# VirtualXAI: A User-Centric Framework for Explainability Assessment Leveraging **GPT-Generated Personas**

1<sup>st</sup> Georgios Makridis Department of Digital Systems Department of Digital Systems Department of Digital Systems University of Piraeus Piraeus, Greece gmakridis@unipi.gr

2<sup>nd</sup> Vasileios Koukos University of Piraeus Piraeus, Greece vkoukos@unipi.gr

3<sup>rd</sup> Georgios Fatouros University of Piraeus Piraeus. Greece gfatouros@unipi.gr

4<sup>th</sup> Dimosthenis Kyriazis University of Piraeus Piraeus, Greece dimos@unipi.gr

Abstract—In today's data-driven era, computational systems generate vast amounts of data that drive the digital transformation of industries, where Artificial Intelligence (AI) plays a key role. Currently, the demand for eXplainable AI (XAI) has increased to enhance the interpretability, transparency, and trustworthiness of AI models. However, evaluating XAI methods remains challenging: existing evaluation frameworks typically focus on quantitative properties such as fidelity, consistency, and stability without taking into account qualitative characteristics such as satisfaction and interpretability. In addition, practitioners face a lack of guidance in selecting appropriate datasets, AI models, and XAI methods —a major hurdle in human-AI collaboration. To address these gaps, we propose a framework that integrates quantitative benchmarking with qualitative user assessments through virtual personas based on the "Anthology" of backstories of the Large Language Model (LLM). Our framework also incorporates a content-based recommender system that leverages dataset-specific characteristics to match new input data with a repository of benchmarked datasets. This yields an estimated XAI score and provides tailored recommendations for both the optimal AI model and the XAI method for a given scenario.

Index Terms—XAI, explainability score, explainability metric, machine learning, deep learning, AI, LLM Anthology, LLM

# I. INTRODUCTION

In today's data-driven environment, advanced computational systems generate vast amounts of data that fuel the digital transformation of industries. This phenomenon has led us into the era of Industry 4.0 [1], where Artificial Intelligence (AI) plays a key role in driving innovation across sectors. At the same time, the demand for eXplainable Artificial Intelligence (XAI) has risen to improve the interpretability, transparency, and trustworthiness of AI models [2], [3]. This is particularly important in safety critical industries, such as [4] and regulated industries ([5], [6]), where explainability and transparency are essential to gain the trust of stakeholders and ensure compliance with legal and ethical requirements.

Recent studies have shown that the most prominent XAI techniques are SHAP [7] and LIME [8]. Although SHAP is well-regarded for its stability and mathematical foundations, LIME is appreciated for its model-agnostic properties despite some noted instability [2]. However, evaluating remains a challenging task, with most approaches relying on quantitative quantitative anecdotald expert opinion [9]. Existing evaluation frameworks have typically focused on properties such as fidelity, consistency, stability, and certainty [7], [8], yet these approaches are often tailored to specific methods [10]. Furthermore, [11], [12] have emphasized the necessity of interpretable models and have outlined the challenges in objectively measuring the quality of explanation. Moreover, the lack of knowledge to select the appropriate AI models and XAI methods has been highlighted as a major hurdle in human-AI collaboration [13].

Large Language Models (LLMs) have recently emerged as powerful tools generating human-readable explanations and bridging the gap between technical algorithms and domain understanding [14]. Their natural language processing capabilities offer significant potential for improving the accessibility and interpretability of complex AI systems, particularly in domains where human judgment and regulatory compliance are paramount or in specialized fields with technical terminology [15]. To address this gap, we propose an XAI scoring framework that integrates quantitative benchmarking with qualitative user assessments through virtual personas based on the GPT-4mini generated Anthology of backstories, following the results of [16]. In addition, our framework incorporates a contentbased recommendation system that uses dataset-specific characteristics to match input data with a repository of benchmarked datasets, thus estimating an XAI score and providing tailored recommendations for both AI and XAI methods.

The contributions of this paper are threefold:

- Development of a XAI Scoring Framework for tabular data that integrates fidelity, simplicity, stability, and accuracy metrics.
- Introduction of an LLM-based qualitative assessment methodology to capture user-centric qualitative assessment
- Creation of a content-based recommender system to assist users in selecting datasets, AI models, and XAI methods by matching dataset characteristics to historical benchmarks.

The remainder of the paper is organized as follows. Section 2 provides a review of existing XAI methods and their evaluation frameworks. In Section 3, we introduce our proposed explainability framework, detailing its mathematical formulation and the integration of various XAI properties. Section 4 presents our experimental evaluation of the explainability metric across diverse datasets. Finally, Section 5 offers conclusions and recommendations for future work.

#### II. RELATED WORK

In recent years, several off-the-shelf frameworks and toolkits have been proposed to facilitate the application of explainability methods to AI models, such as AI Explainability 360 [17] and InterpretML [18]. These frameworks offer guidelines and implementations for various explainability techniques.

Moreover, several surveys and frameworks have been developed to evaluate black-box models [19] and empirically assess the impact of explanation quality on end-user trust and performance [20]. For example, [21] conducted an experimental investigation comparing 14 different metrics across nine state-of-the-art XAI methods. Their findings highlight correlations among certain metrics, and notable limitations in the reliability of these metrics.

[22] introduced a guide to evaluate and rank XAI methods. Their work assesses explanation properties such as robustness, faithfulness, randomization, complexity, and localization across multiple XAI techniques applied to climate data. Similarly, the OpenHEXAI framework [23] offers an open source platform designed for the human-centered evaluation of explainable machine learning. This framework integrates benchmark datasets, pre-trained models, and post hoc explanation methods. Another contribution is the development of the XAI Experience Quality Scale (XEQ) by [24]. Grounded in psychometric theory, the XEQ evaluates user-centered aspects of XAI experiences by measuring dimensions such as learning, utility, fulfillment, and engagement.

In [25], a human-centered approach is presented that explores how users understand explanations generated by machine learning systems. [26] focuses on a specific application by evaluating the explanations provided by a deep learning system for diabetic retinopathy detection. Their user study with medical professionals assessed the impact of explanations on performance

and trust, demonstrating the practical implications of effective explainability in high-stakes decision-making contexts. Similarly, [27] emphasizes not only the generation of high-quality explanations but also their effective communication to users, proposing the quantification of transmitted information using methods from Information Theory. Alongside these efforts, [28] introduces user-centered metrics—including user satisfaction, mental models, curiosity, trust, and the performance of human-AI collaborations—to present a more comprehensive picture of explainability. One step forward towards personalized XAI was made by ehe x-[plAIn] framework. It demonstrates how domain-specific LLMs can democratize XAI accessibility. Developed using ChatGPT Builder, this system adapts explanations to audience expertise levels [29].

In contrary to the emerging need and applications of XAI, quantifying explainability remains a complex task due to its multifaceted nature and the diverse range of stakeholders involved. The inherent subjectivity —where different individuals may have, poses significant challenges in designing a universally applicable explainability metric. One potential solution is to incorporate user feedback into the evaluation process, allowing the metric to adapt to individual preferences via active learning [30]. Moreover, trade-offs often exist between different aspects of explainability, such as simplicity, fidelity, and coverage; a highly accurate and detailed explanation might be more complex and harder for users to understand, whereas a simpler explanation might sacrifice some fidelity for improved comprehensibility [31].

These newer perspectives underscore the need for a holistic evaluation of XAI techniques that encompasses both technical and user-centric aspects. Our work builds on these insights by proposing a methodology that benchmarks XAI techniques against a wide range of criteria, integrating these multifaceted approaches.

# III. METHODOLOGY

The approach is based on the XAI evaluation strategy presented in [13] and further enhanced with qualitative assessment levergaing LLM generated virtual personal based on an anthology of backstories and insights from three surveys [2] [32] [33]. As depicted in Figure 1, the system architecture comprises four primary components: (1) a quantitative evaluation module, (2) a qualitative assessment using virtual personas, (3) a survey-informed evaluation process, and (4) a content-based recommender system for generating an overall XAI score.

# A. System Architecture Overview

 Quantitative Evaluation: Benchmarks XAI methods by measuring fidelity, stability, simplicity, and accuracy/precision. A high-level pipeline for this process is shown in Figure 2.

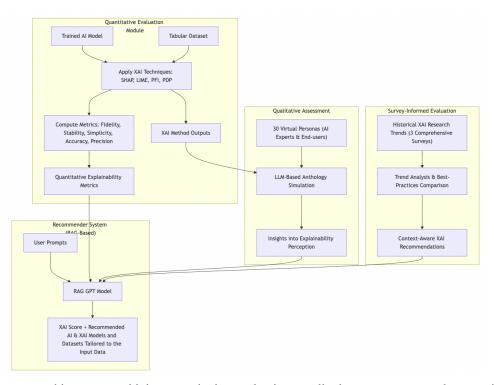


Fig. 1: Integrated system architecture combining quantitative evaluation, qualitative assessment, and survey-informed insights. The recommender system uses these inputs to generate an XAI score and recommend AI and XAI methods based on the dataset's characteristics and user preferences.

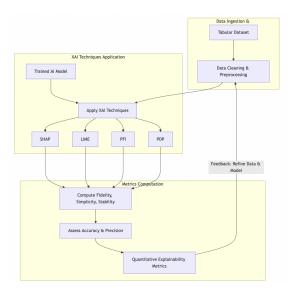


Fig. 2: High-level pipeline for the quantitative evaluation of XAI methods. Data is ingested and preprocessed, then a trained AI model is explained using SHAP, LIME, PFI, or PDP

- 2) Qualitative User Assessment: Uses GPT-4o-mini to generate virtula personas based on an LLM-generatred Anthology of backstories. And these virtual personas respond to a tailored questionaure to capture user preferences and interpretability requirements. This approach, illustrated in Figure 3 (generates virtual personas and aggregates their feedback).
- Survey-Informed Model Evaluation: Integrates findings from large-scale surveys to reflect current best practices and trends across multiple domains.
- Content-Based Recommender System: Leverages dataset characteristics to estimate an XAI score and recommend proper AI and XAI methods.

# B. Quantitative Metrics for Benchmarking XAI Methods

The quantitative evaluation focuses on benchmarking four widely used XAI techniques —SHAP, LIME, PFI, and PDP—against tabular datasets. Following the methodology described in [13], the key metrics for evaluation include:

- Fidelity: Measures the degree to which an explanation accurately reflects the behavior of the underlying model.
- Simplicity: Assesses the interpretability of the explanation by quantifying its complexity.

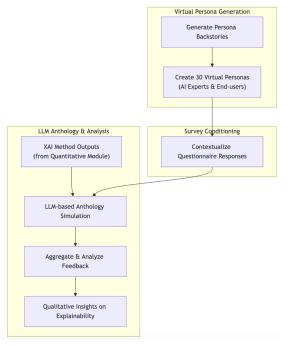


Fig. 3: Overview of the qualitative assessment approach. Virtual personas are generated, and their feedback on the XAI outputs is collected through structured surveys.

- **Stability**: Evaluates the consistency of explanations when minor perturbations are applied to the input data.
- Accuracy and Precision: Quantifies the trade-off between the performance of the model and the quality of its explanations.

For each XAI method, these metrics are computed using a reproducible and robust procedure as detailed in [13].

# C. Qualitative Assessment

The qualitative assessment via virtual personas addresses several challenges inherent in traditional user satisfaction studies. Recruiting real end users can be time-consuming, costly, and prone to biases due to inconsistent participation. By generating virtual personas with carefully designed demographic and professional profiles —as outlined in [16]— our process simulates a broad range of real-world perspectives in a consistent and scalable manner. Figure 3 lustrates the process that involves:

 Virtual Personas Generation: Using GPT-4o-mini (via openai api) we generated 1000 backstories by applying open quesitons. Then we chose in a balance way 100 out of 1000 backstories and based on them the virtual personas are generated using LLMs, each with a unique demographic and professional profile, including factors such as age, profession, level of AI expertise, and specific explainability preferences.

- Survey Conditioning: Each persona's profile conditions a structured questionnaire aimed at evaluating XAI methods.
- **Feedback Analysis**: The responses from these virtual personas are aggregated and analyzed to user satisfaction and preferences, providing qualitative insights into the interpretability of model explanations.

# D. Comprehensive Review of Sector-Specific XAI Applications

To ensure our framework is aligned with current research, we integrate insights from three comprehensive surveys on XAI methodologies, ensuring that our recommendations reflect domain-specific trends and real-world applicability. These surveys provide:

- Detailed breakdowns of XAI adoption across industries such as healthcare and finance.
- Comparative analyses of XAI methods, including their strengths and limitations across different domains.
- Assessments of trends in AI model usage, evaluation strategies, and user preferences.

For illustration purposes, we include example JSON listings that capture key empirical data. Listing 1 shows the frequency of various XAI explanation techniques, and Listing 2 outlines the most frequent XAI models employed across different domains. It is important to note that these listings are examples and not the final versions.

Listing 1: XAI Explanation Techniques Frequency

```
"xai_explanation_techniques_frequency": {
 "SHAP": 175,
  "LIME": 125,
  "Grad-CAM": 75.
 "Decision_Tree": 35,
  "Permutation_Importance": 25,
  "Integrated_Gradients": 20,
 "SmoothGrad": 15,
  "LRP": 10,
  "PDP": 8,
 "Grad-CAM++": 7,
 "Anchors": 6,
  "Logistic_Regression": 5,
 "ALE": 5,
 "CAM": 5,
  "RISE": 5
```

Listing 2: Most Frequent XAI Models in Different Domains

```
"most_frequent_xai_models_per_domain": {
   "healthcare": ["SHAP", "Grad-CAM"],
   "finance": ["SHAP", "LIME"],
   "cybersecurity": ["PFI", "SHAP"],
   "transportation": ["PDP", "Integrated Gradients"
   ],
```

```
"education": ["LIME", "Decision Trees"]
}
```

#### E. Development of a Content-Based Recommender System

To assist end-users in selecting appropriate datasets, AI models, and XAI methods, we propose a content-based recommender system. This leverages the intrinsic characteristics of the input dataset, such as feature distribution, dimensionality, and domain. Its operation comprises two phases:

# 1) Dataset Matching:

- Feature Extraction: Features are extracted from the input dataset (e.g., statistical summaries, feature distributions, and domain indicators).
- Similarity Assessment: The input dataset is compared against a repository of previously benchmarked datasets using content-based filtering techniques to identify similar datasets.

# 2) Recommendation Generation:

- XAI Score Estimation: Based on the performance metrics obtained from matched datasets, an estimated XAI score is computed for the target dataset.
- Method Recommendation: The system recommends optimal AI and XAI methods that have historically yielded high XAI scores on similar datasets.

#### IV. RESULTS

To assess the effectiveness of the proposed XAI Scoring Framework, we conducted a series of benchmarking experiments using a diverse set of tabular datasets and explainability techniques.

#### A. Datasets

We utilized a diverse set of tabular datasets from the UCI repository [34], each representing distinct domains and varying levels of feature complexity. All datasets underwent standardized preprocessing procedures, including categorical encoding and normalization, to ensure consistency. A Random Forest classifier was trained on each dataset, after which the selected XAI techniques were applied to interpret the model predictions. This approach allowed us to evaluate the performance of the explainability methods in a domain-agnostic manner.

#### B. Dataset Distribution Across Domains

The utillized datasets span multiple domains, as illustrated in Figure 4. Notably, the *health and medicine* domain exhibits the largest number of datasets (over 50), followed by *computer science*, *business*, and *physics and chemistry*. This distribution highlights the prominence of medical and clinical use cases in current AI research, a trend corroborated by our survey-based data.

**Algorithm 1** Benchmarking Phase: Quantitative & Qualitative Benchmarking and Information Extraction

# **Require:** • Benchmark datasets $\mathcal{D}$

- Trained AI models M (or training procedure for each D ∈ D)
- XAI techniques  $\mathcal{X} = \{SHAP, LIME, PFI, PDP\}$
- Number of virtual personas P

**Ensure:** Repository  $\mathcal{R}$  containing for each dataset:

- Quantitative metrics  $Q_D$  (fidelity, simplicity, stability, accuracy, precision)
- Qualitative insights  $Q_{qual,D}$
- Dataset characteristics  $C_D$
- 1: for each dataset  $D \in \mathcal{D}$  do
- 2: Preprocess dataset *D* (e.g., data cleaning, normalization, encoding)
- 3: Train AI model  $M_D$  on D (or use an existing model from  $\mathcal{M}$ )
- 4: **for** each XAI technique  $x \in \mathcal{X}$  **do**
- 5: Generate explanation  $E_x$  for  $M_D$  using x
- 6: Compute quantitative metrics:
  - Fidelity: How accurately  $E_x$  reflects  $M_D$
  - Simplicity: Complexity measure of  $E_x$
  - Stability: Consistency of  $E_x$  under input perturbations
  - Accuracy & Precision: Trade-offs between model performance and explanation quality
- 7: Store metrics as  $Q_{x,D}$
- 8: end for
- 9: Aggregate quantitative metrics:  $Q_D \leftarrow \{Q_{x,D} \mid x \in \mathcal{X}\}$
- 10: Extract dataset characteristics  $C_D$  (e.g., feature distributions, dimensionality, sparsity)

# 11: Qualitative Assessment:

- 12: Generate P virtual personas with diverse backstories using GPT-4o-mini
- 13: Condition a structured questionnaire with and apply it to  $E = \{E_x \mid x \in \mathcal{X}\}$
- 14: Aggregate the survey responses to derive qualitative insights  $Q_{qual,D}$
- 15: Store the tuple  $(D, Q_D, Q_{qual,D}, C_D)$  in repository  $\mathcal{R}$
- 16: **end for**
- 17: **return**  $\mathcal{R}$

#### C. Quantitative Benchmarking of XAI Methods

Figure 5 presents the average fidelity of four XAI methods—SHAP, LIME, PFI, and PDP—across multiple domains. The results highlight a clear variation in method performance depending on the domain. In health and medicine, SHAP demonstrates consistently high fidelity, suggesting it aligns well with clinical data. Conversely, PDP exhibits notably higher

**Algorithm 2** Inference Phase: Dataset Matching, XAI Score Estimation and Recommendation

**Require:** • User-uploaded dataset  $D_u$ 

• Repository  $\mathcal{R}$  containing tuples  $(D, Q_D, Q_{qual,D}, C_D)$  from the training phase

**Ensure:** • Estimated XAI score  $S_{XAI}$  for  $D_u$ 

- Recommendations for optimal AI model  $M^{\ast}$  and XAI method  $x^{\ast}$
- 1: Preprocess the uploaded dataset  $D_u$  (cleaning, normalization, encoding)
- 2: Extract dataset characteristics  $C_u$  from  $D_u$
- 3: Compute similarity between  $C_u$  and each stored  $C_D$  in  $\mathcal{R}$  using a similarity metric
- 4: Identify the top k matching datasets  $\{D_{(1)}, \ldots, D_{(k)}\}$  based on similarity scores
- 5: Aggregate the corresponding quantitative metrics and qualitative insights from the matched datasets
- 6: Estimate the XAI score  $S_{XAI}$  for  $D_u$  using the aggregated metrics (e.g., via weighted averaging)
- 7: Determine the optimal AI model and XAI method recommendation  $(M^*, x^*)$  based on historical performance on similar datasets
- 8: **return**  $S_{XAI}$ ,  $M^*$ , and  $x^*$

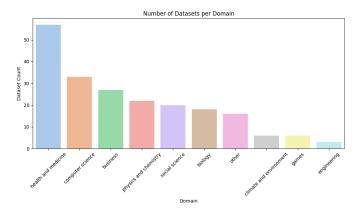


Fig. 4: Number of datasets per domain. The *health and medicine* domain has the highest dataset count, reflecting a significant interest in clinical AI applications.

fidelity in business applications, indicating it may be particularly effective for the kinds of features and relationships found in that domain. LIME and PFI maintain relatively moderate yet steady performance across most domains, with occasional spikes in areas such as biology or computer science. Overall, these findings underscore that no single XAI method dominates in every context, reinforcing the importance of domain-specific considerations when selecting an explainability technique.

Table I provides an additional breakdown for selected

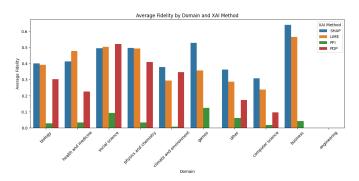


Fig. 5: Average fidelity by domain and XAI method. Higher bars indicate stronger alignment between the explanation and the model's predictions.

datasets. For example, on a Heart Disease dataset within the *health and medicine* domain, SHAP attains a fidelity of 0.82, while LIME exhibits a lower simplicity value (5.1), indicating more concise explanations. PFI excels in stability (0.93), underscoring its consistency under feature perturbations.

Dataset	Method	Fidelity	Simplicity	Stability
Heart Disease	SHAP	0.82	7.3	0.91
	LIME	0.79	5.1	0.88
	PFI	0.76	4.6	0.93
	PDP	0.74	6.0	0.87
Wine Quality	SHAP	0.85	6.8	0.89
	LIME	0.78	5.3	0.85
	PFI	0.74	4.1	0.92
	PDP	0.71	5.7	0.84

TABLE I: Sample of quantitative explainability scores for SHAP, LIME, PFI, and PDP. Fidelity measures alignment with the model, Simplicity indicates fewer features (lower is better), and Stability measures consistency.

# D. Qualitative User Ratings and Interpretability

We measure user-perceived *interpretability* by aggregating virual personas ratings (interpretability, understanding, trust). Figure 6 displays these average interpretability scores by domain. PDP appears to be the highest-scoring method in nearly every domain, indicating that users find its partial dependence approach particularly intuitive or easy to understand. While LIME and SHAP also achieve relatively strong scores in several domains (e.g., business, health and medicine), their performance is slightly outpaced by PDP in most cases. PFI, meanwhile, maintains moderate interpretability levels overall. These results underscore the value of considering multiple XAI methods in practice; while PDP may be a strong default choice in many scenarios, certain domains or user requirements might favor the localized explanations of LIME or the feature-attribution clarity of SHAP.

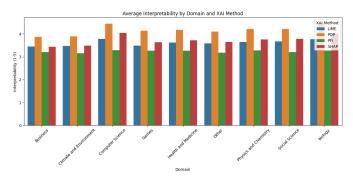


Fig. 6: Average interpretability (1–5 scale) by domain and XAI method. Higher bars indicate that end-users (virtual personas) found the explanations more understandable and trustworthy.

# E. Repository Construction and Domain-Specific Recommendations

All these data sources are integrated into a unified repository keyed by *dataset\_id*.

When a user uploads a new dataset, our system extracts its characteristics (e.g., feature count, numeric/categorical ratio, missing ratio), then computes similarity to benchmark datasets in the repository using cosine similarity. The top-k similar datasets' XAI metrics and user ratings are aggregated to estimate a multidimensional XAI score for each method.

# V. CONCLUSION

Our experimental results demonstrate that:

- 1) Certain domains (*health and medicine*) contain more datasets, reinforcing the strong emphasis on interpretability in clinical settings.
- Quantitative metrics vary significantly by domain and XAI method, suggesting that no single method universally dominates across all contexts.
- End-users' subjective interpretability ratings often diverge from purely technical measures such as fidelity or simplicity, emphasizing the need for user-driven evaluations
- 4) Domain-specific synergy is crucial.

Hence, our approach merges quantitative benchmarking with qualitative persona insights, further refined by domain bonuses from JSON surveys. This fusion significantly advances previous frameworks that relied solely on one dimension of evaluation.

In future work, we plan to expand this approach to other data modalities (e.g., text, images) and refine the synergy model that determines domain bonuses.

#### VI. ACKNOWLEDGEMENT

The research leading to the results presented in this paper has received funding from the Europeans Union's funded Project FAME under grant agreement no 101092639 and HumAIne under grant agreement no 101120218.

#### REFERENCES

- G. Makridis, D. Kyriazis, and S. Plitsos, "Predictive maintenance leveraging machine learning for time-series forecasting in the maritime industry," in 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2020, pp. 1–8.
- [2] M. Saarela and V. Podgorelec, "A comprehensive review of explainable artificial intelligence applications," *Applied Sciences*, vol. 14, no. 8884, pp. 1–25, 2024.
- [3] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins et al., "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information fusion*, vol. 58, pp. 82–115, 2020.
- [4] G. Makridis, P. Mavrepis, and D. Kyriazis, "A deep learning approach using natural language processing and time-series forecasting towards enhanced food safety," *Machine Learning*, vol. 112, no. 4, pp. 1287– 1313, 2023.
- [5] D. Kotios, G. Makridis, S. Walser, D. Kyriazis, and V. Monferrino, "Personalized finance management for smes," in *Big Data and Artificial Intelligence in Digital Finance*. Springer, Cham, 2022, pp. 215–232.
- [6] G. Fatouros, G. Makridis, D. Kotios, J. Soldatos, M. Filippakis, and D. Kyriazis, "Deepvar: a framework for portfolio risk assessment leveraging probabilistic deep neural networks," *Digital Finance*, pp. 1–28, 2022.
- [7] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," arXiv preprint arXiv:1705.07874, 2017.
- [8] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.
- [9] M. Nauta and colleagues, "Xai evaluation: Current challenges and future opportunities," *Journal of Artificial Intelligence Research*, vol. 78, pp. 567–595, 2023.
- [10] I. Lage, E. Chen, J. He, M. Narayanan, B. Kim, S. Gershman, and F. Doshi-Velez, "An evaluation of the human-interpretability of explanation," arXiv preprint arXiv:1902.00006, 2019.
- [11] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," arXiv preprint arXiv:1708.08296, 2017.
- [12] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv preprint arXiv:1702.08608, 2017.
- [13] G. Makridis, V. Koukos, G. Fatouros, D. Kotios, and D. Kyriazis, "A unified framework for explainable ai evaluation," *Journal of Machine Learning*, vol. 45, pp. 876–895, 2023.
- [14] G. Fatouros, K. Metaxas, J. Soldatos, and M. Karathanassis, "Market-senseai 2.0: Enhancing stock analysis through llm agents," arXiv preprint arXiv:2502.00415, 2025.
- [15] G. Fatouros, J. Soldatos, K. Kouroumali, G. Makridis, and D. Kyriazis, "Transforming sentiment analysis in the financial domain with chatgpt," *Machine Learning with Applications*, vol. 14, p. 100508, 2023.
- [16] S. Moon, M. Abdulhai, M. Kang, J. Suh, W. Soedarmadji, E. K. Behar, and D. M. Chan, "Virtual personas for language models via an anthology of backstories," arXiv preprint arXiv:2407.06576, 2024.
- [17] V. Arya, R. K. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilović et al., "Ai explainability 360 toolkit," in Proceedings of the 3rd ACM India joint international conference on data science & management of data (8th ACM IKDD CODS & 26th COMAD), 2021, pp. 376–379.
- [18] H. Nori, S. Jenkins, P. Koch, and R. Caruana, "Interpretml: A unified framework for machine learning interpretability," arXiv preprint arXiv:1909.09223, 2019.
- [19] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM computing surveys (CSUR)*, vol. 51, no. 5, pp. 1–42, 2018.

- [20] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Wortman Vaughan, and H. Wallach, "Manipulating and measuring model interpretability," in *Proceedings of the 2021 CHI conference on human factors in computing systems*, 2021, pp. 1–52.
- [21] J. Stassin *et al.*, "An experimental investigation into explainability metrics for xai methods," *Journal of Explainable AI Research*, vol. 5, no. 3, pp. 123–145, 2023.
- [22] J. Bommer et al., "Evaluating and ranking xai methods: A guide for climate science," Climate Science AI, vol. 7, no. 2, pp. 67–89, 2023.
- [23] L. Ma et al., "Openhexai: A framework for human-centered xai evaluation," Proceedings of the Human-Centered AI Conference, vol. 9, no. 1, pp. 1–15, 2024.
- [24] A. Wijekoon *et al.*, "The xai experience quality (xeq) scale: Evaluating user-centered quality," *Human-AI Interaction Journal*, vol. 10, no. 1, pp. 45–65, 2024.
- [25] M. Narayanan, E. Chen, J. He, B. Kim, S. Gershman, and F. Doshi-Velez, "How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation," arXiv preprint arXiv:1802.00682, 2018.
- [26] E. Beede, E. Baylor, F. Hersch, A. Iurchenko, L. Wilcox, P. Ruamviboonsuk, and L. M. Vardoulakis, "A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy," in *Proceedings of the 2020 CHI conference on human* factors in computing systems, 2020, pp. 1–12.
- [27] D. Pruthi, R. Bansal, B. Dhingra, L. B. Soares, M. Collins, Z. C. Lipton, G. Neubig, and W. W. Cohen, "Evaluating explanations: How much do explanations from the teacher aid students?" *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 359–375, 2022.
- [28] J. Litman, "Measures for explainable ai: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-ai performance," 2023
- [29] P. Mavrepis, G. Makridis, G. Fatouros, V. Koukos, M. M. Separdani, and D. Kyriazis, "Xai for all: Can large language models simplify explainable ai?" arXiv preprint arXiv:2401.13110, 2024.
- [30] A. Holzinger, M. Plass, M. Kickmeier-Rust, K. Holzinger, G. C. Crişan, C.-M. Pintea, and V. Palade, "Interactive machine learning: experimental evidence for the human in the algorithmic loop: A case study on ant colony optimization," *Applied Intelligence*, vol. 49, pp. 2401–2414, 2019.
- [31] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA). IEEE, 2018, pp. 80–89.
- [32] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. Van Keulen, and C. Seifert, "From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai," ACM Computing Surveys, vol. 55, no. 13s, pp. 1–42, 2023.
- [33] M. Nagahisarchoghaei, N. Nur, L. Cummins, N. Nur, M. M. Karimi, S. Nandanwar, S. Bhattacharyya, and S. Rahimi, "An empirical survey on explainable ai technologies: Recent trends, use-cases, and categories from technical and application perspectives," *Electronics*, vol. 12, no. 5, p. 1092, 2023.
- [34] A. Asuncion, D. Newman et al., "Uci machine learning repository," 2007.