XAIport: A Service Framework for the Early Adoption of XAI in AI Model Development

Zerui Wang, Yan Liu, Abishek Arumugam Thiruselvi, Abdelwahab Hamou-Lhadj
{zerui.wang,abishek.arumugamthiruselvi}@mail.concordia.ca;{yan.liu,wahab.hamou-lhadj}@concordia.ca
Concordia University
Montreal, Canada

ABSTRACT

In this study, we propose the early adoption of Explainable AI (XAI) with a focus on three properties: Quality of explanation, the explanation summaries should be consistent across multiple XAI methods; Architectural Compatibility, for effective integration in XAI, the architecture styles of both the XAI methods and the models to be explained must be compatible with the framework; Configurable operations, XAI explanations are operable, akin to machine learning operations. Thus, an explanation for AI models should be reproducible and tractable to be trustworthy. We present XAIport, a framework of XAI microservices encapsulated into Open APIs to deliver early explanations as observation for learning model quality assurance. XAIport enables configurable XAI operations along with machine learning development. We quantify the operational costs of incorporating XAI with three cloud computer vision services on Microsoft Azure Cognitive Services, Google Cloud Vertex AI, and Amazon Rekognition. Our findings show comparable operational costs between XAI and traditional machine learning, with XAIport significantly improving both cloud AI model performance and explanation stability.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; **Artificial intelligence**; • **Applied computing** → **Operations research**.

KEYWORDS

XAI, MLOps, Operational Cost Analysis, Deployment Strategy

ACM Reference Format:
Zerui Wang, Yan Liu, Abishek Arumugam Thiruselvi, Abdelwahab Hamou-Lhadj. 2024. XAIport: A Service Framework for the Early Adoption of XAI in AI Model Development. In *New Ideas and Emerging Results (ICSE-NIER'24), April 14–20, 2024, Lisbon, Portugal.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3639476.3639759

1 INTRODUCTION

Machine Learning Operations (MLOps) is a multidisciplinary approach that includes a set of best practices, concepts, and developments [16]. The major tasks of MLOps include automating the ML lifecycle, such as model development, validation, quality assurance, deployment, monitoring, and governance [32].

The quality assurance of MLOps involves several key components, such as data validation, feature engineering assessment, model training evaluation, cross-validation, performance metrics analysis, fairness, bias evaluation, and model explainability [30]. The explainability of models is necessary in sensitive domains. The lack of model explanation leads to distrust in the AI models [12, 29].

Post-hoc XAI methods provide explanations for complex, already-trained models. The Post-hoc XAI techniques, such as SHAP [21], provide feature attribution as explanations. XAI operations are often considered a post-hoc activity [27], implemented after the model has been trained and verified. The quality assurance [30] of complex software development has shown that incremental development and iterative quality control are efficient and cost-effective. Inspired by this principle, we argue that early adoption of XAI operations enhances the quality assurance of AI models with probing observations at the feature representation level and summarized explanations across datasets and AI models.

We present `XAIport`, an XAI service architecture that allows XAI early adoption across cloud platforms and offers unified open API access. Inconsistent explanations can be misleading to evaluate the AI models. The probing results and derived explanations should be quantified for their stability and consistency across datasets and AI models. The efficiency of applying XAI operations should be measured quantitatively so that the runtime overhead and cost are well balanced in evaluating the benefits of adopting XAIs. We summarize the key considerations for the early adoption of XAI operations as follows:

- **Quality of Explanation.** Explanations generated by XAI methods should adhere to the evaluation metrics, specifically the explanation consistency metrics, defined in an established XAI process [13].
- **Architecture Compatibility.** The XAI service flexibly integrates the AI models and XAI methods within the microservice architecture. This ensures incorporation into existing cloud services via open APIs.
- **Cost-Efficiency in CI/CD.** The adoption of XAI into MLOps should result in proportional cost-efficient operational overhead during the Continuous Integration and Continuous Deployment (CI/CD) phases. In ideal scenarios, the additional complexity XAI introduces is approximately proportional to existing MLOps.

This approach mirrors best practices in software engineering, where early integration of unit tests and quality assurance solidifies the software. Additionally, `XAIport` provides a unified measurement of resource consumption and XAI operation overhead.

ICSE 2024, April 14-20, 2024, Lisbon

ACM ISBN 979-8-4007-0500-7/24/04…$15.00
https://doi.org/10.1145/3639476.3639759

2 RELATED WORKS

We explore the increasing importance of XAI in AI domains, such as healthcare and finance. Subsequently, we review the MLOps workflow. Then, we review XAI methods in the field of computer vision and metrics for the assessment.

Alongside the complexity of AI models, the need for explainability has concurrently risen [1]. Explainability fosters a better understanding of model behavior, facilitates trust and encourages responsible AI usage [3]. In the critical sectors of healthcare, finance, and legal systems, XAI is essential in comprehending the model's decisions and implications and ensuring compliance with legal and ethical protocols [3, 24, 25]. The healthcare domain witnesses an especially pronounced need for XAI due to the growing reliance on AI technologies [18].

MLOps integrate machine learning and operations, emphasizing the importance of explainability or XAI [16]. The process begins with defining system requirements [11]. During data collection, potential biases in the data are often overlooked [35]. Data preprocessing techniques, such as cutmix [37] and puzzlemix [15], aim to improve dataset quality. Feature engineering is central to ML models, and incorporating explainability during feature selection simplifies the model [4, 38]. Traditional model quality assurance metrics are expanded to include explainability, especially in cloud AI services [22, 40]. Consistency evaluations in XAI ensure trustworthy explanations [13]. Deployment in MLOps emphasizes the use of visualization tools for better model understanding [13].

Class Activation Mapping (CAM) [39] emerged as a pioneering approach leveraging the global average pooling layer to localize features within Convolutional Neural Network models. A limitation, however, was its need to adjust the model's fully connected layer. In contrast, Grad-CAM [28] refined this by determining the localization weight through the layer's average gradient, eliminating the need for replacements. Advancing this further, Grad-CAM++ [5] incorporated second-order gradients for enhanced precision. EigenCAM [23] uniquely uses the primary component of activations without class-specific considerations. LayerCAM [14] assigns spatial weights to activations considering only positive gradients, while XGrad-CAM [9] adjusts gradients based on normalized activations. Representing the forefront of XAI methodologies, these techniques have proven their prowess in generating saliency maps for visual-based XAI tasks.

The metrics for evaluating explainability in XAI are essential and gain considerable attention [6]. However, the field is still grappling with several challenges. Vilone et al. [34] list scientific papers for approaches to evaluate the XAI method and point out the lack of consensus in defining unified evaluation metrics. Instead of qualitative metrics, we prefer quantitative metrics to assess XAI methods concretely. A systematic assessment [13] provides clear consistency metrics to XAI feature contributions.

Summary - The XAI operations have been particularly explored in domains such as healthcare and finance. As AI services become available through pre-trained models bundled with elastic cloud computing resources, the operations of XAI in such a context still require thorough architectural-level research.

3 THE CONTEXT OF EARLY ADOPTION OF XAI SERVICES

The goal of XAI adoption is compatible with the objectives of model quality assurance. We design the XAI operations function as a probe into AI models with or without the model's intrinsic structure to provide explanations, for instance, on how features may affect the learning results. Hence, we propose the early adoption of XAI operations revolves around three major `core components`: (1) definition of augmented quality assurance metrics for explanation stability and consistency; (2) compatible architecture styles to integrate with cloud AI service development and deployment; and (3) XAI operations are configurable and measurable in the same manner as cloud AI services. In addition, the adoption of XAI should be cloud-independent and allow cross-validation of multiple XAI methods, AI models, and datasets.

3.1 Augmented Metrics for Explanation Quality Assurance

Several studies [6, 20, 34] have shown that XAI methods do not always offer consistent explanations, especially in experiments involving Post-hoc XAI methods. Beyond the model performance, the adoption of XAI operations should measure consistency to ensure model explainability quality. XAI consistency metrics [13], also shown in Algorithm 1, comprise both `Explanation Stability` and `Explanation Consistency`.

Algorithm 1 Calculation of Explanation Metrics [13]

```
Input: Set of explanation summaries $E = \{\xi^1, \xi^2, \ldots, \xi^m\}$
Output: $f_d^K$ (Stability); $f_d^X$ (Consistency)
    Notations:
    $m$ - Number of summaries in $E$
    $K$ - Combinations, $\binom{m}{2}$
    $\xi^i$ - $i$-th explanation summary
    $\xi^X$ - Summary for XAI method $X$
    $f_d^{[k]}$ - Prediction changes for $k$-th pair
procedure EXPLANATION STABILITY(E)
    $K \leftarrow \binom{m}{2}$
    $f_d^K \leftarrow \frac{1}{K} \sum_{k=1}^{K} f_d^{[k]}(\xi^i, \xi^j)$ where $i \neq j, i, j \leq m$
end procedure
procedure EXPLANATION CONSISTENCY(E, X)
    $f_d^X \leftarrow \frac{1}{m-1} \sum_{k=1}^{m-1} f_d^{[k]}(\xi^X, \xi^i)$ where $i \leq m$
end procedure
```

`Explanation Stability` measures the consistency among explanations from many data samples. To compute this metric, we consider all possible pairs of prediction changes from data samples. We then average all these values to a metric, f_d^K, representing stability. Theoretically, a smaller value of this metric indicates higher consistency among the explanations generated by the XAI method. We employ the `Explanation Stability` in the pilot evaluation.

`Explanation Consistency` measures the consistency between different XAI methods. We calculate the prediction changes among XAI methods and average all to f_d^X. A smaller value indicates that different XAI methods are producing similar explanations.

3.2 Compatible Architecture Styles

The adoption of XAI operations should function in the compatible architecture context of MLOps. As pre-trained ML models on the cloud are available, cloud services become the encapsulation of the models running on elastic computing resources. The communication between XAI methods and AI models thus follows the service orientation. We propose `XAIport`, a service architecture in which the core XAI components are each represented as a microservice with the Open API definition. The APIs for the `XAIport` service are organized according to Open API 3.0 [31] standards and documents on SwaggerHub.

The core architecture includes: `(1) Coordination Center` uses a configuration template to specify pipelines of end-to-end XAI operations from data input, to feature variation, to model inference, to feature contribution explanation, and to evaluation generation; `(2) Data Processing and Storage` is responsible for data preparation and storage intermediate results for explanation generation; `(3) XAI Microservices` encapsulate state-of-the-art XAI methods computing the feature contribution explanation for AI models on a certain dataset; `(4) Evaluation Microservices` computes and visualizes the metrics for XAI explanation. These services produce the answers to questions such as *how is an AI model affected by feature representation?* Figure 1 provides an illustration of integrating XAI operations along with the development of AI models using cloud AI. We assume the development of AI models adopts the best practices and technology supports from MLOps and DevOps [8, 12, 16, 32]. The XAI operations are encapsulated as microservices and deployed on the cloud as well. The communication between XAI operations and cloud AI models is only through the endpoints defined by Open APIs. Thus, the enhanced explanation metrics from XAI provide extra measurements for AI model quality assurance.

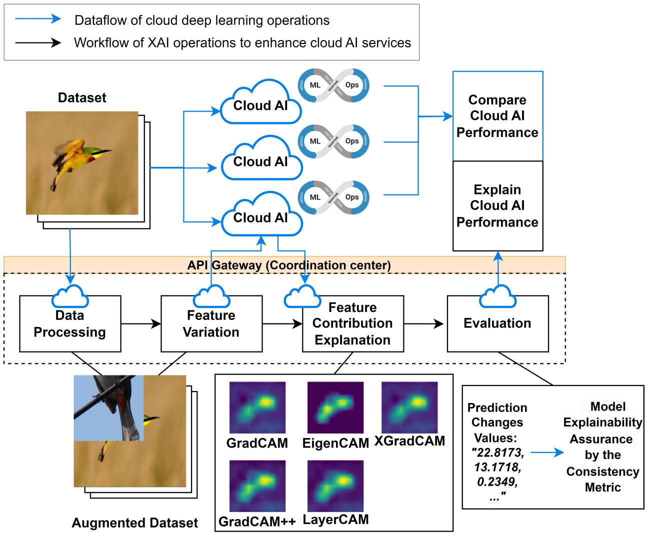

Dataflow of cloud deep learning operations
Workflow of XAI operations to enhance cloud AI services
Dataset
Cloud AI
ML
Ops
Cloud AI
ML
Ops
Cloud AI
ML
Ops
Compare Cloud AI Performance
Explain Cloud AI Performance
API Gateway (Coordination center)
Data Processing
Feature Variation
Feature Contribution Explanation
Evaluation
Augmented Dataset
GradCAM
EigenCAM
XGradCAM
GradCAM++
LayerCAM
Prediction Changes Values: "22.8173, 13.1718, 0.2349, ..."
Model Explainability Assurance by the Consistency Metric

Figure 1: An Illustrating Scenario of Adopting XAI to Multiple Computer Vision Cloud AI Model Development.

Integration with Open Community Pretrained Models. The service-oriented open API architecture `XAIport` is extensible to the AI model development based on open community libraries such as Hugging face [36]. First, the pre-trained models are accessible for trial and testing with open APIs in the same communication model as the cloud AI services in Figure 1. In addition, when the pre-trained models are downloaded for further retraining and fine-tuning on a domain-specific dataset, the model is deployed in a containerized virtual machine that can be run on a cloud or on a proprietary data center. If we assume such a model is further encapsulated with Open API access, then the `XAIport` architecture illustrated in Figure 1 is applicable without any further changes since `XAIport` decouples XAI operations and AI models through only the endpoint communication through Open APIs.

Extension to Support A/B Testing of AI Services. Performing A/B testing on ML models has been adopted by real-world services [19]. The endpoint of an AI model in the form of Open APIs is the operation unit for automated deployment with multiple production variants for A/B testing. As illustrated in Figure 1, the XAI services with Open API as endpoints are capable of either covering a certain variant (such as the data augmentation) or communicating with these A/B testing variant endpoints and derive the explanation to link the model's learning performance and feature contributions.

3.3 Configurable and Measurable XAI Operations

Current XAI methods are in the form of algorithms and disparate library code [1, 3, 13, 21]. We propose the XAI operations should be configurable and measurable. XAI methods, particularly post-hoc techniques, require additional processing power and computational time, thereby increasing the computational overhead [27]. Hardware-wise, the CPU, memory, and GPU requirements may rise to accommodate additional XAI processes. We target to methodically evaluate the operational overhead incurred in the integration and deployment of XAI services across multifarious cloud providers. We focus on measuring the XAI service deployment time and complexity. This exercise entails multiple steps:

Selection of CI/CD Tool. Our methodology commences with the use of cloud pipeline and build tools as the designated continuous integration and continuous deployment (CI/CD) pipeline tool. For instance, Amazon Web Services CodeBuild furnishes essential building facilities for containerized applications, which is imperative for orchestrating a cloud-agnostic XAI milieu, ensuring uniform deployment across diverse cloud platforms.

Measuring XAIport Computational Overhead. We evaluate the computational overhead of diverse AI models and XAI techniques using `CodeCarbon` [17]. This tool, previously applied to several projects [26, 33], is incorporated into `XAIport` to track time, energy, and carbon footprints during XAI activities and AI predictions. Using `CodeCarbon` [17], we differentiate the energy efficiency and time consumption of various XAI operations, guiding the selection of the optimal method for specific use cases.

Measuring XAI Service Deployment Overhead. We assess the effort needed to deploy `XAIport` on multiple cloud providers. We perform Amazon Web Services Elastic Container Service (ECS), Azure Virtual Machines, Azure Container Instances, and Google Kubernetes Engine for their efficiency in deploying AI container applications.

4 A PILOT EVALUATION

We conduct a pilot study using `XAIport` to explore the answer to a data-driven question for cloud-based AI services as follows.

> Can early adoption of XAI improve the learning performance of cloud computer vision services? If any, can the improvement be explained? Can the explanation result be evaluated?

This study uses five visual explanation algorithms, shown in Table 1, applied to three image classification computer vision services, which are Microsoft Azure Cognitive Services [22], Google Cloud Vertex AI [10], and Amazon Rekognition [2]. Cloud platforms offer the following advantages. First, they automate deployment with resource allocation. Second, they offer built-in scalability to manage computational demands efficiently. Third, they deliver monitoring tools for basic model performance. However, These platforms overlook the explainability of models. In this case, we adopt the `XAIport` service framework. Upon evaluating the three cloud AI services, we identify potential areas for further optimization in both model performance and explanation stability. Then, with the integration of Cutmix [37] and Puzzlemix [15] data augmentation techniques in the XAI operation, we enhance both the cloud model performance and explanation stability on these platforms.

4.1 Improving Cloud AI Explanation Metrics with Early XAI Adoption

We apply the ImageNet dataset [7] via the `XAIport` APIs to explore the data-driven question on AI model development. The baseline is the cloud AI service trained only using the original dataset. We adopt five XAI algorithms, which are Grad-CAM [28], Grad-CAM++ [5], EigenCAM [23], LayerCAM [14], and XGrad-CAM [9]. Adoption of these XAI methods takes image data as inputs from the data processing service and generate saliency maps that highlight the focal areas of layers of the model. Then, these data become the inputs for the three cloud AI and return prediction scores. Finally, we use the evaluation service to derive prediction changes and the stability metrics as algorithm 1.

Table 1: Model and XAI Evaluation Results

Service	F1-score	XAI Evaluation				
		GradCAM	GradCAM++	EigenCAM	LayerCAM	XGradCAM
Azure (B)	0.839	22.227	21.211	32.498	20.595	22.229
Google (B)	0.565	22.329	21.233	30.713	21.328	22.327
Amazon (B)	0.807	18.900	17.505	30.119	17.027	18.900
Azure (C)	0.864	4.544	3.773	22.072	0.339	0.339
Google (C)	0.876	4.316	4.437	18.427	5.147	4.316
Amazon (C)	0.818	13.474	11.901	28.623	12.120	13.475
Azure (P)	0.905	0.078	0.107	4.732	0.002	0.002
Google (P)	0.869	10.440	10.246	20.754	10.781	10.440
Amazon (P)	0.828	14.316	13.179	26.105	2.797	3.724

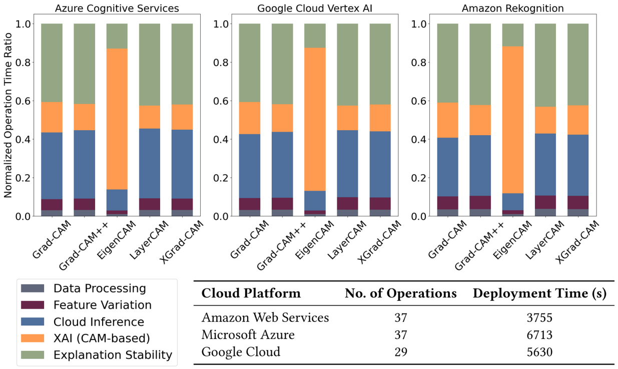
Note: "B" stands for baseline without augmentation, "C" stands for "Cutmix" and "P" stands for "Puzzlemix". The values in the table are XAI stability metrics [13]. (The smaller, the better explanation stability.)

Results and Discussion. Table 1 shows the detailed measurement results. The F1-score shows a subtle enhancement when both CutMix [37] and PuzzleMix [15] techniques are employed. A pronounced improvement is observed in the model's explanation evaluation. The consensus of XAI explanations has alignment with the performance of cloud AI services. Such a consistent explanation result provides a trustworthy view of the contribution of the data-driven technique to the model performance improvement.

These cloud AI services are entirely black-box. There is a lack of access to the model's parameters, the internal network structure, fine-tuned loss functions, and so on. The adoption of XAI through service orientation and open APIs has enabled us to probe the performance and obtain explanations.

4.2 Computational Analysis of XAI Operations and Deployment Across Cloud Services

We record and decompose the time spent on (1) Data Processing, (2) Feature Variation, (3) Cloud Inference, (4) XAI and (5) Explanation Stability. We analyze the operation consumption across the three cloud AI services and the five CAM-based XAI methods.



Cloud Platform	No. of Operations	Deployment Time (s)
Amazon Web Services	37	3755
Microsoft Azure	37	6713
Google Cloud	29	5630

Note: The chart displays the decomposition of average XAI execution time per data sample, derived from 1,000 experiments: Data Processing (0.12s ± 0.03s), Feature Variation (0.23s ± 0.06s), Cloud Inference (Azure 1.39s ± 0.42s, Google 1.26s ± 0.48s, Amazon 1.06s ± 0.28s), XAI methods: GradCAM (0.63s ± 0.12s), GradCAM++ (0.53s ± 0.08s), EigenCAM (9.19s ± 3.38s), LayerCAM (0.46s ± 0.06s), XgradCAM (0.51s ± 0.11s), Explanation Stability (1.60s ± 0.56s).

Figure 2: XAI Operations and Framework Deployment Time

Figure 2 shows the composition of each unit in XAI operations on average per data sample and the framework deployment duration. The model inference time is relatively stable across cloud services. However, XAI methods take different demands and there is a need for optimization in the evaluation metrics.

5 CONCLUSION

This paper outlines the early adoption of XAI operations in the practices of AI model quality assurance. We define the adoption context in three aspects with mature development methods and technology supports. We illustrate a pilot study on adopting XAI operations to answer a data-driven question with regard to improving three cloud AI services. We demonstrate consistent explanation results with measurements of the computation and deployment overhead. We advocate such a context of practice to broad open AI models' quality assurance with XAI to gain trustworthiness. In future work, we aim to further develop the software development toolkit (SDK) based on the `XAIport` framework. The SDK provides the tools for automated deployment of XAI operations along the AI service development using open pre-trained models, dynamic A/B testing of AI services, and validation of new XAI methods.

REFERENCES

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access* 6 (2018), 52138–52160.

[2] Amazon. 2023. *Amazon Web Services*. Amazon. https://aws.amazon.com Accessed: 2023.

[3] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* 58 (2020), 82–115.

[4] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 8 (2013), 1798–1828. https://doi.org/10.1109/TPAMI.2013.50

[5] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. 2018. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 839–847. https://doi.org/10.1109/WACV.2018.00097

[6] Chinu and Urvashi Bansal. 2023. Explainable AI: To Reveal the Logic of Black-Box Models. *New Generation Computing* (2023), 1–35.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[8] Christof Ebert, Gorka Gallardo, Josune Hernantes, and Nicolas Serrano. 2016. DevOps. *IEEE Software* 33, 3 (2016), 94–100. https://doi.org/10.1109/MS.2016.68

[9] Ruigang Fu, Qingyong Hu, Xiaohu Dong, Yulan Guo, Yinghui Gao, and Biao Li. 2020. Axiom-based grad-cam: Towards accurate visualization and explanation of cnns. *arXiv preprint arXiv:2008.02312* (2020).

[10] Google. 2023. *Google Cloud Platform*. https://cloud.google.com Accessed: 2023.

[11] Khan Mohammad Habibullah, Gregory Gay, and Jennifer Horkoff. 2022. Non-functional requirements for machine learning: An exploration of system scope and interest. In *Proceedings of the 1st Workshop on Software Engineering for Responsible AI*. 29–36.

[12] Aspen Hopkins and Serena Booth. 2021. Machine learning practices outside big tech: How resource constraints challenge responsible development. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 134–145.

[13] Jun Huang, Zerui Wang, Ding Li, and Yan Liu. 2022. The Analysis and Development of an XAI Process on Feature Contribution Explanation. In *2022 IEEE International Conference on Big Data (Big Data)*. 5039–5048. https://doi.org/10.1109/BigData55660.2022.10020313

[14] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. 2021. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing* 30 (2021), 5875–5888.

[15] Jang-Hyun Kim, Wonho Choo, and Hyun Oh Song. 2020. Puzzle mix: Exploiting saliency and local statistics for optimal mixup. In *International Conference on Machine Learning*. PMLR, 5275–5285.

[16] Dominik Kreuzberger, Niklas Kühl, and Sebastian Hirschl. 2023. Machine Learning Operations (MLOps): Overview, Definition, and Architecture. *IEEE Access* 11 (2023), 31866–31879. https://doi.org/10.1109/ACCESS.2023.3262138

[17] Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700* (2019).

[18] Philippe Lambin, Ralph T. H. Leijenaar, Timo M. Deist, Jurgen Peerlings, Evelyn E.C. de Jong, Janita E. van Timmeren, Sebastian Sanduleanu, Ruben T H M Larue, Aniek J. G. Even, Arthur Jochems, Yvonka van Wijk, H. Woodruff, Johan van Soest, Tim Lustberg, Erik Roelofs, Wouter van Elmpt, Andre Dekker, Felix M. Mottaghy, Joachim E. Wildberger, and Sean Walsh. 2017. Radiomics: the bridge between medical imaging and personalized medicine. *Nature Reviews Clinical Oncology* 14 (2017), 749–762.

[19] Nicholas Larsen, Jonathan Stallrich, Srijan Sengupta, Alex Deng, Ron Kohavi, and Nathaniel T Stevens. 2023. Statistical challenges in online controlled experiments: A review of a/b testing methodology. *The American Statistician* just-accepted (2023), 1–32.

[20] Ding Li, Yan Liu, Jun Huang, and Zerui Wang. 2023. A Trustworthy View on Explainable Artificial Intelligence Method Evaluation. *Computer* 56, 4 (2023), 50–60. https://doi.org/10.1109/MC.2022.3233806

[21] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[22] Microsoft. 2023. *Azure Cloud Services*. https://azure.microsoft.com Accessed: 2023.

[23] Mohammed Bany Muhammad and Mohammed Yeasin. 2020. Eigen-cam: Class activation map using principal components. In *2020 international joint conference on neural networks (IJCNN)*. IEEE, 1–7.

[24] Jean Jacques Ohana, Steve Ohana, Eric Benhamou, David Saltiel, and Beatrice Guez. 2021. Explainable AI (XAI) models applied to the multi-agent environment of financial markets. In *Explainable and Transparent AI and Multi-Agent Systems: Third International Workshop, EXTRAAMAS 2021, Virtual Event, May 3–7, 2021, Revised Selected Papers 3*. Springer, 189–207.

[25] Urja Pawar, Donna O'Shea, Susan Rea, and Ruairi O'Reilly. 2020. Incorporating Explainable Artificial Intelligence (XAI) to aid the Understanding of Machine Learning in the Healthcare Domain.. In *AICS*. 169–180.

[26] David Rolnick, Priya L Donti, Lynn H Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, et al. 2022. Tackling climate change with machine learning. *ACM Computing Surveys (CSUR)* 55, 2 (2022), 1–96.

[27] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* 1, 5 (2019), 206–215.

[28] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.

[29] Ruey-Kai Sheu and Mayuresh Sunil Pardeshi. 2022. A Survey on Medical Explainable AI (XAI): Recent Progress, Explainability Approach, Human Interaction and Scoring System. *Sensors* 22, 20 (2022), 8068.

[30] Stefan Studer, Thanh Binh Bui, Christian Drescher, Alexander Hanuschkin, Ludwig Winkler, Steven Peters, and Klaus-Robert Müller. 2021. Towards CRISP-ML (Q): a machine learning process model with quality assurance methodology. *Machine learning and knowledge extraction* 3, 2 (2021), 392–413.

[31] Swagger. 2023. *OpenAPI Specification*. https://swagger.io/specification/ Accessed: 2023.

[32] Matteo Testi, Matteo Ballabio, Emanuele Frontoni, Giulio Iannello, Sara Moccia, Paolo Soda, and Gennaro Vessio. 2022. MLOps: A taxonomy and a methodology. *IEEE Access* 10 (2022), 63606–63618.

[33] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature Medicine* 29, 8 (Aug 2023), 1930–1940. https://doi.org/10.1038/s41591-023-02448-8

[34] Giulia Vilone and Luca Longo. 2021. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion* 76 (2021), 89–106.

[35] Steven Euijong Whang and Jae-Gil Lee. 2020. Data collection and quality challenges for deep learning. *Proceedings of the VLDB Endowment* 13, 12 (2020), 3429–3432.

[36] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6

[37] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6023–6032.

[38] Jan Zacharias, Moritz von Zahn, Johannes Chen, and Oliver Hinz. 2022. Designing a feature selection method based on explainable artificial intelligence. *Electronic Markets* 32, 4 (2022), 2159–2184.

[39] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2921–2929.

[40] Mohd Zuhair, Pronaya Bhattacharya, Vivek Kumar Prasad, Manav Barot, and Monil Modi. 2022. Analysis of Boosting Mechanisms in Cloud-based Intrusion Detection Systems. In *2022 5th International Conference on Contemporary Computing and Informatics (IC3I)*. 961–966. https://doi.org/10.1109/IC3I56241.2022.10072683