing broader status-based discrimination (Art. 21(1)). While Article 21(1) aligns with EU anti-discrimination laws, Article 21(2) specifically targets nationality discrimination within the scope of the Treaties.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Non-discrimination	2	1	1.50	1	1	1.50	9.90

Article 22 of the EU Charter emphasises that the European Union must respect and promote the diversity of cultures, religions, and languages across its member states. The recognition of this right underlines the commitment to the coexistence of various cultural identities based on the value of diversity, and ensuring that cultural, religious, or linguistic differences are not a basis for discrimination or exclusion. AI could fail to represent diversity and different cultural nuances.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Cultural, religious and linguistic diversity	1	1	1	1	1	1	9.55

Article 23 of the EU Charter mandates the equality between women and men. This right extends to different areas of life, including employment, work, and pay. It does not only promotes gender equality but also allows for the implementation of positive actions (e.g., specific policies or measures) to benefit the under-represented sex, often women, in situations where historical or structural inequalities exist. AI decision-making could lead to gender-based discrimination.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Equality between men and women	3	1	2	1	1	2	10.25

Article 24 of the EU Charter outlines children's rights to protection, care, and participation in decisions affecting them. It consists of three key principles. First, Children have the right to necessary protection and care for their well-being and freely express their views, which must be considered by age and maturity. Second, in all matters involving children, their best

interests must be a primary consideration for both public authorities and private institutions. Third, every child has the right to maintain regular, personal relationships and contact with both parents, unless this would be contrary to their best interests

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
The rights of the child	1	1	1	1	1	1	9.55

Article 25 of the EU Charter recognises the rights of elderly individuals to live with dignity, maintain their independence, and actively participate in social and cultural life. This article ensures that older people are treated with respect and provided with opportunities to engage fully in society, avoiding exclusion due to age.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
The rights of the elderly	1	1	1	1	1	1	9.55

Article 26 of the EU Charter recognises the right of persons with disabilities to benefit from measures that ensure their independence, social and occupational integration, and active participation in community life. This article underscores the importance of creating an inclusive environment where individuals with disabilities have equal opportunities to engage in all aspects of society.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Integration of persons with disabilities	1	1	1	1	1	1	9.55

CHAPTER 4: SOLIDARITY

Article 27 of the EU Charter establishes the right of workers to information and consultation within their organizations. This right is critical in EU labor law, ensuring that employees or their representatives are informed and consulted on matters affecting the workplace. The scope of this right is dependent on national laws and EU Directives, and it typically applies at various levels such as establishments, undertakings, and company groups. However, Article 27 lacks direct effect unless implemented through specific national or EU legislation.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Workers' right to information and consultation within the undertaking	1	1	1	1	1	1	9.55

Article 28 of the EU Charter guarantees the right of workers and their representatives to engage in collective bargaining and take collective action, including strike action, in defense of their interests. This right applies to the institutions, bodies, offices, and agencies of the EU, as well as to Member States when implementing EU law. However, this right must be exercised in accordance with both Union and national laws, making it subject to limitations based on public interest or economic rights.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Right of collective bargaining and action	1	1	1	1	1	1	9.55

Article 29 of the Charter guarantees the right to access free placement services, facilitating labor mobility and helping individuals find employment within the EU. This right is closely tied to the freedom of movement for workers, ensuring that both EU nationals and workers from member states can benefit from equal access to job opportunities. The role of national and European employment services, including platforms like EURES, is essential in coordinating job vacancies, applications, and labor market data, contributing to a balanced, integrated labor market.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Right of access to placement services	1	1	1	1	1	1	9.55

Article 30 of the Charter protects workers from unjustified dismissal. This right is meant to prevent arbitrary terminations, ensuring that any dismissal follows a lawful procedure, includes appropriate reasons, and provides redress, such as compensation or dispute resolution mechanisms.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Protection in the event of unjustified dismissal	1	1	1	1	1	1	9.55

Article 31 of the Charter of Fundamental Rights of the EU ensures the right to fair and just working conditions. This includes protecting workers' safety, health, and dignity in their working environment. It also guarantees that workers receive adequate rest and paid annual leave.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Fair and just working conditions	1	1	1	1	1	1	9.55

Article 32 of the Charter of Fundamental Rights of the EU ensures the protection of children from economic exploitation and harmful labor practices. It restricts child labor and ensures that any work undertaken by young people is done under safe conditions. It aims to balance the need for vocational opportunities with educational development and safeguarding the health, safety, and well-being of minors.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Prohibition of child labour and protection of young people at work	1	1	1	1	1	1	9.55

Article 33 ensures protection for family life and work-life balance, emphasizing social and economic protection for families.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Family and professional life	1	1	1	1	1	1	9.55

Article 34 of the Charter ensures access to social security, housing, and social assistance. It covers citizens and residents within the EU and supports their right to receive appropriate support, especially during vulnerable times, such as unemployment or illness.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Social security and social assistance	1	1	1	1	1	1	9.55

Article 35 of the Charter ensures access to preventive health care and medical treatment under conditions established by national laws and practices. It mandates that a high level of human health protection be integrated into all Union policies and activities.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Health care	1	1	1	1	1	1	9.55

Article 36 of the Charter concerns access to services of general economic interest (SGEI), ensuring that everyone has the right to access these services, especially those essential for maintaining dignity and ensuring welfare (e.g., electricity, water, healthcare).

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Access to services of general economic interest	1	1	1	1	1	1	9.55

Article 37 of the Charter focuses on integrating environmental protection into all Union policies, aiming for a high level of protection and improvement in environmental quality. However, it does not establish any specific individual right to environmental protection. Instead, it serves as a guiding principle for EU institutions and Member States when implementing Union law, especially concerning policies on climate, biodiversity, pollution, and waste management.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Environmental protection	1	1	1	1	1	1	9.55

Article 38 of the Charter of Fundamental Rights of the European Union ensures that all Union policies and activities provide a high level of consumer protection. It aligns closely with Article 169 TFEU, which promotes consumer health, safety, and economic interests. The article has broad relevance, influencing not only consumer protection but also harmonizing laws affecting trade, free movement, and public policies across the Union.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Consumer protection	1	1	1	1	1	1	9.55

CHAPTER 5: CITIZENS' RIGHTS

Article 39 of the Charter of Fundamental Rights of the EU secures the right for citizens of the Union to vote and stand as candidates in elections to the European Parliament. This right applies to EU institutions and Member States implementing Union law, particularly under the Act on Direct Elections and Council Directive 93/109/EC. It relates closely to Article 40 (municipal election rights) and promotes democratic participation within the EU's political processes.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Right to vote and to stand as a candidate at elections to the European Parliament	1	1	1	1	1	1	9.55

Article 40 guarantees EU citizens the right to vote and stand as a candidate in municipal elections, regardless of their nationality within the Union. This right reflects the EU's commitment to democratic participation and equal treatment, aligning with broader Union values. However, it applies only to local elections, and does not extend to national elections. It ensures non-discriminatory access to political participation at the municipal level for EU citizens residing in other Member States.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Right to vote and to stand as a candidate at municipal elections	1	1	1	1	1	1	9.55

Article 41 of the EU Charter guarantees the right to good administration, ensuring that individuals have the right to fair treatment in dealings with the EU institutions. This includes the right to be heard, access to one's file, and receiving reasons for decisions. While the Article specifically addresses EU institutions, bodies, offices, and agencies, the general principle of good administration also applies to Member States when acting within the scope of EU law, as clarified by case law.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Right to good administration	3	2	2.50	2	2	5	12.35

Article 42 of the Charter grants every citizen of the Union the right to access documents of EU institutions, bodies, offices, and agencies. This right is closely linked to Article 15(3) TFEU and is governed by secondary legislation, most notably Regulation 1049/2001.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Right of access to documents	3	2	2.50	2	2	5	12.35

Article 43 of the Charter guarantees EU citizens and residents the right to complain to the European Ombudsman. This right applies to situations where EU institutions, bodies, offices, or agencies have committed maladministration. If a bank's AI system evaluating creditworthiness is part of an EU-regulated institution, individuals could potentially file a complaint with the Ombudsman if the evaluation process appears unfair or lacks transparency.

The Ombudsman ensures that EU bodies adhere to principles of good administration.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Ombudsman	1	1	1	1	1	1	9.55

Article 44 of the Charter provides the right for every citizen or resident of the EU to petition the European Parliament on matters within the Union's scope of activity. If a bank uses AI to evaluate creditworthiness in a way that violates EU laws or principles, individuals can petition the European Parliament to review the matter. This right serves as an avenue for individuals to raise concerns about transparency or fairness in credit evaluation within the Union's legal framework.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Right to petition	1	1	1	1	1	1	9.55

Article 45 of the Charter grants EU citizens the right to move and reside freely withinthe territory of the Member States. This right mirrors the provisions of Article 20(2)(a)TFEU and is fundamental to the European integration project. It primarily emphasizesthat free movement rights, although central, must align with conditions defined in EUTreaties.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Freedom of movement and of residence	1	1	1	1	1	1	9.55

Article 46 of the Charter grants EU citizens the right to diplomatic and consular protection

from any Member State's embassy or consulate when outside the EU and their own country lacks representation. This provision supports the practical application of EU citizenship and fosters European solidarity. It is directly linked to Article 20 and Article 23 TFEU, which outline similar rights and confer powers for the necessary implementation of protection measures abroad.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Diplomatic and consular protection	1	1	1	1	1	1	9.55

CHAPTER 6: JUSTICE

Article 47 of the EU Charter of Fundamental Rights is centered on the right to an effective remedy and to a fair trial. It encompasses the principles of access to justice, ensuring that any person whose rights under EU law are violated has the right to a fair and just legal remedy. This includes the necessity for transparency and clarity in any judicial or quasijudicial process.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Right to an effective remedy and to a fair trial	1	1	1	1	1	1	9.55

Article 48 of the Charter guarantees the presumption of innocence until proven guilty in legal proceedings. It mirrors Article 6(2) of the European Convention on Human Rights (ECHR) and applies both in criminal law and in procedures where severe sanctions are imposed. This principle also applies to EU competition law and other legal proceedings involving sanctions. The right ensures that the burden of proof lies with the prosecuting authority and promotes fairness in judicial decisions.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Presumption of innocence and right of defence	1	1	1	1	1	1	9.55

Article 49 of the Charter protects the principle that criminal offenses and penalties must be clearly defined by law (legality) and must not be excessive in relation to the offense (proportionality). These principles limit both national and EU powers to create or enforce criminal laws, ensuring fairness throughout the legislative, interpretative, and enforcement stages of EU and national law.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Principles of legality and proportionality of criminal offences and penalties	1	1	1	1	1	1	9.55

Article 50 of the Charter enshrines the principle of ne bis in idem, preventing an individual from being tried or punished again for the same offence within the EU after a final verdict. This principle ensures fairness in criminal proceedings, safeguards finality of legal judgments, and applies broadly across the EU's criminal law framework and administrative sanctions resembling criminal law.

	Severity - Intensity	Severity - Effort of remediation	Severity - Severity Level	Probability of Occurrence (PO) - Likelihood	Probability of Occurrence (PO) - PO Level	Impact Significance (%)	FRIAct Score
Right not to be tried or punished twice in criminal proceedings for the same criminal offence	1	1	1	1	1	1	9.55

Bibliography

[1] European Parliament and Council of the European Union, "Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending certain union legislative acts." `https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689`(), 2024. Official Journal of the European Union, L 1689, 13 June 2024.

[2] European Parliament and Council of the European Union, "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)." `https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679`, 2016. Official Journal of the European Union, L119, 4 May 2016, pp. 1–88.

[3] High-Level Expert Group on Artificial Intelligence (AI HLEG), "Ethics guidelines for trustworthy ai," guidelines, European Commission, 2019. Accessed: 24-10-2025.

[4] Article 29 Working Party, "Guidelines on data protection impact assessment (dpia) and determining whether processing is "likely to result in a high risk" for the purposes of regulation 2016/679," tech. rep., European Commission, 2017. WP248 rev.01, adopted on 4 October 2017, endorsed by the European Data Protection Board on 25 May 2018. Accessed: 2025-10-24.

[5] G. Migliorelli, L. Romeo, and E. Frontoni, "FAITH-FDSS: Fiscal AI for Trustworthy and High-compliance Federated Decision Support Systems." Manuscript under submission, 2025.

[6] G. Migliorelli, L. Romeo, and E. Frontoni, "COHESIA: a Cohesive Impact Assessment Framework for DPIA and FRIA with Applications to Four Case Studies." Manuscript under submission, 2025.

[7] M. Di Cosmo, G. Migliorelli, F. P. Villani, M. Francioni, A. Muçaj, E. Frontoni, S. Moccia, and M. C. Fiorentino, "FedStenoNet X-ray Coronary Angiography Dataset for Stenosis Detection," Oct. 2025.

[8] M. C. Fiorentino, G. Migliorelli, F. P. Villani, E. Frontoni, and S. Moccia, "Contrastive prototype federated learning against noisy labels in fetal standard plane detection," *International Journal of Computer Assisted Radiology and Surgery*, vol. 20, no. 7, pp. 1431–1439, 2025.

[9] G. Migliorelli, R. Pietrini, A. Galdelli, E. Frontoni, and M. Paolanti, "FashionDSS: A Human-in-the-Loop Decision Support System for forecasting production efficiency and predicting item defects in the luxury fashion industry." Manuscript under submission, 2025.

[10] G. Migliorelli, R. Pietrini, A. Galdelli, E. Frontoni, and M. Paolanti, "FashionSight: A Human-in-the-Loop Decision Support System for Sales Forecasting in the Luxury Fashion Industry." Manuscript under submission, 2025.

[11] M. Di Cosmo, G. Migliorelli, M. Francioni, A. Muçaj, A. Maolo, A. Aprile, E. Frontoni, M. C. Fiorentino, and S. Moccia, "A federated learning framework for stenosis detection," in *Image Analysis and Processing - ICIAP 2023 Workshops* (G. L. Foresti, A. Fusiello, and E. Hancock, eds.), (Cham), pp. 211–222, Springer Nature Switzerland, 2024.

[12] M. Francioni, A. Muçaj, A. Maolo, A. Aprile, R. Manfredi, M. D. Cosmo, M. C. Fiorentino, G. Migliorelli, E. Frontoni, S. Moccia, and M. Marini, "Clinical insights into coronary angiography heterogeneity: a federated learning investigation." Manuscript being revised for resubmission, 2025.

[13] European Council, "European council meeting (19 october 2017) – conclusions." EUCO 14/17, CO EUR 17, CONCL 5, 2017. (OR. en).

[14] E. C. H.-L. E. G. on Artificial Intelligence, "Policy and investment recommendations for trustworthy ai," tech. rep., European Commission, 2019. Accessed: 2025-10-24.

[15] E. Commission, "White paper on artificial intelligence: A european approach to excellence and trust," tech. rep., European Commission, 2020. Accessed: 2025-10-24.

[16] "Assessment list for trustworthy ai (altai)," tech. rep., European Commission, High-Level Expert Group on Artificial Intelligence, 2020. Accessed: 24-October-2025.

[17] "Regulation (eu) 2018/1807 of the european parliament and of the council of 14 november 2018 on a framework for the free flow of non-personal data in the european union." Official Journal of the European Union, L 303, 28.11.2018, pp. 59–68, Nov. 2018. Available at: `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32018R1807`.

[18] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, 2017.

[19] P. Gazzola, D. Grechi, I. Iliashenko, and R. Pezzetti, "The evolution of digitainability in the fashion industry: a bibliometric analysis," *Kybernetes*, vol. 53, no. 13, pp. 101–126, 2024.

[20] M. Pavan and L. Samant, "Digitalization and e-commerce trends in the industry," in *Consumption and Production in the Textile and Garment Industry: A Comparative Study Among Asian Countries*, pp. 253–272, Springer, 2024.

[21] A. M. Pereira, J. A. B. Moura, E. D. B. Costa, T. Vieira, A. R. Landim, E. Bazaki, and V. Wanick, "Customer models for artificial intelligence-based decision support in fashion online retail supply chains," *Decision Support Systems*, vol. 158, p. 113795, 2022.

[22] L. Chen, H. Halepoto, C. Liu, X. Yan, and L. Qiu, "Research on influencing mechanism of fashion brand image value creation based on consumer value co-creation and experiential value perception theory," *Sustainability*, vol. 14, no. 13, p. 7524, 2022.

[23] K. Lysenko-Ryba and D. Zimon, "Customer behavioral reactions to negative experiences during the product return," *Sustainability*, vol. 13, no. 2, p. 448, 2021.

[24] S. Chakraborty, M. Moore, and L. Parrillo-Chapman, "Automatic defect detection for fabric printing using a deep convolutional neural network," *International Journal of Fashion Design, Technology and Education*, vol. 15, no. 2, pp. 142–157, 2022.

[25] P. Zajec, J. M. Rožanec, S. Theodoropoulos, M. Fontul, E. Koehorst, B. Fortuna, and D. Mladenić, "Few-shot learning for defect detection in manufacturing," *International Journal of Production Research*, pp. 1–20, 2024.

[26] M. Alkahtani, A. Choudhary, A. De, and J. A. Harding, "A decision support system based on ontology and data mining to improve design using warranty data," *Computers & industrial engineering*, vol. 128, pp. 1027–1039, 2019.

[27] M. Eldessouki, "Computer vision and its application in detecting fabric defects," in *Applications of computer vision in fashion and textiles*, pp. 61–101, Elsevier, 2018.

[28] F. Barrera, M. Segura, and C. Maroto, "Multiple criteria decision support system for customer segmentation using a sorting outranking method," *Expert Systems with Applications*, vol. 238, p. 122310, 2024.

[29] M. Wang, X. Li, Y. Liu, P. Chau, and Y. Chen, "A contrast-composition-distraction framework to understand product photo background's impact on consumer interest in e-commerce," *Decision Support Systems*, vol. 178, p. 114124, 2024.

[30] F. Caro and A. S. de Tejada Cuenca, "Believing in analytics: Managers' adherence to price recommendations from a dss," *Manufacturing & Service Operations Management*, vol. 25, no. 2, pp. 524–542, 2023.

[31] A. G. Oskouei, M. A. Balafar, and C. Motamed, "Rdeic-lfw-dss: Resnet-based deep embedded image clustering using local feature weighting and dynamic sample selection mechanism," *Information Sciences*, vol. 646, p. 119374, 2023.

[32] A. L. Loureiro, V. L. Miguéis, and L. F. Da Silva, "Exploring the use of deep neural networks for sales forecasting in fashion retail," *Decision Support Systems*, vol. 114, pp. 81–93, 2018.

[33] M.-A. Filz, J. P. Bosse, and C. Herrmann, "Digitalization platform for data-driven quality management in multi-stage manufacturing systems," *Journal of Intelligent Manufacturing*, vol. 35, no. 6, pp. 2699–2718, 2024.

[34] L. Wang and Z. Liu, "Data-driven product design evaluation method based on multi-stage artificial neural network," *Applied Soft Computing*, vol. 103, p. 107117, 2021.

[35] A. C. Silva, J. Machado, and P. Sampaio, "Predictive quality model for customer defects," *The TQM Journal*, vol. 36, no. 9, pp. 155–174, 2024.

[36] A. Nikseresht, S. Shokouhyar, E. B. Tirkolaee, E. Nikookar, and S. Shokoohyar, "An intelligent decision support system for warranty claims forecasting: Merits of social media and quality function deployment," *Technological Forecasting and Social Change*, vol. 201, p. 123268, 2024.

[37] M. Babakmehr, S. Baumanns, A. Chehade, T. Hochkirchen, M. Kalantari, V. Krivtsov, and D. Schindler, "Data-driven framework for warranty claims forecasting with an application for automotive components," *Engineering Reports*, vol. 6, no. 5, p. e12764, 2024.

[38] R. Rasheed, F. Qazi, A. Ahmed, A. Asif, H. Shams, *et al.*, "Machine learning approaches for in-vehicle failure prognosis in automobiles: A review," *VFAST Transactions on Software Engineering*, vol. 12, no. 1, pp. 169–182, 2024.

[39] 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.

[40] A. Avogaro, L. Capogrosso, A. Toaiari, F. Fummi, and M. Cristani, "New fashion products performance forecasting: A survey on evolutions, models and emerging trends," *arXiv preprint arXiv:2501.10324*, 2025.

[41] S. Anitha and R. Neelakandan, "Demand forecasting new fashion products: A review paper," *Journal of Forecasting*, vol. 44, no. 2, pp. 270–280, 2025.

[42] M. Sousa, A. Loureiro, and V. Miguéis, "Predicting demand for new products in fashion retailing using censored data," *Expert Systems with Applications*, vol. 259, p. 125313, 2025.

[43] M. Rajendran and B. Hong, "Autoregressive multimodal transformer for zero-shot sales forecasting of fashion products with exogenous data," *Applied Intelligence*, vol. 55, no. 2, pp. 1–16, 2025.

[44] B. Bindi, R. Bandinelli, V. Fani, and M. E. P. Pero, "Supply chain strategy in the luxury fashion industry: impacts on performance indicators," *International Journal of Productivity and Performance Management*, vol. 72, no. 5, pp. 1338–1367, 2023.

[45] B. Shen, T. Zhang, X. Xu, H.-L. Chan, and T.-M. Choi, "Preordering in luxury fashion: Will additional demand information bring negative effects to the retailer?," *Decision Sciences*, vol. 53, no. 4, pp. 681–711, 2022.

[46] A. Dhaliwal, D. P. Singh, and J. Paul, "The consumer behavior of luxury goods: A review and research agenda," *Journal of Strategic Marketing*, vol. 33, no. 1, pp. 66–92, 2025.

[47] M. Arribas-Ibar, N. Arimany-Serrat, and P. A. Nylund, "Cognitive biases in innovation ecosystems: global luxury fashion on the eve of post covid-19 recovery," *International Journal of Business and Globalisation*, vol. 40, no. 2, pp. 107–129, 2025.

[48] L. Hu, M. Olivieri, M. Giovannetti, and E. Cedrola, "The retail strategies of luxury fashion firms in the metaverse: Enhancing brand experiences," *Journal of Retailing and Consumer Services*, vol. 84, p. 104202, 2025.

[49] A. R. Chowdhury, R. Paul, and F. Z. Rozony, "A systematic review of demand forecasting models for retail e-commerce enhancing accuracy in inventory and delivery planning," *International Journal of Scientific Interdisciplinary Research*, vol. 6, no. 1, pp. 01–27, 2025.

[50] B. Lindemann, T. Müller, H. Vietz, N. Jazdi, and M. Weyrich, "A survey on long short-term memory networks for time series prediction," *Procedia Cirp*, vol. 99, pp. 650–655, 2021.

[51] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[52] W. Liao, S. Wang, D. Yang, Z. Yang, J. Fang, C. Rehtanz, and F. Porté-Agel, "Timegpt in load forecasting: A large time series model perspective," *Applied Energy*, vol. 379, p. 124973, 2025.

[53] S. Ren, H.-L. Chan, and P. Ram, "A comparative study on fashion demand forecasting models with multiple sources of uncertainty," *Annals of Operations Research*, vol. 257, pp. 335–355, 2017.

[54] T. Van Nguyen, L. Zhou, A. Y. L. Chong, B. Li, and X. Pu, "Predicting customer demand for remanufactured products: A data-mining approach," *European Journal of Operational Research*, vol. 281, no. 3, pp. 543–558, 2020.

[55] Y. Dai and J. Huang, "A sales forecast method for products with no historical data," in *2021 IEEE 6th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA)*, pp. 229–233, IEEE, 2021.

[56] V. L. Miguéis, A. Pereira, J. Pereira, and G. Figueira, "Reducing fresh fish waste while ensuring availability: Demand forecast using censored data and machine learning," *Journal of cleaner Production*, vol. 359, p. 131852, 2022.

[57] C. Giri, S. Thomassey, J. Balkow, and X. Zeng, "Forecasting new apparel sales using deep learning and nonlinear neural network regression," in *2019 International Conference on Engineering, Science, and Industrial Applications (ICESI)*, pp. 1–6, IEEE, 2019.

[58] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

[59] R. Van Steenbergen and M. R. Mes, "Forecasting demand profiles of new products," *Decision support systems*, vol. 139, p. 113401, 2020.

[60] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Information Sciences*, vol. 622, pp. 178–210, 2023.

[61] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[62] N. Meinshausen and G. Ridgeway, "Quantile regression forests.," *Journal of machine learning research*, vol. 7, no. 6, 2006.

[63] A. Avogaro, L. Capogrosso, F. Fummi, and M. Cristani, "Mdiff: Exploiting multimodal score-based diffusion models for new fashion product performance forecasting," in *European Conference on Computer Vision*, pp. 337–351, Springer, 2025.

[64] J. Lin and H. Li, "A short-term pv power forecasting method using a hybrid kmeans-gra-svr model under ideal weather condition," *Journal of Computer and Communications*, vol. 8, no. 11, pp. 102–119, 2020.

[65] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[66] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[67] O. Sagi and L. Rokach, "Ensemble learning: A survey," *Wiley interdisciplinary reviews: data mining and knowledge discovery*, vol. 8, no. 4, p. e1249, 2018.

[68] A. Suilin, "Kaggle-web-traffic," in *GitHub*, 2017.

[69] S. Kim and H. Kim, "A new metric of absolute percentage error for intermittent demand forecasts," *International Journal of Forecasting*, vol. 32, no. 3, pp. 669–679, 2016.

[70] S. Larson and M. Overton, "Modeling approach matters, but not as much as preprocessing: Comparison of machine learning and traditional revenue forecasting techniques," *Public Finance Journal*, vol. 1, 03 2024.

[71] Z. Lin, "Tax share analysis and prediction of kernel extreme learning machine optimized by vector weighted average algorithm." `https://doi.org/10.21203/rs.3.rs-6763689/v1`, 2025. Preprint, Research Square.

[72] L. Zhu and W.-T. Pan, "Methodology and application of fiscal and tax forecasting analysis based on multi-source big data fusion," *Mathematical Problems in Engineering*, vol. 2022, pp. 1–12, June 2022.

[73] C. Swenson, "Using machine deep learning ai to improve forecasting of tax payments for corporations," *Forecasting*, vol. 6, pp. 968–984, 10 2024.

[74] D. Byrd and A. Polychroniadou, "Differentially private secure multi-party computation for federated learning in financial applications," in *Proceedings of the First ACM International Conference on AI in Finance*, ICAIF '20, (New York, NY, USA), Association for Computing Machinery, 2021.

[75] M. I. Rawanda, K. Nisa, S. Mufti, S. Ansarullah, S. Ikhlaq, and T. Yousuf, *Artificial Intelligence in Tax Compliance: Transforming Taxpayer Behavior and System Efficiency*, pp. 251–270. IGI Global, 05 2025.

[76] K. Komarudin and P. Hermawan, "A study of trust base voluntary tax compliance through tax administration digital transformation in indonesia," *International Journal of Current Science Research and Review*, vol. 05, 08 2022.

[77] A. Lopo Martinez, "Artificial intelligence in tax administration: Enhancing compliance, transparency, and ethical governance," *SSRN Electronic Journal*, 01 2025.

[78] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Aguera y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS 2017)*, vol. 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282, PMLR, 2017.

[79] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proceedings of the 3rd Machine Learning and Systems Conference (MLSys 2020)*, 2020.

[80] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proceedings of the 3rd Theory of Cryptography Conference (TCC '06)*, pp. 265–284, Springer, 2006.

[81] C. Wei, W. Li, G. Chen, and W. Chen, "Differentially private sgd with dynamic clipping through gradient norm distribution estimation," *arXiv preprint arXiv:2503.22988*, 2025.

[82] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, (Red Hook, NY, USA), p. 4768–4777, Curran Associates Inc., 2017.

[83] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *Proceedings of the 33rd International Conference on Machine Learning (ICML 2016)*, pp. 1050–1059, PMLR, 2016.

[84] Y. Gal, *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge, 2016.

[85] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," 2017.

[86] P. S. Foundation, *Python Language Reference, version 3.8.10*, 2023.

[87] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[88] D. J. Beutel, T. Topal, A. Mathur, X. Qiu, J. Fernandez-Marques, Y. Gao, L. Sani, H. L. Kwing, T. Parcollet, P. P. d. Gusmão, and N. D. Lane, "Flower: A friendly federated learning research framework," *arXiv preprint arXiv:2007.14390*, 2020.

[89] A. Yousefpour, I. Shilov, A. Sablayrolles, D. Testuggine, K. Prasad, M. Malek, J. Nguyen, S. Ghosh, A. Bharadwaj, J. Zhao, G. Cormode, and I. Mironov, "Opacus: User-friendly differential privacy library in pytorch," *arXiv preprint arXiv:2109.12298*, 2021.

[90] I. Mironov, "Rényi differential privacy," in *2017 IEEE 30th computer security foundations symposium (CSF)*, pp. 263–275, IEEE, 2017.

[91] 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.

[92] M. A.-u.-d. Khan, M. F. Uddin, and N. Gupta, "Seven v's of big data understanding big data to extract value," in *Proceedings of the 2014 Zone 1 Conference of the American Society for Engineering Education*, pp. 1–5, 2014.

[93] V. V. Danilov, K. Y. Klyshnikov, O. M. Gerget, A. G. Kutikhin, V. I. Ganyukov, A. F. Frangi, and E. A. Ovcharenko, "Real-time coronary artery stenosis detection based on modern neural networks," *Scientific Reports*, vol. 11, no. 1, p. 7582, 2021.

[94] M. Popov, A. Amanturdieva, N. Zhaksylyk, A. Alkanov, A. Saniyazbekov, T. Aimyshev, E. Ismailov, A. Bulegenov, A. Kuzhukeyev, A. Kulanbayeva, *et al.*, "Dataset for automatic region-based coronary artery disease diagnostics using x-ray angiography images," *Scientific Data*, vol. 11, no. 1, p. 20, 2024.

[95] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, 2019.

[96] G. Hinton and S. Roweis, "Stochastic neighbor embedding," in *Proceedings of the 16th International Conference on Neural Information Processing Systems*, NIPS'02, (Cambridge, MA, USA), p. 857–864, MIT Press, 2002.

[97] D. Han, J. Liu, Z. Sun, Y. Cui, Y. He, and Z. Yang, "Deep learning analysis in coronary computed tomographic angiography imaging for the assessment of patients with coronary artery stenosis," *Computer Methods and Programs in Biomedicine*, vol. 196, p. 105651, 2020.

[98] M. Penso, S. Moccia, E. G. Caiani, G. Caredda, M. L. Lampus, M. L. Carerj, M. Babbaro, M. Pepi, M. Chiesa, and G. Pontone, "A token-mixer architecture for CAD-RADS classification of coronary stenosis on multiplanar reconstruction CT images," *Computers in Biology and Medicine*, vol. 153, p. 106484, 2023.

[99] L. Bolognese and M. R. Reccia, "Computed tomography to replace invasive coronary angiography? the discharge trial," *European Heart Journal Supplements*, vol. 24, no. Supplement_I, pp. I25–I28, 2022.

[100] A. Garavand, A. Behmanesh, N. Aslani, H. Sadeghsalehi, and M. Ghaderzadeh, "Towards diagnostic aided systems in coronary artery disease detection: a comprehensive multiview survey of the state of the art," *International Journal of Intelligent Systems*, vol. 2023, pp. 1–19, 2023.

[101] J. S. Lawton, J. E. Tamis-Holland, S. Bangalore, E. R. Bates, T. M. Beckie, J. M. Bischoff, J. A. Bittl, M. G. Cohen, J. M. DiMaio, C. W. Don, *et al.*, "2021 ACC/AHA/SCAI guideline for coronary artery revascularization: executive summary: a report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines," *Circulation*, vol. 145, no. 3, pp. e4–e17, 2022.

[102] C. Cong, Y. Kato, H. D. Vasconcellos, J. Lima, and B. Venkatesh, "Automated stenosis detection and classification in x-ray angiography using deep neural network," in *2019 IEEE International Conference on Bioinformatics and Biomedicine*, pp. 1301–1308, IEEE, 2019.

[103] J. H. Moon, W. C. Cha, M. J. Chung, K.-S. Lee, B. H. Cho, J. H. Choi, *et al.*, "Automatic stenosis recognition from coronary angiography using convolutional neural networks," *Computer Methods and Programs in Biomedicine*, vol. 198, p. 105819, 2021.

[104] E. Ovalle-Magallanes, J. G. Avina-Cervantes, I. Cruz-Aceves, and J. Ruiz-Pinales, "Improving convolutional neural network learning based on a hierarchical bezier generative model for stenosis detection in x-ray images," *Computer Methods and Programs in Biomedicine*, vol. 219, p. 106767, 2022.

[105] D. Zhang, G. Yang, S. Zhao, Y. Zhang, H. Zhang, and S. Li, "Direct quantification for coronary artery stenosis using multiview learning," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pp. 449–457,

Springer, 2019.

[106] C.-K. Lee, J.-W. Hong, C.-L. Wu, J.-M. Hou, Y.-A. Lin, K.-C. Huang, and P.-H. Tseng, "Real-time coronary artery segmentation in cag images: A semi-supervised deep learning strategy," *Artificial Intelligence in Medicine*, p. 102888, 2024.

[107] A. Rauniyar, D. H. Hagos, D. Jha, J. E. Håkegård, U. Bagci, D. B. Rawat, and V. Vlassov, "Federated learning for medical applications: A taxonomy, current trends, challenges, and future research directions," *IEEE Internet of Things Journal*, 2023.

[108] D. C. Nguyen, Q.-V. Pham, P. N. Pathirana, M. Ding, A. Seneviratne, Z. Lin, O. Dobre, and W.-J. Hwang, "Federated learning for smart healthcare: A survey," *ACM Computing Surveys (CSUR)*, vol. 55, no. 3, pp. 1–37, 2022.

[109] H. Guan and M. Liu, "Domain adaptation for medical image analysis: A survey," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 3, pp. 1173–1185, 2022.

[110] M. Jiang, H. Yang, C. Cheng, and Q. Dou, "Iop-fl: Inside-outside personalization for federated medical image segmentation," *IEEE Transactions on Medical Imaging*, 2023.

[111] Q. Liu, C. Chen, J. Qin, Q. Dou, and P.-A. Heng, "Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1013–1023, 2021.

[112] R. Zhang, Z. Fan, Q. Xu, J. Yao, Y. Zhang, and Y. Wang, "Grace: A generalized and personalized federated learning method for medical imaging," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 14–24, Springer, 2023.

[113] M. Di Cosmo, G. Migliorelli, M. Francioni, A. Muçaj, A. Maolo, A. Aprile, E. Frontoni, M. C. Fiorentino, and S. Moccia, "A federated learning framework for stenosis detection," in *International Conference on Image Analysis and Processing*, pp. 211–222, Springer, 2023.

[114] R. Mineo, A. Sorrenti, and F. Proietto Salanitri, "Fedetr: A federated approach for stenosis detection in coronary angiography," in *International Conference on Image Analysis and Processing*, pp. 189–200, Springer, 2023.

[115] Y. Wang, R. Zhang, S. Zhang, M. Li, Y. Xia, X. Zhang, and S. Liu, "Domain-specific suppression for adaptive object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9603–9612, 2021.

[116] J. Chu, J. Chen, X. Chen, W. Dong, J. Shi, and Z. Huang, "Knowledge-aware multi-center clinical dataset adaptation: Problem, method, and application," *Journal of Biomedical Informatics*, vol. 115, p. 103710, 2021.

[117] V. Vs, V. Gupta, P. Oza, V. A. Sindagi, and V. M. Patel, "Mega-cda: Memory guided attention for category-aware unsupervised domain adaptive object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4516–4526, 2021.

[118] S. Kim, J. Choi, T. Kim, and C. Kim, "Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6092–6101, 2019.

[119] G. Zhao, G. Li, R. Xu, and L. Lin, "Collaborative training between region proposal localization and classification for domain adaptive object detection," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pp. 86–102, Springer, 2020.

[120] C. Chen, Z. Zheng, X. Ding, Y. Huang, and Q. Dou, "Harmonizing transferability and discriminability for adapting object detectors," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8869–8878, 2020.

[121] P. Mateus, J. Moonen, M. Beran, E. Jaarsma, S. M. van der Landen, J. Heuvelink, M. Birhanu, A. G. Harms, E. Bron, F. J. Wolters, *et al.*, "Data harmonization and federated learning for multi-cohort dementia research using the omop common data model: A netherlands consortium of dementia cohorts case study," *Journal of biomedical informatics*, p. 104661, 2024.

[122] Y. Chen, H. Wang, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, "Scale-aware domain adaptive faster r-cnn," *International Journal of Computer Vision*, vol. 129, no. 7, pp. 2223–2243, 2021.

[123] F. Yu, D. Wang, Y. Chen, N. Karianakis, T. Shen, P. Yu, D. Lymberopoulos, S. Lu, W. Shi, and X. Chen, "Sc-uda: Style and content gaps aware unsupervised domain adaptation for object detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 382–391, 2022.

[124] T. Kim, M. Jeong, S. Kim, S. Choi, and C. Kim, "Diversify and match: A domain adaptive representation learning paradigm for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12456–12465, 2019.

[125] S. Cao, D. Joshi, L.-Y. Gui, and Y.-X. Wang, "Contrastive mean teacher for domain adaptive object detectors," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 23839–23848, 2023.

[126] J. Deng, D. Xu, W. Li, and L. Duan, "Harmonious teacher for cross-domain object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23829–23838, 2023.

[127] J. Ma, "Histogram matching augmentation for domain adaptation with application to multi-centre, multi-vendor and multi-disease cardiac image segmentation," in *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges: 11th International Workshop, STACOM 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers 11*, pp. 177–186, Springer, 2021.

[128] B. Pu, L. Wang, J. Yang, G. He, X. Dong, S. Li, Y. Tan, M. Chen, Z. Jin, K. Li, *et al.*, "M3-uda: A new benchmark for unsupervised domain adaptive fetal cardiac structure detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11621–11630, 2024.

[129] J. Liang, R. He, and T. Tan, "A comprehensive survey on test-time adaptation under distribution shifts," *arXiv preprint arXiv:2303.15361*, 2023.

[130] S. Winther, S. Schmidt, L. D. Rasmussen, L. J. J. Orozco, F. H. Steffensen, H. Bøtker, J. Knuuti, and M. Bøttcher, "Validation of the european society of cardiology pre-test probability model for obstructive coronary artery disease.," *European heart journal*, 2020.

[131] G. Sianos, M.-A. Morel, A. P. Kappetein, M.-C. Morice, A. Colombo, K. Dawkins, M. van den Brand, N. Van Dyck, M. E. Russell, F. W. Mohr, *et al.*, "The syntax score: an angiographic tool grading the complexity of coronary artery disease," *EuroIntervention*, vol. 1, no. 2, pp. 219–227, 2005.

[132] M. R. Patel, J. H. Calhoon, G. J. Dehmer, J. A. Grantham, T. M. Maddox, D. J. Maron, and P. K. Smith, "Acc/aats/aha/ase/asnc/scai/scct/sts 2017 appropriate use criteria for coronary revascularization in patients with stable ischemic heart disease: a report of the american college of cardiology appropriate use criteria task force, american association for thoracic surgery, american heart association, american society of echocardiography, american society of nuclear cardiology, society for cardiovascular angiography and interventions, society of cardiovascular computed tomography, and society of thoracic surgeons," *Journal of the American College of Cardiology*, vol. 69, no. 17, pp. 2212–2241, 2017.

[133] I. Ben-Dor, M. Mahmoudi, T. Deksissa, A. B. Bui, M. A. Gaglia, M. A. Gonzalez, G. Maluenda, G. Sardi, R. Romaguera, A. Laynez-Carnicero, *et al.*, "Correlation between fractional flow reserve and intravascular ultrasound lumen area in intermediate coronary artery stenosis," *Journal of the American College of Cardiology*, vol. 57, no. 14, p. E1855, 2011.

[134] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *13th European Conference*, pp. 740–755, Springer, 2014.

[135] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 1137–1149, 2015.

[136] P. V. Stralen, D. L. Rodrigues, A. L. Oliveira, M. N. Menezes, and F. J. Pinto, "Stenosis detection in x-ray coronary angiography with deep neural networks leveraged by attention mechanisms," in *Proceedings of the 9th International Conference on Bioinformatics Research and Applications*, pp. 123–128, 2022.

[137] Y. Li, T. Yoshimura, Y. Horima, and H. Sugimori, "A preprocessing method for coronary artery stenosis detection based on deep learning," *Algorithms*, vol. 17, no. 3, p. 119, 2024.

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

[139] M. Zhang, S. Levine, and C. Finn, "Memo: Test time robustness via adaptation and augmentation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 38629–38642, 2022.

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

[141] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.

[142] T. Han, D. Ai, X. Li, J. Fan, H. Song, Y. Wang, and J. Yang, "Coronary artery stenosis detection via proposal-shifted spatial-temporal transformer in x-ray angiography," *Computers in Biology and Medicine*, p. 106546, 2023.

[143] K. Pang, D. Ai, H. Fang, J. Fan, H. Song, and J. Yang, "Stenosis-DetNet: Sequence consistency-based stenosis detection for x-ray coronary angiography," *Computerized Medical Imaging and Graphics*, vol. 89, p. 101900, 2021.

[144] E. M. Nwanosike, B. R. Conway, H. A. Merchant, and S. S. Hasan, "Potential applications and performance of machine learning techniques and algorithms in clinical practice: a systematic review," *International journal of medical informatics*, vol. 159, p. 104679, 2022.

[145] R. P. Evans, L. D. Bryant, G. Russell, and K. Absolom, "Trust and acceptability of data-driven clinical recommendations in everyday practice: a scoping review," *International Journal of Medical Informatics*, vol. 183, p. 105342, 2024.

[146] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 702–703, 2020.

[147] F. Garcea, A. Serra, F. Lamberti, and L. Morra, "Data augmentation for medical imaging: A systematic literature review," *Computers in Biology and Medicine*, vol. 152, p. 106391, 2023.

[148] C. Jones, D. C. Castro, F. De Sousa Ribeiro, O. Oktay, M. McCradden, and B. Glocker, "A causal perspective on dataset bias in machine learning for medical imaging," *Nature Machine Intelligence*, vol. 6, no. 2, pp. 138–146, 2024.

[149] B. Koçak, A. Ponsiglione, A. Stanzione, C. Bluethgen, J. Santinha, L. Ugga, M. Huisman, M. E. Klontzas, R. Cannella, R. Cuocolo, *et al.*, "Bias in artificial intelligence for medical imaging: fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects," *Diagn Interv Radiol*, 2024.

[150] M. C. Fiorentino, F. P. Villani, R. Benito Herce, M. A. González Ballester, A. Mancini, and K. López-Linares Román, "An intensity-based self-supervised domain adaptation method for intervertebral disc segmentation in magnetic resonance imaging," *International Journal of Computer Assisted Radiology and Surgery*, pp. 1–9, 2024.

[151] M. C. Fiorentino, S. Moccia, E. Cipolletta, E. Filippucci, and E. Frontoni, "A learning approach for informative-frame selection in us rheumatology images," in *New Trends in Image Analysis and Processing–ICIAP 2019: ICIAP International Workshops, BioFor, PatReCH, e-BADLE, DeepRetail, and Industrial Session, Trento, Italy, September 9–10, 2019, Revised Selected Papers 20*, pp. 228–236, Springer, 2019.

[152] G. Migliorelli, M. C. Fiorentino, M. Di Cosmo, F. P. Villani, A. Mancini, and S. Moccia, "On the use of contrastive learning for standard-plane classification in fetal ultrasound imaging," *Computers in Biology and Medicine*, vol. 174, p. 108430, 2024.

[153] M. C. Fiorentino, F. P. Villani, M. Di Cosmo, E. Frontoni, and S. Moccia, "A review on deep-learning algorithms for fetal ultrasound-image analysis," *Medical Image Analysis*, p. 102629, 2022.

[154] J. Bai, Z. Zhou, Z. Ou, G. Koehler, R. Stock, K. Maier-Hein, M. Elbatel, R. Martí, X. Li, Y. Qiu, *et al.*, "PSFHS challenge report: Pubic symphysis and fetal head segmentation from intrapartum ultrasound images," *Medical Image Analysis*, p. 103353, 2024.

[155] T. Kiserud, G. Piaggio, G. Carroli, M. Widmer, J. Carvalho, L. Neerup Jensen, D. Giordano, J. G. Cecatti, H. Abdel Aleem, S. A. Talegawkar, *et al.*, "The world health organization fetal growth charts: a multinational longitudinal study of ultrasound biometric measurements and estimated fetal weight," *PLoS medicine*, vol. 14, no. 1, p. e1002220, 2017.

[156] T. B. Krishna and P. Kokil, "Standard fetal ultrasound plane classification based on stacked ensemble of deep learning models," *Expert Systems with Applications*, vol. 238, p. 122153, 2024.

[157] T. B. Krishna and P. Kokil, "Automated classification of common maternal fetal ultrasound planes using multi-layer perceptron with deep feature integration," *Biomedical Signal Processing and Control*, vol. 86, p. 105283, 2023.

[158] A. Jiménez-Sánchez, M. Tardy, M. A. G. Ballester, D. Mateus, and G. Piella, "Memory-aware curriculum federated learning for breast cancer classification," *Computer Methods and Programs in Biomedicine*, vol. 229, p. 107318, 2023.

[159] C. Sendra-Balcells, V. M. Campello, J. Torrents-Barrena, Y. A. Ahmed, M. Elattar, B. Ohene-Botwe, P. Nyangulu, W. Stones, M. Ammar, L. N. Benamer, *et al.*, "Generalisability of fetal ultrasound deep learning models to low-resource imaging settings in five african countries," *Scientific Reports*, vol. 13, no. 1, p. 2728, 2023.

[160] T. B. Krishna and P. Kokil, "A deep convolutional neural network with adaptive channel weight technique for automated identification of standard fetal biometry planes," *IEEE Transactions on Instrumentation and Measurement*, 2024.

[161] B. Liu, N. Lv, Y. Guo, and Y. Li, "Recent advances on federated learning: A systematic survey," *Neurocomputing*, p. 128019, 2024.

[162] N. Orlando, I. Gyacskov, D. J. Gillies, F. Guo, C. Romagnoli, D. D'Souza, D. W. Cool, D. A. Hoover, and A. Fenster, "Effect of dataset size, image quality, and image type on deep learning-based automatic prostate segmentation in 3D ultrasound," *Physics in Medicine & Biology*, vol. 67, no. 7, p. 074002, 2022.

[163] R. Yan, L. Qu, Q. Wei, S.-C. Huang, L. Shen, D. L. Rubin, L. Xing, and Y. Zhou, "Label-efficient self-supervised federated learning for tackling data heterogeneity in medical imaging," *IEEE Transactions on Medical Imaging*, vol. 42, no. 7, pp. 1932–1943, 2023.

[164] J. Xu, Z. Chen, T. Q. Quek, and K. F. E. Chong, "Fedcorr: Multi-stage federated learning for label noise correction," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10184–10193, 2022.

[165] Z. Chen, W. Li, X. Xing, and Y. Yuan, "Medical federated learning with joint graph purification for noisy label learning," *Medical Image Analysis*, vol. 90, p. 102976, 2023.

[166] X. P. Burgos-Artizzu, D. Coronado-Gutiérrez, B. Valenzuela-Alcaraz, E. Bonet-Carne, E. Eixarch, F. Crispi, and E. Gratacós, "Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes," *Scientific Reports*, vol. 10, no. 1, pp. 1–12, 2020.

[167] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning*, pp. 1597–1607, PMLR, 2020.

[168] Y. Wang, X. Ma, Z. Chen, Y. Luo, J. Yi, and J. Bailey, "Symmetric cross entropy for robust learning with noisy labels," in *IEEE/CVF International Conference on Computer Vision*, pp. 322–330, 2019.

[169] C. Tan, J. Xia, L. Wu, and S. Z. Li, "Co-learning: Learning from noisy labels with self-supervision," in *29th ACM International Conference on Multimedia*, pp. 1405–1413, 2021.

[170] European Union, "Charter of fundamental rights of the european union." `https://eur-lex.europa.eu/eli/treaty/char_2012/oj/eng`(), 2012. Official Journal of the European Union, C 326, 26 October 2012, pp. 391–407.

[171] H. J. Pandit and T. Rintamäki, "Towards an automated ai act fria tool that can reuse gdpr's dpia," in *CLAIRvoyant: Convention on Artificial Intelligence Regulation Workshop*, (Dublin City University, Dublin, Ireland), 2024.

[172] Italian Republic, "Gazzetta ufficiale della repubblica italiana, serie generale n. 269, 19 november 2018." Official Gazette of the Italian Republic, 2018. General Series No. 269 of 19 November 2018.

[173] P. Ceravolo *et al.*, "Hh4ai: A methodological framework for ai human rights impact assessment under the eu ai act," in *HH4AI Workshop*, (Università degli Studi di Milano, Deloitte Financial Advisory S.r.l. S.B., Milan, Italy), 2024.

[174] CNIL, "Pia tool (v 3.0, june 2021)." `https://www.cnil.fr/en/privacy-impact-assessment-cnil-releases-version-30-its-pia-software`. accessed 21 October 2025.

[175] A. Cosentini *et al.*, "Assessing the impact of artificial intelligence systems on fundamental rights," *Media Laws (FRIAct)*, 2025.

[176] Republic of Italy, "Law no. 241 of 7 august 1990: New rules on administrative procedure and the right of access to administrative documents." `https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:legge:1990-08-07;241!vig=`, 1990. Published in the *Gazzetta Ufficiale No. 192*, 18 August 1990.

[177] R. A. Byrne, X. Rossello, J. Coughlan, E. Barbato, C. Berry, A. Chieffo, M. J. Claeys, G.-A. Dan, M. R. Dweck, M. Galbraith, *et al.*, "2023 esc guidelines for the management of acute coronary syndromes: developed by the task force on the

management of acute coronary syndromes of the european society of cardiology (esc)," *European Heart Journal: Acute Cardiovascular Care*, vol. 13, no. 1, pp. 55–161, 2024.

[178] M. Chen, M. Jiang, Q. Dou, Z. Wang, and X. Li, "Fedsoup: Improving generalization and personalization in federated learning via selective model interpolation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 318–328, Springer, 2023.

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

[180] J. Oh, S. Kim, and S.-Y. Yun, "Fedbabu: Toward enhanced representation for federated image classification," in *International Conference on Learning Representations*, 2021.

[181] L. Qu, N. Balachandar, M. Zhang, and D. Rubin, "Handling data heterogeneity with generative replay in collaborative learning for medical imaging," *Medical Image Analysis*, vol. 78, p. 102424, 2022.

[182] X. Li, Y. Gu, N. Dvornek, L. H. Staib, P. Ventola, and J. S. Duncan, "Multi-site fmri analysis using privacy-preserving federated learning and domain adaptation: Abide results," *Medical Image Analysis*, vol. 65, p. 101765, 2020.

[183] Z. Yan, J. Wicaksana, Z. Wang, X. Yang, and K.-T. Cheng, "Variation-aware federated learning with multi-source decentralized medical image data," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 7, pp. 2615–2628, 2020.

[184] M. Cossio, "Augmenting medical imaging: A comprehensive catalogue of 65 techniques for enhanced data analysis," *arXiv preprint arXiv:2303.01178*, 2023.

[185] P. S. Jois, A. Manjunath, and T. Fevens, "Boosting segmentation performance across datasets using histogram specification with application to pelvic bone segmentation," in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1364–1369, IEEE, 2021.

[186] R. Szeliski, "Computer vision: Algorithms and applications," 2010.

[187] A. Sabater, L. Montesano, and A. C. Murillo, "Robust and efficient post-processing for video object detection," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10536–10542, IEEE, 2020.

[188] R. Shad, J. P. Cunningham, E. A. Ashley, C. P. Langlotz, and W. Hiesinger, "Designing clinically translatable artificial intelligence systems for high-dimensional medical imaging," *Nature Machine Intelligence*, vol. 3, no. 11, pp. 929–935, 2021.

[189] M. Vaduganathan, G. A. Mensah, J. V. Turco, V. Fuster, and G. A. Roth, "The global burden of cardiovascular diseases and risk: A compass for future health," 2022.

[190] E. Braunwald, "How to live to 100 before developing clinical coronary artery disease: a suggestion," 2022.

[191] E. J. Benjamin, P. Muntner, A. Alonso, M. S. Bittencourt, and et al., "Heart disease and stroke statistics—2019 update: A report from the American Heart Association," *Circulation*, vol. 139, no. 10, 2019.

[192] A. Saraste and J. Knuuti, "ESC 2019 guidelines for the diagnosis and management of chronic coronary syndromes: Recommendations for cardiovascular imaging," *Herz*, vol. 45, no. 5, p. 409, 2020.

[193] A. Tejero-de Pablos, K. Huang, H. Yamane, Y. Kurose, Y. Mukuta, J. Iho, Y. Tokunaga, M. Horie, K. Nishizawa, Y. Hayashi, *et al.*, "Texture-based classification of significant stenosis in ccta multi-view images of coronary arteries," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pp. 732–740, Springer, 2019.

[194] P. I. Ngam, C. C. Ong, P. Chai, S. S. Wong, C. R. Liang, and L. L. San Teo, "Computed tomography coronary angiography-past, present and future," *Singapore medical journal*, vol. 61, no. 3, p. 109, 2020.

[195] M. Zreik, R. W. Van Hamersvelt, J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Išgum, "A recurrent CNN for automatic detection and classification of coronary artery plaque and stenosis in coronary CT angiography," *IEEE transactions on medical imaging*, vol. 38, no. 7, pp. 1588–1598, 2018.

[196] T. Han, D. Ai, Y. Wang, Y. Bian, R. An, J. Fan, H. Song, H. Xie, and J. Yang, "Recursive centerline-and direction-aware joint learning network with ensemble strategy for vessel segmentation in x-ray angiography images," *Computer Methods and Programs in Biomedicine*, vol. 220, p. 106787, 2022.

[197] C. Zhao, A. Vij, S. Malhotra, J. Tang, H. Tang, D. Pienta, Z. Xu, and W. Zhou, "Automatic extraction and stenosis evaluation of coronary arteries in invasive coronary angiograms," *Computers in Biology and Medicine*, vol. 136, p. 104667, 2021.

[198] E. Ovalle-Magallanes, J. G. Avina-Cervantes, I. Cruz-Aceves, and J. Ruiz-Pinales, "Hybrid classical–quantum convolutional neural network for stenosis detection in x-ray coronary angiography," *Expert Systems with Applications*, vol. 189, p. 116112, 2022.

[199] W. Wu, J. Zhang, H. Xie, Y. Zhao, S. Zhang, and L. Gu, "Automatic detection of coronary artery stenosis by convolutional neural network with temporal constraint," *Computers in Biology and Medicine*, vol. 118, p. 103657, 2020.

[200] 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.

[201] S. Nazir and M. Kaleem, "Federated learning for medical image analysis with deep neural networks," *Diagnostics*, vol. 13, no. 9, p. 1532, 2023.

[202] F. Ślazyk, P. Jabłecki, A. Lisowska, M. Malawski, and S. Płotka, "CXR-FL: Deep learning-based chest x-ray image analysis using federated learning," in *International Conference on Computational Science*, pp. 433–440, Springer, 2022.

[203] Y. Wu, D. Zeng, Z. Wang, Y. Shi, and J. Hu, "Distributed contrastive learning for medical image segmentation," *Medical Image Analysis*, vol. 81, p. 102564, 2022.

[204] M. Y. Lu, R. J. Chen, D. Kong, J. Lipkova, R. Singh, D. F. Williamson, T. Y. Chen, and F. Mahmood, "Federated learning for computational pathology on gigapixel whole slide images," *Medical Image Analysis*, vol. 76, p. 102298, 2022.

[205] J. Xiao, L. Yu, Z. Zhou, Y. Bai, L. Xing, A. Yuille, and Y. Zhou, "CateNorm: Categorical normalization for robust medical image segmentation," in *MICCAI Workshop on Domain Adaptation and Representation Transfer*, pp. 129–146, Springer, 2022.

[206] Y. Li, S. Xie, X. Chen, P. Dollar, K. He, and R. Girshick, "Benchmarking detection transfer learning with vision transformers," *arXiv preprint arXiv:2111.11429*, 2021.

[207] R. Myrzashova, S. H. Alsamhi, A. V. Shvetsov, A. Hawbani, and X. Wei, "Blockchain meets federated learning in healthcare: A systematic review with challenges and opportunities," *IEEE Internet of Things Journal*, 2023.

[208] Q. Yang, J. Zhang, W. Hao, G. P. Spell, and L. Carin, "Flop: Federated learning on medical datasets using partial networks," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 3845–3853, 2021.

[209] Y. Wang, Q. Shi, and T.-H. Chang, "Batch normalization damages federated learning on non-iid data: Analysis and remedy," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2023.

[210] L. Zhao and L. Wang, "Task-specific inconsistency alignment for domain adaptive object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14217–14226, 2022.

[211] H.-K. Hsu, C.-H. Yao, Y.-H. Tsai, W.-C. Hung, H.-Y. Tseng, M. Singh, and M.-H. Yang, "Progressive domain adaptation for object detection," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 749–757, 2020.

[212] Z. Fang, J. Lu, F. Liu, and G. Zhang, "Semi-supervised heterogeneous domain adaptation: Theory and algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 1087–1105, 2022.

[213] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.

[214] S. F. Aktar and S. Andrei, "Bias, ethical concerns, and explainable decision-making in medical imaging research," in *Data Driven Approaches on Medical Imaging*, pp. 179–205, Springer, 2023.

[215] N. Wang, Y. Deng, W. Feng, S. Fan, J. Yin, and S.-K. Ng, "One-shot sequential federated learning for non-iid data by enhancing local model diversity," in *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 5201–5210, 2024.

[216] H. Wu, B. Zhang, C. Chen, and J. Qin, "Federated semi-supervised medical image segmentation via prototype-based pseudo-labeling and contrastive learning," *IEEE Transactions on Medical Imaging*, vol. 43, no. 2, pp. 649–661, 2023.

[217] Y. Zheng, P. Tang, T. Ju, H. Wang, W. Qiu, and J. C. Rajapakse, "Federated semi-supervised learning for medical image segmentation with intra-client and inter-client consistency," in *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 4054–4059, IEEE, 2024.

[218] L. Qiu, J. Cheng, H. Gao, W. Xiong, and H. Ren, "Federated semi-supervised learning for medical image segmentation via pseudo-label denoising," *IEEE journal of biomedical and health informatics*, vol. 27, no. 10, pp. 4672–4683, 2023.

[219] F. Li, P. Li, X. Wu, P. Zeng, G. Lyu, Y. Fan, P. Liu, H. Song, and Z. Liu, "Fhuspnet: a multi-task model for fetal heart ultrasound standard plane recognition and key anatomical structures detection," *Computers in Biology and Medicine*, vol. 168, p. 107741, 2024.

[220] K. G. Sindhu and R. Annamalai, "Enhanced multi-class fetal plane detection with limb localization in ultrasound images," in *2024 IEEE International Conference on Contemporary Computing and Communications*, vol. 1, pp. 1–6, IEEE, 2024.

[221] 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.

[222] H.-Y. Chen and W.-L. Chao, "On bridging generic and personalized federated learning for image classification," *arXiv preprint arXiv:2107.00778*, 2021.

[223] A. Lasala, M. C. Fiorentino, A. Bandini, and S. Moccia, "Fetalbrainawarenet: Bridging gans with anatomical insight for fetal ultrasound brain plane synthesis," *Computerized Medical Imaging and Graphics*, p. 102405, 2024.

[224] E. Verna, G. Genta, M. Galetto, and F. Franceschini, "Zero defect manufacturing: a self-adaptive defect prediction model based on assembly complexity," *International Journal of Computer Integrated Manufacturing*, vol. 36, no. 1, pp. 155–168, 2023.

[225] G. Fouch, "A guide to zero defects. quality and reliability assurance handbook 4155.12-h," 1965.

[226] H.-T. Chen, "Quality function deployment in failure recovery and prevention," *The service industries Journal*, vol. 36, no. 13-14, pp. 615–637, 2016.

[227] F. Psarommatis, G. May, P.-A. Dreyfus, and D. Kiritsis, "Zero defect manufacturing: state-of-the-art review, shortcomings and future directions in research," *International journal of production research*, vol. 58, no. 1, pp. 1–17, 2020.

[228] G. May and D. Kiritsis, "Zero defect manufacturing strategies and platform for smart factories of industry 4.0," in *Proceedings of the 4th International Conference on the Industry 4.0 Model for Advanced Manufacturing: AMP 2019 4*, pp. 142–152, Springer, 2019.

[229] E. Commission, D.-G. for Research, Innovation, A. Renda, S. Schwaag Serger, D. Tataj, A. Morlet, D. Isaksson, F. Martins, M. Mir Roca, C. Hidalgo, A. Huang, S. Dixson-Declève, P. Balland, F. Bria, C. Charveriat, K. Dunlop, and E. Giovannini, *Industry 5.0, a transformative vision for Europe : governing systemic transformations towards a sustainable industry*. Publications Office of the European Union, 2022.

[230] P. K. Wan and T. L. Leirmo, "Human-centric zero-defect manufacturing: State-of-the-art review, perspectives, and challenges," *Computers in Industry*, vol. 144, p. 103792, 2023.

[231] T. Y. Pang, T. S. T. Lo, T. Ellena, H. Mustafa, J. Babalija, and A. Subic, "Fit, stability and comfort assessment of custom-fitted bicycle helmet inner liner designs, based on 3d anthropometric data," *Applied ergonomics*, vol. 68, pp. 240–248, 2018.

[232] N. Liu, P.-S. Chow, and H. Zhao, "Challenges and critical successful factors for apparel mass customization operations: recent development and case study," *Annals of Operations Research*, vol. 291, pp. 531–563, 2020.

[233] Y. Wang, H.-S. Ma, J.-H. Yang, and K.-S. Wang, "Industry 4.0: a way from mass customization to mass personalization production," *Advances in manufacturing*, vol. 5, pp. 311–320, 2017.

[234] M. P. Pessôa and J. J. Becker, "Smart design engineering: a literature review of the impact of the 4th industrial revolution on product design and development," *Research in engineering design*, vol. 31, no. 2, pp. 175–195, 2020.

[235] M. Yuan, H. Yu, J. Huang, and A. Ji, "Reconfigurable assembly line balancing for cloud manufacturing," *Journal of Intelligent Manufacturing*, vol. 30, pp. 2391–2405, 2019.

[236] R. Dou, Y. Zhang, and G. Nan, "Customer-oriented product collaborative customization based on design iteration for tablet personal computer configuration," *Computers & Industrial Engineering*, vol. 99, pp. 474–486, 2016.

[237] P. Pourhejazy, P. Thamchutha, and T.-i. Namthip, "A dea-based decision analytics framework for product deletion in the luxury goods and fashion industry," *Decision Analytics Journal*, vol. 2, p. 100019, 2022.

[238] J.-G. Lee, T. Kim, K. W. Sung, and S. W. Han, "Automobile parts reliability prediction based on claim data: The comparison of predictive effects with deep learning," *Engineering Failure Analysis*, vol. 129, p. 105657, 2021.

[239] A. Theissler, J. Pérez-Velázquez, M. Kettelgerdes, and G. Elger, "Predictive maintenance enabled by machine learning: Use cases and challenges in the automotive industry," *Reliability engineering & system safety*, vol. 215, p. 107864, 2021.

[240] M. Bertolini, D. Mezzogori, M. Neroni, and F. Zammori, "Machine learning for industrial applications: A comprehensive literature review," *Expert Systems with Applications*, vol. 175, p. 114820, 2021.

[241] K. Thiel and S. Postlethwaite, "Human-centric research of skills and decision-making capacity in fashion garment manufacturing to support robotic design tool development," *Human Aspects of Advanced Manufacturing*, pp. 20–30, 2023.

[242] R. D. Hisrich and M. Soltanifar, "Unleashing the creativity of entrepreneurs with digital technologies," *Digital Entrepreneurship: Impact on Business and Society*, pp. 23–49, 2021.

[243] M. Hassenzahl, M. Burmester, and F. Koller, "User experience is all there is: twenty years of designing positive experiences and meaningful technology," *i-com*, vol. 20, no. 3, pp. 197–213, 2021.

[244] L. Sun and L. Zhao, "Technology disruptions: Exploring the changing roles of designers, makers, and users in the fashion industry," *International Journal of Fashion Design, Technology and Education*, vol. 11, no. 3, pp. 362–374, 2018.

[245] M. Murzyn-Kupisz and D. Hołuj, "Fashion design education and sustainability: towards an equilibrium between craftsmanship and artistic and business skills?," *Education Sciences*, vol. 11, no. 9, p. 531, 2021.

[246] I. Mironov, K. Talwar, and L. Zhang, "Rényi differential privacy of the sampled gaussian mechanism," *arXiv preprint arXiv:1908.10530*, 2019.

[247] 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.

[248] G. Andrew, O. Thakkar, B. McMahan, and S. Ramaswamy, "Differentially private learning with adaptive clipping," *Advances in Neural Information Processing Systems*, vol. 34, pp. 17455–17466, 2021.

[249] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[250] X. Wu, Y. Chen, H. Yu, and Z. Yang, "Privacy-preserving federated learning based on noise addition," *Expert Systems with Applications*, vol. 267, p. 126228, 2025.

[251] M. Li, Y. Tian, J. Zhang, D. Fan, and D. Zhao, "The trade-off between privacy and utility in local differential privacy," in *2021 International Conference on Networking and Network Applications (NaNA)*, pp. 373–378, 2021.

Funded by the European Union – NextGenerationEU

Mission 4

Component 2

CUP D83C22000770001