Humanizing AI: A Human-Centered Architecture to Developing Trustworthy Intelligent Systems

Muhammad Uzair Akmal[1(✉)], Selvine George Mathias[1], Saara Asif[1], Leonid Koval[1], Simon Knollmeyer[1], and Daniel Grossmann[2]

[1] AImotion Bavaria
Technische Hochschule Ingolstadt, 85049 Ingolstadt, Germany

[2] Faculty of Computer Science and Data Processing
Technische Hochschule Ingolstadt, 85049 Ingolstadt, Germany

Abstract. The lack of trust and fairness in artificial intelligence (AI) systems driven by biases, misclassified data, lack of transparency, and limited interoperability, raises significant ethical concerns and socioeconomic impacts. This study presents a reference architecture for an AI pipeline aligned with Industry 5.0 principles, focusing on human-centered design, sustainability, social responsibility, and resilience. It enhances human-AI collaboration by involving four user types (data scientists, domain experts, organizations, and end users) who share decision-making responsibilities during the AI system development process. The architecture incorporates Active Learning (AL) to address data bias and misclassification issues and Transfer Learning (TL) to ensure model reusability in resource-constrained environments. Post-modeling Explainability gives stakeholders insight into model behavior and outcomes, fostering transparency and trust. Additionally, two user-ranked custom validation metrics evaluate the architecture and calculate Mean Average Precision (MAP) for Rankings. These metrics ensure the architecture design and outcomes adhere to ethical AI principles while promoting collaborative, responsible, and sustainable AI development.

Keywords: Artificial intelligence, Human-centric AI, Active learning, Transfer learning, Explainable AI, Intelligent systems, Industry 5.0.

1 Introduction

The recent industrial revolution known as Industry 5.0, has driven the focus of corporations to change their business strategies from purely economic to promoting social values and well-being [1]. Industry 5.0 introduces a new era of industrialization where human-AI collaboration is expected to drive workplace processes towards optimization [2]. Previously, Industry 4.0, focused on industrial digitalization with minimal human intervention and prioritized automation by relying deliberately on AI systems (robots, intelligent models, etc.) for decision-making and task completion [3]. Continuing with the current pace of technological advancement which is highly inclined towards implementing intelligent systems, however, lacks transparency, control, and trust, making it challenging for AI's future. The limited involvement of humans in the decision-making and AI system's development process from inception to completion triggers trust and fairness issues that result in lower acceptance of AI systems. Some negative impacts of AI dominance are unemployment, economic inequality, reduced human creativity and productivity, ethical and privacy concerns, security and safety risks, bias and discrimination, and the fear of AI misuse or exploitation [5]. Therefore, a shift is required that motivates the developers to master AI rather than relying completely on AI.

David C. Wyld et al. (Eds): SEC, DMA, AIAPP, CYBI, CoSIT, SIGL, NLPML – 2025
pp. 27-44, 2025. CS & IT - CSCP 2025 DOI: 10.5121/csit.2025.150802

At present, most AI systems are developed using either a model-centric or data-centric approach, focusing on improving the models or data quality without considering societal values and well-being [6]. However, recent developments have shifted the focus from designing AI systems prioritizing AI dominance to a more human-centered approach. This new approach, rather than replacing humans, aims to empower them, enabling them to complete tasks more actively and efficiently with the collaboration of AI. The objective is to create business strategies for AI systems that are human-centric [8] and adhere to human-in-the-loop (HITL) [11] principles by encompassing social values, transparency, and responsibility. The High-Level Expert Group on AI of the European Union presented Ethics Guidelines for Trustworthy AI in 2019 [7], that suggested AI systems must be accountable, explainable, unbiased, and must adhere to three core principles:

- **Lawful** by following laws and regulations.
- **Ethical** by following ethical principles and values.
- **Robust** by being adaptive, reliable, fair, and trustworthy in terms of technical aspects while considering its social environment.

1.1 Motivation

This study is aimed at presenting a reference architecture for the future intelligent systems by integrating human-centered approaches and HITL principles throughout the AI pipeline. The architecture is designed to address key challenges in AI development and deployment:

- Improvement of Data Quality and Collaboration: An active learning approach is integrated into the data preprocessing step to improve data quality and promote human-AI collaboration from the beginning.
- Enhancement of Model Reusability and Accessibility: In the model selection step, we apply transfer learning techniques to significantly improve the model reusability and accessibility across diverse applications and domains.
- Increased Interpretability and Trust: In the evaluation step, we incorporate post-modeling explainability to enhance the system's interpretability and build user trust.

This reference architecture is expected to bridge the gap between AI systems and human operators. It represents a significant step towards creating AI systems that are not only technologically advanced but also aligned with human values. By doing so, we aim to:

- Promote collaboration and sharing of responsibilities between humans and AI.
- Ensure informed and equitable human participation during decision-making processes.
- Bridge the gap between next-generation AI technologies and their practical, ethical application in real-world scenarios.

1.2 Contributions

The major contributions of our study are listed below:

- Design of a conceptual human-centric AI (HCAI) architecture.
- Definition of customized validation metrics to verify the architecture design and outcomes against effectiveness and ethical compliance.
- Definition of an evaluation criteria for evaluating the architecture through Mean Average Precision (MAP) for Rankings.

The remainder of the paper is organized as follows: Section 2 provides a brief overview of the foundational aspects used in the proposed architecture. Section 3 presents the related work covering the recent advancements in foundational aspects on which the architecture is based. Section 4 covers the architecture design in detail and a use case-based application is given in Section 5. In Section 6, we define customized validation metrics to verify the design and outcomes of the proposed architecture. Section 7 covers a brief discussion on the proposed architecture followed by its limitations. Finally, in Section 8, we conclude the study by highlighting the current research gaps and suggesting directions for future work.

2 Key Aspects of the Proposed Architecture

The architecture is designed based on the principle of Trustworthy AI [4] and Industry 5.0 [3]. Table 1 defines the factors corresponding to the principles of Trustworthy AI development. By incorporating these factors, the architecture is aimed to address the various challenges associated with AI adoption, while promoting responsible AI development practices [4].

Table 1. Factors to incorporate for Trustworthy AI development [4].

Factors	**Definition**
Trust	Users should have confidence in the AI system to perform its tasks reliably, ethically, and transparently.
Fairness	Equal treatment of all users by AI systems by avoiding discrimination based on race, gender, age, or other characteristics.
Transparency	The tasks and decision-making processes of the AI system should be explainable, understandable, and allow users to see how these decisions are being made.
Interoperability	AI systems' decisions and outcomes should be understandable and interpretable for users.
Ethical	AI systems should conform to ethical guidelines and standards and ensure that they comply with user privacy needs and societal norms.
Responsibility	AI systems should be accountable for their actions and decisions. The identification of who is responsible for AI's outcomes and its deployment should be clear.

2.1 Industry 5.0

Industry 5.0 prioritizes a human-centered approach based on societal values rather than economic values [1]. The idea behind Industry 5.0 is to implement a collaborative environment where humans and AI systems work together to achieve tasks with maximum efficiency and optimize the manufacturing industry. It involves designing AI systems that are transparent, sustainable, robust, and efficient through human-machine collaboration and active human involvement in the decision-making process [2]. Recent advancements in Industry 5.0 [1–3] include the development of advanced manufacturing robots and industrial automation systems, where robots can assist by taking over repetitive tasks, thereby enhancing overall productivity and sustainability by promoting systems powered by renewable energy sources [19]. An overview of the foundational aspects of the proposed architecture designed under Industry 5.0 is shown in Fig. 1.

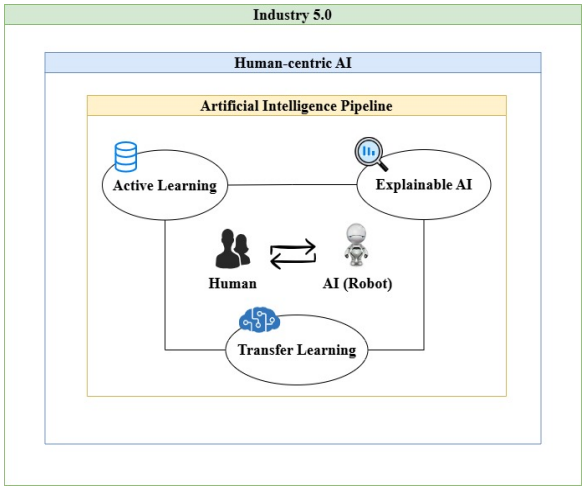

Fig. 1. An abstract-level overview of the architecture designed under the Industry 5.0 umbrella.

2.2 Human-centric AI

Human-centric AI (HCAI) emphasizes human-machine collaboration and follows HITL principles by ensuring that technology should be used for the betterment of society and to support and enhance human capabilities [8]. It focuses on designing AI systems that work collaboratively with humans towards shared goals unlike traditional AI, which is a fully autonomous decision-making system and mainly focuses on maximizing efficiency or performance. HCAI is an iterative process that requires continuous monitoring, feedback, and refinement to ensure that the design and implementation of AI systems are aligned with human needs, social values, and well-being [12]. The underlying characteristics of human-centric AI on which the proposed architecture is based, are briefly defined in Table 2.

Table 2. Definition of the key characteristics of Human-centric AI [9, 10].

Factors	**Definition**
Human-in-the-Loop	HITL refers to informed and equitable human involvement in AI systems, emphasizing shared responsibilities where AI is designed to support and enhance human decision -making rather than replace it.
Transparency	Transparency refers to building user trust and enabling users to rely confidently on AI. It ensures that users and stakeholders can understand how the AI system operates and why it produces specific outcomes.
Fairness	Fairness refers to designing AI systems that treat everyone equally, reduce biases, and promote inclusion across diverse user groups and societal contexts.
Ethics	Ethics refers to designing AI systems around human needs, values, and experiences. It ensures that ethical guidelines are followed throughout the development process while ensuring user privacy needs and societal norms.
Sustainability	Sustainability refers to the development and deployment of AI systems while focusing on responsible use of resources, following ethical practices, and ensuring AI supports long-term social and environmental well-being.
Usability	Usability refers to the designing of AI systems that are easy to use and people can interact effectively with them to achieve their goals. It emphasizes on creating simple, user-friendly designs that anyone can understand and use effectively.

2.3 Active Learning

Active learning is a data annotation approach where an algorithm (active learner) interactively queries users to label data with the desired outputs [17]. The learner proactively selects the subset of data to be labeled next from the pool of unlabeled data, embeds it in a query, and passes the request to the oracle (human annotator) for labeling [17] as shown in Fig. 2. The involvement of human annotators in learning and utilizing their expert knowledge to improve the labels makes active learning part of the HITL paradigm [18]. The labeled data is then used to train an ML model that predicts the labels for the remaining unlabeled dataset. Some of the most used and recent techniques for active learning are Deep Active Learning [20, 21], Adversarial Active Learning [22], Diversity-Based Approaches [23], Query by Committee (QBC) [24], Uncertainty Sampling [25], and hybrid approaches by combining multiple techniques [26].

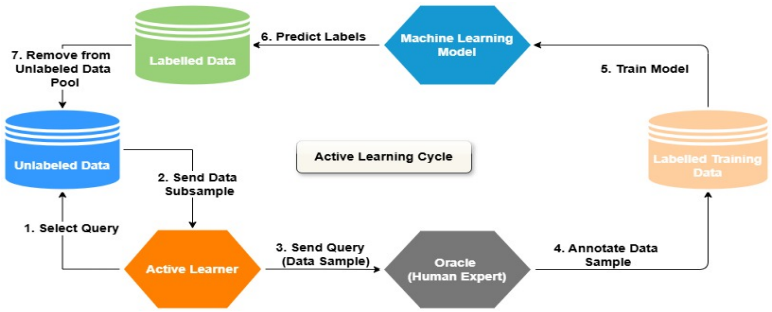

Fig. 2. A typical life-cycle of Active Learning approach [17].

2.4 Transfer Learning

The knowledge of a pre-trained machine learning model is applied to a different but related problem in the transfer learning approach [27]. The general idea is to utilize the knowledge and patterns a model has learned from a task with a considerable volume of labeled training data for a new task with limited data as shown in Fig. 3. Transfer learning has seen significant advancements in recent years, driven by the rapid evolution of deep learning and the growing availability of pre-trained models. Some of the most used and recent techniques in transfer learning are domain adaptation, feature extraction, self-supervised learning, unsupervised learning, and meta-learning [28–30].

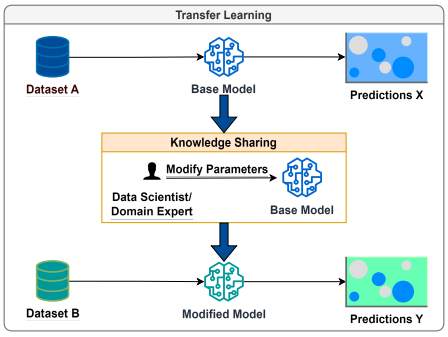

Fig. 3. A typical life-cycle of Transfer Learning approach [27].

2.5 Explainable AI

Explainable Artificial Intelligence (XAI) focuses on making AI models' decisions understandable to humans. XAI is a set of processes and methods that allows human users to comprehend and trust the results and output created by machine learning algorithms [31]. XAI implements specific techniques and procedures to ensure that each decision made during the ML process can be traced and explained as shown in Fig. 4. Some of the most used and recent techniques in explainable AI are model-agnostic or model-specific explanation methods, interpretable models or post-hoc interpretation methods, interactive and visual explanation tools, and multi-modal explanations [32–34].

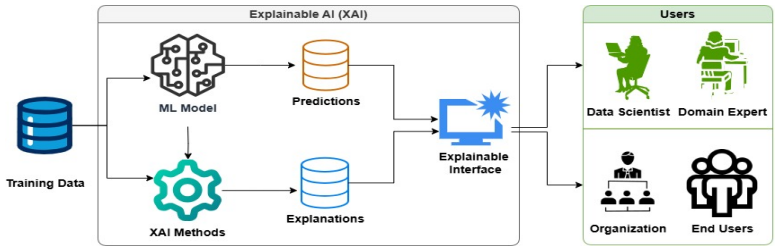

Fig. 4. An overview of the Explainable AI pipeline [31].

3 Related Work

3.1 State-of-the-Art

The current literature explains Industry 5.0 and its recent advancements by recognizing the importance of integrating HITL approaches in AI tasks, emphasizing human involvement in automation processes. The authors of [1–3] focused on the recent shift toward Industry 5.0 and its enabling technologies that extend Industry 4.0 by including collaborative robots, blockchain, and advanced data analytics. A balanced technological advancement with human needs, environmental concerns, and artificial intelligence is the need of the hour focusing on personalization, sustainability, and human-machine interaction. With a widespread application across domains like healthcare, manufacturing, and supply chain, the authors highlight the challenges associated with the paradigm shift.

A review to identify, evaluate, and analyze human-centered AI papers is carried out by the authors of [8]. After studying multiple human-centered AI frameworks, the authors stress that machines should augment human effort rather than replace it, and the key goal of HCAI should be to produce reliable, safe, and trustworthy systems. Moreover, it emphasizes the importance of recognizing AI as a product of human values and the need to explicitly guide its development to benefit all stakeholders and society at large. Whereas the authors of [12] argue that AI is already human-centric by highlighting that AI technology is evidence of human activity, as it is designed by humans to help humans. Later the authors present two community-centered frameworks for developing and deploying AI systems that are more inclusive of humans and aligned with human values. The author of [13] focuses on prioritizing reliability, safety, and trustworthiness, and advocates for a human-integrated approach to developing AI systems keeping humans in mind and prioritizing human needs and values. The study introduced a framework to balance human control and AI autonomy during AI systems development. The study [14] presents

a systematic literature review of 162 publications from 2016 to 2020, where the authors explore the integration of deep learning techniques within human-centered machine learning (HCML). By categorizing the work based on adaptability and usability, the authors analyzed the machine learning systems that were developed by prioritizing human needs. The authors have highlighted the challenges and opportunities associated with the HCML field by emphasizing the need for more human-understandable and collaborative AI systems. The proposition is to introduce features like explainability and interoperability for user engagement and control during the AI development cycle.

The authors of [15] introduced TagLab, a software developed with a key focus on the HITL scheme, with AI-generated segmentation and classification for underwater imagery. The process of incorporating human intervention and feedback shows significant improvement in annotation speed and accuracy showcasing an active learning approach as compared to traditional manual methods. Similarly, the authors of [16] designed a Human4ML framework that looks at the need to have HITL to ensure effective labeling, proper data collection, consistent data quality, effective feature space construction and incorporating trust and transparency into AI applications. To validate this, authors from [18] also highlighted the role of HITL approaches in optimizing ML processes including annotation, active learning, and transfer learning techniques. The paper [11] offers a thorough overview of the state-of-the-art in HITL for a machine learning pipeline beginning from the labeling and annotation process, followed up with interactive machine learning through human guidance eventually leading to explainable AI for better comprehension of the pipeline and the underlying AI strategy, thereby making it easier to identify and rectify errors or biases in AI systems. Although the authors link various approaches for handling multi-modal data with human involvement, the overall picture remains unclear.

3.2 Research Gap

In the given state-of-the-art, scattered indications of a formal process are apparent, but no evidence of a unified architecture is visible. Nevertheless, some of these frameworks are very well structured and perfectly suitable for comparison. Human4ML [16], for instance, provides a lifecycle perspective in three phases: human-guided data preparation, human-assisted feature construction and model learning, and interactive model assessment and explanation. Such phases clearly relate with the data analysis, modelling, and evaluation and deployment steps similar to our work but leave critical gaps. The fundamental collaboration models and mechanisms remain unclear. Additionally, there is a need for a deeper understanding of human biases in HITL AI systems. Building on this foundation, our work identifies the key users involved in the development process and provides a clear outline of their roles, responsibilities, and the areas for collaboration. We are examining three key methodologies: Active Learning, Transfer Learning, and Explainable AI and analyzing how these methodologies are collectively applied and implemented within the AI pipeline. By effectively integrating these methodologies into an architecture, we aim to illustrate their role in stimulating human-centricity, which contributes to creating more efficient, trustworthy, transparent, adaptable, and ethically aligned AI systems. The current literature emphasizes the significance of human involvement in AI processes but fails to address critical questions such as *Who to involve? Where to involve users in the AI pipeline? What is the role of each user? What tasks will each user perform? How will user involvement improve the processes? How to measure the improvement? What can be trusted? Who will be responsible for AI actions?*

To bridge these gaps, we propose a comprehensive roadmap as a unified reference archi-

tecture for AI system development that systematically incorporates users and integrates advanced methodologies. Our architecture aims to enhance the inclusivity and sustainability of AI systems. It outlines specific points in the AI pipeline where user involvement is essential, identifies relevant stakeholders, and defines their respective roles and responsibilities. The architecture ensures that user participation is facilitated and optimized to improve the overall effectiveness, transparency, and ethical alignment of AI processes. This approach promises a more collaborative and sustainable future for AI development.

4 Human-centered Architecture for AI

The architecture is designed to ensure responsibility in AI systems and enable equitable involvement of both humans and AI in all major milestones and decision-making processes during the development of AI systems. The primary objective is integrating human-centric characteristics mentioned in Table 2 into the AI pipeline. The architecture is detailed in Fig. 5, outlining the various phases that compose it.

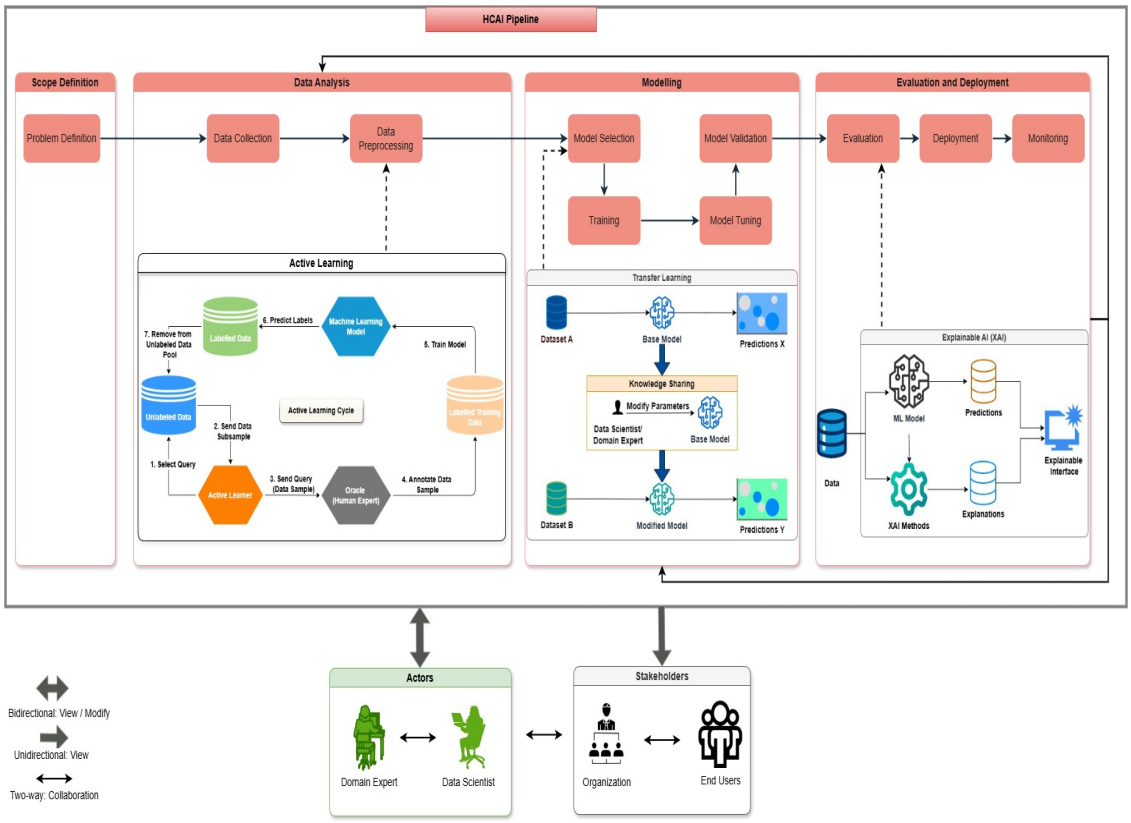

Fig. 5. A detailed overview of the proposed architecture(HCAI Pipeline) with added modalities in the state-of-the-art AI pipeline.

4.1 Identification of Actors and Key Stakeholders

To incorporate responsibility in AI systems and make it easier to trace who is responsible for AI actions and deployment, it is important to identify the responsible actors that are involved in designing and implementing the AI system. To serve this purpose, the proposed

architecture includes data scientists and domain experts as the main actors. Additionally, to keep humans in the loop for decision-making processes, it is important to recognize the stakeholders involved such as organizations, and end users (e.g., medical practitioners, production employees, etc.) that will be directly benefited or affected by the AI system. The users involved in the architecture, their activity, collaboration, and associated goals are shown in Fig. 6.

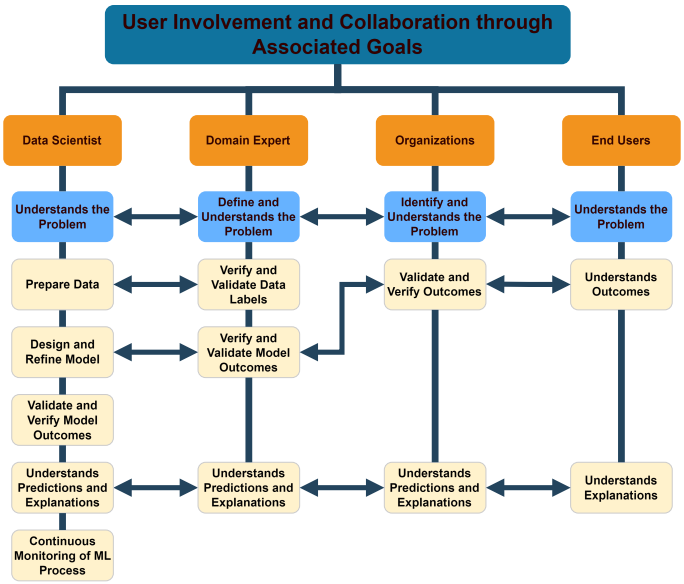

Fig. 6. An overview of the users involved in the architecture and their collaboration through associated goals.

4.2 Integration of Human-centric Approaches in the AI Pipeline

We enhance a state-of-the-art AI pipeline by integrating three techniques to promote trust, ensure fairness, and increase human-AI collaboration to achieve shared goals. We are explaining below how active learning, transfer learning, and explainable AI are integrated into the AI pipeline keeping the stakeholders involved.

Scope Definition: The project's scope involves a clear definition of the problem that needs to be solved and the objectives of the AI system. To ensure human feedback and decisions are essentially incorporated into the process, project scope definitions and goals are decided through close collaboration between actors and stakeholders. Once an agreement is reached, the actors proceed towards the data analysis step.

Data preprocessing through Active Learning Approach: We introduce an active learning approach based on human-centric principles to address the challenge of limited labeled data. This approach automates the repetitive data labeling task, enhances labeling outcomes, and reduces misclassification through equitable collaboration between AI and domain experts/labelers. Providing data insights to stakeholders ensures transparency and allows them to identify and correct any bias or misrepresentation. Providing data insights to stakeholders is crucial for confirming against any bias or misrepresentation, thereby ensuring transparency in the process.

Model Selection through Transfer Learning: Transfer learning promotes knowledge sharing and model reusability within the same domain and problem space. It allows for rapidly adapting AI models to similar problems, facilitating quicker iteration and collaboration between human developers and AI systems. Although transfer learning is not inherently human-centric, incorporating it into the architecture provides sufficient transparency for users. It enables stakeholders to understand the problem, the handling approach, and the critical parameters involved, which helps in validating the results effectively.

Evaluation through Post-Modeling Explainability: The explainability mechanism is integrated into the evaluation step of the AI pipeline against model predictions to offer interpretable insights and explanations against the model decisions. The explainable interface enables users to answer critical questions such as: *Do users understand why or why not the model made a prediction? Do users understand when the model is successful or when it is a failure? Do users know why the predictions are correct or incorrect? Do users know when to trust the model? Do users know why the model might have erred?* This transparency fosters trust and allows users to engage with the system more effectively.

4.3 Deployment and Continuous Monitoring

This modular architecture enhances the existing AI pipeline by integrating additional techniques that improve processes, optimize outcomes, and increase the adaptability of AI systems. These added modalities build upon the state-of-the-art pipeline without altering its core structure, focusing instead on refining and extending its capabilities. The modalities can be added or removed without disrupting the overall AI pipeline. The architecture provides flexibility in selecting the most suitable methods for implementing these techniques. Continuous monitoring in the architecture is achieved through an integrated feedback loop, where the predictions are verified and validated, and any misclassifications are used to iteratively update and improve the model. This iterative process enhances the robustness, scalability, transparency, and responsibility of the pipeline, ultimately ensuring the ethical integrity of the AI system.

5 Architecture Application for Pet and Stray Dog Recognition

Applying the architecture to a high-level example of a pet and stray dog recognition system illustrates a conceptual overview. The idea is to recognize stray and pet dogs to provide rabies vaccinations. We explain the overall structure, important steps, a broad understanding of how things fit together, and how our architecture will approach this problem of building a pet and stray dog recognition system.

5.1 Problem Definition

Use Case: Scheduling of vaccination against rabies for dogs identified as stray.

Identify Actors and Stakeholders:

- Actors: The *data scientist* is responsible for developing the system in collaboration with *domain experts* who validate and verify decisions and outcomes based on their expertise and knowledge.
- Stakeholders: *Organizations* responsible for vaccinating the dogs and *end users* such as dog owners or people infected with rabies.

Objective: Identify a dog as a pet or stray based on the characteristics mentioned below.

- Environmental context i.e. *location* where the dog is found and *companionship* whether the dog is accompanied by a human.
- Identification marks i.e. the dog has *collars, tags, or microchips.*

The objectives are finalized based on a close collaboration between organizations, domain experts, and data scientists. The objectives are communicated to end users to ensure trust and transparency during the development of the recognition system.

5.2 Dataset Collection

Identification of Data Sources: Data can be collected from a combination of sources such as pet owners, social media, animal shelters, street cameras, existing datasets, and public databases.

5.3 Data Preprocessing

Active Learning Cycle: A small balanced dataset (100 labeled images of dogs, consisting of 50 pets and 50 strays) labeled by data scientists and validated by domain experts can be used initially to train a model that may not be very accurate but serves as a starting point. For this classification problem, the uncertainty sampling technique is used. The model identifies the most "uncertain" data points that it struggles to classify and then queries for their labels to data scientists. This conforms to the HITL step. The data scientist labels these data points and the domain experts validate them to improve the labeling and model learning efficiency. Once labeled, these uncertain instances are added to the training set, and the model is retrained with the newly labeled data. The model is expected to improve performance, especially in difficult-to-classify instances. The process is repeated and the retrained model is again used to predict labels for the remaining unlabeled data, identify the most uncertain samples, query them for labels, and retrain until the performance meets the desired threshold. Once the model's accuracy is satisfactory, it automatically labels the remaining unlabeled data. A subset of these labels is still manually validated to ensure the model's predictions are accurate. This step is crucial to verify that the model maintains high accuracy across different subsets of the data and does not introduce significant bias.

5.4 Model Selection

The idea behind transfer learning is to use a pre-trained model originally trained on a large dataset for a related task and adapt it for a specific, smaller task like recognizing pet dogs. To apply transfer learning, we utilize a pre-trained convolutional neural network (CNN) [35] and customize the model to fit our problem. The first step is selecting an appropriate pre-trained model. Since our task is related to image recognition, we can start with a model such as EfficientNet, ResNet, Inception, or VGG, which have been trained on large datasets like cats and dogs [36] or Stanford Dog dataset [37] that include many animal classes, including dogs. These models are already mature in extracting important features (such as shapes, textures, etc.) from images, which we can leverage. The lower layers of the model are frozen and capture generic features like edges and textures whereas, the upper layers are modified and retrained to learn the specific differences between the two categories. To distinguish between a stray dog and a pet dog the presence of a collar or specific information such as location, environment, etc. is used by the model.

5.5 Model Evaluation

We can measure the model's accuracy on unseen test data samples to ensure it can correctly classify pets from stray dogs. Numerous, evaluation metrics such as accuracy, precision, and recall, etc. can be used to evaluate the model's performance based on the true and false classes. A visual representation can be best illustrated using a confusion matrix for stakeholders' understanding of the model outcomes.

5.6 Post-Modelling Explainability

The purpose of including post-modeling explainability is to enable users and stakeholders to understand the model's outcomes and provide a logical explanation as to why the model identifies a dog as a pet or stray. To achieve this a combination of explainability techniques such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) [32, 33] can be employed. These techniques are then integrated into a user-friendly interface that provides interactive visual explanations of predictions and feature importance by utilizing visualization tools like Tableau [40, 41] and TensorBoard [38, 39]. Moreover, validating these explanations with stakeholders involved in the process ensures their relevance and alignment with the desired goals. This transparency and interoperability are expected to create trust in the AI model's predictions while enhancing user adaptability and confidence in the AI system.

5.7 Deployment and Monitoring

A suitable approach for deploying a trained and tested AI model is to utilize Docker [44] to ensure consistency and portability across multiple platforms. The AI pipeline can be modularized into independent components such as data preprocessing, model inference, and evaluation. These components can then be deployed as a web service using modern frameworks like Flask [42] or FastAPI [43], where an image is given as input to the web service and it processes it through the pipeline and returns a prediction indicating whether the dog is a pet or a stray. For handling challenges like data drift, continuous monitoring is integrated into the architecture to detect and adapt to new data types or features not present in the original training dataset. A feedback loop is implemented to validate and verify predictions, allowing data scientists and domain experts to identify any misclassifications, which is then used to refine and update the model, ensuring continuous improvements and efficiency.

6 Evaluation

The evaluation of the architecture is carried out in two stages: first, the design of the architecture and later the outcomes generated by the architecture will be assessed. We have introduced two validation metrics, each comprising multiple factors. The metrics for validating the design of the architecture are based on the factors defined in Table 3. The metrics for validating the outcomes (predictions) from the architecture are based on the factors defined in Table 4. For both architecture design and architecture outcomes, the users such as data scientists, domain experts, stakeholders, and end users rank each factor as positive or negative. The rank is binary, with *positive* being *1*, and *negative* being *0*. Finally, to calculate the effectiveness of both metrics, we are using *Mean Average Precision (MAP)* [45] for rankings [46], enabling a comprehensive assessment of how well the architecture aligns with the evaluation criteria.

Table 3. Metrics and Factors for validating the architecture design.

Metric	Factors
Collaboration	Cooperation: Does the design include human-AI cooperation to achieve a shared goal? Communication: Does the design include an active feedback system? Coordination: Do the users interact for important decision-making?
Adaptability	Modularity: Is it possible to test individual components and add or remove modalities without affecting the overall ML pipeline? Reconfigurability: Is it possible for the architecture design to adapt to changing requirements or functionalities through minimal modifications? Flexibility: Does the architecture design allow for the integration of new techniques or methods with minimal disruptions? Robustness: Is the performance being maintained under varied conditions like unforeseen challenges or disruptions?
Usability	User-Centered: Is the architecture easy to use and incorporates user needs and satisfactions while developing AI systems? Learnability: Does the design enable users to quickly understand its features and functions by reducing the training time? Reusability: Is it possible for the architecture to accommodate different use cases to meet varying requirements without significant modifications?
Sustainability	Socially Responsible: Is it possible to ensure fairness, transparency, and inclusivity in AI systems, following the architecture design? Energy Consumption: Does the architecture design involve unnecessary computations that increase power consumption? Memory Consumption: Does the architecture design introduce unnecessary and repetitive computations that lead to increased memory consumption?
Scalability	Complexity: Does the architecture design have high component inter-dependencies, where modifications in one part affect the complete ML process? Integration: Are the added modalities in the architecture design easily compatible with the existing ML pipeline? Time and Cost Effectiveness: Does the architecture design help in optimizing costs and minimizing rework? Security and Compliance: Does the architectural design comply with defined regulations and standards?

Table 4. Metrics and Factors for validating the architecture outcomes.

Metric	Factors
Understandability	Clarity: Are the predictions meaningful for end users? Correctness: Do the predictions align with true values? Consistency: Are the predictions free from deviations? Relevance: Do the predictions align with the desired outcomes or objectives of the system?
Fairness	Unbiased: Are the model predictions equal to the expected true values over the relevant data? Authenticity: Are the predictions transparent, based on quality data, and accurately reflect the real world?
Ethical	Conformity: Are the predictions fair and aligned with societal values and defined ethical standards? Reliability: Are the predictions dependable even for new and unforeseen scenarios? Responsibility: Are the predictions free from harmful biases and support positive real-world implications?
Explainability	Interoperabillity: Are the predictions clear, accurately understandable, and provide actionable insights across different platforms and user groups? Reasoning: Do the predictions enable users in decision-making and allow them to understand how and why specific outcomes are reached?

6.1 Mean Average Precision (MAP) for Rankings

The evaluation of the architecture based on user ranks by computing MAP against hypothetical values is shown in Table 5. Where, each user involved in the AI pipeline has ranked the factors f_y as either 1 or 0. The precision corresponding to the individual factors based on positive and negative ranks is calculated by (1).

Table 5. A hypothetical evaluation based on user rankings to calculate MAP for the architecture.

	Metric	Factors	Data Scientist	Domain Expert	Stakeholders	End Users	P(f_y)	AP(M_x)	$(MAP)_{AD}$	$(MAP)_{OUT}$
Architecture Design	Collaboration	Cooperation	1	1	1	1	1	0.83	0.85	**N/A**
		Communication	1	1	1	0	0.75			
		Coordination	1	1	1	0	0.75			
	Adaptability	Modularity	1	1	1	1	1	0.81		
		Reconfigurability	1	1	1	1	1			
		Flexibility	1	1	0	1	0.75			
		Robustness	1	1	0	0	0.50			
	Usability	User-Centered	1	1	1	1	1	1		
		Learnability	1	1	1	1	1			
		Reusability	1	1	1	1	1			
	Sustainability	Socially Responsible	1	1	1	1	1	0.92		
		Energy Efficient	1	1	1	1	1			
		Memory Efficient	1	1	0	0	0.75			
	Scalability	Complexity	1	1	1	1	1	0.69		
		Integration	1	1	1	0	0.75			
		Time and Cost Effective	1	0	0	0	0.25			
		Security and Compliance	1	1	1	0	0.75			
Architecture Outcomes	Understandability	Clarity	1	0	0	0	0.25	0.75	**N/A**	0.71
		Correctness	1	1	1	1	1			
		Consistency	1	1	0	1	0.75			
		Relevance	1	1	1	1	1			
	Fairness	Unbiased	1	1	1	1	1	0.75		
		Authenticity	1	1	0	0	0.50			
	Ethical	Conformity	1	1	1	0	0.75	0.58		
		Reliability	1	1	0	0	0.50			
		Responsibility	1	1	0	0	0.50			
	Explainability	Interoperability	1	1	1	0	0.75	0.75		
		Reasoning	1	1	1	0	0.75			

$$P(f_y) = \frac{\texttt{Total PR}}{\texttt{Total PR} + \texttt{Total NR}} \tag{1}$$

where, *Total PR* corresponds to the total positive ranks and *Total NR* corresponds to the total negative ranks.

Average precision for metric M_x, based on precision per factor f_y, is calculated by (2).

$$AP(M_x) = \frac{1}{N} \sum_{y=1}^{N} P(f_y) \tag{2}$$

where, N is the total number of factors for metric x.

MAP for the architecture design MAP_{AD} based on 5 metrics is calculated by (3).

$$(MAP)_{AD} = \frac{1}{5} \sum_{x=1}^{5} AP(M_x) \tag{3}$$

MAP for the architecture outcomes MAP_{OUT} based on 4 metrics is calculated by (4).

$$(MAP)_{OUT} = \frac{1}{4} \sum_{x=1}^{4} AP(M_x) \tag{4}$$

A generic formula for calculating MAP is given by (5).

$$MAP = \frac{1}{K} \sum_{x=1}^{K} AP(M_x) \tag{5}$$

where, K is set to 5 for architecture design and 4 for architecture outcomes depending on the total number of metrics.

7 Discussion and Limitations

The architecture is designed to incorporate human-centricity in AI system development. We have tried to strengthen AI trust, transparency, safety, and adaptability by involving HITL in decision-making and enabling human-machine collaboration. The active learning approach embedded in the architecture optimizes the data annotation process and improves model performance with limited labeled data by reducing the costly and repetitive data labeling efforts. To increase the reusability of models, transfer learning is introduced in the architecture to enable the rapid adaptation of AI models to new problems. However, it is important to recognize its potential risks, especially in the context of fairness. Pre-trained models can carry biases from their source domain to the target domain as highlighted in the studies [48, 49]. This transfer of bias can undermine efforts to ensure fairness, potentially introducing new forms of inequity or amplifying existing ones. Therefore, to mitigate potential biases from pre-trained models, we involve domain experts Using an active learning approach, AI and human experts are working together to improve data quality. After training, we are using explainable AI tools to identify any unfair patterns or biases in the model's predictions. Additionally, we have domain experts and stakeholders review and validate the model to ensure it is fair and ethically aligned. By incorporating these techniques, we can possibly make transfer learning unbiased. Explainable AI introduced in the architecture provides understandability and enhances user engagement and trust which as a result makes the AI systems more adaptable. Moreover, the architecture ensures flexibility and scalability by enabling seamless integration and removal of modalities without disrupting the core functionality of the AI pipeline.

To demonstrate the practical applications of the architecture, we have applied the architecture to a hypothetical problem, navigating it through the steps of the AI pipeline using a synthetic working example, as detailed in Section 5. However, the architecture can be applied in various domains, including healthcare and autonomous systems. In the automotive industry for Advanced Driver-Assistance Systems (ADAS), HCAI can help develop systems that enhance driver safety and comfort [50]. These systems adapt to individual driving styles and provide intuitive assistance, ensuring a balanced interaction between humans and machines. In medical imaging applications, explainable AI techniques can help interpret complex deep learning models, enabling doctors to understand how decisions are made in tasks like tumor identification and organ segmentation [51]. Moreover, it can be utilized to provide personalized treatment through AI systems that can analyze an individual patient's health and genetic data to identify the most effective treatment options. In addition, in diagnostic support, AI algorithms can analyze patient data to help diagnose diseases more accurately and quickly, improving patient outcomes and experiences [52].

This study is focused mainly on defining the key characteristics of human-centric AI and using these principles to design a human-centered architecture for developing future AI

systems. However, it is important to consider that utilizing active learning and explainable AI in real-world human-centered systems has several challenges. AL requires frequent model retraining, and especially for large-scale models, it can be time-consuming and resource-intensive, making AL a computational overhead [53]. Moreover, it is difficult to embed explainability into complex models like deep neural networks, which often provide better accuracy but are harder to understand and interpret, which as a result leads to end-user resistance in adapting such systems [54].

8 Conclusion and Future Directions

This study highlights the importance of Industry 5.0, where more human intervention is appreciated and work has to be carried out by equally incorporating humans and artificial intelligent (AI) systems such as robots, AI agents, and intelligent models. The overall goal is to balance technological advancement with human needs and environmental concerns, addressing the limitations of Industry 4.0 by focusing on personalization, sustainability, and human-machine interaction. In response to this, our study introduces a reference architecture designed to integrate human-centric methodologies such as active learning, transfer learning, and explainable AI, thereby enhancing the trustworthiness, transparency, and adaptability of AI systems and ensuring that they are developed to support humans rather than replacing them and utilized for the benefit of society. The techniques integrated into the architecture have proven their ability and effectiveness already across various AI related problems [13, 18, 34, 47]. Additionally, a detailed application-level overview has been presented in the study which explains how the architecture handles a problem in different phases of the AI pipeline. Finally, a validation metric is introduced which can be utilized to investigate, whether the design and outcomes from the architecture conform with human-centered and ethical AI principles. Ultimately, this study provides a human-centered architecture for designing and developing future intelligent systems that are ethically and morally aligned with human and societal values.

Continuation of our work will involve implementing the proposed architecture with real-world datasets to solve real-world problems and deploying it to assess its feasibility across multiple industrial domains. The main focus will be to evaluate its impact on the overall ML pipeline and analyze key metrics such as cost, time, energy efficiency, and resource utilization. The outcomes will then help to understand and quantify the complexities involved and determine the viability of the proposed architecture in diverse industrial applications.

Acknowledgments

Declaration of generative AI and AI-assisted technologies in the writing process

While preparing this work, we used ChatGPT [55] to correct sentence structure, improve text clarity, and solve grammatical errors. We would like to highlight that the tool is not used to generate complete text or paragraphs to populate the sections, for data analysis, code generation, or any ideas related to the proposed architecture and concept. After using this tool, we carefully reviewed and edited the generated content as needed and took full responsibility for the content of the published article.

References

1. Adel, A. Future of industry 5.0 in society: human-centric solutions, challenges and prospective research areas. *Journal Of Cloud Computing.* **11** (2022,9)

2. Demir, K. & Cicibaş, H. The Next Industrial Revolution: Industry 5.0 and Discussions on Industry 4.0. (2019,1)
3. Barata, J. & Kayser, I. Industry 5.0 – Past, Present, and Near Future. *Procedia Computer Science*. **219** pp. 778-788 (2023), https://www.sciencedirect.com/science/article/pii/S1877050923003605, CENTERIS – International Conference on ENTERprise Information Systems / ProjMAN – International Conference on Project MANagement / HCist – International Conference on Health and Social Care Information Systems and Technologies 2022
4. Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J. & Zhou, B. Trustworthy AI: From Principles to Practices. *ACM Comput. Surv.*. **55** (2023,1), https://doi.org/10.1145/3555803
5. Cheng-Tai, M. The impact of artificial intelligence on human society and bioethics. *Tzu-Chi Medical Journal*. **32** pp. 339 - 343 (2020), https://api.semanticscholar.org/CorpusID:222125998
6. Hamid, O. From Model-Centric to Data-Centric AI: A Paradigm Shift or Rather a Complementary Approach?. *2022 8th International Conference On Information Technology Trends (ITT)*. pp. 196-199 (2022)
7. AI, H. High-level expert group on artificial intelligence. *Ethics Guidelines For Trustworthy AI*. **6** (2019)
8. Domfeh, E., Weyori, B., APPIAHENE, P., Mensah, J., Awarayi, N. & Afrifa, S. Human-Centered Artificial Intelligence, a review. *Authorea Preprints*. (2022)
9. Lepri, B., Oliver, N. & Pentland, A. Ethical machines: The human-centric use of artificial intelligence. *IScience*. **24** (2021)
10. Schmager, S. & Vassilakopoulou, P. Defining Human-Centered AI: A Comprehensive Review of HCAI Literature. (2023,9)
11. Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J. & Fernández-Leal, Á. Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*. **56**, 3005-3054 (2023,4), https://doi.org/10.1007/s10462-022-10246-w
12. Taylor, R., O'Dell, B. & Murphy, J. Human-centric AI: philosophical and community-centric considerations. *AI SOCIETY*. (2023,5), https://doi.org/10.1007/s00146-023-01694-1
13. Shneiderman, B. Human-Centered Artificial Intelligence: Reliable, Safe Trustworthy. (2020), https://arxiv.org/abs/2002.04087
14. Kaluarachchi, T., Reis, A. & Nanayakkara, S. A Review of Recent Deep Learning Approaches in Human-Centered Machine Learning. *Sensors*. **21** (2021), https://www.mdpi.com/1424-8220/21/7/2514
15. Pavoni, G., Corsini, M., Ponchio, F., Muntoni, A. & Cignoni, P. TagLab: A human-centric AI system for interactive semantic segmentation. *ArXiv Preprint ArXiv:2112.12702*. (2021)
16. Wang, J., Guo, B. & Chen, L. Human-in-the-loop Machine Learning: A Macro-Micro Perspective. (2022), https://arxiv.org/abs/2202.10564
17. Settles, B. Active Learning. *Synthesis Lectures On Artificial Intelligence And Machine Learning*. **6** (2012,6)
18. Monarch, R. Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI. (Simon,2021)
19. Martos, V., Ahmad, A., Cartujo, P. & García, J. Ensuring Agricultural Sustainability through Remote Sensing in the Era of Agriculture 5.0. *Applied Sciences*. **11** (2021,6)
20. Li, D., Wang, Z., Chen, Y., Jiang, R., Ding, W. & Okumura, M. A Survey on Deep Active Learning: Recent Advances and New Frontiers. (2024), https://arxiv.org/abs/2405.00334
21. Ren, P., Xiao, Y., Chang, X., Huang, P., Li, Z., Gupta, B., Chen, X. & Wang, X. A survey of deep active learning. *ACM Computing Surveys (CSUR)*. **54**, 1-40 (2021)
22. Sinha, S., Ebrahimi, S. & Darrell, T. Variational adversarial active learning. *Proceedings Of The IEEE/CVF International Conference On Computer Vision*. pp. 5972-5981 (2019)
23. Jin, Q., Yuan, M., Qiao, Q. & Song, Z. One-shot active learning for image segmentation via contrastive learning and diversity-based sampling. *Knowledge-Based Systems*. **241** pp. 108278 (2022)
24. Kee, S., Del Castillo, E. & Runger, G. Query-by-committee improvement with diversity and density in batch active learning. *Information Sciences*. **454** pp. 401-418 (2018)
25. Zhu, J., Wang, H., Tsou, B. & Ma, M. Active Learning With Sampling by Uncertainty and Density for Data Annotations. *IEEE Transactions On Audio, Speech, And Language Processing*. **18**, 1323-1331 (2010)
26. Wu, X., Chen, C., Zhong, M. & Wang, J. HAL: Hybrid active learning for efficient labeling in medical domain. *Neurocomputing*. **456** pp. 563-572 (2021)
27. Hosna, A., Merry, E., Gyalmo, J., Alom, Z., Aung, Z. & Azim, M. Transfer learning: a friendly introduction. *Journal Of Big Data*. **9**, 102 (2022,10), https://doi.org/10.1186/s40537-022-00652-w
28. Iman, M., Arabnia, H. & Rasheed, K. A Review of Deep Transfer Learning and Recent Advancements. *Technologies*. **11** (2023), https://www.mdpi.com/2227-7080/11/2/40
29. Usman, M., Zia, T. & Tariq, A. Analyzing transfer learning of vision transformers for interpreting chest radiography. *Journal Of Digital Imaging*. **35**, 1445-1462 (2022)

30. Chughtai, I., Naseer, A., Tamoor, M., Asif, S., Jabbar, M. & Shahid, R. Content-based image retrieval via transfer learning. *Journal Of Intelligent Fuzzy Systems*. **44**, 8193-8218 (2023)
31. A., S. & R., S. A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends. *Decision Analytics Journal*. **7** pp. 100230 (2023), https://www.sciencedirect.com/science/article/pii/S277266222300070X
32. Letrache, K. & Ramdani, M. Explainable Artificial Intelligence: A Review and Case Study on Model-Agnostic Methods. *2023 14th International Conference On Intelligent Systems: Theories And Applications (SITA)*. pp. 1-8 (2023)
33. Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G. & Ranjan, R. Explainable AI (XAI): Core Ideas, Techniques, and Solutions. *ACM Comput. Surv.*. **55** (2023,1), https://doi.org/10.1145/3561048
34. Rawal, A., McCoy, J., Rawat, D., Sadler, B. & Amant, R. Recent Advances in Trustworthy Explainable Artificial Intelligence: Status, Challenges, and Perspectives. *IEEE Transactions On Artificial Intelligence*. **3**, 852-866 (2022)
35. Asif, S., Uzair Akmal, M., Koval, L., Knollmeyer, S., Mathias, S. & Grossmann, D. Supervised Anomaly Detection for Production Line Images using Data Augmentation and Convolutional Neural Network. *2024 IEEE 29th International Conference On Emerging Technologies And Factory Automation (ETFA)*. pp. 1-8 (2024)
36. Parkhi, O., Vedaldi, A., Zisserman, A. & Jawahar, C. Cats and dogs. *2012 IEEE Conference On Computer Vision And Pattern Recognition*. pp. 3498-3505 (2012)
37. Khosla, A., Jayadevaprakash, N., Yao, B. & Fei-Fei, L. Novel Dataset for Fine-Grained Image Categorization. *First Workshop On Fine-Grained Visual Categorization, IEEE Conference On Computer Vision And Pattern Recognition*. (2011,6)
38. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M. & Zheng, X. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. (2016,3)
39. Spinner, T., Schlegel, U., Schäfer, H. & El-Assady, M. explAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning. *IEEE Transactions On Visualization And Computer Graphics*. **26**, 1064-1074 (2020)
40. Beard, L. & Aghassibake, N. Tableau (version 2020.3). *Journal Of The Medical Library Association: JMLA*. **109**, 159 (2021)
41. Pala, S. Advance Analytics for Reporting and Creating Dashboards with Tools like SSIS, Visual Analytics and Tableau. (IJOPE,2017)
42. Grinberg, M. Flask web development: developing web applications with python. (" O'Reilly Media, Inc.",2018)
43. Lubanovic, B. FastAPI. (" O'Reilly Media, Inc.",2023)
44. Merkel, D. Docker: lightweight linux containers for consistent development and deployment. *Linux Journal*. **2014**, 2 (2014)
45. Beitzel, S., Jensen, E. & Frieder, O. MAP. *Encyclopedia Of Database Systems*. pp. 1691-1692 (2009), https://doi.org/10.1007/978-0-387-39940-9
46. Hast, A. Consensus ranking for increasing mean average precision in keyword spotting. *VIPERC 2020, 2nd International Workshop On Visual Pattern Extraction And Recognition For Cultural Heritage Understanding. Bari, Italy, 29 January, 2020.*. **2602** pp. 46-57 (2020)
47. Bhattacharya, M., Penica, M., O'Connell, E., Southern, M. & Hayes, M. Human-in-Loop: A Review of Smart Manufacturing Deployments. *Systems*. **11** (2023), https://www.mdpi.com/2079-8954/11/1/35
48. Salmani, P. & Lewis, P. Transfer Learning Can Introduce Bias. (2024,10)
49. Salman, H., Jain, S., Ilyas, A., Engstrom, L., Wong, E. & Madry, A. When does Bias Transfer in Transfer Learning?. (2022), https://arxiv.org/abs/2207.02842
50. Bellet, T., Banet, A., Petiot, M., Richard, B. & Quick, J. Human-centered AI to support an adaptive management of human-machine transitions with vehicle automation. *Information*. **12**, 13 (2020)
51. Chaddad, A., Peng, J., Xu, J. & Bouridane, A. Survey of explainable AI techniques in healthcare. *Sensors*. **23**, 634 (2023)
52. Chen, Y., Clayton, E., Novak, L., Anders, S. & Malin, B. Human-centered design to address biases in artificial intelligence. *Journal Of Medical Internet Research*. **25** pp. e43251 (2023)
53. Nenno, S. Potentials and Limitations of Active Learning: For the Reduction of Energy Consumption During Model Training. *Weizenbaum Journal Of The Digital Society*. **4** (2024)
54. Joshi, G., Walambe, R. & Kotecha, K. A Review on Explainability in Multimodal Deep Neural Nets. *IEEE Access*. **9** pp. 59800-59821 (2021)
55. Welsby, P. & Cheung, B. ChatGPT. *Postgraduate Medical Journal*. **99** pp. 1047-1048 (2023)